

UCLA

UCLA Electronic Theses and Dissertations

Title

A Unified Framework with Benchmarks for Human-like Visual and Relational Reasoning in the Real World

Permalink

<https://escholarship.org/uc/item/93h7n0ng>

Author

Ma, Xiaojian

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Unified Framework with Benchmarks  
for Human-like Visual and Relational Reasoning in the Real World

A dissertation submitted in partial satisfaction  
of the requirements for the degree  
Doctor of Philosophy in Computer Science

by

Xiaojian Ma

2023



© Copyright by

Xiaojian Ma

2023

# ABSTRACT OF THE DISSERTATION

A Unified Framework with Benchmarks  
for Human-like Visual and Relational Reasoning in the Real World

by

Xiaojian Ma

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2023

Professor Song-Chun Zhu, Chair

*Cogito, ergo sum.* Building machines that can think and reason like humans is a long-standing goal of AI. Despite the tremendous progress in AI we witnessed in recent years, it is still not clear whether these learning machines at scale can solve problems that require sophisticated thinking and reasoning, especially when the problems are also tied to ontologies (entities, relations) in the real world and raw sensory observations, *i.e.* visual and relational reasoning. Further, human-level reasoning and thinking also call for the capability of generalizing what the machine has learned to problem instances with their novel forms and combinations. We anticipate such generalization should be possible even with few data as well as on diverse modalities, *e.g.* vision, text, embodied 3D scenes, etc, which creates a significant gap between humans and machines.

This dissertation studies human-like visual and relational reasoning in the real world, aiming at closing the aforementioned gap between humans and machines. The first part of this thesis focuses on deepening the current understanding of the limitations of existing ML-based reasoning systems when compared to humans. To this end, a series of benchmarks are

developed in hope of examining the full spectrum of anticipated capabilities of these systems, including zero-shot, few-shot generalization, and adaptation to difficult modalities including embodied 3D scenes. Based off these new quests for AI reasoning, thorough evaluations are conducted with recently proposed reasoning systems, and their limitations are discussed.

The second part of this dissertation introduces a unified framework by drawing inspiration from the human language system, which is grounded, entity-centric, semantically rich, and could be the key to human-level generalization in reasoning [Fod75, Cho06, And18]. Specifically, the problem of learning language-like representations from a generative learning perspective is investigated. The resulting models can facilitate learning object-centric representations from images and discrete-continuous hybrid representations from text using an energy-based formulation. Further, intuitive and scalable inductive biases are developed to leverage the semantic supervision from the English language to learn object-centric and relational representations, to directly tackle the challenging zero-shot systematic generalization problem in visual and relational reasoning. Finally, what could be the next major move in the field is highlighted.

The dissertation of Xiaojian Ma is approved.

Cho-Jui Hsieh

Yizhou Sun

Ying Nian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2023

*Eureka.*

## TABLE OF CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Benchmarking Real World Human-like Visual and Relational Reasoning	3
1.2	A Unified Framework For Human-like Visual and Relational Reasoning in the Real World	4
<b>I</b>	<b>Benchmarking Real World Human-like Visual and Relational Reasoning</b>	<b>8</b>
<b>2</b>	<b>Bongard-HOI: Human-like Few-shot Visual Reasoning in the Wild</b>	<b>9</b>
2.1	Introduction	9
2.2	Bongard-HOI Benchmark	14
2.3	Possible Models for Bongard-HOI	18
2.4	Experiments	23
2.5	Related Work	26
2.6	Conclusion	28
2.A	More details on the Bongard-HOI Benchmark	29
2.B	More details on the oracle model	34
<b>3</b>	<b>Reconciling the Quest of Embodied AI and Scene Understanding: the SQA3D benchmark</b>	<b>38</b>
3.1	Introduction	38
3.2	The Situated Question Answering in 3D Scenes (SQA3D) Dataset	42
3.3	Models for SQA3D	46

3.4	Experiments . . . . .	49
3.5	Related Work . . . . .	54
3.6	Conclusion . . . . .	55
3.A	Data collection . . . . .	56
3.B	Dataset details . . . . .	57
3.C	Model details . . . . .	60
3.D	Additional empirical results . . . . .	65

## **II A Unified Framework for Human-like Visual and Relational Reasoning in the Real World 68**

<b>4</b>	<b>Unsupervised Object-Centric Learning using Deep Region Competition 69</b>
4.1	Introduction . . . . . 70
4.2	Related Work . . . . . 72
4.3	Methodology . . . . . 75
4.4	Experiments . . . . . 82
4.5	Conclusion . . . . . 89
4.A	Details on Models and Hyperparameters . . . . . 90
<b>5</b>	<b>Object-centric and Relational Representation Learning with RelViT . . 93</b>
5.1	Introduction . . . . . 93
5.2	Methodology . . . . . 97
5.3	Experiments . . . . . 101
5.4	Related Work . . . . . 109
5.5	Conclusion . . . . . 111

5.A	A formal description of learning in RelViT . . . . .	111
5.B	Additional details on RelViT . . . . .	112
5.C	Additional details on the datasets . . . . .	114
5.D	Additional results . . . . .	117
<b>6</b>	<b>Learning Hybrid Latent Representations with LDEBM . . . . .</b>	<b>120</b>
6.1	Introduction . . . . .	120
6.2	Preliminaries: Symbol-Vector Coupling Energy-Based Model (EBM) . . . . .	123
6.3	Latent Diffusion Energy-Based Model . . . . .	124
6.4	Experiments . . . . .	132
6.5	Discussions and Related Work . . . . .	140
6.6	Conclusion and Future Works . . . . .	143
6.A	Extra Experiment Details and Discussion . . . . .	144
<b>7</b>	<b>Conclusion . . . . .</b>	<b>150</b>
	<b>References . . . . .</b>	<b>152</b>



## LIST OF FIGURES

1.1	Human-level thinking and reasoning. . . . .	2
1.2	A representation learning perspective of <i>Plato’s allegory of the cave</i> . . . . .	5
1.3	An overview of our unified framework for visual and relational reasoning. . . . .	6
2.1	Illustration of a few-shot learning instance from our Bongard-HOI benchmark. . . . .	10
2.2	Examples of different actions with the same object. . . . .	11
2.3	Illustration of our four separate test sets for different types of generalization. . . . .	12
2.4	Class-agnostic (objectness) detections. . . . .	20
2.5	Illustration of our oracle model. . . . .	22
2.6	Samples of annotations where curations are needed. . . . .	30
2.7	Illustration of the context-dependent reasoning property of the Bongard problems (few-shot instances) in our Bongard-HOI benchmark. . . . .	33
2.8	Illustration of our oracle model. . . . .	34
2.9	Illustration of our oracle model. . . . .	35
2.10	A failure of our oracle model. . . . .	37
3.1	Task illustration of <u>S</u> ituated <u>Q</u> uestion <u>A</u> nswering in <u>3D</u> Scenes (SQA3D). . . . .	38
3.2	Examples from SQA3D. . . . .	40
3.3	Data collection pipeline of SQA3D. . . . .	43
3.4	Word cloud of $s^{\text{txt}}$ in SQA3D. . . . .	46
3.5	Question distribution in SQA3D . . . . .	46
3.6	Potential models for SQA3D. . . . .	47
3.7	Qualitative results. . . . .	53

3.8	Dataset collection Web UI for Stage I. . . . .	58
3.9	Dataset collection Web UI for Stage II. . . . .	59
3.10	Additional instruction set to the Amazon MTurk (AMT) participants in Stage II.	60
3.11	Dataset collection Web UI for Stage III. . . . .	61
3.12	Answer distribution (organized by question prefixes) before balancing. . . . .	62
3.13	Answer distribution (organized by question prefixes) after balancing. . . . .	62
3.15	Additional qualitative results. . . . .	66
3.16	Failure mode. . . . .	67
4.1	Overview of Deep Region Competition (DRC). . . . .	69
4.2	Pixel re-assignment. . . . .	78
4.3	Foreground extraction results for each dataset. . . . .	85
5.1	An overview of our method. . . . .	93
5.2	Results on HICO. . . . .	95
5.3	Histogram of reasoning hops over GQA training questions. . . . .	105
5.4	Ablation study on HICO. . . . .	108
5.5	Visual illustrations of image features against HOI categories on the HICO test set via t-SNE. . . . .	109
5.6	Visualization of correspondence. . . . .	109
5.7	Histograms of concepts in GQA training set. . . . .	117
6.1	Graphical illustration of the latent diffusion process. . . . .	121
6.2	Evaluation on 2D synthetic data. . . . .	133
6.3	Visualization of color-coded data points. . . . .	134

## LIST OF TABLES

2.1	An overview of different benchmark datasets covering HOI detection, few-shot learning, and abstract visual reasoning. . . . .	15
2.2	Quantitative results on the Bongard-HOI benchmark. . . . .	24
2.3	Number of concepts and few-shot instances in the validation and test sets. . . .	32
3.1	An overview of the different benchmark datasets covering grounded 3D scene understanding. . . . .	42
3.2	SQA3D dataset statistics. . . . .	46
3.3	Quantitative results on the SQA3D benchmark. . . . .	50
4.1	Foreground extraction results on training data measured in IoU and Dice. . . . .	84
4.2	Ablation study on Birds. . . . .	87
4.3	Performance of DRC on training and held-out testing data. . . . .	88
4.4	Performance of DRC on unseen testing categories. . . . .	89
4.5	Dimension of latent variables on each dataset. . . . .	90
4.6	Architecture of the generators, LEBMs and auxiliary classifiers. . . . .	91
4.7	Architecture of the generators, LEBMs and auxiliary classifiers (Cont'd). . . . .	92
5.1	Results on HICO dataset. . . . .	103
5.2	Results on GQA dataset. . . . .	105
5.3	Hyperparameters for RelViT. . . . .	113
5.4	Key details about the loss implementation in baselines and RelViT . . . . .	114
5.5	Statistics of the splits of HICO dataset. . . . .	114
5.6	Statistics of the splits of GQA dataset. . . . .	115

5.7	Examples of semantics (reasoning hops) in GQA dataset. . . . .	116
5.8	Statistics of concepts in GQA training set. . . . .	118
5.9	Results with larger ViT models on HICO. . . . .	119
5.10	Results with larger ViT models on GQA. . . . .	119
6.1	Results of language generation on PTB dataset. . . . .	135
6.2	Sentence completion on JerichoWorld dataset. . . . .	136
6.3	Results of interpretable text modeling on Daily Dialog (DD). . . . .	137
6.4	Dialog evaluation results on Stanford Multi-Domain Dialog (SMD) and DD. . .	137
6.5	Samples of unsupervisedly discovered action categories and corresponding utterances on SMD. . . . .	138
6.6	Dialog cases generated by Latent Diffusion Energy-Based Model (LDEBM) given the context. . . . .	139
6.7	Accuracy of sentence attribute control on Yelp. . . . .	140
6.8	Generated positive and negative reviews on Yelp. . . . .	140
6.9	Accuracy on AGNews. . . . .	141
6.10	Network architecture for the LDEBM prior. . . . .	147
6.11	Network architecture for the LDEBM prior (Cont'd). . . . .	148
6.12	Dialog evaluation results on SMD and DD. . . . .	149

## ACKNOWLEDGMENTS

First and foremost, I would like to give my deepest thanks to Professor Song-Chun Zhu, my advisor, for onboarding me to this legendary effort of building general artificial intelligence. Indeed, it's been such a fascinating and rewarding journey full of challenges, surprises, and growth. Your passion for AI research and grand vision for the future of this community have been my inspirations throughout the Ph.D. journey and will also guide me through the next chapter. I'm grateful for all the insightful ideas and suggestions you share.

I would also like to express my gratitude to Professor Ying Nian Wu, who is always there to chat with me and helps me navigate through myriad research challenges and career questions. I appreciate your wisdom and your dedication to pure research. Your pursuit of simplicity and elegance in statistics and AI have been invaluable lessons to me. The coffee chat we had in the office would be among the most precious moments in my 20s.

My committee members, Professor Cho-Jui Hsieh, and Professor Yizhou Sun, thank you for your generous support along the way. I've been a loyal reader of the research papers you authored. They always show me the most cutting-edge advancements in machine learning research and foster my multidisciplinary knowledge.

My Ph.D. research career wouldn't have been possible without the help from folks I've been fortunate enough to work with. Sirui Xie, Peiyu Yu, and Professor Yixin Zhu: It's always been my great pleasure to work with you. Especially during the unsettling days of the hype of the pandemic, it was the progress we made that kept me going and I'm proud of everything we achieved together. Siyuan Huang, Zilong Yong, Zilong Zheng, Qing Li, Professor Yitao Liang, Anji Liu, Zihao Wang, Shaofei Cai, Jiangyong Huang, William Zhu, Baoxiong Jia, Professor Wenjuan Han, Rujie Wu, Wei Wang, Zhenliang Zhang, Yixin Chen: thank you for supporting me during the collaborations with BIGAI, the projects we worked as a team are highlights of my Ph.D., you are the most talented group of people I have ever been together with. Chi Zhang, Lifeng Fan, Pan Lu, Ruiqi Gao, Shuwen Qiu, Yining

Hong, Muzhi Han, Feng Gao, Ran Gong, Mark Edmonds, Xu Xie, Zeyu Zhang, Ziyuan Jiao: thanks for all the chats and discussions we had in the past few years and I cannot appreciate more on how much I learned from the knowledge, ideas, and insights you shared.

I enjoyed the industry research experiences I had working with great minds on eye-opening projects. I thank Pannag Sanketi and Laura Graesser for their support during the days at Google Brain Robotics and the fascinating large-scale vision-based RL project. I also thank Professor Anima Anandkumar, Professor Yuke Zhu, Weili Nie, Professor Huaizu Jiang, Zhiding Yu, and Professor Chaowei Xiao from NVIDIA Research for their help and constructive advice on our large-scale transformer for the visual reasoning project. Last but not least, my last internship with DeepMind is full of great memories and I thank Kory Mathewson, Professor Doina Precup, Timothy Lillicrap, Owen He, Adam Santoro, Peter Humphreys, Zhitao Gong, Rui Zhu, Daniel Kasenberg and folks from both Montreal office and the interactive learning team in London office for their tremendous support and offerings. I appreciate everything I learned from you all.

Finally, I thank my parents for their unconditional love and support in every stage of my life. This thesis is dedicated to you.

## VITA

- 2019      B.Eng. in Computer Science and Technology, Tsinghua University
- 2020      Research Intern, Google Brain
- 2020      M.S. in Computer Science, UCLA
- 2021      Research Intern, NVIDIA Research
- 2021      Ph.D. Candidate in Computer Science, UCLA
- 2022      Research Scientist Intern, DeepMind

## PUBLICATIONS

*(Only publications mentioned in this dissertation are listed)*

*SQA3D: Situated Question Answering in 3D Scenes*, **Xiaojian Ma** and Silong Yong *et al.*, in International Conference on Learning Representations (ICLR), 2023

*Latent Diffusion Energy-Based Model for Interpretable Text Modeling*, Peiyu Yu, Sirui Xie and **Xiaojian Ma** *et al.*, in International Conference on Machine Learning (ICML), 2022

*Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions*, **Xiaojian Ma** and Huaizu Jiang *et al.*, in IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2022

*RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning*, **Xiaojian Ma** *et al.*, in International Conference on Learning Representations (ICLR), 2022

*Unsupervised Foreground Extraction via Deep Region Competition*, Peiyu Yu, Sirui Xie and **Xiaojian Ma** *et al.*, in Neural Information Processing Systems, 2021



# CHAPTER 1

## Introduction

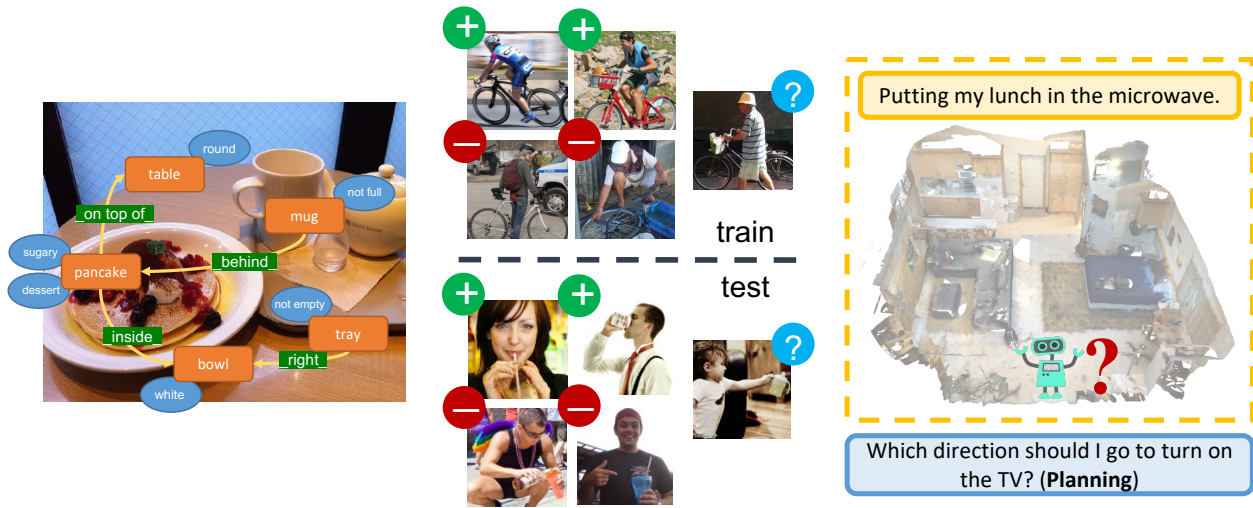
I propose to consider the question: ‘Can machine think?’

— A. M. Turing, 1950 [TUR50]

Building a machine that can think and reason like humans is a long-standing goal of AI. Since the earliest proposal of the Turing Test in 1950, AI has come a long way and it now can perform tasks that are previously thought to be exclusive to humans. However, despite the tremendous progress in AI we’ve witnessed in the past decade, the quest of building a thinking machine is still far from over. Even with the latest advancement in the large-scale transformer neural networks [VSP17a] that require massive web-scale data to train tens of billions of parameters [KMK20, BMR20, ADL22], it is still not clear to which extent these learning systems at scale can replicate human’s extraordinary capability of solving challenging problems with sophisticated thinking and reasoning [MIB23].

When it comes to thinking and reasoning, what do humans have that machines are still struggling with? Some studies rooted in cognitive science and developmental psychology [Kar94, Car00, Che11, AGP15] have pinpointed some key capabilities drawing the line between humans and machines. Specifically, humans are capable of:

- Reasoning over complex entities and relations among them that are grounded to the real world, from raw sensory observations, *i.e.*, **visual and relational reasoning**.
- Generalizing what they have learned (*e.g.* concepts, knowledge) to their novel forms and combinations, and solve the new reasoning problems accordingly with no or very



(a) Reasoning over complex objects and their relations. (b) Generalization with zero-shot and few-shot learning. (c) Adaptation to diverse modalities, *e.g.* embodied 3D scenes.

Figure 1.1: Humans can perform extensive thinking and reasoning in the real world, posing huge challenges to AI (zoom in for a better view).

few data, *i.e.*, **systematic generalization** and **zero-shot and few-shot learning**.

- Adapting the above capabilities to **diverse modalities**, *e.g.* images, text, embodied 3D scenes, *etc.*

Figure 1.1 demonstrates these capabilities, which indeed, create a *valley of despair* between humans and machines. Nonetheless, building machines that can mimic these capabilities could be crucial to not only understanding how human intelligence emerges from a developmental perspective but also developing general AI and putting it into many real-world applications & teaming with humans.

The fundamental research question this thesis tries to address is to what extent can we shrink this gap and therefore pave the way towards **building human-level thinking and reasoning machines**. It calls for a joint effort of **computer vision**, **natural language processing**, **machine learning**, and **cognitive science** so we can have a better understanding of the challenges, and design models that could accommodate how our human think

and reason by drawing inspirations from these disciplines. We will describe how we conduct these efforts in the following sections.

## 1.1 Benchmarking Real World Human-like Visual and Relational Reasoning

[computers] are tyrants. They insist on being spoken to in special computer languages, and act as though they can't understand a simple English sentence.

— Terry Winograd, 1971 [Win71]

We would like to start with a better understanding of the current AI reasoning through benchmarking. Despite the importance of filling this gap between humans and machines in terms of the capabilities illustrated in Figure 1.1, unfortunately, many of these key aspects still remain largely untouched in the current computer vision, NLP, multimodal ML, and reasoning research in general. Let's break down the literature based off the aforementioned capabilities. 1) visual and relational reasoning in the real world: A majority of visual and relational reasoning tasks are still using synthetic images [ZGJ19a, NYM20a, SLY17, JHV17], which makes their generality to the real-world questionable. Some recent work including VQA and NLVRv2 [AAL15a, GKS17a, SZZ18, MRF19] pioneer real-world evaluations but they lack complex and abstract concepts & relations required to solving their tasks. GQA [HM19] combines the best of both worlds with real-world images and sophisticated relations to reason over. However, it does not offer generalization challenges as many counterparts. 2) generalization and zero-shot/few-shot learning: many zero-shot and few-shot learning tasks [VBL16, TZD20, KLG18, HPQ20a] only vaguely mention generalization in their evaluation protocols and fail to establish a clear benchmark on what exactly is expected to be generalized; 3) adaptation to diverse modalities: most of the existing visual reasoning benchmarks [GKS17a, SZZ18, HM19] are only evaluated on images, which limits their applicability to other modalities such as 3D scenes, *etc.*

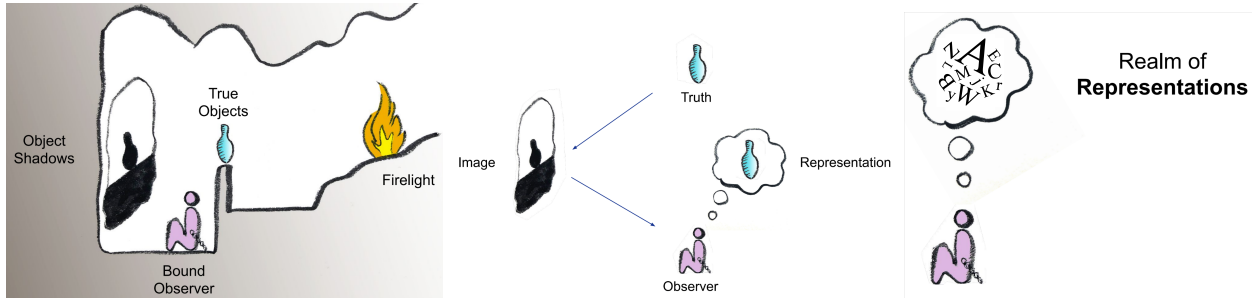
Motivated by the aforementioned deficiencies of existing works, in Chapter 2 we introduce Bongard-HOI [JNY22], the first benchmark that combines the best of both worlds: real-world images, reasoning over complex entities and their relations (human-object interactions, *i.e.* HOIs), few-shot learning and explicitly-required zero-shot generalization of the underlying HOI of the problem, all encapsulated in a minimalist format of the seminal Bongard problems [Bon68]. Our benchmarking demonstrates the gap between humans and machines on few-shot visual and relational reasoning and zero-shot generalization, and further suggests possible future directions including representation learning. Next, we bring visual reasoning to embodied 3D scenes in Chapter 3, aiming at benchmarking the reasoning capabilities of embodied agents. The resulting SQA3D dataset [MYZ22] reconciles 3D scene understanding and embodied AI via the proposed *embodied scene understanding and reasoning* tasks. The situation-aware, knowledge-intensive reasoning problems impose huge challenges to the current state-of-the-art reasoning systems, including powerful LLMs like GPT-3 [BMR20].

## 1.2 A Unified Framework For Human-like Visual and Relational Reasoning in the Real World

Solving a problem simply means representing it so as to make the solution transparent.

— Herbert A. Simon, 1969 [Sim69]

Our intuition to shrinking the human-machine gap in reasoning is to build a unified framework that draws inspiration from how our humans perceive the world & think and reason about it. There is rich philosophy, neuroscience, and cognitive science literature on explaining human thinking [BJ96, DA05, Kar94, AGP15], here we focus on a representation learning perspective of the infamous *Plato’s allegory of the cave* [EHR20] and the language of thought hypothesis [Fod75] (an illustration can be found in Figure 1.2). Frankly, humans are capable of developing representations out of raw sensory observations of the real-world.



(a) An illustration of Plato’s alle- (b) An interpretation from the per- (c) How will the representa-  
 gory of the cave. spective of representations. tions be like?

Figure 1.2: A representation learning perspective of *Plato’s allegory of the cave*. Humans can build mental representations just from raw sensory observations (“shadows”) of the real-world, in service of thinking, achieving goals, *etc.* We believe such representations are deeply shaped by our mental languages [Fod75] and might be the key to human-level reasoning. Graphics are adapted from [EHR20] (zoom in for a better view).

Such representations are highly structured, arguably shaped by our *mental language*, which is believed to closely connect to the language we use. Therefore, language-alike structures are very likely to be established within our mental representations of the world, empowering us with extraordinary capabilities of thinking and reasoning as discussed before. In short, we hypothesize that the representations are:

- **Entity-centric.** Just like how sentences are composed of words, the representations are composed by entities abstracted from the raw sensory observations, which provide the basis for mental processing later on.
- **Relational.** Upon the grounded entities consolidated in the representations, the properties of these entities (*unary relations*) and how they interact with each other (*binary and n-ary relations*) should also be infused. This is similar to how the sentences are enriched with more details.
- **Hybrid.** There is no doubt that our brain is analog and therefore continuous. But at the same time, we also anticipate some discrete or categorical aspects to form a

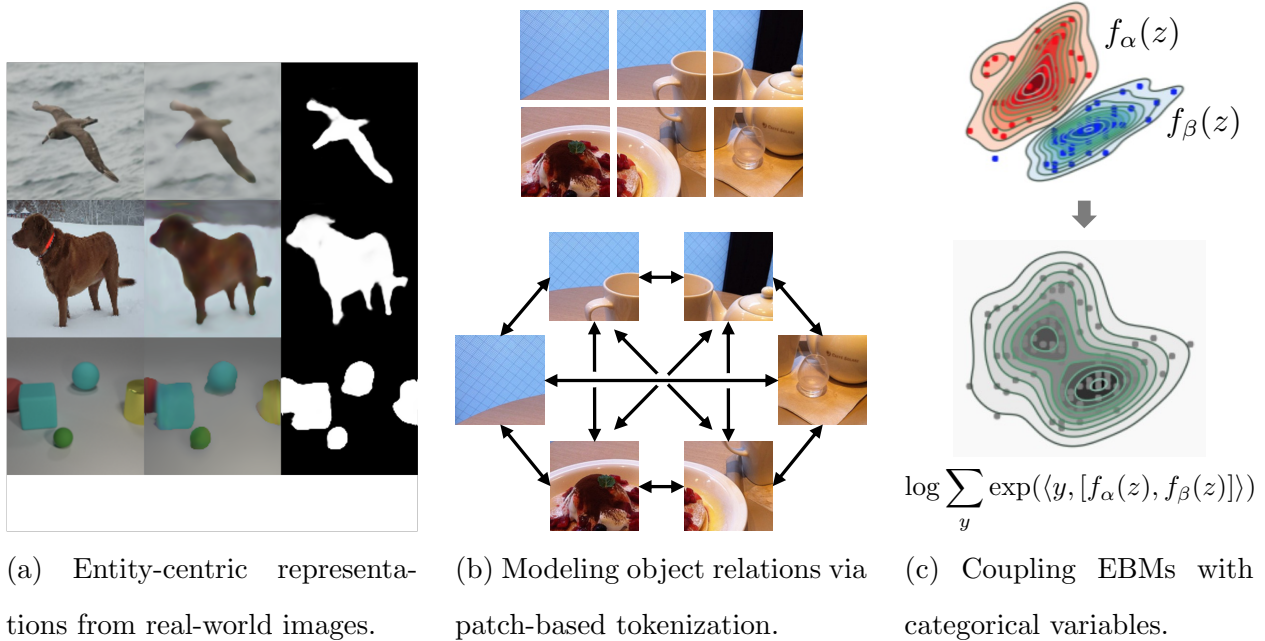


Figure 1.3: Our unified framework aims at mimicking the suggested human mental representations. The framework is composed of three components: 1) learning entity-centric representations from real-world images; 2) learning object relations via patch-based tokenization and relational inductive biases; 3) learning hybrid representations using symbol-vector coupling EBMs.

hybrid system. This effectively reflects how our discrete language system works – gluing various continuous parts in our brain and body.

The inverse engineering of human mental representation for reasoning ultimately leads to our unified framework. As demonstrated in Figure 1.3, our framework suggests learning these representations in a bottom-up fashion. In Chapter 4, we propose an energy-based generative learning formulation to extract object-centric representation from natural images. Fueling with a novel inductive bias that helps distinguish the background, the resulting DRC model [YXM21] robustly identifies the foreground objects and can generalize to new datasets in a zero-shot fashion. Next, we devise concept-guided contrastive learning [MNY22] based off vision transformers [DBK20] that boosts the emergence of better object-centric and relational representations in Chapter 5. Our model reaches state-of-the-art performances in

challenging visual and relational reasoning datasets without pretraining and further demonstrates promising results on the hard generalization tests. Finally in Chapter 6, we go back to the generative perspective of representation learning and tackle the problem of learning hybrid representations via symbol-vector coupling EBMs [PHN20]. By reconciling the powerful diffusion model [HJA20] with the original formulation, we achieve much better performances in downstream tasks with superior inference efficiency.

The dissertation is structured by the aforementioned two parts of understanding and tackling the problem of human-like visual and relational reasoning. We summarize our contributions and envision the possible next moves of this community in the last chapter.

Part I

**Benchmarking Real World  
Human-like Visual and Relational  
Reasoning**



## CHAPTER 2

# Bongard-HOI: Human-like Few-shot Visual Reasoning in the Wild

### 2.1 Introduction

In recent years, great strides have been made on visual recognition benchmarks, such as ImageNet [DDS09] and COCO [PH16]. Nonetheless, there remains a considerable gap between machine-level pattern recognition and human-level cognitive reasoning. Current image understanding models typically require a large amount of training data yet struggle to generalize beyond the visual concepts seen during training. In contrast, humans can reason about new visual concepts in a compositional manner from just a few examples [LST15]. To march towards human-level visual cognition, we need to depart from conventional benchmarks on closed-vocabulary recognition tasks and aim to systematically examine compositional and few-shot learning of novel visual concepts.

While existing benchmarks such as miniImageNet [VBL16], Meta-Dataset [TZD20], and ORBIT [TZD20] have been dedicated to studying few-shot visual learning, they focus on recognizing object categories instead of the compositional structures of visual concepts, *e.g.*, visual relationships. A parallel line of research aims at building benchmarks for abstract reasoning by taking inspiration from cognitive science such as RPM (Raven-style Progressive Matrices) [BHS18a, TWC20a] and Bongard-LOGO [Bon68, NYM20b]. In these benchmarks, a model has to learn concept induction rules from a few examples and the concepts are context-dependent in each task. However, they use simple synthetic images [BHS18a, NYM20b] or

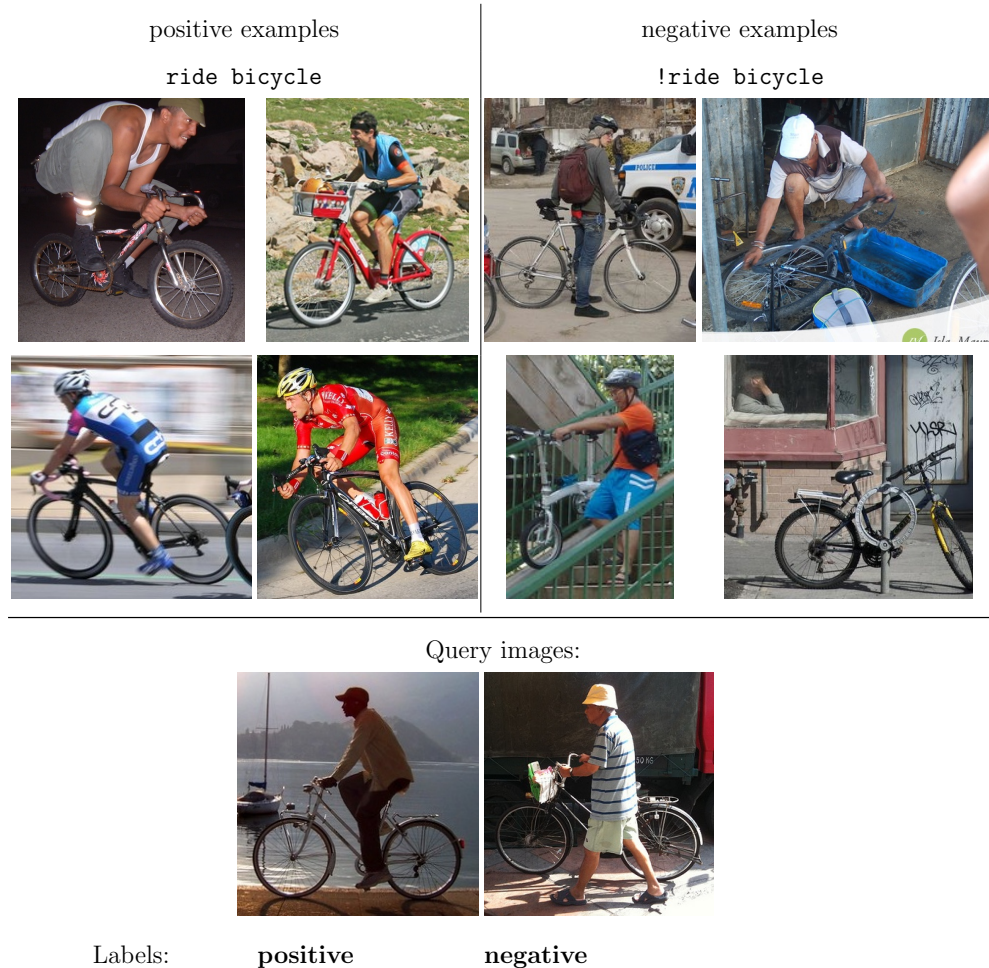


Figure 2.1: **Illustration of a few-shot learning instance from our Bongard-HOI benchmark.** The positive images in the top left part follow the visual relationship of riding a bike between the person and objects while such a relationship does not exist in the negative examples. Note that an actual problem in Bongard-HOI contains 6 images of positive examples, 6 negative examples, and 1 query image, which is different from the illustration here.

focus on basic object-level properties, such as shapes and categories [TWC20a].

**Our new benchmark:** In this paper, we introduce **Bongard-HOI**, a new benchmark for compositional visual reasoning with natural images. It studies human-object interactions (HOIs) as the visual concepts, requiring explicit compositional reasoning of object-level concepts. Our Bongard-HOI benchmark inherits two important characteristics of the classic



Figure 2.2: **Examples of different actions with the same object.** From top to bottom, left to right: washing, walking, and feeding dogs; eating, squeezing, and peeling oranges. To differentiate these images, we need compositional understanding on both the actions and the objects. We exploit this to select *hard negatives* in Bongard-HOI: negative images contain the same object as the positives, but the actions are different.

Bongard problems (BPs) [Bon68]: 1) *few-shot binary prediction*, where a visual concept needs to be induced from just six positive and six negative examples and 2) *context-dependent reasoning*, where the label of an image may be interpreted differently under different contexts.

Furthermore, Bongard-HOI upgrades the original BPs from synthetic graphics to natural images. Our benchmark contains rich visual stimuli featuring large intra-class variance, cluttered background, diverse scene layouts, etc. In Bongard-HOI, a single few-shot binary prediction instance, referred to as BP, contains a set of six positive images and a set of six negative images, along with query images (see Fig. 2.1 for examples). The task is making binary predictions on the query images.



Figure 2.3: **Illustration of our four separate test sets for different types of generalization.** We show a few HOI concepts in the training and test sets in the top and bottom row, respectively. We use the red fonts to denote an object or action class that is available in the training set and blue fonts indicate those held-on unseen ones in the test set.

We construct the few-shot instances in Bongard-HOI on top of the HAKE dataset [LXH19, LLL20]. To encourage the explicit reasoning of visual relationships, we use *hard negatives* to construct few-shot instances. The hard negatives consist of negatives that contain objects from the same categories as those contained in the positive sets but with different action labels. Fig. 2.2 presents some examples of these images. Since both positive and negative examples contain object instances from the same categories, mere recognition of object categories is insufficient to complete the tasks. Rather, reasoning about visual relationships between person and objects is required to solve these few-shot binary prediction problems. The existence of such hard negatives distinguishes our benchmark from existing visual abstract reasoning counterparts [BHS18a, TWC20a, NYM20b]. Comparisons with different benchmarks can be found in Table 2.1.

We carefully curate the annotations in HAKE when constructing the few-shot instances. Recall the visual concept contained in the positive images should not appear in any of the negative ones. Thus, we have to carefully select the images in both sets. We employ high-quality annotators from the Amazon Mechanical Turk platform to curate the test set to

further remove ambiguously and wrongly labeled few-shot instances. In this process, 2.5% of the few-shot instances in the test set are discarded. We end up with 23K and 15K few-shot instances in disjoint training and test sets, respectively.

An important goal of the Bongard-HOI benchmark is to *systematically* study the generalization of machine learning models for real-world visual relationship reasoning. To this end, we introduce four separate test sets to investigate different types of generalization, depending on whether the action and object classes are seen in the training set. Fig. 2.3 illustrates their design. This way, we have full control of the overlap between the concepts (*i.e.*, HOIs) between training and test of few-shot instances. It enables us to carefully examine the generalization of visual learning models. Ideally, a learning model should be able to generalize beyond the concepts it has seen during training. Even for unseen HOI concepts, the model should be able to learn *how to induce* the underlying visual relationship from just a few examples.

**Establishing baselines:** In our experiments, we first examine the state-of-the-art HOI detection models’ performance on this new task, we trained an oracle model with HOITrans [ZWH21b] on all the HOI categories, *including those in the test sets of our Bongard-HOI benchmark*, and output binary prediction on the query image via a majority vote based on HOI detections. Its accuracy is only 62.46% (with a chance performance of 50%), demonstrating the challenge of our visual reasoning tasks. We then evaluate state-of-the-art few-shot learning approaches, including non-episodic and meta-learning methods. We show that the current learning models struggle to solve the Bongard-HOI problems. Compared to amateur human testers’ 91.42% overall accuracy, who have access to a few examples of visual relationships before working on solving our problems, the state-of-the-art few-shot learning model [CWL20] only has 55.82% accuracy.

The results above lead to this question: *why do they perform so poorly?* To this end, we offer a detailed analysis of the results and propose several conjectures. The first one is a lack of holistic perception and reasoning systems, since models that have only good pattern

recognition performances, *e.g.* HOITrans, are likely to fail on our benchmarks. Moreover, we believe there is a need for additional representation learning, *e.g.* pre-training, since currently we only train on binary labels of few-shot instances. Nonetheless, we believe much effort is still needed to further investigate the challenges brought by our benchmark.

To sum up, this paper makes the following contributions:

- We introduce Bongard-HOI, a new benchmark for few-shot visual reasoning with human-object interactions, aiming at combining the best of few-shot learning, compositional reasoning, and challenging real-world scenes.
- We carefully curate Bongard-HOI with hard negatives, making mere recognition of object categories insufficient to complete our tasks. We also introduce multiple test sets to systematically study different types of generalization.
- We analyze state-of-the-art few-shot learning and HOI detection methods. However, experimental results show their inability on achieving good results on Bongard-HOI. Our conjectures suggest future research in models with holistic perception-reasoning systems and better representations.

## 2.2 Bongard-HOI Benchmark

For a few-shot binary prediction instance in Bongard-HOI, it has a set of positive examples  $\mathcal{P}$ , a set of negative samples  $\mathcal{N}$ , and a query image  $I_q$ . Images in  $\mathcal{P}$  depict a certain visual concept (*e.g.*, `ride bicycle` in Fig. 2.1), while images in  $\mathcal{N}$  do not. In each task, there are only six images in both  $\mathcal{P}$  and  $\mathcal{N}$ . As a result, a human tester or machine learning model needs to induce the underlying concept from just a few examples. Given the query image  $I_q$ , a binary prediction needs to be made: whether the certain visual concept depicted in  $\mathcal{P}$  is available in  $I_q$  or not. Later, we will detail how to construct these few-shot instances.



Table 2.1: **An overview of different benchmark datasets covering HOI detection, few-shot learning, and abstract visual reasoning.** In the first row, the abbreviation *ctx* denotes context; *generalization types* indicates if a benchmark includes multiple test splits to examine different types of generalization. \*We consider the concept of object counts as compositional while others such as object attributes and categories not [TWC20a]).

	concept	compositional concept	natural image	few- shot	ctx-dependent reasoning	hard negatives	generalization types	#concepts	#tasks
HAKE [LXH19, LLL20]	HOI	✓	✓	✗	✗	✓	✗	600	122.6K
Omniglot [LSG11]	shape	✗	✗	✓	✓	✗	✗	50	1.62K
miniImageNet [VBL16]	image label	✗	✓	✗	✓	✗	✗	100	60K
Meta-Dataset [TZD20]	image label	✗	✓	✓	✗	✗	✗	4,934	52.8M
ORBIT [MZB21]	frame label	✗	✓	✓	✗	✗	✗	486	2.69M
RPM [BHS18a]	shape	✗	✗	✓	✓	✗	✓	50	11.36M
V-PROM [TWC20a]	attributes & counts	✓*	✓	✓	✓	✗	✓	478	235K
Bongard-LOGO [NYM20b]	shape	✗	✗	✓	✓	✗	✓	627	12K
Bongard-HOI (ours)	HOI	✓	✓	✓	✓	✓	✓	242	53K

### 2.2.1 Constructing Bongard Problems

Few-shot instances in Bongard-HOI are constructed with natural images. We choose to use visual relationships as underlying visual concepts. In our early experiments, we also studied visual attributes to construct few-shot instances, for example, color and shape of bird parts [WBW11], facial attributes [LLW15]. But such visual attributes annotations either require too much domain knowledge for human annotators or are too noisy to curate. Another option we investigated is scene graph [KOJ17], which is a combination of both visual relationships and visual attributes. However, there could be too many convoluted visual concepts in a single image, resulting in ambiguous few-shot instances.

In this paper, we construct few-shot instances on top of the HAKE dataset [LXH19, LLL20] focusing on human-object interactions. It provides unified annotations following the annotation protocol in HICO [CWH15a] for a set of datasets widely used for HOI detection, including HICO [CWH15a], V-COCO [GM15], OpenImages [KRA20], HCVRD [ZWS18],

and PIC [LLW20]. HAKE has 80 object categories, which are consistent with the vocabulary defined in the standard COCO dataset [LMB14]. It also has 117 action labels, leading to 600 human-object relationships<sup>1</sup>.

Denote a concept  $c = \langle s, a, o \rangle$  as a visual relationship triplet, where  $s, a, o$  are the class labels of subject, action, and object, respectively. In this paper,  $s$  is always **person**. We start with selecting a set of positive images  $\mathcal{I}_c = \{I_1, \dots\}$  from HAKE that depict such a relationship. We also need negative images, where the visual concept  $c$  is not contained by them. In specific, we collect another set of images  $\mathcal{I}_{\bar{c}}$  with concept  $\bar{c} = \langle s, \bar{a}, o \rangle$ , where  $\bar{a} \neq a$ , meaning that we select *hard negatives*. As a result, images from both  $\mathcal{I}_c$  and  $\mathcal{I}_{\bar{c}}$  contain the same categories of objects and the only differences are the action labels, *making it impossible to trivially distinguish positive images from the negatives by doing visual recognition of object categories only*. Rather, detailed visual reasoning about the interactions of human and objects are desired. Fig. 2.2 illustrates the difficulties introduced by the hard negatives. Finally, as an entire image may contain multiple HOI instances, we use image regions (crops) around each HOI instance instead of the original image to ensure only a single HOI instance is presented in a single image.

Next, we need to sample few-shot instances from the positive images  $\mathcal{I}_c$  and the negatives  $\mathcal{I}_{\bar{c}}$ . We randomly sample images to form  $\mathcal{P}, \mathcal{N}$ , and a query image  $I_q$ . Two parameters control the sampling process:  $M$ , the number of images in  $\mathcal{P}$  and  $\mathcal{N}$  ( $M = 6$  in Bongard-HOI), and the overlap threshold  $\tau$ , indicating the maximum number of overlapped images between two few-shot instances. We want to sample as many few-shot instances as possible, but we also need to avoid significant image overlap between few-shot instances, which limits the diversity of the data. We end up setting  $\tau = 3$  and  $\tau = 2$  for training and test sets, respectively. More details can be found in the supplementary material.

---

<sup>1</sup>Some combinations of objects and actions are infeasible.



### 2.2.2 Data Curation

Although the HAKE dataset [LXH19, LLL20] has provided high-quality annotations, we found that curations are still needed to construct few-shot instances. Recall, to sample negative images, we assume a particular action is not depicted in them. In HAKE, an image region may have multiple action labels. Naively relying on the provided annotations is problematic as the action labels are either not manually exclusive or not exhaustively annotated. We hire high-quality testers on the Amazon Mechanical Turk (MTurk) platform, who maintain a good job approval record, to curate existing HOI annotations. We discuss the data curation process in detail and show visual examples in detail in the supplementary material.

After the aforementioned data curations, each image region is assigned to a single action label, describing the most salient visual relationship. With the curated annotations, action labels between a person and objects of a certain category are mutually exclusive so that we can significantly reduce the ambiguity when constructing few-shot instances. Finally, we hire high-quality testers on the MTurk platform to further remove the ambiguous few-shot instances in the test set. Every single few-shot instance is assigned to three independent testers. We compare their responses with the ground-truth labels and discard about 2.5% few-shot instances where none of the three testers correctly classifies the query images. In the end, we report the accuracy of human testers on those left unambiguous few-shot instances as a human study to examine human-level performance on our Bongard-HOI benchmark, where the average accuracy is 91.42%.

### 2.2.3 Generalization Tests

Transferring the knowledge that an agent has seen and learned is a hallmark of visual intelligence, which is a long-stand goal for the entire AI community. It is also a core focus of the Bongard-HOI benchmark. Following [BHS18a], we provide multiple test splits to investigate different types of generalization, aiming at a systematic understanding of how the

tested models generalize on our benchmark. Specifically, the visual concept we consider in Bongard-HOI is an HOI triplet  $\langle s, a, o \rangle$  and we have two variables of freedom: action  $a$  and object  $o$ . Therefore, by controlling whether an action or object is seen during training, we can study generalization to unseen actions, unseen objects, or a combination of two. We end up introducing four separate test sets, as shown in Fig. 2.3. We provide detailed statistics on our training and test sets in the supplementary material.

Ideally, after learning from examples of `sit_on bed`, a machine learning model can quickly grasp the concept `sit_on bench`. More importantly, such a model should learn *how to learn* from just a few examples, so that they can still induce the correct concept (visual relationship) in the most challenging cases, where both actions and objects are not seen during training (*e.g.*, `shear sheep`).

## 2.3 Possible Models for Bongard-HOI

There are many possible ways of tackling Bongard-HOI, such as few-shot learning, conventional HOI detection, etc. We are particularly interested in investigating few-shot learning methods, as our benchmark requires the learner to identify the visual concept with very few samples (positive and negative images in  $\mathcal{P}$  and  $\mathcal{N}$ , respectively). To further improve the few-shot learning methods, we consider encoding the images with Relation Network [SRB17], aiming at better compositionality in the learned representations. Finally, we introduce an oracle model to testify whether Bongard-HOI can be trivially solved using state-of-the-art HOI detection models.

### 2.3.1 Few-shot Learning in Bongard-HOI

We start with a formal definition of the few-shot learning problem in Bongard-HOI. Specifically, each task includes multiple few-shot *instance* with  $N = 2$  classes and  $2M$  samples, *i.e.*, the model learns from a training set  $\mathcal{S} = \mathcal{P} \cup \mathcal{N} = \{(I_1^P, 1), \dots, (I_M^P, 1), (I_1^N, 0), \dots, (I_M^N, 0)\}$

and is evaluated on a query image  $(I_q, y_q)$ . Each example  $(I, y)$  includes an image  $I \in \mathbb{R}^{H \times W \times 3}$  and a class label  $y \in \{0, 1\}$ , indicating whether  $I$  contains the visual concepts depicted in  $\mathcal{P}$ . In Bongard-HOI, we set  $M = 6$  as our default parameter and therefore each few-shot instance is “2-way, 6-shot”. Following [TZD20], we propose to solve these few-shot prediction instances with the following two families of approaches: **Non-episodic methods**. In these methods, a simple classifier is trained to map all the images in a few-shot instance (including images in  $\mathcal{P}$ ,  $\mathcal{N}$ , and the query image) to the class of the query. The classifier can be parameterized as a neural network over some learned image embeddings, *i.e.* representations produced by convolutional neural networks (CNNs). In other words, we view each few-shot instance as a single training sample  $(\bigcup_{i=1}^{2M+1} I_i, y_q)$  rather than a few-shot instance with multiple training samples  $(I, y)$ . Our experiments cover two different ways to encode the images: CNN and Wide Relational Network (WReN) [BHS18a, NYM20b]. **Meta-learning methods**. These methods adopt the episodic learning setting, *i.e.*, they learn to train a classifier using  $2M$  samples from  $\mathcal{S}$  and evaluate their trained classifier on the query  $(I_q, y_q)$ . In general, their objective (also called *meta-objective*) is to minimize the prediction error on the query. Different meta-learning methods have their own ways to build the classifier and optimize the meta-objective. In our experiments, we consider the following state-of-the-art methods: 1) *ProtoNet* [SSZ17], a metric-based method; 2) *MetaOptNet* [LMR19] and *ANIL* [RRB20], two optimization-based approaches. Moreover, we also use a strong baseline meta-learning model, *Meta-Baseline* [CWL20], which reports competitive results in many few-shot prediction tasks. We refer readers to the related papers for more details.

### 2.3.1.1 Image Encoding with Relational Network

As mentioned above, representation learning of the input images can be crucial to the success of few-shot learning methods on Bongard-HOI. As our benchmark demands learning compositional concepts (HOIs), simply feeding an image into a Convolutional Neural Network

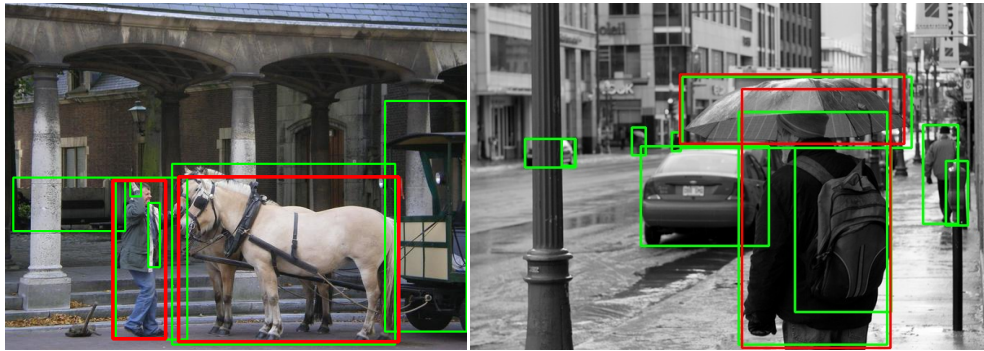


Figure 2.4: **Class-agnostic (objectness) detections.** We show the detections from our class-agnostic detector (in green) and ground-truth human and object boxes (in red).

(CNN) is not optimal. To this end, we propose to use the Relational Network [SRB17], which shows promising compositional reasoning accuracy on a Visual Question Answering (VQA) benchmark [JHM17], to explicitly encode the compositionality of visual relationships. In specific, the feature representations of the image  $I$  is computed as

$$\text{RN}(I) = f_{\phi} \circ \sum_{i,j} g_{\psi} (\text{concat}(h_{\theta}(o_i, I), h_{\theta}(o_j, I))),$$

where  $o_i$  and  $o_j$  are two detected objects of the image  $I$ , provided by ground truth object annotations or a pre-trained object detector like Faster R-CNN [RHG17].  $h_{\theta}$  denotes the RoI Pooled features of  $o_i$  from a ResNet backbone [HZR16] followed by a MLP (multi-layer perceptron) [RHG17], which is parameterized by  $\theta$ .  $g_{\psi}$  and  $f_{\phi}$  are two additional MLPs.

A challenge we are facing is the unseen object categories in the test sets. Since the object detector has to be pre-trained on a dataset without the unseen object categories, it is likely to fail on our test set where images could contain objects belonging to these categories. To tackle this issue, we train a binary class-agnostic (objectness) detection model instead to get  $o_i$  and  $o_j$ . Class-agnostic object detections are shown in Fig. 2.4. As we can see, all objects of interest have been successfully detected. But at the same time, there are a lot of other distracting ones, such as the bench and the wagon in the left image of Fig. 2.4. This is a unique challenge of dealing with visual reasoning over real-world images. We devote discussions to it in the experiment section.

### 2.3.2 Oracle

One may wonder if our Bongard-HOI benchmark could be trivially solved using the state-of-the-art HOI detection model. To address this concern, we develop an oracle model resorting to the HOITrans [ZWH21b], which is based on the Transformer model [VSP17b] and reports state-of-the-art accuracy on the HICO [CWH15a] and V-COCO [GM15] benchmarks. In specific, let’s denote the HOI detections in the  $\mathcal{P}$  and  $\mathcal{N}$  as  $\mathcal{D}^P$  and  $\mathcal{D}^N$ , respectively.  $\mathcal{D}^P$  contains the detections from all of the images in the  $\mathcal{P}$ , defined as  $\mathcal{D}^P = \{c_i^P\}_{i=1}^{N_P}$ , where  $c_i^P$  is a HOI triplet introduced in Section 2.2.1.  $N_P$  is the total number of detections. Note that there may be multiple or no detections for a single image. Similarly,  $\mathcal{D}^N$  is defined as  $\mathcal{D}^N = \{c_i^N\}_{i=1}^{N_N}$ . According to the property of Bongard-HOI, the visual concept  $c_P$  should only appear in the  $\mathcal{P}$ , not in the  $\mathcal{N}$ . We, therefore, compute  $c_P$  as

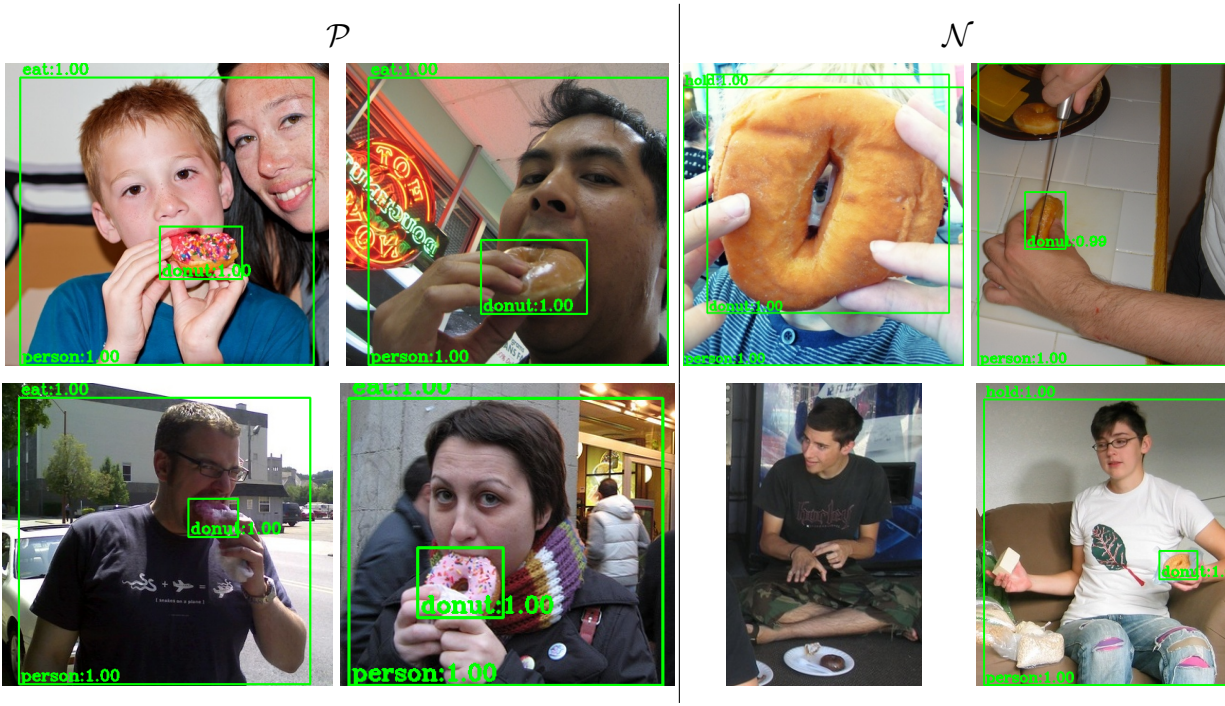
$$c_P = \text{majority\_vote}(\mathcal{D}^P - \mathcal{D}^N),$$

where  $-$  is the set operator for set subtraction. Here we first exclude the HOIs detected in  $\mathcal{N}$  from  $\mathcal{D}^P$ , then the majority of the remaining HOIs will be viewed as the visual concept  $c_P$ . Given the detections  $\mathcal{D}^q = \{c_i^q\}_{i=1}^{N_q}$  for the query image  $I_q$ , our prediction  $y$  becomes

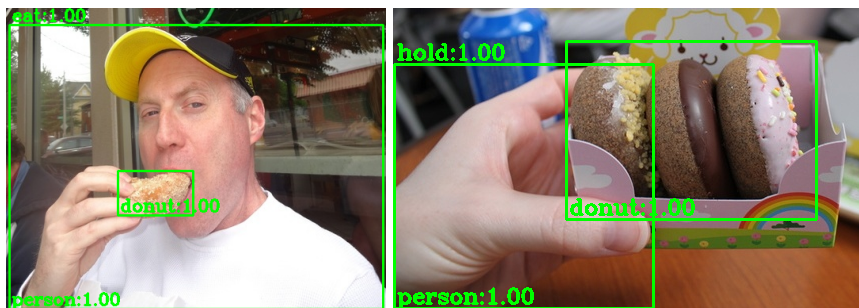
$$y = \begin{cases} 1, & \text{if } c_P \in \mathcal{D}^q, \\ 0, & \text{otherwise.} \end{cases}$$

Discussions of how to deal with the corner cases, *e.g.*, `majority_vote` returns more than 1 concept,  $\mathcal{D}^q$  is empty, etc, are provided in the supplementary material. We illustrate how this model works in Fig. 2.5, where we show HOI detections in each image.

We call it our oracle model as it has privileged information, *i.e.*, the entire HOI action & object vocabulary, including those held-out ones in the test set. As we shall we in Section 2.4, such an oracle model still struggles on our Bongard-HOI benchmark, achieving only 62.46% accuracy on average, which is far below the human-level performance of 91.42%. It suggests that our Bongard-HOI benchmark is not trivial to solve.



Query images:



Predictions:

positive

negative

Figure 2.5: **Illustration of our oracle model.** We first generate some detections for all the images using HOITrans [ZWH21b]. Note that some images may not have any detection at all. According to the detections in the  $\mathcal{P}$  and  $\mathcal{N}$ , the common concept is eat donut. As a result, in the bottom row, the first query image is considered to be positive as its HOI detections contain eat donat. The second query image is negative. Zoom in for the best view.



## 2.4 Experiments

### 2.4.1 Implementation Details

We benchmark the models introduced in Section 2.3 on Bongard-HOI to test their performance on human-level few-shot visual reasoning. We use a ResNet50 [HZR16] as an encoder for the input images. We consider different pre-training strategies: 1) no pre-training at all (scratch), 2) pre-trained on the ImageNet dataset with manual labels [DDS09], and 3) latest self-supervised approach [CFG20] pre-trained on ImageNet but without manual labels. We train an Faster R-CNN [RHG17] class-agnostic objectness detection model on the COCO dataset [PH16] using a ResNet101 [HZR16] pre-trained on ImageNet [DDS09] as the backbone. We use the RoIPool operation [RHG17] to get feature representations for each bounding box. We also use ground-truth bounding boxes provided in HAKE [LXH19] as input to diagnose the effectiveness of the visual perception. In addition to RoIPooled region features, we also concatenate each bounding box’s normalized coordinates (center and spatial dimensions) as spatial information to the Relational Network encoder introduced in Section 2.3.1.1.

### 2.4.2 Quantitative Results

The quantitative results of different models on our Bongard-HOI benchmark can be found in Table 2.2. We make the following observations: First of all, despite the overall difficulties brought by our benchmarks, most models perform worse on the challenging test splits, where actions and/or object categories are completely unseen during training. This observation aligns well with our hypothesis, *i.e.* existing machine learning approaches can be limited in terms of generalizing beyond training concepts. It also echos the findings in Bongard-LOGO [NYM20b], a dataset studying a similar problem with synthetic images. Second, meta-learning approaches generally tend to perform better than non-episodic counterparts, which can be on par with or even worse than random guesses (50% chance). We hypothesize

Table 2.2: **Quantitative results on the Bongard-HOI benchmark.** All the models use a ResNet50 as the image encoder. For the input of bounding boxes (bbox), we consider two options: from an object detection model (det) and ground-truth annotations (gt). For the ResNet50 encoder, we experiment with different pre-training strategies: no pre-training at all (scratch), pre-trained on the ImageNet dataset with manual labels (IN), and state-of-the-art self-supervised approach MoCoV2 [CFG20]. (\* denotes that we are unable to get meaningful results; # indicates that the trained model has a run-time error during the inference stage since the condition of the QP solver can not be satisfied).

	bbox	pre-train	test set				avg.
			seen act., seen obj.	seen act., unseen obj.	unseen act., seen obj.	unseen act., unseen obj.	
CNN-Baseline [NYM20b]	-	scratch	50.03	49.89	49.77	50.01	49.92
WReN-BP [BHS18a, NYM20b]	-	IN	50.31	49.72	49.97	49.01	49.75
ProtoNet* [SSZ17]	det	IN	-	-	-	-	-
ProtoNet [SSZ17]	gt	IN	58.90	58.77	57.11	58.34	58.28
MetaOptNet# [LMR19]	det	IN	-	-	-	-	-
MetaOptNet [LMR19]	gt	IN	58.60	58.28	58.39	56.59	57.97
ANIL [RRB20]	det	IN	50.18	50.13	49.81	48.83	49.74
ANIL [RRB20]	gt	IN	52.73	50.11	49.55	48.19	50.15
Meta-Baseline [CWL20]	det	scratch	54.61	53.79	54.58	53.94	54.23
Meta-Baseline [CWL20]	det	MoCoV2	55.23	54.54	54.32	53.11	54.30
Meta-Baseline [CWL20]	det	IN	56.45	56.02	55.60	55.21	55.82
Meta-Baseline [CWL20]	gt	IN	58.82	58.75	58.56	57.04	58.30
HOITrans [ZWH21b] (oracle)	-	-	59.50	64.38	63.10	62.87	62.46
Human (Amateur)	-	-	87.21	90.01	93.61	94.85	91.42

the reason to be the focus on *learning to learn* in these methods, which is essentially required to solve the few-shot instances in the Bongard-HOI benchmark, especially for the challenging test splits with novel categories. Similar observations have also been made in Bongard-



LOGO. Moreover, some meta-learning models are distracted by bounding boxes provided by an object detection model. We will discuss this issue in the next section.

Surprisingly, the oracle model (*HOITrans*) also struggles on our tests with an averaged accuracy of 62.46%, albeit being trained with direct HOI supervision and all action&object categories. It suggests a clear gap between the existing HOI detection datasets, *e.g.* HAKE [LXH19] and Bongard-HOI, where the latter one requires capabilities beyond perception, *e.g.* HOI recognition. Rather, a model might also need context-dependent reasoning, learning-to-learn from very few examples, etc., to perform well on our benchmarks.

Finally, machine learning models still largely fall behind amateur human testers (*e.g.*, 55.82% of Meta-Baseline vs 91.42%). While we only give human testers a couple of examples about visual relationships before they start working on solving Bongard-HOI, they can quickly learn how to induce visual relationships from just a few examples, reporting an average 91.42% accuracy on our Bongard-HOI benchmark. Particularly, there are no significant differences for the different subsets of the test set. We hope our findings will foster more research efforts on closing this gap.

### 2.4.3 Discussions

**We need holistic perception and reasoning.** Our work suggests that the significant challenges in current visual reasoning systems lie in both the reliability of perception and the intricacy of the reasoning task itself. Models that have only good pattern recognition performances are likely to fail on our benchmarks. Rather, an ideal learner needs to integrate visual perception in natural scenes and detailed cognitive reasoning as a whole. This marks our key motivation to propose Bongard-HOI as the first step towards studying these two problems holistically.

**Pre-training improves performances.** Intuitively, models for Bongard-HOI might need additional representation learning, *e.g.* pre-training, since currently we only train on binary labels of few-shot instances. We can see from Table 2.2 that *pre-training is very helpful*.

Compared to no pre-training, using either manual labels or self-supervision leads to a performance boost. In particular, the self-supervised pre-training [CFG20] does not use any manual labels for supervision. Yet it can produce better results than learning from scratch.

**Visual perception matters in Bongard-HOI.** Finally, an imperfect perception could still be a major obstacle here. Different from Bongard-LOGO [NYM20b] which uses synthetic shapes instead, Bongard-HOI studies visual reasoning on natural scenes, which often contain rich visual stimuli, issuing such as large intra-class variance and cluttered background also present challenges to reliable visual perception on which reasoning is based. In our case, bounding boxes produced by an object detection model can be inevitably noisy. Some meta-learning models, including ProtoNet [SSZ17], have difficulties inducing the true visual relationships. For MetaOptNet [LMR19], although we can finish training, we constantly encounter run-time errors where the condition of the QP solver is not satisfied during the inference stage. Instead, when taking clean ground-truth bounding boxes as input, all of these approaches produce better accuracy. Note that using ground-truth bounding boxes only serves as an oracle, which does not indicate the models’ authentic performance.

## 2.5 Related Work

**Visual relationship detection benchmarks.** Various benchmarks are also dedicated for visual relationship recognition and detection, particularly for human-centric relationships (*i.e.*, HOI). In the seminal work of Visual Genome [KOJ17], scene graph annotations, including relationships of different objects, are provided. A subset of the annotations is used in VRD [LKB16] to focus on visual relationship detection. In a recent effort, large-scale visual relationships are provided in the Open Images dataset [KRA20]. HOI, is of particular interest to understand the interactions of humans and other objects. A lot of HOI benchmarks, such as HICO [CWH15a], COCO-a [RP15], vCOCO [GM15], and HOI-COCO [HBQ21], are built on top of the object categories provided in the COCO dataset [LMB14]. The MECCANO [RFL21] dataset focuses on human-object interactions in egocentric settings

and industrial scenarios. Ambiguous-HOI [LLL20] is part of the HAKE project [LXH19], where the focus is human activity understanding with a large-scale knowledge base and visual reasoning.

Although our Bongard-HOI benchmark is built on top of the dataset HAKE [LXH19], it differs from the existing visual relationship and HOI benchmarks, since we focus on human-level cognitive reasoning instead of recognition. To solve Bongard-HOI, one might not need to explicitly name the underlying visual relationship but does need to induce the HOI from a few images and perform context-dependent reasoning. Our results also suggest that Bongard-HOI cannot be trivially solved by the state-of-the-art models on these datasets, *e.g.* HOITrans [ZWH21b].

**Few-shot and meta learning models.** Few-shot learning aims at learning from a limited number of training samples [Fe 03, KZS15]. With the goal of extracting the generic knowledge across tasks and generalizing to a new task using task-specific information, meta-learning (or learning-to-learn) [HYC01] becomes one of the leading approaches to deal with the few-shot learning problems. In general, meta-learning methods are divided into three categories: 1) memory-based methods, such as MANN [SBB16] and SNAIL [MRC18], 2) metric-based methods, such as Matching Networks [VBL16] and ProtoNet [SSZ17], and 3) optimization-based methods, such as MetaOptNet [LMR19] and ANIL [RRB20].

These meta-learning methods have been evaluated on several commonly used few-shot learning benchmarks, including miniImageNet [VBL16] and tieredImageNet [RTR18]. Although state-of-the-art meta-learning algorithms have achieved excellent performance on these standard few-shot image classification benchmarks, whether these approaches can generalize to tasks where the concepts to learn (in a few-shot manner) are compositional, *e.g.* visual relationships rather than simple object categories is unknown [KLG18, HPQ20a]. In other words, existing benchmarks fail to account for the challenging problem of generalizing to new compositional concepts in few-shot learning. Therefore, with a focus on the more challenging visual concepts of visual relationships, we propose Bongard-HOI to serve as a

new benchmark for the few-shot learning methods. We believe that our benchmark can foster the development of new few-shot learning, especially meta-learning algorithms to achieve better performances on learning and generalizing with compositional concepts. **Abstract visual reasoning benchmarks.** Inspired by cognitive studies, several benchmarks have been built for abstract reasoning, highlighting cognitive abstract reasoning. Notable examples include compositional question answering [JHM17, MNY22], physical reasoning [BMJ19, YGL20], math problems [SGH19], and general artificial intelligence [Cho19, XMY21a]. The most relevant to our benchmark are RPM [BHS18a, ZGJ19a], its variant with natural images [TWC20a], and Bongard problems with synthetic shapes [NYM20b] and physical problems [WR12]. While most of them consider synthetic images [BHS18a, NYM20b, WR12], our Bongard-HOI benchmark studies cognitive reasoning on natural images, which impose unique challenges due to the difficulty of visual perception. Moreover, we use human-object interaction as the underlying concepts to construct few-shot instances, which require explicit compositional concept learning in a few-shot manner, compared to the object categories and shapes [TWC20a]. Moreover, the existence of hard negatives in the few-shot instances makes our benchmark more challenging.

## 2.6 Conclusion

In this paper, we introduced the Bongard-HOI benchmark focusing on the few-shot learning and the generalization with compositional concepts in real-world visual relationship reasoning. Drawing inspirations from the classic Bongard problems [Bon68], we constructed few-shot instances using the visual relationships between humans and objects as the underlying concepts. Our benchmark is built on top of an existing HOI dataset, HAKE [LXH19], where we carefully curated the provided annotations to construct the few-shot instances. We benchmarked state-of-the-art few-shot learning methods, including both non-episodic and meta-learning approaches. Our findings suggested that current machine learning models still struggle to generalize beyond concepts that they have seen during the training process.

Moreover, natural images in our benchmark contain rich stimuli, imposing great challenges to the machine learning models in the real-world visual relationship reasoning tasks. By building the Bongard-HOI benchmark, we hope to foster research efforts in real-world visual relationship reasoning, especially in holistic perception-reasoning systems and better representation learning.

## 2.A More details on the Bongard-HOI Benchmark

### 2.A.1 Constructing Bongard Problems

Given positive images  $\mathcal{I}_c$  that depict a certain relationship  $c = \langle s, a, o \rangle$  and negative images  $\mathcal{I}_{\bar{c}}$  that does not, we need to sample few-shot instances from them. We randomly sample images to form  $\mathcal{P}$ ,  $\mathcal{N}$ , and a query image  $I_q$ . Two parameters control the sampling process:  $M$ , the number of images in  $\mathcal{P}$  and  $\mathcal{N}$  ( $M = 6$  in Bongard-HOI), and the overlap threshold  $\tau$ , indicating the maximum number of overlapped images between two few-shot instances. We want to sample as many few-shot instances as possible, but we also need to avoid significant image overlap between few-shot instances, which limits the diversity of the data. We set  $\tau = 3$  and  $\tau = 2$  for training and test sets, respectively.

### 2.A.2 Data Curation

Although the HAKE dataset [LXH19] has provided high-quality annotations, we found that curations are still needed to construct the Bongard problems (few-shot instances) for our Bongard-HOI benchmark. Recall, to sample negative images, we assume a particular action is not depicted in them. In HAKE, an image region may have multiple action labels. Naively relying on the provided annotations is problematic as the action labels are either not manually exclusive or not exhaustively annotated. We show different cases of data curations in Fig. 2.6 and discuss them in details as follows.

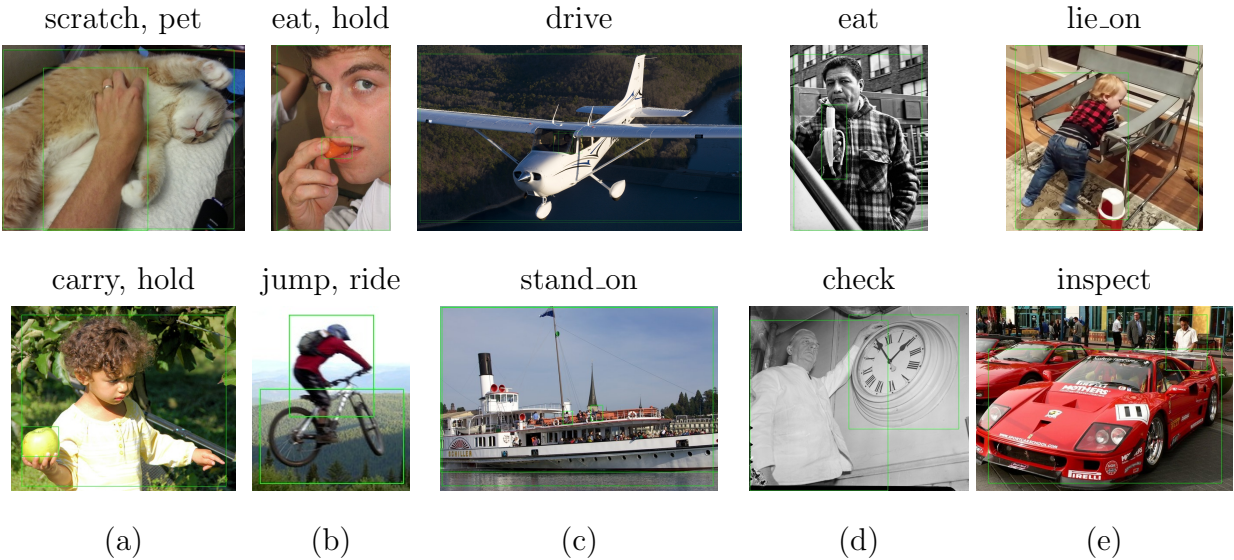


Figure 2.6: **Samples of annotations where curations are needed.** For each image region, its annotated action labels are shown on its top and bounding boxes corresponding to the person and object are shown for visualization purpose. From left to right: (a) similar actions, (b) hierarchical annotations, (c) hard-to-see objects, (d) extrapolating annotations, and (e) inaccurate or confusing annotations.

**Similar actions.** Although some action labels may convey different semantic meanings, for some certain object categories, they look visually similar and indistinguishable. As shown in Fig. 2.6(a), `scratch cat` and `pet cat` are hard to differentiate visually. If we simply use images of `scratch cat` as negatives to construct few-shot instances for `pet cat`, such few-shot instances are ambiguous, as it violates the basic assumption that the visual concept depicted in the Set  $\mathcal{A}$  is not available in the Set  $\mathcal{B}$ . We therefore simply merge such similar action labels to reduce the visual ambiguity.

**Hierarchical actions.** Action labels are inherently hierarchical. For example, as shown in Fig. 2.6(b), `eat carrot` very likely also means `hold carrot` visually. There are two problems to construct few-shot instances with multiple hierarchical action labels associated with the same image region. First of all, as we previously explained, using images of `eat carrot` as negatives for `hold carrot` may cause ambiguity. More importantly, there is the

*visual specificity* issue. People tend to focus on capturing the most salient actions in an image, which are usually the parent actions (`eat carrot` in this case). In our preliminary experiments, images of `eat carrot` were used as positives for `hold carrot` to construct few-shot instances. We found that it caused a lot of confusion for human testers. To this end, we merge such hierarchical action labels for the same region, keeping the parent action labels only.

**Hard-to-see objects.** In some cases, the person or the objects in image regions are hard to see. For example, in Fig. 2.6(c), the person with the action label `stand_on boat` is hard to see clearly. On the one hand, it causes significant challenges for a visual perception system (*e.g.*, [HGD20]) to accurately localize the meaningful objects. At the same time, it also imposes difficulty for annotators to accurately annotate the image region. We simply discard all image regions with hard-to-see objects.

**Extrapolating actions.** Actions are continuous. As a result, annotators tend to *extrapolate* the action label given a single image, instead of describing the current state the action. For example, as we can see in the top row of Fig. 2.6(d), the `eat` action is about to happen. Yet, the action is different from a normal `hold banana` without any indication of `eat`. To distinguish different scenarios, we introduce `hold_not_about_to_eat banana`, `hold_and_about_to_eat banana`, and `eat banana`. In this way, all the actions are mutually exclusive. We can sample image regions for form few-shot instances without worrying about causing ambiguity.

**Inaccurate or confusing actions.** In some rare cases, the annotations in HAKE are inaccurate or confusing, as shown in Fig. 2.6(e). We modify the action labels if such a image region depicts a clear action label. Otherwise we discard such regions to avoid introducing ambiguity to sampled few-shot instances.

**MTurk data curation.** After performing the aforementioned data curations, each image region is assigned to a single action label, describing the most salient content. Such action labels are mutually exclusive so that we can significantly reduce the ambiguity when con-

	seen object	unseen object		seen object	unseen object
seen action	99 / 5008	36 / 5002	seen action	102 / 4476	27 / 4562
unseen action	20 / 3402	12 / 3775	unseen action	21 / 3291	16 / 1612

(a) validation set

(b) test set

Table 2.3: **Number of concepts and few-shot instances in the validation and test sets.**

Depending on whether an action and object is seen during the training, we divide the validation and test sets into four categories, where we can study the systematic generalization of machine learning models. For each category, we show number of concepts (combinations of action and object) and number of few-shot instances.

structuring few-shot instances. Finally, we hire high-quality testers on the Amazon Mechanical Turk (MTurk) platform, who maintain a good job approval record, to curate the testing set to further remove the ambiguous few-shot instances. Every single BP is assigned to three independent testers. We compare their responses with the ground-truth labels and discard about 2.5% few-shot instances where none of the three testers correctly classifies the query images.

### 2.A.3 Dataset statistics

Our Bongard-HOI benchmark provides disjoint training, validation, and testing sets. In specific, there are 118 concepts (visual relationships) and 21,956 few-shot instances in the training set. There are 17,184 and 13,941 few-shot instances in the validation and testing set, respectively, corresponding to 167 and 166 visual concepts. Detailed distribution of concepts and few-shot instances among different generalization types are provided in Table 2.3.



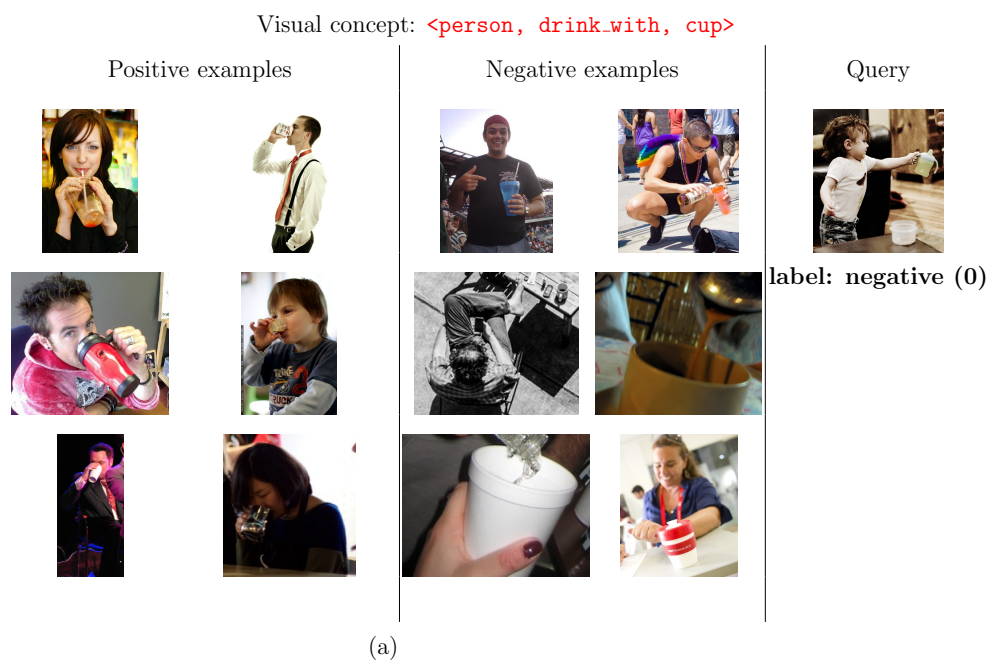


Figure 2.7: **Illustration of the context-dependent reasoning property of the Bongard problems (few-shot instances) in our Bongard-HOI benchmark.** Two instances are shown here with their underlying visual concepts (relationships) displayed on top with red color. The same query image receives two different labels (negative in the top and positive in the bottom) among different context (*i.e.*, positive and negative examples).

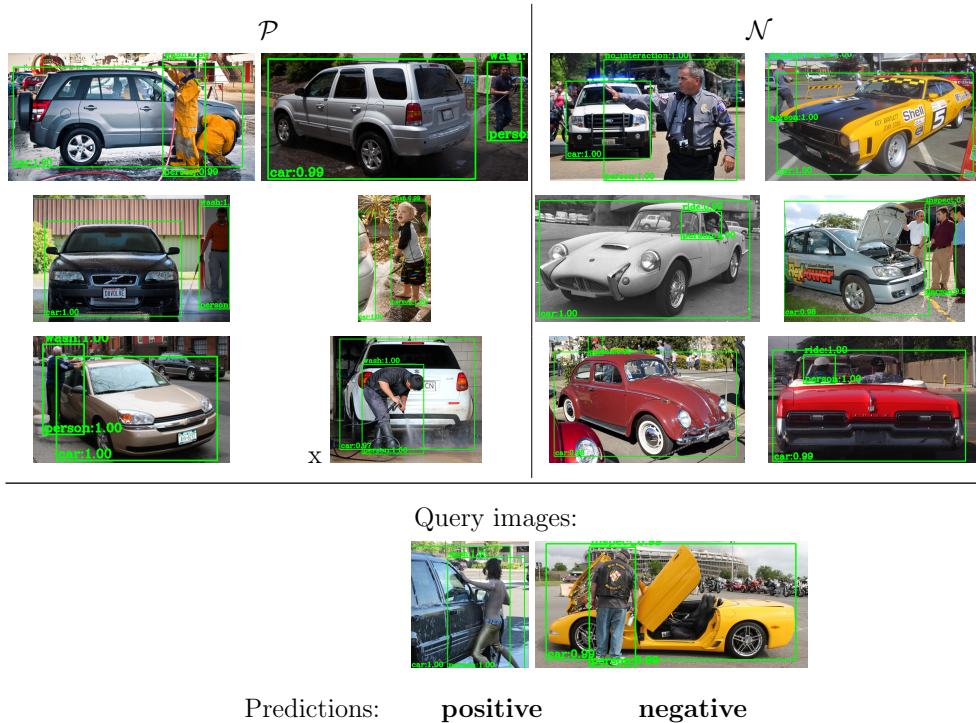


Figure 2.8: **Illustration of our oracle model.** The concept in  $\mathcal{P}$  is wash car.

#### 2.A.4 Illustration about the Context-Dependent Reasoning Property

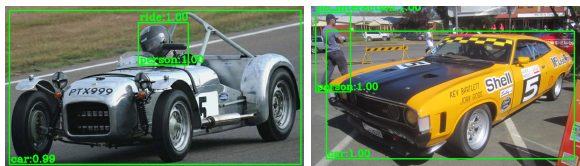
Two Bongard problems (few-shot instances) are shown in Fig. 2.7. For the same query image, among different context (*i.e.*, positive and negative examples), it receives different classification labels. This context-dependent reasoning property distinguishes our Bongard-HOI benchmark from other few-shot learning ones, where an image always has a fixed label.

## 2.B More details on the oracle model

We first review how our oracle model works. Denoting the HOI detections in the  $\mathcal{P}$  and  $\mathcal{N}$  as  $\mathcal{D}^P$  and  $\mathcal{D}^N$ , respectively.  $\mathcal{D}^P$  contains the detections from all of the images in the  $\mathcal{P}$ , defined as  $\mathcal{D}^P = \{c_i^P\}_{i=1}^{N_P}$ , where  $c_i^P$  is a HOI triplet.  $N_P$  is the total number of detections. Note that there may be multiple or no detections for a single image. Similarly,  $\mathcal{D}^N$  is defined as  $\mathcal{D}^N = \{c_i^N\}_{i=1}^{N_N}$ . According to the property of Bongard-HOI, the visual concept  $c_P$  should



Query images:



Predictions:                      **positive**                      **negative**

Figure 2.9: **Illustration of our oracle model.** The concept in  $\mathcal{P}$  is ride car.

only appear in the  $\mathcal{P}$ , not in the  $\mathcal{N}$ . We, therefore, compute  $c_P$  as

$$c_P = \text{majority\_vote}(\mathcal{D}^P - \mathcal{D}^N),$$

where  $-$  is the set operator for set subtraction. Given the detections  $\mathcal{D}^q = \{c_i^q\}_{i=1}^{N_q}$  for the query image  $I_q$ , our prediction  $y$  becomes

$$y = \begin{cases} 1, & \text{if } c_P \in \mathcal{D}^q, \\ 0, & \text{otherwise.} \end{cases}$$

We now discuss some possible corner cases where the main paper does not cover.

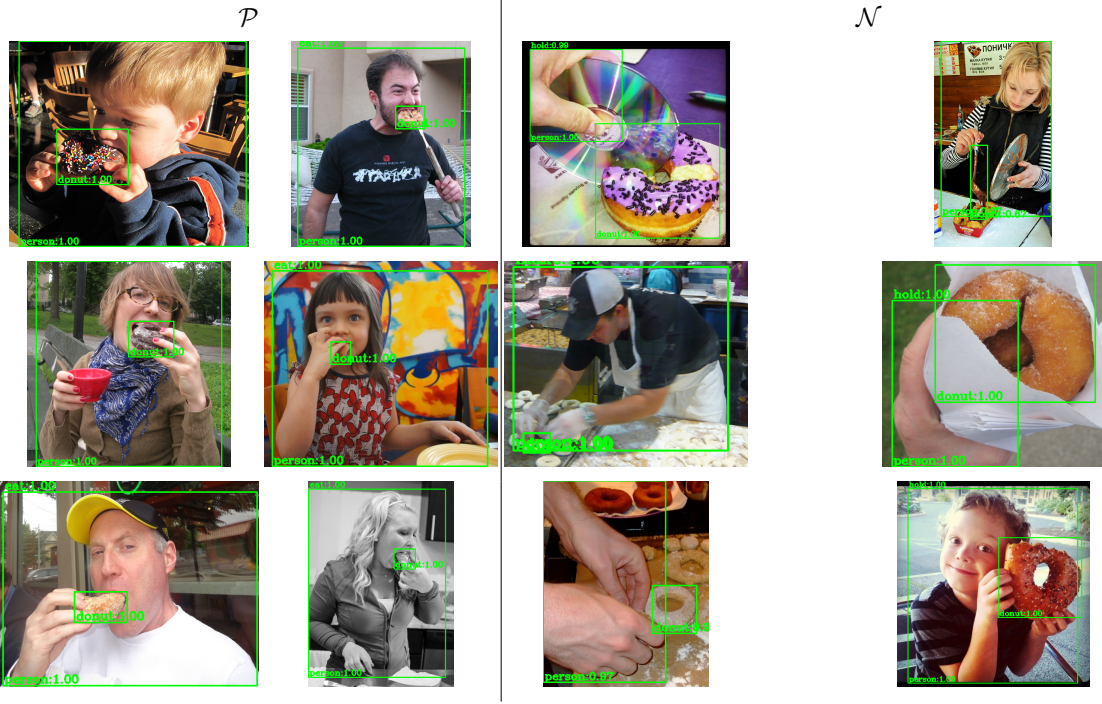
**What if majority\_vote return multiple concepts?** In this case, we simply enumerate

each of them when making predictions for  $y$ . The predicted  $y$  will be 1 as long as at least one returned concepts present in  $\mathcal{D}^q$ ; otherwise it will be 0.

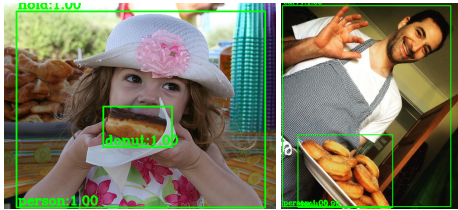
**What if  $\mathcal{D}^P$ ,  $\mathcal{D}^N$  or  $\mathcal{D}^q$  is empty?** In case when  $\mathcal{D}^P$  is empty, we view this example as an failure case for our oracle model, as it does not induce the right concept as expected. On the contrary, it's totally fine that  $\mathcal{D}^N$ , meaning that no detection need to be removed from  $\mathcal{D}^P$ . Finally, how we handle the case when  $\mathcal{D}_q$  is empty depends on the true label  $y^*$ . If  $y^*$  is 1, then we view this example as an failure case. But we will make the prediction an automatic success if  $y^*$  is 0, since our oracle model finds there is no ground truth concept presenting in the query, which should be the right prediction.

We show successful cases of our oracle model in Fig. 2.8 and Fig. 2.9. A failure case is shown in Fig. 2.10.





Query images:



Predictions:            negative (**wrong**)            negative

Figure 2.10: **A failure of our oracle model.** The concept in  $\mathcal{P}$  is eat cake. The *HOITrans* model [ZWH21a] incorrectly recognizes the first query image as hold cake (which should be eat cake). As a result, it makes a wrong prediction for the first query image.

## CHAPTER 3

# Reconciling the Quest of Embodied AI and Scene

## Understanding: the SQA3D benchmark

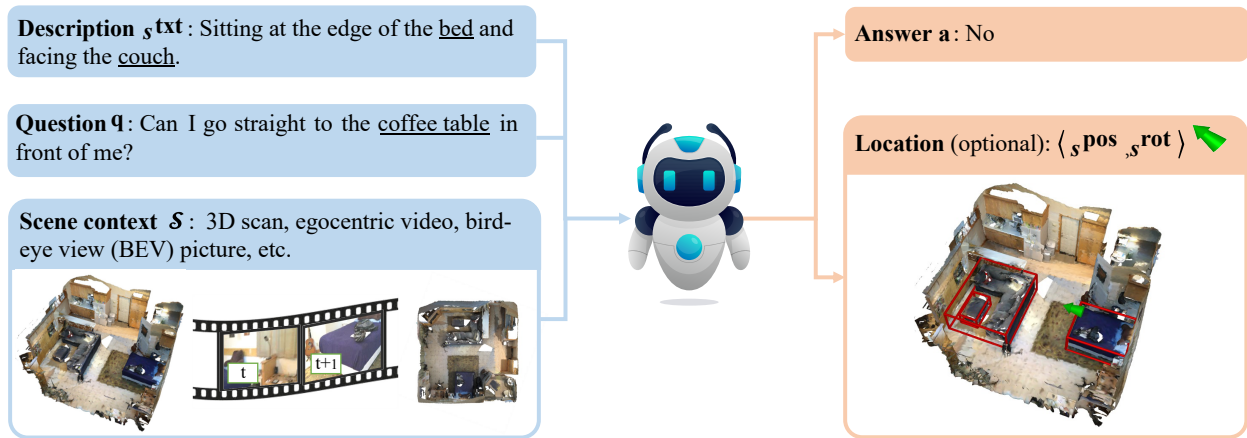


Figure 3.1: Task illustration of Situated Question Answering in 3D Scenes (SQA3D). Given scene context  $\mathcal{S}$  (*e.g.*, 3D scan, egocentric video, bird-eye view picture), SQA3D requires an agent to first comprehend and localize its **situation** (position, orientation, *etc.*) in the 3D scene from a textual description  $s^{\text{txt}}$ , then answer a question  $q$  under that situation. **Note that understanding the situation and imagining the corresponding egocentric view correctly is necessary to accomplish our task.** We provide more example questions in Figure 3.2.

### 3.1 Introduction


In recent years, the endeavor of building intelligent embodied agents has delivered fruitful achievements. Robots now can navigate [AWT18] and manipulate objects [LML19, SKM19,

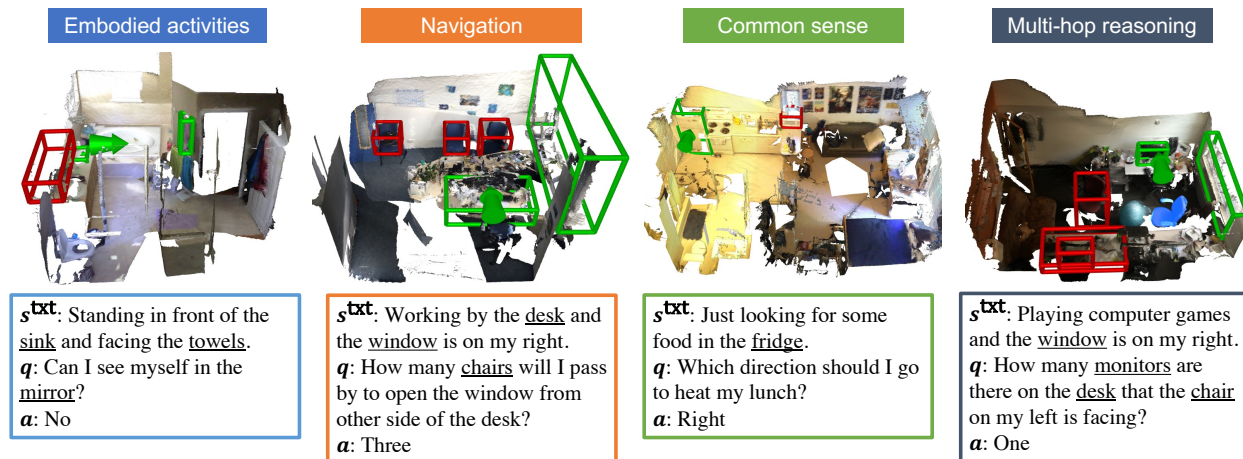
[SMF22, ABB22] following natural language commands or dialogues. Albeit these promising advances, their actual performances in real-world embodied environments could still fall short of human expectations, especially in generalization to different situations (scenes and locations) and tasks that require substantial, knowledge-intensive reasoning. To diagnose the fundamental capability of realistic embodied agents, we investigate the problem of **embodied scene understanding**, where the agent needs to understand its situation and the surroundings in the environment from a *dynamic* egocentric view, then perceive, reason, and act accordingly, to accomplish complex tasks.

**What is at the core of embodied scene understanding?** Drawing inspirations from situated cognition [Gre98, AGR00], a seminal theory of embodiment, we anticipate it to be two-fold:

- **Situation understanding.** The ability to imagine what the agent will see from arbitrary situations (position, orientations, *etc.*) in a 3D scene and understand the surroundings anchored to the situation, therefore generalize to novel positions or scenes;
- **Situated reasoning.** The ability to acquire knowledge about the environment based on the agents’ current situation and reason with the knowledge, therefore further facilitates accomplishing complex action planning tasks.

To step towards embodied scene understanding, we introduce **SQA3D**, a new task that reconciles the best of both parties, situation understanding, and situated reasoning, into embodied 3D scene understanding. Figure 3.1 sketches our task: given a 3D scene context (*e.g.*, 3D scan, ego-centric video, or bird-eye view (BEV) picture), the agent in the 3D scene needs to first comprehend and localize its situation (position, orientation, *etc.*) from a textual description, then answer a question that requires substantial situated reasoning from that perspective. We crowd-sourced the situation descriptions from AMT, where participants are instructed to select diverse locations and orientations in 3D scenes. To systematically examine the agent’s ability in situated reasoning, we collect questions that cover a wide spectrum of knowledge, ranging from spatial relations to navigation, common sense reasoning, and

Figure 3.2: **Examples from SQA3D**. We provide some example questions and the corresponding situations ( $s^{\text{txt}}$  and ) and 3D scenes. The categories listed here do not mean to be exhaustive and a question could fall into multiple categories. The **green boxes** indicate relevant objects in situation description  $s^{\text{txt}}$  while **red boxes** are for the questions  $q$ .



multi-hop reasoning. In total, SQA3D comprises 20.4k descriptions of 6.8k unique situations collected from 650 ScanNet scenes and 33.4k questions about these situations. Examples of SQA3D can be found Figure 3.2.

Our task closely connects to the recent efforts on 3D language grounding [DCS17, CCN20, CGN21, HLZ21, AAX20, WCL22, AMK22]. However, most of these avenues assume observations of a 3D scene are made from some third-person perspectives rather than an embodied, egocentric view, and they primarily inspect *spatial understanding*, while SQA3D examines scene understanding with a wide range of knowledge, and the problems have to be solved using an (imagined) first-person view. Embodied QA [DDG18, WDM19] draws very similar motivation as SQA3D, but our task adopts a simplified protocol (QA only) while still preserving the function of benchmarking embodied scene understanding, therefore allowing more complex, knowledge-intensive questions and a much larger scale of data collection. Comparisons with relevant tasks and benchmarks are listed in Table 3.1.

**Benchmarking existing baselines:** In our experiments, we examine state-of-the-art multi-modal reasoning models, including ScanQA from [AMK22] that leverages 3D scan



data, ClipBERT [LLZ21] and MCAN [YYC19a] that exploits egocentric videos and BEV pictures. However, the results unveil that both models still largely fall behind human performances by a large margin (47.2% of the best model vs. 90.06% of amateur human testers). To understand the failure modes, we conduct experiments on settings that could alleviate the challenges brought by situation understanding. The improvement of these models confirms that the current models are indeed struggling with situation understanding, which is pivotal for embodied scene understanding. Finally, we explore whether powerful Large Language Models (LLMs) like GPT-3 [BMR20] and Unified QA [KMK20] could tackle our tasks by converting the multi-modal SQA3D problems into single-modal surrogates using scene captioning. However, our results read that these models can still be bottlenecked by the lack of spatial understanding and accurate captions.

Our contributions can be summarized as follow:

- We introduce SQA3D, a new benchmark for embodied scene understanding, aiming at reconciling the challenging capabilities of situation understanding and situated reasoning and facilitating the development of intelligent embodied agents.
- We meticulously curate the SQA3D to include diverse situations and interesting questions. These questions probe a wide spectrum of knowledge and reasoning abilities of embodied agents, ranging from spatial relation comprehension to navigation, common sense reasoning, and multi-hop reasoning.
- We perform extensive analysis on the state-of-the-art multi-modal reasoning models. However, experimental results indicate that these avenues are still struggling on SQA3D. Our hypothesis suggests the crucial role of proper 3D representations and the demand for better situation understanding in embodied scene understanding.

Table 3.1: **An overview of the different benchmark datasets covering grounded 3D scene understanding.** In general, we consider semantic grounding, language-driven navigation, and question-answering in photo-realistic 3D scenes. In the first row, *situated* indicates whether the benchmark task is supposed to be completed by a “situated” agent with its egocentric perspective. *navigation*, *common sense*, and *multi-hop reasoning* show whether the task requires a certain capability or knowledge level of 3D understanding. \*Rather than observing a complete 3D scan of the scene, the learner needs to navigate in a simulator to perceive the 3D scene incrementally.

dataset	task	situated?	3D type	text collection	navigation?	common sense?	multi-hop reasoning?	#scenes	#tasks
ScanNet [DCS17]	seg.	✗	scan	n/a	✗	✗	✗	800 rooms	1.5k
ScanRefer [CCN20]	det.	✗	scan	human	✗	✗	✗	800 rooms	52k
ReferIt3D [AAX20]	det.	✗	scan	human	✗	✗	✗	707 rooms	41k
ScanQA [AMK22]	q.a.	✗	scan	template	✗	✗	✗	800 rooms	41k
3D-QA [YCH21]	q.a.	✗	scan	human	✗	✗	✗	806 rooms	5.8k
CLEVR3D [YYD21]	q.a.	✗	scan	template	✗	✗	✓	478 rooms	60k
MP3D-R2R [AWT18]	nav.	✓	*nav.	human	✓	✗	✗	190 floors	22k
MP3D-EQA [WDM19]	q.a.	✓	*nav.	template	✓	✗	✗	146 floors	1.1k
SQA3D (Ours)	q.a.	✓	scan	human	✓	✓	✓	650 rooms	33.4k

## 3.2 The SQA3D Dataset

A problem instance in SQA3D can be formulated as a triplet  $\langle \mathcal{S}, s, q \rangle$ , where  $\mathcal{S}$  denotes the scene context, *e.g.*, 3D scan, egocentric video, bird-eye view (BEV) picture, *etc.*;  $s = \langle s^{\text{txt}}, s^{\text{pos}}, s^{\text{rot}} \rangle$  denotes a situation, where the textual situation description  $s^{\text{txt}}$  (*e.g.*, “*Sitting at the edge of the bed and facing the couch*” in Figure 3.1) depicts the position  $s^{\text{pos}}$  and orientation  $s^{\text{rot}}$  of an agent in the scene;  $q$  denotes a question. The task is to retrieve the correct answer from the answer set  $a = \{a_1, \dots, a_N\}$ , while optionally predicting the ground truth location  $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$  from the text. The additional prediction of location could help alleviate the challenges brought by situation understanding. The following subsections will

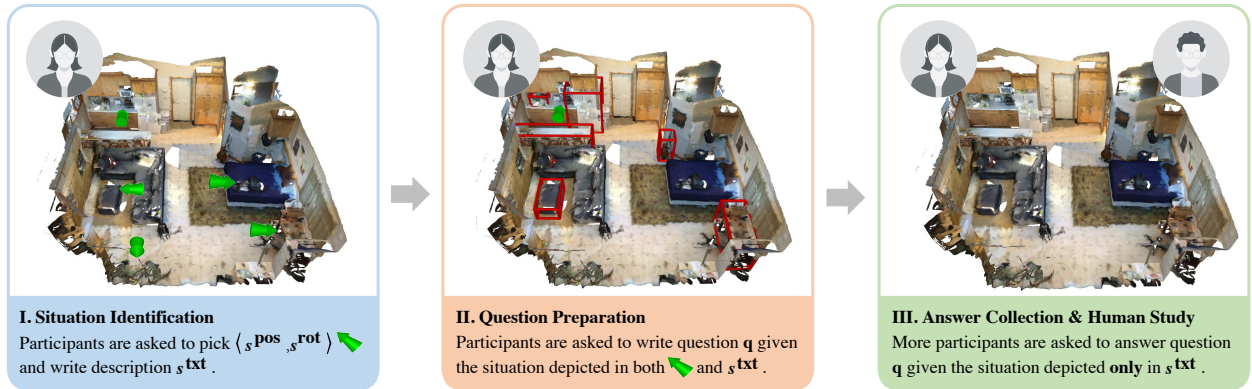


Figure 3.3: **Data collection pipeline of SQA3D.** Since our dataset comprises multiple types of annotations (situations and their descriptions, questions, answers, *etc.*), we found it more manageable to break down a single annotation task into three sub-tasks: i) Situation Identification; ii) Question Preparation; iii) Answer Collection & Human Study, where the participants recruited on AMT only need to focus on a relatively simple sub-task at a time.


detail how to collect and curate the data and then build the benchmark.


### 3.2.1 Data Formation


The 3D indoor scenes are selected from the ScanNet [DCS17] dataset. We notice that some scenes could be too crowded/sparse, or overall tiny, making situations and questions collection infeasible. Therefore, we first manually categorize these scenes based on the richness of objects/layouts and the space volume. We end up retaining 650 scenes after dropping those that failed to meet the requirement. We then develop an interactive web-based user interface (UI) to collect the data. Details of UI design can be found in *appendix*. All the participants are recruited on AMT.

Compared to counterparts, the annotation load of a single SQA3D problem instance could be significantly heavier as participants need to explore the scene, pick a situation, make descriptions, and ask a few questions. All these steps also require dense interaction with the 3D scene. To ensure good quality, we introduce a **multi-stage collection** pipeline, which

breaks down the load into more manageable sub-tasks. Figure 3.3 delineates this process:

**I. Situation Identification.** We ask the workers to pick 5 situations by changing the location  $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$  of a virtual avatar  in a ScanNet scene  $\mathcal{S}$ . The workers are then instructed to write descriptions  $s^{\text{txt}}$  that can **uniquely** depict these situations in the scene. We also use examples and bonuses to encourage **more natural sentences** and the **use of human activities** (e.g., “*I’m waiting for my lunch to be heated in front of the microwave*”). All the collected situations are later manually curated to ensure diversity and the least ambiguity. If necessary, we would augment the data with more situations to cover different areas of the scene.

**II. Question Preparation.** We collect a set of questions w.r.t. each pair of the 3D scene  $\mathcal{S}$ , and the situation description  $s^{\text{txt}}$  (the virtual avatar  is also rendered at  $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$ ). To help prepare questions that require **substantial situated reasoning**, we tutor the workers before granting them access to our tasks. They are instructed to follow the rules and learn from good examples. We also remove & penalize the responses that do not depend on the current situation, e.g. “*How many chairs are there in the room?*”.

**III. Answer Collection & Human Study.** In addition to the answers collected alongside the questions, we send out the questions to more workers and record their responses. These workers are provided with the same interface as in stage **II** except showing  in the scene to ensure consistency between question and answer collection. There is also **mandatory scene familiarization** in all three steps before the main job starts and we find it extremely helpful especially for more crowded scenes. More details can be found in *appendix*.

### 3.2.2 Curation, Data Statistics, and Metrics

**Curation.** Our multi-stage collection ends up with around 21k descriptions of 6.8k unique situations and 35k questions. Although the aforementioned prompt did yield many high-

quality annotations, some of them are still subject to curation. We first apply a basic grammar check to clean up the language glitches. Then we follow the practices in VQAv2 [GKS17b] and OK-VQA [MRF19] to further eliminate low-effort descriptions and questions. Specifically, we eliminate & rewrite template-alike descriptions (*e.g.*, repeating the same sentence patterns) and questions that are too simple or do not require looking at the scene. We also notice the similar answer bias reported in [MRF19] where some types of questions might bias toward certain answers. Therefore, we remove questions to ensure a more uniform answer distribution. A comparison of answer distribution before and after the balancing can be found in *appendix*. As a result, our final dataset comprises 20.4k descriptions and 33.4k diverse and challenging questions. Figure 3.2 demonstrates some example questions in SQA3D.

**Statistics.** Compared to most counterparts with template-based text generation, SQA3D is crowd-sourced on AMT and therefore enjoys more naturalness and better diversity. To the best of our knowledge, SQA3D is the **largest** dataset of grounded 3D scene understanding with the human-annotated question-answering pairs (a comparison to the counterparts can be found in Table 3.1). Table 3.2, Figure 3.4, and Figure 3.5 illustrate the basic statistics of our dataset, including the word cloud of situation descriptions and question distribution based on their prefixes. It can be seen that descriptions overall meet our expectations as human activities like “sitting” and “facing” are among the most common words. Our questions are also more diverse and balanced than our counterparts, where those starting with “What” make up more than half of the questions and result in biased questions [AMK22]. More statistics like distributions over answers and length of the text can be found in *appendix*.

**Dataset splits and evaluation metric.** We follow the practice of ScanNet and split SQA3D into *train*, *val*, and *test* sets. Since we cannot access the semantic annotations in ScanNet *test* set, we instead divide the ScanNet validation scenes into two subsets and use them as our *val* and *test* sets, respectively. The statistics of these splits can be found in Table 3.2. Following the protocol in VQAv2 [GKS17b], we provide a set of 706 “top-K”



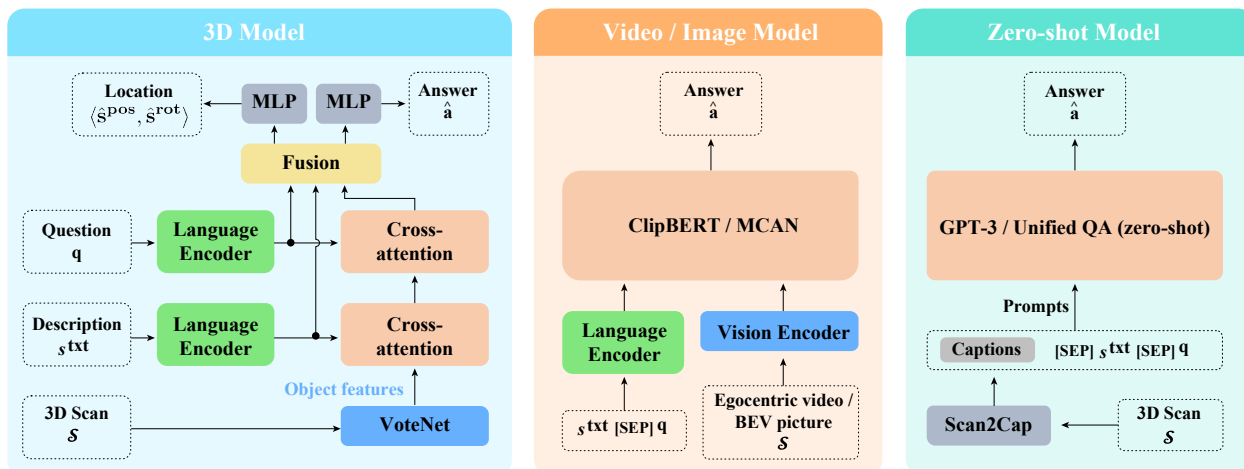


Figure 3.6: **Potential models for SQA3D.** We split the considered models into three groups: 3D model, video / image model, and zero-shot model. The 3D model is modified from the ScanQA model [AMK22] and maps 3D scan input to the answer. While the video / image models are effectively borrowed from canonical video QA and VQA tasks but we augment them with the additional situation input. The zero-shot model explores the potential of large pre-trained LLMs on our tasks. But they have to work with an additional 3D caption model that converts the 3D scene into text.

ibility with the protocol in SQA3D. To further improve this model, we consider including some auxiliary tasks during training [MNY22]. For other types of 3D scene context, *e.g.* egocentric video clips and BEV pictures, we employ the corresponding state-of-the-art models. Finally, we explore the potential of recently-introduced LLMs like GPT-3 [BMR20] and Unified QA [KMK20] on solving SQA3D in a zero-shot fashion. An overview of these models can be found in Figure 3.6.

**3D model.** We use the term *3D model* to refer a modified version of the ScanQA model [AMK22], depicted in the blue box of Figure 3.6. It includes a VoteNet [QLH19]-based 3D perception module that extracts object-centric features, LSTM-based language encoders for processing both questions  $q$  and situation description  $s^{\text{txt}}$ , and some cross-attention transformer blocks [VSP17b]. The object-centric feature tokens attend to the language tokens of  $s^{\text{txt}}$  and  $q$  successively. Finally, these features will be fused and mapped to predict the answer.



Optionally, we can add one head to predict the location  $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$  of the agent. Since the VoteNet module is trained from scratch, we also employ an object detection objective (not shown in the figure).

**Auxiliary task.** As we mentioned before, situation understanding plays a crucial role in accomplishing SQA3D tasks. To encourage a better understanding of the specified situation, we introduce two auxiliary tasks: the model is required to make predictions about the  $s^{\text{pos}}$  and  $s^{\text{rot}}$  of the situation. We use mean-square-error (MSE) loss for these tasks. The overall loss for our problem therefore becomes  $\mathcal{L} = \mathcal{L}_{\text{ans}} + \alpha\mathcal{L}_{\text{pos}} + \beta\mathcal{L}_{\text{rot}}$ , where  $\mathcal{L}_{\text{ans}}$ ,  $\mathcal{L}_{\text{pos}}$ , and  $\mathcal{L}_{\text{rot}}$  depicts the losses of the main and auxiliary tasks,  $\alpha$  and  $\beta$  are balancing weights.

**Video and Image-based model.** The orange box in the middle of Figure 3.6 demonstrates the models for video and image-based input. SQA3D largely resembles a video question answering or visual question answering problem when choosing to represent the 3D scene context  $\mathcal{S}$  as egocentric video clips or BEV pictures. However, SQA3D also requires the model to take both question  $q$  and the newly added situation description  $s^{\text{txt}}$  as input. We, therefore, follow the practice in the task of context-based QA [RJL18] and prepend  $s^{\text{txt}}$  to the question as a *context*. For the model, we use the state-of-the-art video QA system ClipBERT [LLZ21] and VQA system MCAN [YYC19a]. We adopt most of their default hyper-parameters and the details can be found in *appendix*.

**Zero-shot model.** We explore to which extent the powerful LLMs like GPT-3 [BMR20] and Unified QA [KMK20] could tackle our tasks. Following prior practices that apply GPT-3 to VQA [CKS22, GPT22], we propose to convert the 3D scene into text using an emerging technique called 3D captioning [CGN21]. We provide the caption,  $s^{\text{txt}}$ , and  $q$  as part of the prompt and ask these models to complete the answer. For GPT-3, we further found providing few-shot examples in the prompt helpful with much better results. Minor post-processing is also needed to ensure answer quality. We provide more details on prompt engineering in the *appendix*.



## 3.4 Experiments

### 3.4.1 Setup

We benchmark the models introduced in Section 3.3 to evaluate their performances on SQA3D. As mentioned before, we examine three types of scene context  $\mathcal{S}$ : 3D scan (point cloud), egocentric video, and BEV picture. Both the 3D scan and egocentric video for each scene are provided by ScanNet [DCS17]. However, we down-sample the video to allow more efficient computation per the requirement of the ClipBERT model [LLZ21]. The BEV pictures are rendered by placing a top-down camera on top of the scan of each 3D scene. We also conduct additional experiments that investigate factors that could contribute to the results, *e.g.*, situation and auxiliary tasks. In our early experiments, we found that the 3D model overall performs better than the video or image-based models. Therefore we only conduct these additional experiments with the variants of our 3D model due to the limit of computational resources. We use the official implementation of ScanQA, ClipBERT, and MCAN and include our modifications for SQA3D. For the zero-shot models, we extract 3D scene captions from two sources: ScanRefer [CCN20] and ReferIt3D [AAX20]. Considering the limit on the length of the input prompt, these 3D captions are also down-sampled. The Unified QA model weights are obtained from its Huggingface official repo. All the models are tuned using the validation set and we only report results on the test set. More details on model implementation can be found in *appendix*.

### 3.4.2 Quantitative Results

We provide the quantitative results of the considered models (detailed in Section 3.3) on our SQA3D benchmark in Table 3.3. The findings are summarized below:

**Question types.** In Table 3.3, we demonstrate accuracy on six types of questions based on their prefixes. Most models tend to perform better on the “Is” and “Can” questions

	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	<b>69.53</b>	37.89	43.41	43.65
ScanQA (w/o $s^{\text{txt}}$ )	3D scan	VQ→A	28.58	65.03	<b>47.31</b>	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	<b>69.53</b>	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	<b>66.10</b>	42.37	<b>69.53</b>	43.02	<b>46.40</b>	<b>47.20</b>
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA <sub>Large</sub>	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	<b>45.17</b>	41.11	41.00
Unified QA <sub>Large</sub>	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	<b>39.67</b>	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Table 3.3: **Quantitative results on the SQA3D benchmark.** Results are presented in accuracy (%) on different types of questions. In the “Format” column: V = 3D visual input  $\mathcal{S}$ ; S = situation description  $s^{\text{txt}}$ ; Q = question  $q$ ; A = answer  $a$ ; L = location  $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$ . In ScanQA, *aux. task* indicates the use of both  $\mathcal{L}_{\text{pos}}$  and  $\mathcal{L}_{\text{rot}}$  as additional losses. We use the *Large* variant as Unified QA [KMK20] as it works better.

while delivering worse results on “What” questions, likely due to a smaller number of answer candidates – most questions with binary answers start with “Is” and “Can”, offering a better chance for the random guess. Moreover, we observe the hugest gap between the blind test (model w/o 3D scene context input) and our best model on the “What” and “Which” categories, suggesting the need for more visual information for these two types of questions. This also partially echoes the finding reported in [LYB18].

**Situation understanding and reasoning.** At the heart of SQA3D benchmark is the requirement of situation understanding and reasoning. As we mentioned in Section 3.2.1,

the model will be more vulnerable to wrong answer predictions if ignoring the situation that the question depends on (*e.g.* “*What is in front of me*” could have completely different answers under different situations). In Table 3.3, removing situation description  $s^{\text{txt}}$  from the input leads to worse results, while adding the auxiliary situation prediction tasks boosts the overall performance, especially on the challenging “What” questions. The only exception is “How” questions, where a majority of them are about counting. We hypothesize that most objects in each ScanNet scene only have a relatively small number of instances, and the number could also correlate to the object category. Therefore, guessing/memorization based on the question only could offer better results than models with the situation as input if the situation understanding & reasoning are still not perfect yet. Additionally, we also provide an inspection of the relation between situation understanding and QA using attention visualization in Section 3.4.3.

**Representations of 3D scenes.** Indeed, SQA3D does not limit the input to be 3D scan only, as we also offer options of egocentric videos and BEV pictures. Compared to models with the 3D scan as input, the tested models with other 3D representations (*i.e.*, MCAN and ClipBERT) deliver much worse results, implying that the 3D scan so far could still be a better representation for the 3D scene when the reasoning models are probed with questions that require a holistic understanding of the scene. On the other hand, MCAN and ClipBERT are general-purpose QA systems, while ScanQA is designed for 3D-language reasoning tasks. The generalist-specialty trade-off could also partially account for the gap. Finally, the poor results of BEV and egocentric videos based models compared to the blind test could also be due to the additional “vision-bias” when the visual input is provided [AAL15b]. Note that the vision-bias can be mitigated with better visual representations [WXT21], implying that ScanQA, which seems to suffer less from the vision-bias than the counterparts using BEV and egocentric videos, is fueled by better visual representations in terms of combating the dataset bias.

**Zero-shot vs. training from scratch.** The success of pre-trained LLMs like GPT-3 on

myriads of challenging reasoning tasks [WWS22, WTB22] suggests that these models could possibly also understand embodied 3D scenes with language-only input [LJ93]. However, SQA3D imposes a grand challenge to these models. The powerful Unified QA (*Large* variant) and GPT-3 both fail to deliver reasonable results on our tasks. Further, we hypothesize the bottleneck could also be on the 3D captions, as the results verify the consistent impact on model performances brought by a different source of captions (ScanRefer→ReferIt3D). However, we still believe these models have great potential. For example, one zero-shot model (GPT-3 + ScanRefer) do pretty well on the challenging “What” questions (39.67%), even better than the best ScanQA variant.

**Human vs. machine.** Finally, all the machine learning models largely fall behind amateur human participants (47.2% of ScanQA + aux. task vs. 90.06%). Notably, we only offer a limited number of examples for the testers before sending them the SQA3D problems. Our participants promptly master how to interact with the 3D scene, understand the situation from the textual description, and answer the challenging questions. The human performance also shows no significant bias for different question types.

### 3.4.3 Qualitative Results

Finally, we offer some qualitative results of the variants of our 3D model in Figure 3.7. We primarily focus on visualizing both the answer predictions and the transformer attention over the object-centric feature tokens (bounding boxes) generated by the VoteNet [QLH19] backbone. We highlight the most-attended bounding box among all the predictions by the transformer-based model, in the hope of a better understanding of how these models perceive the 3D scene to comprehend the situations and answer the questions. In Figure 3.7, the correct predictions are always associated with attention over relevant objects in the situation description  $s^{\text{txt}}$  and questions. Moreover, in case there are multiple instances of the same object category, it is also crucial to identify the correct instance. For example, only ScanQA + aux. task makes the correct prediction for the first question and also attends

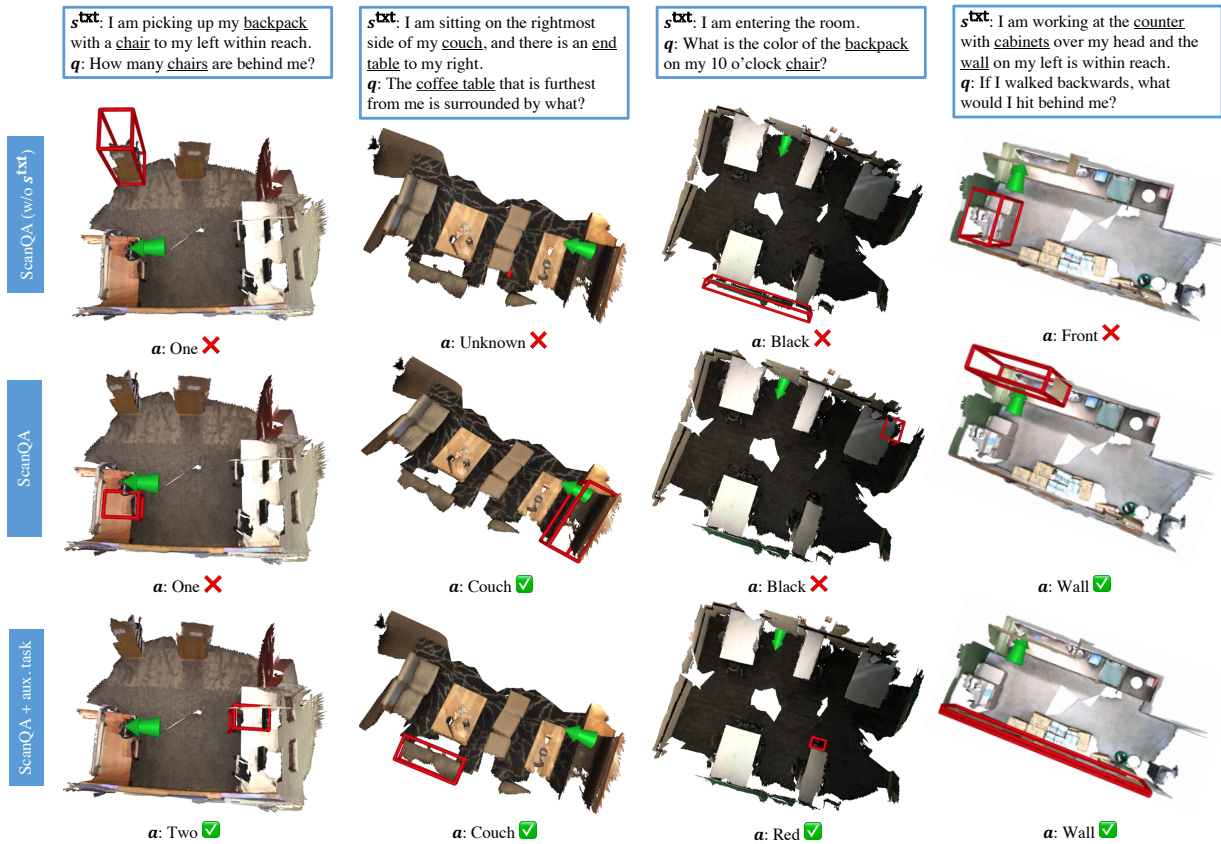



Figure 3.7: **Qualitative results.** We show the predicted answer and **bbox** with highest attention for the variants of ScanQA [AMK22] models. We anticipate the **bbox** to indicate the object that situation description  $s^{\text{txt}}$  or question  $q$  refers to. We observe that better situation understanding (via comprehension on  $s^{\text{txt}}$  or auxiliary tasks) could result in more reasonable attention over objects, which positively correlates to more robust answer prediction.

to the right chair behind , while ScanQA focuses on a wrong instance. These results confirm our findings in Section 3.4.2 about the critical role of situation understanding. We also provide some failure modes in *appendix*.

### 3.5 Related Work

**Embodied AI.** The study of embodied AI [Bro90] emerges from the hypothesis of “*ongoing physical interaction with the environment as the primary source of constraint on the design of intelligent systems*”. To this end, researchers have proposed a myriad of AI tasks to investigate whether intelligence will emerge by acting in virtual or photo-realistic environments. Notable tasks including robotic navigation [DDG18, AWT18, SKM19, CSM19, WKM19a, QWA20, DVH22] and vision-based manipulation [KMH17, PRB18, XLZ19, STG20, SYC20, SMF22]. These tasks are made more challenging as instructions or natural-dialogues are further employed as conditions. Sophisticated models have also been developed to tackle these challenges. Earlier endeavors usually comprise multi-modal fusion [TF96, PSD18] and are trained from scratch [WXW18, FHC18, WHC19], while recent efforts would employ pre-trained models [PSS21, HWQ21, SGT21]. However, the agents still suffer from poor generalization to novel and more complex testing tasks [STG20] compared to results on training tasks. More detailed inspection has still yet to be conducted and it also motivates our SQA3D dataset, which investigates one crucial capability that the current embodied agents might need to improve: **embodied scene understanding**.

**Grounded 3D understanding.** Visual grounding has been viewed as a key to connecting human knowledge, which is presumably encoded in our language, to the visual world, so as enable the intelligent agent to better understand and act in the real environment. It is natural to extend this ability to 3D data as it offers more immersive representations of the world. Earlier work has examined word-level grounding with detection and segmentation tasks on 3D data [GAM13, SX14, DCS17, CDF17]. Recent research starts to cover sentence-level grounding with complex semantics [CCN20, AAX20, CGN21]. More recently, new benchmarks introduce complex visual reasoning to 3D data [AMK22, YCH21, YYD21]. However, these tasks mostly assume a passive, third-person’s perspective, while our SQA3D requires problem-solving with an egocentric viewpoint. This introduces both challenges and

chances for tasks that need a first-person’s view, *e.g.* embodied AI.

**Multi-modal question answering.** Building generalist question answering (QA) systems has long been a goal for AI. Along with the progress in multi-modal machine learning, VQA [AAL15b, ZGB16] pioneers the efforts of facilitating the development of more human-like, multi-modal QA systems. It has been extended with more types of knowledge, *e.g.* common sense [ZBF19] and factual knowledge [MRF19]. Recent research has also introduced QA tasks on video [LYB18, JCH20, JLZ22, GKA21, WYC21, DDC22], and 3D data [YCH21, AMK22, YYD21]. We propose the SQA3D benchmark also in hope of facilitating multi-modal QA systems with the ability of embodied scene understanding. Notably, models for SQA3D could choose their input from a 3D scan, egocentric video, or BEV picture, which makes our dataset compatible with a wide spectrum of existing QA systems.

### 3.6 Conclusion

We’ve introduced SQA3D, a benchmark that investigates the capability of embodied scene understanding by combining the best of situation understanding and situated reasoning. We carefully curate our dataset to include diverse situations and interesting questions while preserving the relatively large scale (20.4k situation descriptions and 33.4k questions). Our questions probe a wide spectrum of knowledge and reasoning abilities of embodied agents, notably navigation, common sense, and multi-hop reasoning. We examine many state-of-the-art multi-modal reasoning systems but the gap between the best ML model and human performances so far is still significant. Our findings suggest the crucial role of proper 3D representations and better situation understanding. With SQA3D, we hope of fostering research efforts in developing better embodied scene understanding methods and ultimately facilitate the emergence of more intelligent embodied agents.

## 3.A Data collection

### 3.A.1 Data collection Web UI

We present the Web UI of our data collection in Figure 3.8 (Stage I), Figure 3.9 (Stage II) and Figure 3.11 (Stage III) respectively. We developed our UI based on [CCN20]. These UIs share some common components: a 3D scene viewer, where the user can drag, rotate, and zoom in/out the scene; clickable objects/tags, where users might click on either the object mesh directly or the tag on the sidebar to highlight it in the scene; and an instruction set that guide the user through the task. Users may also switch between a full scene or object mesh only to focus on the tasks. The users are also required to submit multiple responses with the same scene.

Notably, we create detailed tutorials for each stage (not shown in the UI) with examples and animated demonstrations. We found tutorials and instruction sets with clear criteria on **rejection** and **bonus** (*e.g.* Figure 3.10) helpful with high-quality data. Finally, all the testers need to pass a test before the qualification for our task is granted.

### 3.A.2 Data post-processing

There are two major data post-processing steps in SQA3D: **cleaning** and **balancing**. For cleaning, we primarily focus on grammatical correction. We adopt both rule-based cleaning and an ML-based tool called GECToR [OAC20] in our grammatical correction pipeline. We adjust the correction threshold based on human judgment over the corrected data samples.

In the balancing step, our goal is to reduce the question-answer bias in the dataset. Therefore we follow the practice in [AAL15b, MRF19] and re-sample the questions based on their prefixes and answer type, in hope of a more balanced answer distribution. We provide answer distribution before and after balancing in Section 3.B.1.



### 3.A.3 More MTurk details

We provide the detailed MTurk job settings below:

**Region.** We enable access to our tasks in the following countries/regions:

US, DE, GB, AU, CA, SG, NZ, NO, SE, FI, DK, IE
--

**Approval rate & Number of approved jobs.** The testers are required to have at least a 95% approval rate and have completed more than 1000 tasks. However, we relax this requirement to a 90% approval rate for Stage III as it is simpler than the other annotation tasks.

**Reward.** The participants will be rewarded \$0.5 for each task in Stage I and II, and \$0.2 for the QA tasks in Stage III, with a possibility of a bonus depending on the overall quality. We actively monitor the response quality and send bonuses/rejections daily. Note that we collected 5 responses for each task in all three stages.

**Task lifetime.** We set the lifetime as 10 days for tasks in Stage I and 20 days for those in Stage II and III. However, we found most of the tasks can be completed in less than 7 days.

## 3.B Dataset details

### 3.B.1 More statistics

We provide the histogram of the answer distribution before & after balancing in Figure 3.12 and Figure 3.13, respectively. It can be seen that we manage to ensure there is no single answer that dominates any type of question (categorized by their prefixes). However, we do acknowledge that prefix-based balancing might still not be sufficient since models could also learn to use the n-grams pattern. A more effective avenue is collecting more questions with less-frequent answers, which we leave as future work.



Figure 3.8: Dataset collection Web UI for Stage I.

In Figure 3.14a and Figure 3.14b, we show the histogram of the length of situation description  $s^{\text{txt}}$  and question  $q$ . Overall most of the descriptions and questions are middle-length sentences (10-20 words).

### 3.B.2 Details on egocentric video and BEV image

For egocentric videos, we uniformly downsample the frames of the original ScanNet [DCS17] video by using the first frame of every 20 frames. Afterward, we resize all the frames to



Figure 3.9: Dataset collection Web UI for Stage II.

$224 \times 224$  to create the video used for training ClipBERT[LLZ21]. Blender is used for rendering all BEV images. We compute the radius of the bounding sphere of the scene and put the camera at the top of the scene with a distance of 7 times the radius to the center of the bounding sphere. Images of size  $1920 \times 1080$  are rendered for clarity while the input to the MCAN[YYC19a] model is the resized version of the images to  $224 \times 224$ .

**Assignment with questions that can be answered without considering the context will be rejected.**

-- How many chairs are there in the room? ✘

-- Is the amount of monitors on the desk I'm facing at odd or even? ✔

**Assignment with questions that can be answered by merely reading the description (no need to look at the 3D scene) will be rejected.**

**You may ask at most 3 "simple" questions ((you need to ask 5 questions in total) as we encourage creative questions; otherwise your assignment will be rejected. Questions below are viewed as "simple":**

- Questions about simple object category/property or counting, ex. Is there a chair on my 6 o'clock direction?; What is the color of the table on my right?; How many chairs are there behind me?

- Questions that repeat the same pattern

- Questions that can be answered with "Yes" or "No" (Note: some creative questions can also have answer "yes" or "no", but you still need to control the overall amount of questions with answer "yes" or "no" in your assignment)

Figure 3.10: Additional instruction set to the AMT participants in Stage II.

## 3.C Model details

### 3.C.1 Input pipeline

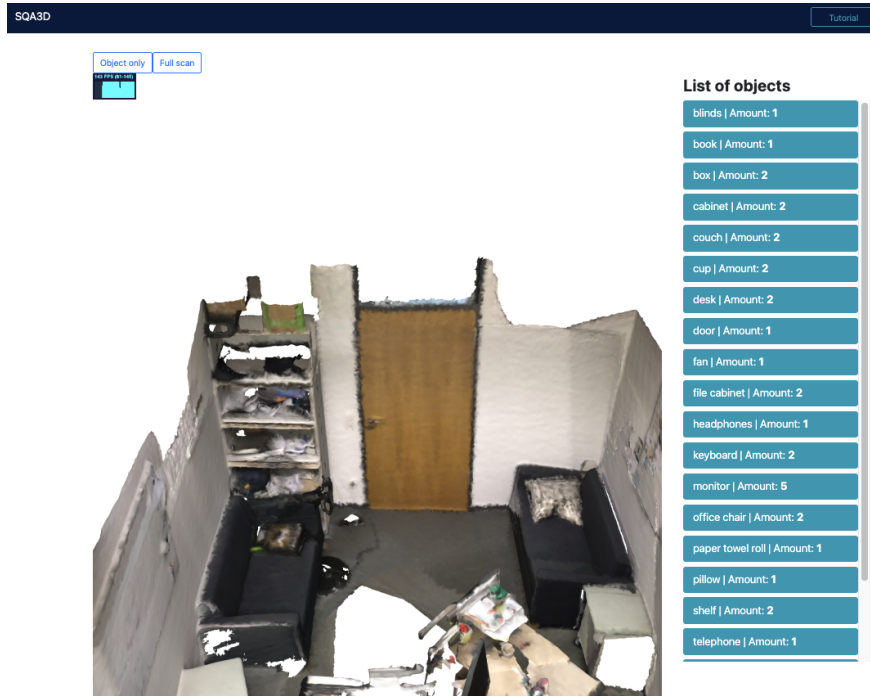
We follow the input pipeline in ScanQA[AMK22] without further modification. As for MCAN, we only transform the images to fit the ImageNet-pretrained encoder. In ClipBERT, we randomly sample 8 clips with each clip consisting of 2 frames of the video to feed into the model as the scene representation. Note that each frame is resized to  $1000 \times 1000$  following the practice of original ClipBERT[LLZ21].

### 3.C.2 Hyper-parameters

We provide the hyper-parameters of the considered models in Table 5.3.

### 3.C.3 Additional details on zero-shot models

We uniformly sample 30 sentences from our 3D caption sources for both models. When testing with the Unified QA<sub>Large</sub> model, we employ a simple greedy sampling method and the following prompt:



**Step 1** Explore the 3D scene above. Click 'Tutorial' for details on how to interact with the scene and objects.

**Step 2** The following describes your current **status**:

I'm sitting on the office chair facing the desks with the one window in my 1 o'clock direction and one in my 3 o'clock direction.

Based on those (we name them "**context**") above, answer the following question, then click 'Submit'.

How many monitors on my desk?

**Answer** --- It is expected to be a **simple word**.

three

Submit

Figure 3.11: Dataset collection Web UI for Stage III.

$\{s^{\text{txt}}\}$

Q:  $\{q\}$

A:

, where  $\{s^{\text{txt}}\}$  and  $\{q\}$  are replaced by the situation description and question. For GPT-3, we use the `text-davinci-002` variant and the following prompt:

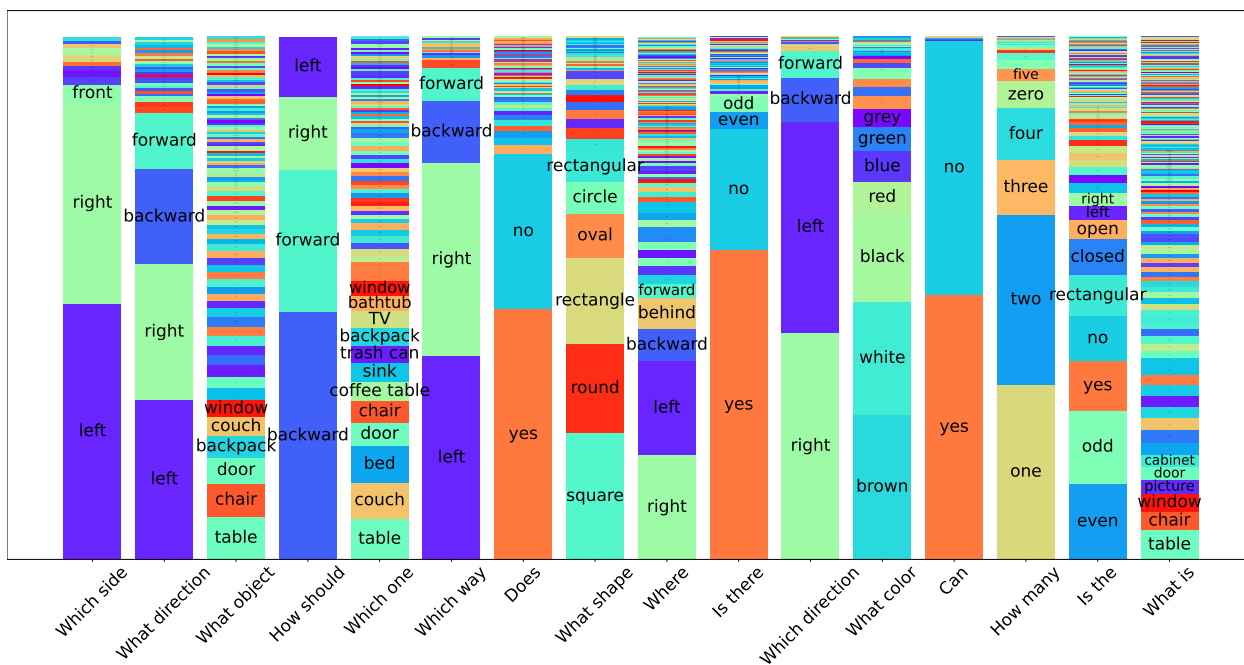


Figure 3.12: Answer distribution (organized by question prefixes) before balancing.

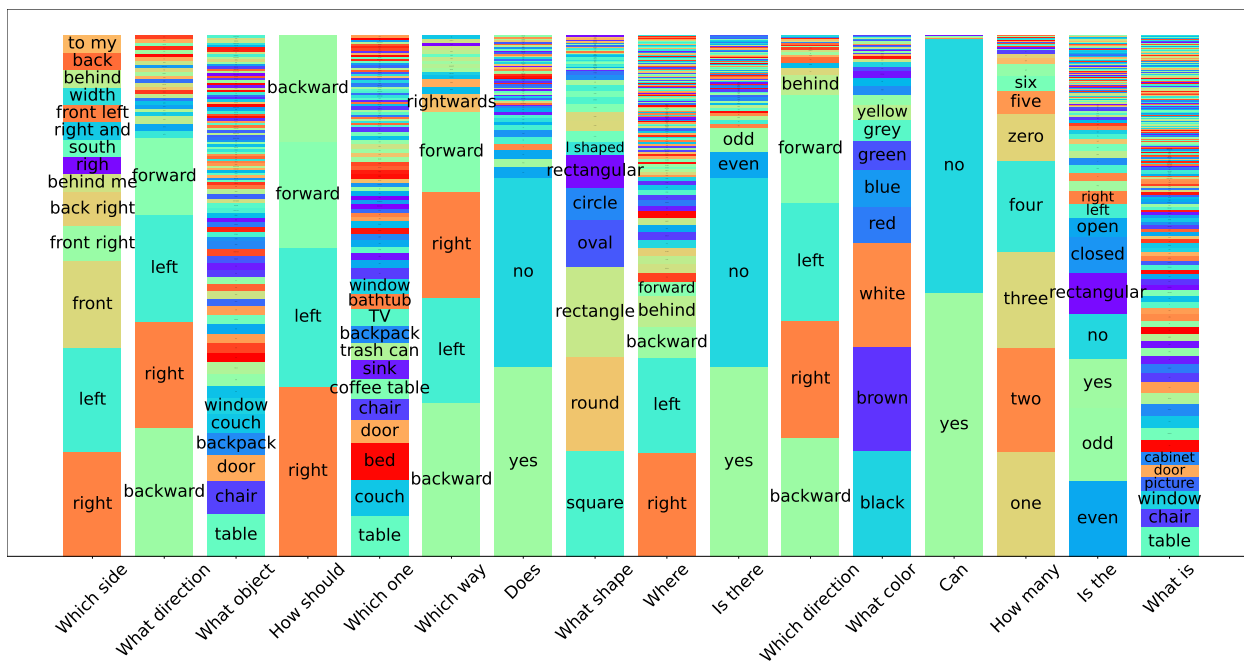
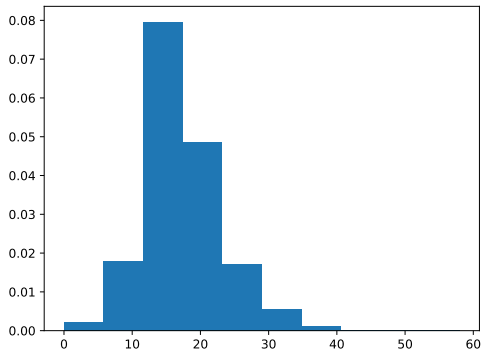
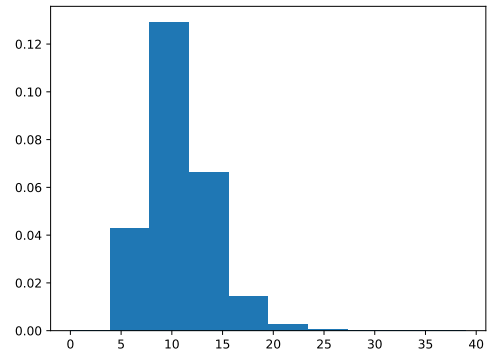


Figure 3.13: Answer distribution (organized by question prefixes) after balancing.



(a) Histogram of description  $s^{\text{txt}}$  length.



(b) Histogram of question  $q$  length.

Context: There is a book on the desk. A laptop with a green cover is to the left of the book.

Q: I'm working by the desk. What is on the desk beside the book?

A: laptop

Context:  $\{s^{\text{txt}}\}$

Q:  $\{q\}$

A:

, where we use a 1-shot example to demonstrate the format of our task. Interestingly, we found only GPT-3 would benefit from few-shot examples.

### 3.C.4 Additional details on SCANQA/MCAN/CLIPBERT

**ScanQA** [AMK22]. We slightly modify the original ScanQA code base to make it fit our task better. The original reference branch is discarded and the supervision signal for the language classification branch is changed to make use of it as a regression branch. More specific details can be found below.

- The original data loader only outputs the question as a whole (meaning that the situation is concatenated before the question), while our version split the two sentences.
- The original model takes language as 1 input, while we feed situation and question separately into the model.
- The original model uses 1 self-attention block and 1 cross-attention block for the fusion of language and visual features, while our version uses 2 self-attention blocks and 2 cross-attention blocks to treat situations and questions separately.
- The original model uses additive operation to fuse language & visual features, while our version uses concatenation for fusion.
- To conduct the ablation experiment of blind test, we simply discard the output feature of VoteNet and only feed the situation feature and question feature into the QA head.
- To conduct the ablation experiment of w/o  $s^{\text{txt}}$ , we replace situation with several  $\langle unk \rangle$  tokens to make a fair comparison.
- To add an auxiliary task into training, we change the supervision of the language classification head from Cross Entropy to MSE Loss to make it a regression head.

**MCAN** [YYC19a]. We use the code base from ReViT [MNY22] since its implementation of MCAN could take raw images as input while the original one cannot. The default training setting is kept except for learning rate decay. We cancel it to make a fair comparison with the other baselines. We concatenate the situation before the question to make them as a whole and use this new sentence as the question that MCAN requires.

**ClipBERT** [LLZ21]. We use the official repository of ClipBERT and follow the instruction to transform our data into the format ClipBERT takes. The configuration file for MSR-VTT QA [XMY16] is used for generality as we find all the configuration files to be almost identical. The evaluated question types are changed since our focus is different from



MSR-VTT. We turn off mixed precision training as we observe instability when using it. We concatenate the situation before the question to make them as a whole and use this new sentence as the question that ClipBERT requires.

### **3.D Additional empirical results**

We provide additional qualitative results and failure modes in Figure 3.15 and Figure 3.16.

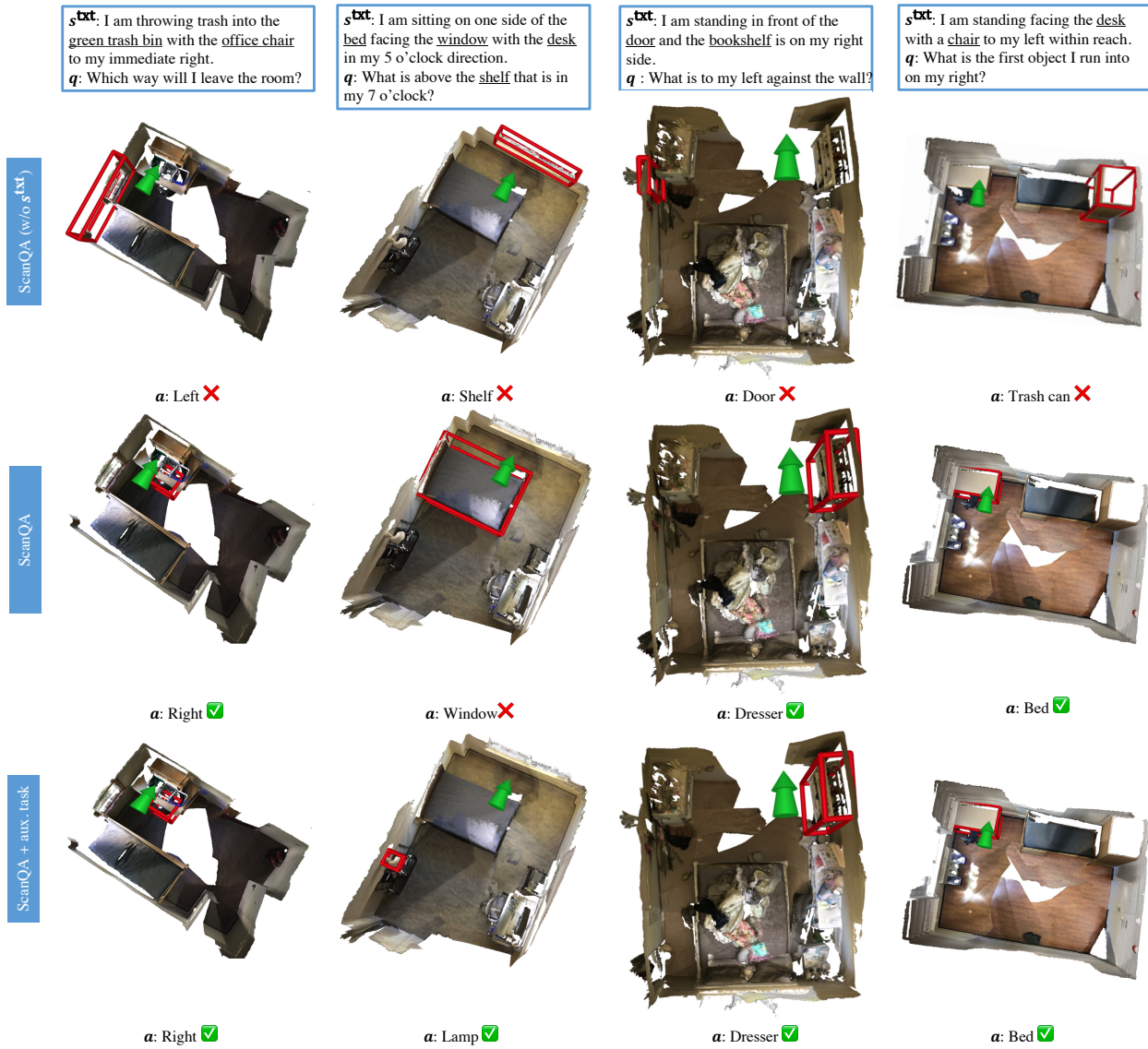


Figure 3.15: Additional qualitative results.

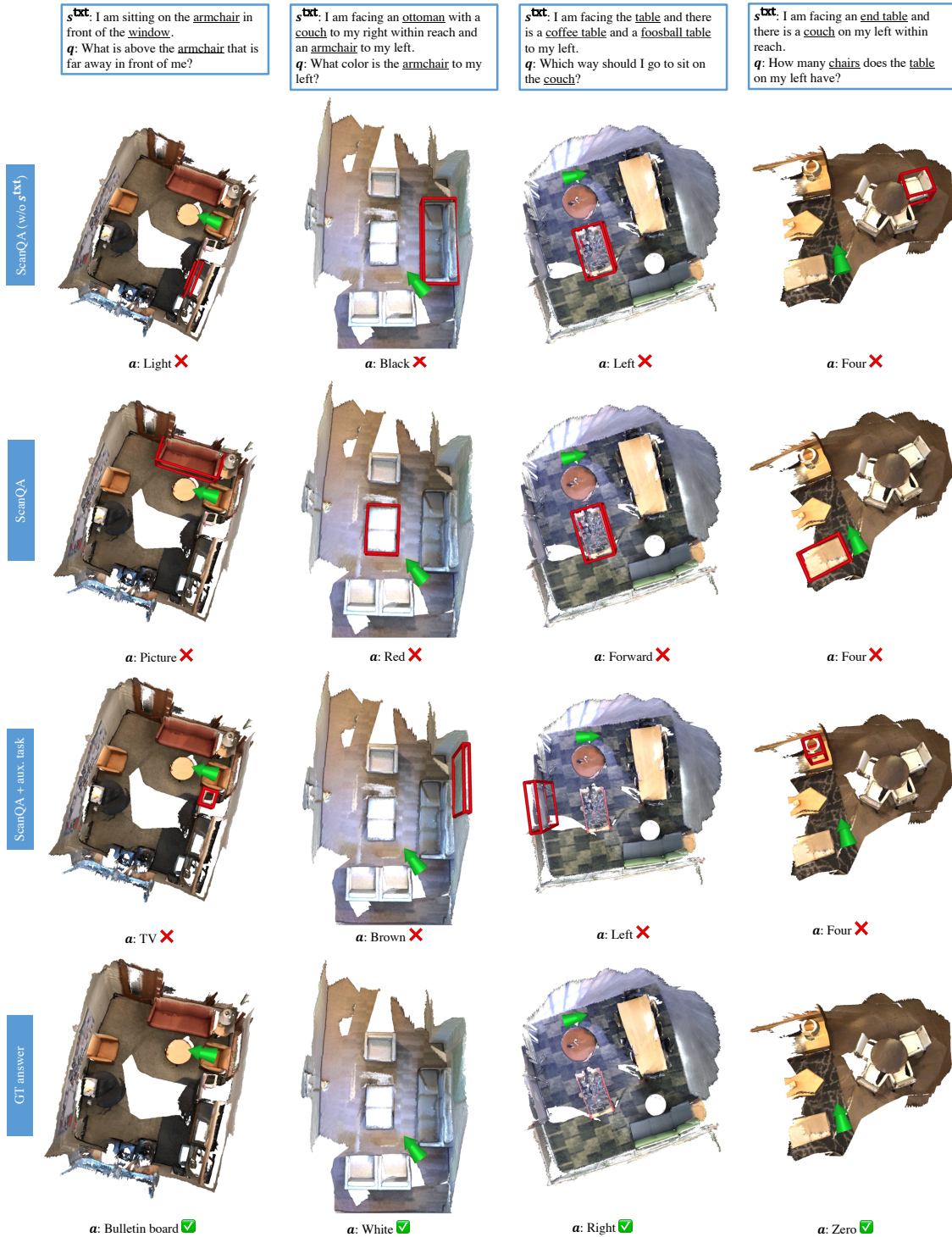


Figure 3.16: **Failure mode.** Models are likely to predict the wrong answers when they do not attend to relevant objects.

Part II

**A Unified Framework for Human-like  
Visual and Relational Reasoning in  
the Real World**

## CHAPTER 4

# Unsupervised Object-Centric Learning using Deep Region Competition

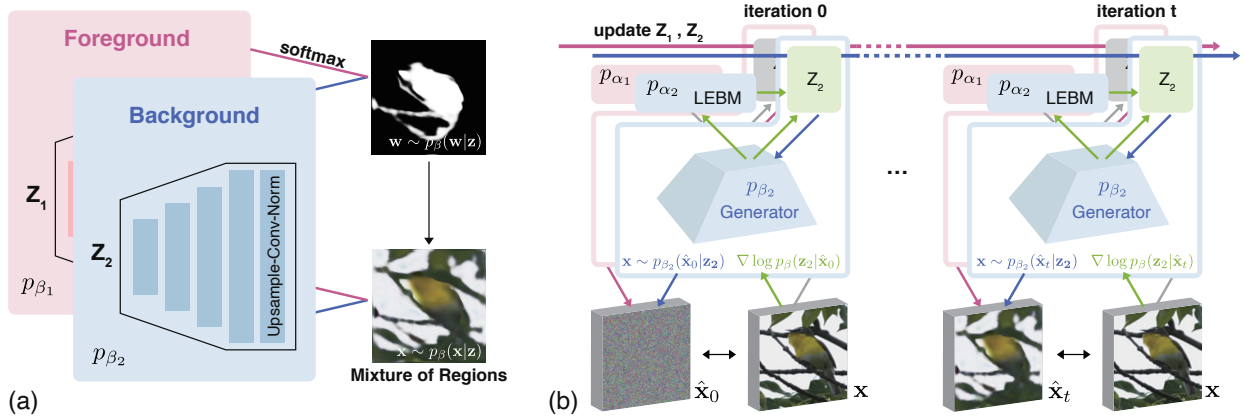


Figure 4.1: **Overview of DRC.** (a) The model generates foreground and background regions using sampled latent variables  $\mathbf{z} = \{z_1, z_2\}$ .  $p_{\beta_k}$ ,  $k = 1, 2$  represents the generator for each region. Of note, the pixel re-assignment function is absorbed in the background generator; see section 4.3.2 for details. (b) DRC samples the latent variables  $\mathbf{z}$  in an iterative manner. Let  $\mathbf{x}$  denote the observed image; we use  $\hat{\mathbf{x}}_t$ ,  $t = 0, 1, \dots$  to represent the image generated by  $p_{\beta}(\mathbf{x}|\mathbf{z})$  at the  $t$ -th sampling step. DRC has a two-step workflow for learning unsupervised foreground extractors that resembles the E- and M-step in the classic Expectation-Maximization (EM) algorithm. In the E-step, it employs gradient-based MCMC sampling to infer the latent variables  $\mathbf{z}$  as shown in (b). Of note, only the latent variables  $\mathbf{z}$  are updated in this step. In the M-step, the sampled latent variables  $\mathbf{z}$  are fed into the model for image generation as shown in (a), where the generators are updated to minimize the reconstruction error.

## 4.1 Introduction

Foreground extraction, being a special case of generic image segmentation, aims for a binary partition of the given image with specific semantic meaning, *i.e.*, a foreground that typically contains identifiable objects and the possibly less structural remaining regions as the background. There is a rich literature on explicitly modeling and representing a given image as foreground and background (or more general visual regions), such that a generic inference algorithm can produce plausible segmentations ideally for any images without or with little supervision [ZY96, SM00, TZ02, BJ01, RKB04, CMH14, JWY13, ZLW14]. However, such methods essentially rely on low-level visual features (*e.g.*, edges, color, and texture), and some further require human intervention at initialization [BJ01, RKB04], which largely limits their practical performance on modern datasets of complex natural images with rich semantic meanings [LMB14, EVW10]. These datasets typically come with fine-grained semantic annotations, exploited by supervised methods that learn representation and inference algorithm as one monolithic network [ZSQ17, LSD15, BKC17, CPK17, RFB15, HGD17]. Despite the success of densely supervised learning, the unsupervised counterpart is still favored due to its resemblance to how humans perceive the world [CTY06, SK01].

Attempting to combine unsupervised or weakly supervised learning with modern neural networks, three lines of work surge recently for foreground extraction: (1) deep networks as feature extractors for canonical segmentation algorithms, (2) GAN-based foreground-background disentanglement, and (3) compositional latent variable models with slot-based object modeling. Despite great successes of these methods, the challenge of unsupervised foreground extraction remains largely open.

Specifically, the first line of work trains designated deep feature extractors for canonical segmentation algorithms or metric networks as learned partitioning criteria [XK17, Kan18, JHV19]. These methods (*e.g.*, W-Net [XK17]) define foreground objects' properties using learned features or criteria and are thus generally bottle-necked by the selected post-

processing segmentation algorithm [AMF10, ASS12]. As a branch of pioneering work that moves beyond these limitations, [YLS19, YLS21] have recently proposed a general contextual information separation principle and an efficient adversarial learning method that is generally applicable to unsupervised segmentation, separation and detection. GAN-based models [GPM14, YKB17, CAD19, OSL18, SOL19, BW20] capture the foreground objectness with oversimplified assumptions or require additional supervision to achieve foreground-background disentanglement. For example, the segmentation model in ReDO [CAD19] is trained by redrawing detected objects, which potentially limits its application to datasets with diverse object shapes. OneGAN [BW20] and its predecessors [OSL18, SOL19], though producing impressive results on foreground extraction, require a set of background images without foreground objects as additional inputs. Lastly, compositional latent variable models [GRB16, EHW16b, GSS17, SCG18, BMW19, GKK19a, LWU20a, EKJ20, LWP20a] include the background as a “virtual object” and induce the independence of object representations using an identical generator for all object slots. Although these methods exhibit strong performance on synthetic multi-object datasets with simple backgrounds and foreground shapes, they may fail on complex real-world data or even synthetic datasets with more challenging backgrounds [GKK19a, LWU20a]. In addition, few unsupervised learning methods have provided explicit identification of foreground objects and background regions. While they can generate valid segmentation masks, most of these methods do not specify which output corresponds to the foreground objects. These deficiencies necessitate rethinking the problem of unsupervised foreground extraction. We propose to confront the challenges in formulating (1) a generic inductive bias for modeling foreground and background regions that can be baked into neural generators, and (2) an effective inference algorithm based on a principled criterion for foreground-background partition.

Inspired by Region Competition [ZY96], a seminal approach that combines optimization-based inference [KWT88, Coh91, AB94] and probabilistic visual modeling [ZWM98, GZW07] by minimizing a generalized Bayes criterion [Lec89], we propose to solve the foreground ex-



traction problem by reconciling energy-based prior [PHN20] with generative image modeling in the form of Mixture of Experts (MoE) [JJN91, JJ94]. To generically describe background regions, we further introduce the learned pixel re-assignment as the essential inductive bias to capture their regularities. Fueled by our modeling, we propose to find the foreground-background partition through Expectation-Maximization (EM). Our algorithm effectively exploits the interaction between the mixture components during the partitioning process, echoing the intuition described in Region Competition [ZY96]. We therefore coin our method Deep Region Competition (DRC). We summarize our **contributions** as follows:

1. We provide probabilistic foreground-background modeling by reconciling energy-based prior with generative image modeling in the form of MoE. With this modeling, the foreground-background partition can be naturally produced through EM. We further introduce an inductive bias, *pixel re-assignment*, to facilitate foreground-background disentanglement.
2. In experiments, we demonstrate that DRC exhibits more competitive performances on complex real-world data and challenging multi-object scenes compared with prior methods. Furthermore, we empirically show that using learned pixel re-assignment as the inductive bias helps to provide explicit identification for foreground and background regions.
3. We find that DRC can potentially generalize to novel foreground objects even from categories unseen during training, which may provide some inspiration for the study of out-of-distribution (OOD) generalization in more general unsupervised disentanglement.

## 4.2 Related Work

A typical line of methods frames unsupervised or weakly supervised foreground segmentation within a generative modeling context. Several methods build upon generative adversarial



networks (GAN) [GPM14] to perform foreground segmentation. LR-GAN [YKB17] learns to generate background regions and foreground objects separately and recursively, which simultaneously produces the foreground objects mask. ReDO (ReDrawing of Objects) [CAD19] proposes a GAN-based object segmentation model, based on the assumption that replacing the foreground object in the image with a generated one does not alter the distribution of the training data, given that the foreground object is correctly discovered. Similarly, SEIGAN [OSL18] learns to extract foreground objects by recombining the foreground objects with the generated background regions. FineGAN [SOL19] hierarchically generates images (*i.e.*, first specifying the object shape and then the object texture) to disentangle the background and foreground object. [BW20] further hypothesize that a method solving an ensemble of unsupervised tasks altogether improves the model performance compared with the one that solves each individually. Therefore, they train a complex GAN-based model (OneGAN) to solve several tasks simultaneously, including foreground segmentation. Although LR-GAN and FineGAN do produce masks as part of their generative process, they cannot segment a given image. Despite SEIGAN and OneGAN achieving decent performance on foreground-background segmentation, these methods require a set of clean background images as additional inputs for weak supervision. ReDO captures the foreground objectness with possibly oversimplified assumptions, limiting its application to datasets with diverse object shapes.

On another front, compositional generative scene models [GRB16, EHW16b, GSS17, SCG18, BMW19, GKK19a, LWU20a, EKJ20, LWP20a], sharing the idea of scene decomposition stemming from DRAW [GDG15], learn to represent foreground objects and background regions in terms of a collection of latent variables with the same representational format. These methods typically exploit the spatial mixture model for generative modeling. Specifically, IODINE [GKK19a] proposes a slot-based object representation method and models the latent space using iterative amortized inference [MYM18]. Slot-Attention [LWU20a], as a step forward, effectively incorporates the attention mechanism into the slot-based object rep-

resentation for flexible foreground object binding. Both methods use fully shared parameters among individual mixture components to entail permutation invariance of the learned multi-object representation. Alternative models such as MONet [BMW19] and GENESIS [EKJ20] use multiple encode-decode steps for scene decomposition and foreground object extraction. Although these methods exhibit strong performance on synthetic multi-object datasets with simple background and foreground shapes, they may fail when dealing with complex real-world data or even synthetic datasets with more challenging background [GKK19a, LWU20a].

More closely related to the classical methods, another line of work focuses on utilizing image features extracted by deep neural networks or designing energy functions based on data-driven methods to define the desired property of foreground objects. [PDS18] and [SHK12] obtain impressive results when depth images are accessible in addition to conventional RGB images, while such methods are not directly applicable for data with RGB images alone. W-Net [XK17] extracts image features via a deep auto-encoder jointly trained by minimizing reconstruction error and normalized cut. The learned features are further processed by CRF smoothing to perform hierarchical segmentation. [Kan18] proposes to employ a neural network as part of the partitioning criterion (inspired by [UVL20]) to minimize the chosen intra-region pixel distance for segmentation directly. [JHV19] propose to use Invariant Information Clustering as the objective for segmentation, where the network is trained to be part of the learned distance. As an interesting extension, one may also consider adapting methods that automatically discover object structures [LBM19] to foreground extraction. Though being pioneering work in image segmentation, the aforementioned methods are generally bottle-necked by the selected post-processing segmentation algorithm or require extra transformations to produce meaningful foreground segmentation masks. [YLS19, YLS21] in their seminal work propose an information-theoretical principle and adversarial contextual model for unsupervised segmentation and detection by partitioning images into maximally independent sets, with the objective of minimizing the predictability of one set by the other sets. Additional efforts have also been devoted to weakly supervised foreground segmentation us-

ing image classification labels [PCM15, PKD15, HWW18], bounding boxes [DHS15, KBH17], or saliency maps [OBK17, ZZL19, VMB21].

## 4.3 Methodology

Foreground extraction performs a binary partition for the image  $\mathbf{I}$  to extract the foreground region. Without explicit supervision, we propose to use learned pixel re-assignment as a generic inductive bias for background modeling, upon which we derive an EM-like partitioning algorithm. Compared with prior methods, our algorithm can handle images with more complex foreground shapes and background patterns, while providing explicit identification of foreground and background regions.

### 4.3.1 Preliminaries

Adopting the language of EM algorithm, we assume that for the observed sample  $\mathbf{x} \in \mathbb{R}^D$ , there exists  $\mathbf{z} \in \mathbb{R}^d$  as its latent variables. The complete-data distribution is

$$p_{\theta}(\mathbf{z}, \mathbf{x}) = p_{\alpha}(\mathbf{z})p_{\beta}(\mathbf{x}|\mathbf{z}), \quad (4.1)$$

where  $p_{\alpha}(\mathbf{z})$  is the prior model with parameters  $\alpha$ ,  $p_{\beta}(\mathbf{x}|\mathbf{z})$  is the top-down generative model with parameters  $\beta$ , and  $\theta = (\alpha, \beta)$ .

The prior model  $p_{\alpha}(\mathbf{z})$  can be formulated as an energy-based model, which we refer to as the Latent-space Energy-Based Model (LEBM) [PHN20] throughout the paper:

$$p_{\alpha}(\mathbf{z}) = \frac{1}{Z_{\alpha}} \exp(f_{\alpha}(\mathbf{z})) p_0(\mathbf{z}), \quad (4.2)$$

where  $f_{\alpha}(\mathbf{z})$  can be parameterized by a neural network,  $Z_{\alpha}$  is the partition function, and  $p_0(\mathbf{z})$  is a reference distribution, assumed to be isotropic Gaussian prior commonly used for the generative model. The prior model in Eq. (4.2) can be interpreted as an energy-based correction or exponential tilting of the original prior distribution  $p_0$ .

The LEBM can be learned by Maximum Likelihood Estimation (MLE). Given a training sample  $\mathbf{x}$ , the learning gradient for  $\alpha$  is derived as shown by [PHN20],

$$\delta_\alpha(\mathbf{x}) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})} [\nabla_\alpha f_\alpha(\mathbf{z})] - \mathbb{E}_{p_\alpha(\mathbf{z})} [\nabla_\alpha f_\alpha(\mathbf{z})]. \quad (4.3)$$

In practice, the above expectations can be approximated by Monte-Carlo average, which requires sampling from  $p_\theta(\mathbf{z}|\mathbf{x})$  and  $p_\alpha(\mathbf{z})$ . This step can be done with stochastic gradient-based methods, such as Langevin dynamics [WT11] or Hamiltonian Monte Carlo [BGJ11].

An extension to LEBM is to further couple the vector representation  $\mathbf{z}$  with a symbolic representation  $\mathbf{y}$  [PW21]. Formally,  $\mathbf{y}$  is a  $K$ -dimensional one-hot vector, where  $K$  is the number of possible  $\mathbf{z}$  categories. Such symbol-vector duality can provide extra entries for auxiliary supervision; we will detail it in section 4.3.4.

### 4.3.2 Generative Image Modeling

#### Mixture of Experts (MoE) for Image Generation.

Inspired by the regional homogeneity assumption proposed by [ZY96], we use separate priors and generative models for foreground and background regions, indexed as  $\alpha_k$  and  $\beta_k$ ,  $k = 1, 2$ , respectively; see figure 4.1. This design leads to the form of MoE [JJN91, JJ94] for image modeling, as shown below.

Let us start by considering only the  $i$ -th pixel of the observed image  $\mathbf{x}$ , denoted as  $\mathbf{x}_i$ . We use a binary one-hot random variable  $\mathbf{w}_i$  to indicate whether the  $i$ -th pixel belongs to the foreground region. Formally, we have  $\mathbf{w}_i = [w_{i1}, w_{i2}]$ ,  $w_{ik} \in \{0, 1\}$  and  $\sum_{k=1}^2 w_{ik} = 1$ . Let  $w_{i1} = 1$  indicate that the  $i$ -th pixel  $\mathbf{x}_i$  belongs to the foreground, and  $w_{i2} = 1$  indicate the opposite.

We assume that the distribution of  $\mathbf{w}_i$  is prior-dependent. Specifically, the mixture parameter  $\pi_{ik}$ ,  $k = 1, 2$ , is defined as the output of a gating function  $\pi_{ik} = p_\beta(w_{ik} = 1|\mathbf{z}) = \text{Softmax}(l_{ik})$ ;  $l_{ik} = h_{\beta_k}(\mathbf{z}_k)$ ,  $k = 1, 2$  are the logit scores given by the foreground and

background generative models respectively;  $\beta = \{\beta_1, \beta_2\}$ ,  $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2\}$ . Taken together, the joint distribution of  $\mathbf{w}_i$  is

$$p_\beta(\mathbf{w}_i|\mathbf{z}) = \prod_{k=1}^2 \pi_{ik}^{w_{ik}}. \quad (4.4)$$

The learned distribution of foreground and background contents are

$$p_\beta(\mathbf{x}_i|w_{ik} = 1, \mathbf{z}_k) = p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k) \sim \mathbf{N}(g_{\beta_k}(\mathbf{z}_k), \sigma^2\mathbf{I}), \quad k = 1, 2 \quad (4.5)$$

where we assume that the generative model for region content,  $p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)$ ,  $k = 1, 2$ , follows a Gaussian distribution parameterized by the generator network  $g_{\beta_k}$ . As in VAE,  $\sigma$  takes an assumed value. We follow the common practice and use a shared generator for parameterizing  $\pi_{ik}$  and  $p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)$ . We use separate branches only at the output layer to generate logits and contents.

Generating  $\mathbf{x}_i$  based on  $\mathbf{w}_i$ 's distribution involves two steps: (1) sample  $\mathbf{w}_i$  from the distribution  $p_\beta(\mathbf{w}_i|\mathbf{z})$ , and (2) choose either the foreground model (*i.e.*,  $p_{\beta_1}(\mathbf{x}_i|\mathbf{z}_1)$ ) or the background model (*i.e.*,  $p_{\beta_2}(\mathbf{x}_i|\mathbf{z}_2)$ ) to generate  $\mathbf{x}_i$  based on the sampled  $\mathbf{w}_i$ . As such, this distribution of  $\mathbf{x}_i$  is a MoE,

$$p_\beta(\mathbf{x}_i|\mathbf{z}) = \sum_{k=1}^2 p_\beta(w_{ik} = 1|\mathbf{z})p_\beta(\mathbf{x}_i|w_{ik} = 1, \mathbf{z}_k) = \sum_{k=1}^2 \pi_{ik}p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k), \quad (4.6)$$

wherein the posterior responsibility of  $w_{ik}$  is

$$\gamma_{ik} = p(w_{ik} = 1|\mathbf{x}_i, \mathbf{z}) = \frac{\pi_{ik}p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)}{\sum_{m=1}^2 \pi_{im}p_{\beta_m}(\mathbf{x}_i|\mathbf{z}_m)}, \quad k = 1, 2. \quad (4.7)$$

Using a fully-factorized joint distribution of  $\mathbf{x}$ , we have  $p_\beta(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^D \sum_{k=1}^2 \pi_{ik}p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)$  as the generative modeling of  $\mathbf{x} \in \mathbb{R}^D$ .

### Learning Pixel Re-assignment for Background Modeling.

We use pixel re-assignment in the background generative model as the essential inductive bias for modeling the background region. This is partially inspired by the concepts of “texture” and “texton” by Julez [GZW07, Jul81], where the textural part of an image may

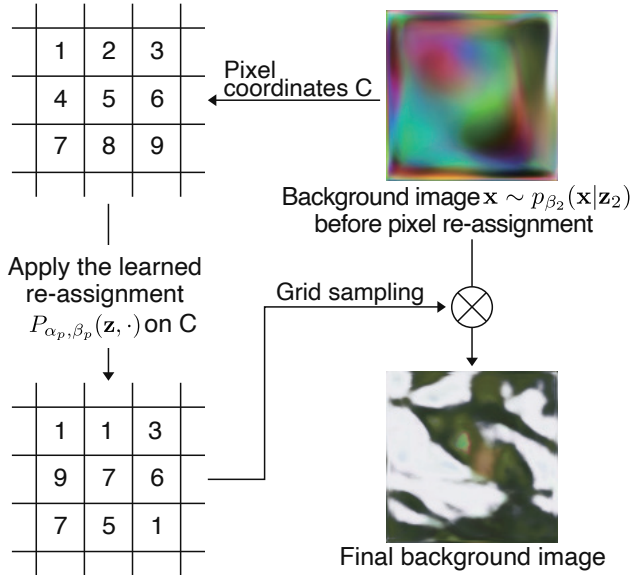


Figure 4.2: **Pixel re-assignment.** The output of  $\beta_p$  can be viewed as a learned re-assignment of the original background pixels that follows the mapped grid  $P_{\alpha_p, \beta_p}(z, C)$ . Note that the re-assignment function  $P_{\alpha_p, \beta_p}(z, \cdot)$  might not be injective. The final background image is generated via grid sampling.

contain fewer structural elements in preattentive vision, which coincides with our intuitive observation of the background regions.

We use a separate pair of energy-based prior model  $\alpha_{\text{pix}}$  and generative model  $\beta_{\text{pix}}$  to learn the re-assignment. For simplicity, we absorb  $\alpha_{\text{pix}}$  and  $\beta_{\text{pix}}$  in the models for background modeling, *i.e.*,  $\alpha_2$  and  $\beta_2$ , respectively. In practice, the re-assignment follows the output of  $\beta_{\text{pix}}$ , a shuffling grid with the same size of the image  $\mathbf{x}$ . Its values indicate the re-assigned pixel coordinates; see figure 4.2. We find that shuffling the background pixels using the learned re-assignment facilitates the model to capture the regularities of the background regions. Specifically, the proposed model with this essential inductive bias learns to constantly give the correct mask assignment, whereas most previous fully unsupervised methods do not provide explicit identification of the foreground and background regions; see discussion in section 4.4.1 for more details.

### 4.3.3 Deep Region Competition: from Generative Modeling to Foreground Extraction

The complete-data distribution from the image modeling is

$$\begin{aligned}
p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) &= p_\beta(\mathbf{x}|\mathbf{w}, \mathbf{z})p_\beta(\mathbf{w}|\mathbf{z})p_\alpha(\mathbf{z}) \\
&= \left( \prod_{i=1}^D \prod_{k=1}^2 p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)^{w_{ik}} \right) \left( \prod_{i=1}^D \prod_{k=1}^2 \pi_{ik}^{w_{ik}} \right) p_\alpha(\mathbf{z}) \\
&= p_\alpha(\mathbf{z}) \prod_{i=1}^D \prod_{k=1}^2 (\pi_{ik} p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k))^{w_{ik}},
\end{aligned} \tag{4.8}$$

where  $p_\alpha(\mathbf{z}) = p_{\alpha_1}(\mathbf{z}_1)p_{\alpha_2}(\mathbf{z}_2)$  is the prior model given by LEBMs.  $\alpha = \{\alpha_1, \alpha_2\}$ , and  $\theta = \{\alpha, \beta\}$ .  $\mathbf{w}$  is the vector of  $(\mathbf{w}_i), i = 1, \dots, D$ , whose joint distribution is assumed to be fully-factorized.

Next, we derive the complete-data log-likelihood as our learning objective:

$$\mathcal{L}(\theta) = \log p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) = \log p_\alpha(\mathbf{z}) + \sum_{i=1}^D \sum_{k=1}^2 w_{ik} (\log \pi_{ik} + \log p_{\beta_k}(\mathbf{x}_i|\mathbf{z}_k)). \tag{4.9}$$

Of note,  $\mathbf{w}$  and  $\mathbf{z}$  are unobserved variables in the modeling, which makes it impossible to learn the model directly through MLE. To calculate the gradients of  $\theta$ , we instead optimize  $\mathbf{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{x}), \mathbf{w} \sim p(\mathbf{w}|\mathbf{x}, \mathbf{z})}[\mathcal{L}(\theta)]$  based on the fact that underlies the EM algorithm:

$$\begin{aligned}
\nabla_\theta \log p_\theta(\mathbf{x}) &= \int_{\mathbf{z}} p_\theta(\mathbf{z}|\mathbf{x}) d\mathbf{z} \int_{\mathbf{w}} p_\theta(\mathbf{w}|\mathbf{z}, \mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{w}) d\mathbf{w} \\
&= \mathbf{E}_{\mathbf{z} \sim p_\theta(\mathbf{z}|\mathbf{x}), \mathbf{w} \sim p_\theta(\mathbf{w}|\mathbf{x}, \mathbf{z})}[\nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{w})].
\end{aligned} \tag{4.10}$$

Therefore, the derived surrogate learning objective becomes

$$\max_{\theta} \mathbf{E}_{\mathbf{z} \sim p_\theta(\mathbf{z}|\mathbf{x})} [\mathcal{J}(\theta)], \text{ s.t. } \forall i, \sum_{k=1}^2 \pi_{ik} = 1, \tag{4.11}$$

$$\mathcal{J}(\theta) = \underbrace{\log p_\alpha(\mathbf{z})}_{\text{objective for LEBM}} + \underbrace{\sum_{i=1}^D \sum_{k=1}^2 \gamma_{ik} \log \pi_{ik}}_{\text{foreground-background partitioning}} + \underbrace{\sum_{i=1}^D \sum_{k=1}^2 \gamma_{ik} \log p_{\theta_k}(\mathbf{x}_i|\mathbf{z}_k)}_{\text{objective for image generation}}, \tag{4.12}$$

where  $\mathcal{J}(\theta) = \mathbf{E}_{\mathbf{w} \sim p_\theta(\mathbf{w}|\mathbf{x}, \mathbf{z})} [\mathcal{L}(\theta)]$  is the conditional expectation of  $\mathbf{w}$ , which can be calculated in closed form; see the supplementary material for additional details.

Eq. (4.11) has an intuitive interpretation. We can decompose the learning objective into three components as in Eq. (4.12). In particular, the second term  $\sum_{i=1}^D \sum_{k=1}^2 \gamma_{ik} \log \pi_{ik}$  has a similar form to the cross-entropy loss commonly used for supervised segmentation task, where the posterior responsibility  $\gamma_{ik}$  serves as the target distribution. It is as if the foreground and background generative models compete with each other to fit the distribution of each pixel  $\mathbf{x}_i$ . If the pixel value at  $\mathbf{x}_i$  fits better to the distribution of foreground,  $p_{\beta_1}(\mathbf{x}_i|\mathbf{z}_1)$ , than to that of background,  $p_{\beta_2}(\mathbf{x}_i|\mathbf{z}_2)$ , the model tends to assign that pixel to the foreground region (see Eq. (4.7)), and vice versa. This mechanism is similar to the process derived in [ZY96], which is the reason why we coin our method Deep Region Competition (DRC).

Prior to our proposal, several methods [ZY96, GKK19a, LWU20a] also employ mixture models and competition among the components to perform unsupervised foreground or image segmentation. The original Region Competition [ZY96] combines several families of image modeling with Bayesian inference but is limited by the expressiveness of the pre-specified probability distributions. More recent methods, including IODINE [GKK19a] and Slot-attention [LWU20a], learn amortized inference networks for latent variables and induce the independence of foreground and background representations using an identical generator. Our method combines the best of the two worlds, reconciling the expressiveness of learned generators with the regularity of generic texture modeling under the framework of LEBM.

To optimize the learning objective in Eq. (4.11), we approximate the expectation by sampling from the prior  $p_\alpha(\mathbf{z})$  and posterior model  $p_\theta(\mathbf{z}|\mathbf{x}) \propto p_\alpha(\mathbf{z})p_\beta(\mathbf{x}|\mathbf{z})$ , followed by calculating the Monte Carlo average. We use Langevin dynamics [WT11] to draw persistent MCMC samples, which iterates

$$\mathbf{z}_{t+1} = \mathbf{z}_t + s \nabla_{\mathbf{z}} \log Q(\mathbf{z}_t) + \sqrt{2s} \epsilon_t, \quad (4.13)$$

where  $t$  is the Langevin dynamics's time step,  $s$  the step size, and  $\epsilon_t$  the Gaussian noise.  $Q(\mathbf{z})$



is the target distribution, being either  $p_\alpha(\mathbf{z})$  or  $p_\theta(\mathbf{z}|\mathbf{x})$ .  $\nabla_{\mathbf{z}} \log Q(\mathbf{z}_t)$  is efficiently computed via automatic differentiation in modern learning libraries [PGM19]. We summarize the above process in Algorithm 1.

---

**Algorithm 1 Learning models of DRC via EM.**

---

**Input:** Learning iterations  $T$ , initial parameters for LEBMs  $\alpha^{(0)} = \{\alpha_1^{(0)}, \alpha_2^{(0)}\}$  and generators  $\beta^{(0)} = \{\beta_1^{(0)}, \beta_2^{(0)}\}$ ,  $\theta^{(0)} = \{\alpha^{(0)}, \beta^{(0)}\}$ , learning rate  $\eta_\alpha$  for LEBMs,  $\eta_\beta$  for foreground and background generators, observed examples  $\{\mathbf{x}^{(i)}\}_{i=1}^N$ , batch size  $M$ , and initial latent variables  $\{\mathbf{z}_-^{(i)} = \{\mathbf{z}_{1-}^{(i)}, \mathbf{z}_{2-}^{(i)}\} \sim p_0(\mathbf{z})\}_{i=1}^N$  and  $\{\mathbf{z}_+^{(i)} = \{\mathbf{z}_{1+}^{(i)}, \mathbf{z}_{2+}^{(i)}\} \sim p_0(\mathbf{z})\}_{i=1}^N$ .

**Output:**  $\theta^{(T)} = \{\alpha_1^{(T)}, \beta_1^{(T)}, \alpha_2^{(T)}, \beta_2^{(T)}\}$ .

**for**  $t = 0 : T - 1$  **do**

Sample a minibatch of data  $\{\mathbf{x}^{(i)}\}_{i=1}^M$ ;

**Prior sampling for learning LEBMs:** For each  $\mathbf{x}^{(i)}$ , update  $\mathbf{z}_-^{(i)}$  using Eq. (4.13), with target distribution  $\pi(\mathbf{z}) = p_{\alpha^{(t)}}(\mathbf{z})$ ;

**Posterior sampling for foreground and background generation:** For each  $\mathbf{x}^{(i)}$ , update  $\mathbf{z}_+^{(i)}$  using Eq. (4.13), with target distribution  $Q(\mathbf{z}) = p_{\theta^{(t)}}(\mathbf{z}|\mathbf{x})$ ;

**Update LEBMs:**  $\alpha^{(t+1)} = \alpha^{(t)} + \eta_\alpha \frac{1}{m} \sum_{i=1}^m [\nabla_{\alpha} f_{\alpha^{(t)}}(\mathbf{z}_+^{(i)}) - \nabla_{\alpha} f_{\alpha^{(t)}}(\mathbf{z}_-^{(i)})]$ ;

**Update foreground and background generators:**

$\beta^{(t+1)} = \beta^{(t)} + \eta_\beta \frac{1}{m} \sum_{i=1}^m \nabla_{\beta} \log p_{\beta^{(t)}}(\mathbf{x}^{(i)}|\mathbf{z}_+^{(i)})$ ;

**end for**

---

During inference, we initialize the latent variables  $\mathbf{z}$  for MCMC sampling from Gaussian white noise and run only the posterior sampling step to obtain  $\mathbf{z}_+$ . The inferred mask and region images are then given by the outputs of generative models  $p_{\beta_k}(\mathbf{w}|\mathbf{z}_+)$  and  $p_{\beta_k}(\mathbf{x}|\mathbf{z}_+)$ ,  $k = 1, 2$ , respectively.

#### 4.3.4 Technical Details

##### Pseudo label for additional regularization.

Although the proposed DRC explicitly models the interaction between the regions, it is still possible that the model converges to a trivial extractor, which treats the entire image as

the foreground or background region, leaving the other region null. We exploit the symbolic vector  $\mathbf{y}$  emitted by the LEBM (see section 4.3.1) for additional regularization. The strategy is similar to the mutual information maximization used in InfoGAN [CDH16]. Specifically, we use the symbolic vector  $\mathbf{y}$  inferred from  $\mathbf{z}$  as the pseudo-class label for  $\mathbf{z}$  and train an auxiliary classifier jointly with the above models; it ensures that the generated regions  $\mathbf{x}_k$  contain similar symbolic information for  $\mathbf{z}_k$ . Intuitively, this loss prevents the regions from converging to null since the symbolic representation  $\mathbf{y}_k$  would never be well retrieved if that did happen.

### Implementation.

We adopt a similar architecture for the generator as in DCGAN [RMC15] throughout the experiments and only change the dimension of the latent variables  $\mathbf{z}$  for different datasets. The generator consists of a fully connected layer followed by five stacked upsample-conv-norm layers. We replace the batch-norm layers [IS15] with instance-norm [UVL16] in the architecture. The energy-term in LEBM is parameterized by a 3-layered MLP. We adopt orthogonal initialization [SMG14] commonly used in generative models to initialize the networks and orthogonal regularization [BLR16] to facilitate training. In addition, we observe performance improvement when adding Total-Variation norm [ROF92] for the background generative model. More details, along with specifics of the implementation used in our experiments, are provided in the supplementary material.

## 4.4 Experiments

We design experiments to answer three questions: (1) How does the proposed method compare to previous state-of-the-art competitors? (2) How do the proposed components contribute to the model performance? (3) Does the proposed method exhibit generalization on images containing unseen instances (*i.e.*, same category but not the same instance) and even objects from novel categories?

To answer these questions, we evaluate our method on five challenging datasets in two groups: (1) Caltech-UCSD Birds-200-2011 (Birds) [WBM10], Stanford Dogs (Dogs) [KJY11], and Stanford Cars (Cars) [KSD13] datasets; (2) CLEVR6 [JHM17] and Textured Multi-dSprites (TM-dSprites) [MHH17] datasets. The first group of datasets covers complex real-world domains, whereas the second group features environments of the multi-object foreground with challenging spatial configurations or confounding backgrounds. As to be shown, the proposed method is generic to various kinds of input and produces more competitive foreground-background partition results than prior methods.

#### 4.4.1 Results on Foreground Extraction

##### Single object in the wild.

In the first group of datasets, there is typically a single object in the foreground, varying in shapes, texture, and lighting conditions. Unsupervised foreground extraction on these datasets requires much more sophisticated visual cues than colors and shapes. Birds dataset consists of 11,788 images of 200 classes of birds annotated with high-quality segmentation masks, Dogs dataset consists of 20,580 images of 120 classes annotated with bounding boxes, and Cars dataset consists of 16,185 images of 196 classes. The latter two datasets are primarily made for fine-grained categorization. To evaluate foreground extraction, we follow the practice in [BW20], and approximate ground-truth masks for the images with Mask R-CNN [HGD17], pre-trained on the MS COCO [LMB14] dataset with a ResNet-101 [HZR16] backend. The pre-trained model is acquired from the detectron2 [WKM19b] toolkit. This results in 5,024 dog images and 12,322 car images with a clear foreground-background setup and corresponding masks.

On datasets featuring a single foreground object, we use the 2-slot version of IODINE and Slot-attention. Since ReDO, IODINE, and Slot-Attention do not distinguish foreground and background in output regions, we choose the best-matching scores from the permutation of foreground and background masks as in [CAD19]. We observe that the proposed method

Model	Single Object						Multi-Object			
	Birds		Dogs		Cars		CLEVR6		TM-dSprites	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
W-Net*	24.8	38.9	47.7	62.1	52.8	67.6	-	-	-	-
GrabCut	30.2	42.7	58.3	70.9	61.3	73.1	19.0	30.5	61.9	71.0
ReDO <sup>§</sup>	46.5	60.2	55.7	70.3	52.5	68.6	18.6	31.0	9.4	17.2
OneGAN <sup>*†</sup>	55.5	69.2	71.0	81.7	71.2	82.6	-	-	-	-
IODINE <sup>§</sup>	30.9	44.6	54.4	67.0	51.7	67.3	19.9	32.4	7.3	12.8
Slot-Attn. <sup>§</sup>	35.6	51.5	38.6	55.3	41.3	58.3	83.6	90.7	7.3	13.5
Ours	<b>56.4</b>	<b>70.9</b>	<b>71.7</b>	<b>83.2</b>	<b>72.4</b>	<b>83.7</b>	<b>84.7</b>	<b>91.5</b>	<b>78.8</b>	<b>87.5</b>

Table 4.1: **Foreground extraction results on training data measured in IoU and Dice.** Higher is better in all scores. \*Results of W-Net and OneGAN are provided by [BW20]. Of note, results of these two models on Dogs and Cars datasets may **not** be directly comparable to other listed methods, as the data used for training and evaluation could be different. We include these results as a rough reference since no official implementation or pretrained model are publicly available. § indicates unfair baseline results obtained using extra ground-truth information, *i.e.*, we choose the best-matching scores from the permutation of foreground and background masks. †OneGAN is a strong **weakly supervised** baseline, which requires clean background images to provide additional supervision. We include this model as a potential upper bound of the fully unsupervised methods.

and Grabcut are the only two methods that provide explicit identification of foreground objects and background regions. While the Grabcut algorithm actually requires a predefined bounding box as input that specifies the foreground region, our method, thanks to the learned pixel re-assignment (see section 4.3.2), can achieve this in a fully unsupervised manner. Results in table 4.1 show that our method outperforms all the unsupervised baselines by a large margin, exhibiting comparable performance even to the weakly supervised baseline that requires additional background information as inputs [BW20]. We provide samples of foreground extraction results as well as generated background and foreground regions in figure 4.3. Note that our final goal is not to synthesize appealing images but to learn foreground extractors in a fully unsupervised manner. As the limitation of our method, DRC generates foreground and background regions less realistic than those generated by state-of-the-art GANs, which hints a possible direction for future work. More detailed discussions of the limitation can be found in supplementary material.

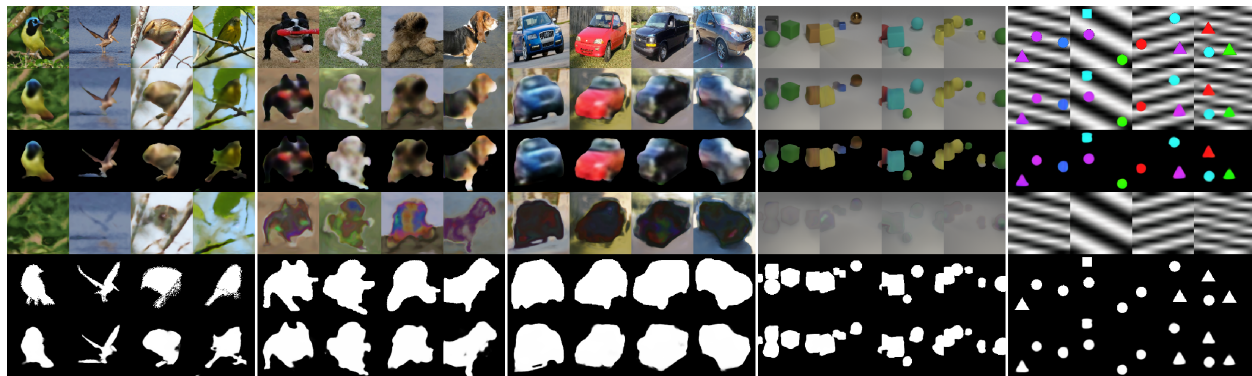


Figure 4.3: **Foreground extraction results for each dataset.**; zoom in for better visibility. From top to bottom: (i) observed images, (ii) generated images, (iii) masked generated foregrounds, (iv) generated backgrounds, (v) ground-truth foreground masks, and (vi) inferred foreground masks. More samples and results of baselines can be found in the supplementary material.

### Multi-object scenes.

The second group of datasets contains images with possibly simpler foreground objects but more challenging scene configurations or background parts. Visual scenes in the CLEVR6 dataset contain various objects and often with partial occlusions and truncations. Following

the evaluation protocol in IODINE and Slot-attention, we use the first 70K samples from CLEVR [JHM17] and filter the samples for scenes with at most 6 objects for training and evaluation, *i.e.*, CLEVR6. The TM-dSprites dataset is a variant of Multi-dSprites [MHH17] but has strongly confounding backgrounds borrowed from Textured MNIST [GRB16]. We generate 20K samples for the experiments. Similar to [GKK19a] and [LWU20a], we evaluate on a subset containing 1K samples for testing. Note that IODINE and Slot-attention are designed for segmenting complex multi-object scenes using slot-based object representations. Ideally, the output of these models consists of masks for each individual object, while the background is viewed as a virtual “object” as well. In practice, however, it is possible that the model distributes the background over all the slots as mentioned in [LWU20a]. We therefore propose two corresponding approaches (see the supplementary material for more details) to convert the output object masks into a foreground-background partition and report the best results of these two options for IODINE and Slot-attention in table 4.1.

On the CLEVR6 dataset, we use the publicly available pretrained model for IODINE, which achieves a reasonable ARI (excluding background pixels) of 94.4 on the testing data, close to the testing results in [GKK19a]. We observe that IODINE distributes the background over all the slots for some of the testing samples, resulting in much lower IoU and Dice scores. We re-train the Slot-attention model using the official implementation on CLEVR6, as no pretrained model is publicly available. The re-trained model achieves a foreground ARI of 98.0 on the 1K testing samples, which we consider as a sign of valid re-implementation. Results in table 4.1 demonstrate that the proposed method can effectively process images of challenging multi-object scenes. To be specific, our method demonstrates competitive performance on the CLEVR6 dataset compared with the SOTA object discovery method. Moreover, as shown empirically in figure 4.3, the proposed method can handle the strongly confounding background introduced in [GRB16], whereas previous methods are distracted by the background and mostly fail to capture the foreground objects.

Model	IoU	Dice
amortized inference*	-	-
w/o pix. re-assign.	21.8	35.3
w/o pseudo label	48.7	64.2
w/o TV-norm reg.	53.0	68.1
w/o ortho. reg.	54.3	69.2
short-run chain <sup>†</sup>	52.5	67.7
Full model	56.4	70.9

Table 4.2: **Ablation study on Birds.** \*We replace the LEBM with encoders to perform amortized inference for the latent variables  $\mathbf{z}$  within a variational framework as in VAE [KW13]. <sup>†</sup>We explore the possibility of using short-run MCMC [NHZ19] instead of persistent chain sampling.

#### 4.4.2 Ablation Study

We provide ablation studies on the Birds dataset to inspect the contribution of each proposed component in our model. As shown in table 4.2, we observe that replacing the LEBMs in the foreground and background models with amortized inference networks significantly harms the performance of the proposed method. In particular, the modified model fails to generate any meaningful results (marked as - in table 4.2). We conjecture that LEBM benefits from the low-dimensionality of the latent space [PHN20] and therefore enjoys more efficient learning. However, the inference networks need to learn an extra mapping from the high-dimensional image space to the latent space and require more elaborate architecture and tuning for convergence. Furthermore, we observe that the model that does not learn pixel re-assignment for background can still generate meaningful images but only vaguely captures masks for foreground extraction.

#### 4.4.3 Generalizable Foreground Extraction

**Extracting novel foreground objects from training categories.**

We show results on generalizing to novel objects from the training classes. To evaluate our method, we split the Birds dataset following [CAD19], resulting in 10K training images and 1K testing images. On Dogs and Cars datasets, we split the dataset based on the original train-test split [KJY11, KSD13]. This split gives 3,286 dog images and 6,218 car images for training, and 1,738 dog images and 6,104 car images for testing, respectively. As summarized in table 4.3, our method shows superior performances compared with baselines; the performance gap between training and testing is constantly small over all datasets.

Model	Birds		Dogs		Cars	
	IoU	Dice	IoU	Dice	IoU	Dice
	Tr.—Te.	Tr.—Te.	Tr.—Te.	Tr.—Te.	Tr.—Te.	Tr.—Te.
GrabCut*	30.2—30.3	42.7—42.8	58.3—57.9	70.8—70.5	60.9—61.6	72.7—73.5
ReDO	46.8—47.1	61.4—61.7	54.3—52.8	69.2—67.9	52.6—52.5	68.7—68.6
Ours	<b>54.8—54.6</b>	<b>69.5—69.4</b>	<b>71.6—72.3</b>	<b>83.2—83.6</b>	<b>71.9—70.8</b>	<b>83.3—82.5</b>

Table 4.3: **Performance of DRC on training and held-out testing data.** \*Note that GrabCut is a deterministic method that does not require training. We report the results of GrabCut [RKB04] on these splits only for reference. Tr. indicates the performance on training data, and Te. indicates the performance on testing data.

### Extracting novel foreground objects from unseen categories.

To investigate how well our method generalizes to categories unseen during training, we evaluate the models trained in real-world single object datasets on the held-out testing data from different categories. We use the same training and testing splits on these datasets as in the previous experiments. table 4.4 shows that our method outperforms the baselines on the Birds dataset when the model has trained on Dogs or Cars dataset, which have quite different distributions in foreground object shapes. Competitors like ReDO also exhibit such out-of-distribution generalization but only to a limited extent. Similar results are observed when using Dogs or Cars as the testing set. We can see that when the model is trained on Dogs and evaluated on Cars or vice versa, it still maintains comparable performances w.r.t.



Test	Train	GrabCut		ReDO		Ours	
		IoU	Dice	IoU	Dice	IoU	Dice
	Birds*			47.1	61.7	54.6	69.4
Birds	Dogs	30.3	42.8	22.2	35.3	<b>41.3</b>	<b>57.4</b>
	Cars			16.4	27.7	39.2	55.3
	Dogs*			52.8	67.9	72.3	83.6
Dogs	Cars	57.9	70.5	44.5	61.2	<b>67.8</b>	<b>80.4</b>
	Birds			44.0	60.3	53.6	69.1
	Cars*			52.5	68.6	70.8	82.5
Cars	Dogs	61.6	73.5	51.6	67.1	<b>68.6</b>	<b>81.0</b>
	Birds			41.8	58.6	45.1	61.7

Table 4.4: **Performance of DRC on unseen testing categories.** \*We include the testing results of models trained with data from the same categories for reference.

those are trained&tested on the same class. We hypothesize that these two datasets have similar distributions in foreground objects and background regions. In the light of this, we can further entail that the distribution of Dogs is most similar to that of Cars and less similar to that of Birds according to the results, which is consistent with our intuitive observation of the data. We provide a preliminary analysis of the statistics of these datasets in the supplementary material.

## 4.5 Conclusion

We have presented the Deep Region Competition, an EM-based fully unsupervised foreground extraction algorithm fueled by energy-based prior and generative image modeling. We propose learned pixel re-assignment as an inductive bias to capture the background

regularities. Experiments demonstrate that DRC exhibits more competitive performances on complex real-world data and challenging multi-object scenes. We show empirically that learned pixel re-assignment helps to provide explicit identification for foreground and background regions. Moreover, we find that DRC can potentially generalize to novel foreground objects even from categories unseen during training. We hope our work will inspire future research along this challenging but rewarding research direction.

## 4.A Details on Models and Hyperparameters

**Architecture.** As mentioned in the paper, we use the same overall architecture for different datasets (while the size of latent variables may vary). The details for the generators and LEBMs are summarized in the table 4.5 and table 4.7.

Dataset	Foreground	Background	Pixel Re-assignment
Birds	256	256	512
Dogs	256	256	512
Cars	256	192	512
CLEVR6	256	2	256
TM-dSprites	256	4	1024

Table 4.5: Dimension of latent variables on each dataset.

**Hyperparameters and Training Details.** For the Langevin dynamics sampling [WT11], we use  $K_0$  and  $K_1$  to denote the number of prior and posterior sampling steps with step sizes  $s_0$  and  $s_1$  respectively. Our hyperparameter choices are:  $K_0 = 60, K_1 = 40, s_0 = 0.4$  and  $s_1 = 0.1$ . These are identical across different datasets. During testing, we set the posterior sampling steps to 300 for Dogs and Cars, and 2.5K, 5K and 5K for Birds, CLEVR6 and TM-dSprites respectively. The parameters of the generators and LEBMs are initialized with orthogonal initialization [SMG14]. The gain is set to 1.0 for all the models. We use the

Layers	In-Out size	Comment
LEBM for Foreground/Background Models		
Input: $\mathbf{z}$	$D^*$	
Linear, LReLU	200	
Linear, LReLU	200	
Linear	$K^\dagger$	
LEBM for Pixel Re-assignment Model		
Input: $\mathbf{z}$	$D^*$	
Linear, LReLU	200	
Linear, LReLU	200	
Linear, LReLU	200	
Linear	1	
Generator for Foreground/Background Model and Re-assignment Model		
Input: $\mathbf{z}$	$D^*$	
Linear, LReLU	$4 \times 4 \times 128$	reshaped output
UpConv3x3Norm, LReLU	$8 \times 8 \times 1024$	stride 1 & padding 1
UpConv3x3Norm, LReLU	$16 \times 16 \times 512$	stride 1 & padding 1
UpConv3x3Norm, LReLU	$32 \times 32 \times 256$	stride 1 & padding 1
UpConv3x3Norm, LReLU	$64 \times 64 \times 128$	stride 1 & padding 1
UpConv3x3Norm, LReLU	$128 \times 128 \times 64$	stride 1 & padding 1
Conv3x3	$128 \times 128 \times (3 + 1)$	RGB & Mask
	$128 \times 128 \times 2$	Re-assignment grid

Table 4.6: Architecture of the generators, LEBMs and auxiliary classifiers.

Auxiliary classifier for Foreground/Background Model		
Input: $\mathbf{x}$	$128 \times 128 \times 3$	generated image
Conv4x4Norm, LReLU	$64 \times 64 \times 64$	stride 2 & padding 1
Conv4x4Norm, LReLU	$32 \times 32 \times 128$	stride 2 & padding 1
Conv4x4Norm, LReLU	$16 \times 16 \times 256$	stride 2 & padding 1
Conv4x4Norm, LReLU	$8 \times 8 \times 512$	stride 2 & padding 1
Conv4x4Norm, LReLU	$4 \times 4 \times 1024$	stride 2 & padding 1
Conv4x4	$1 \times 1 \times K^\dagger$	

Table 4.7: Architecture of the generators, LEBMs and auxiliary classifiers (Cont’d). Up-Conv3x3Norm denotes a Upsampling-Convolutional-InstanceNorm layer with a convolution kernel size of 3. Similarly, Conv4x4Norm denotes a Convolutional-InstanceNorm layer with a kernel size of 4. LReLU denotes the Leaky-ReLU activation function. The leak factor for LReLU is 0.2 in LEBMs and auxiliary classifiers, and 0.01 in generators.  $*D$  represents the dimensions of the latent variables for different datasets; see table 4.5.  $\dagger K$  represents the pre-specified category number for latent variables. We use 200 for both the foreground and background LEBMs on real-world datasets, and 30 and 10 in the foreground and background LEBMs on multi-object datasets respectively.

ADAM optimizer [KB14] with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ . Generators are trained with a constant learning rate of 0.0001, and LEBMs with 0.00002. We run experiments on a single V100 GPU with 16GB of RAM and with a batch size of 48. We set the maximum training iterations to 10K and run for at most 48hrs for each dataset.

# CHAPTER 5

## Object-centric and Relational Representation Learning with RelViT

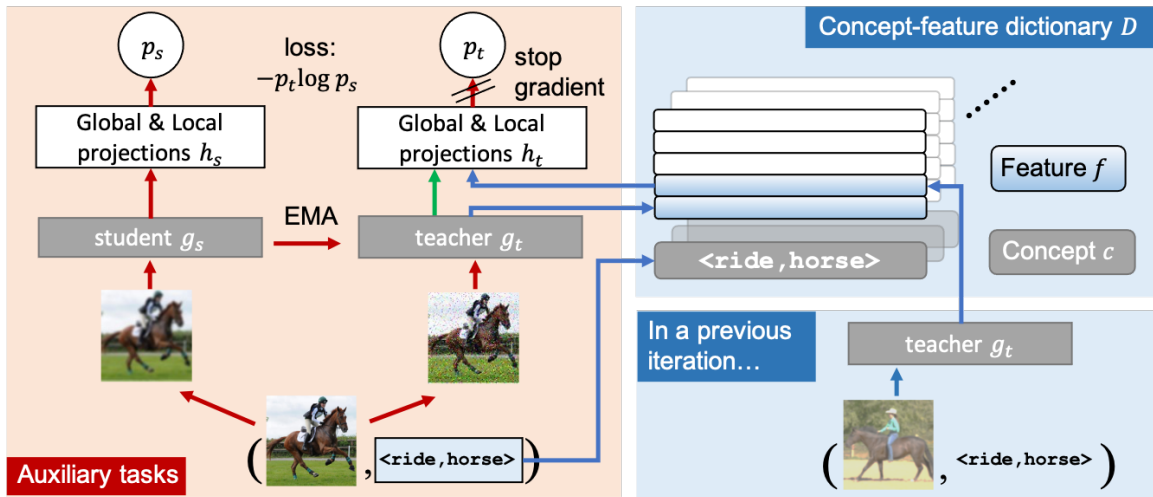


Figure 5.1: An overview of our method. **Red+Green**: the learning pipeline of DINO [CTM21] and EsViT [LYZ21]; **Red+Blue**: our pipeline. We introduce a *concept-feature dictionary*, where the key is a concept  $c$  and its value is a queue of image features  $f$  with the same concept, to allow flexible feature retrieval with the concept keys. With the proposed dictionary, we further develop our concept-guided global and local tasks. EMA denotes the exponential moving average.

### 5.1 Introduction

Deep neural networks have achieved great success in visual recognition. However, their ability for visual relational reasoning, *i.e.* reasoning with entities and their relationships in a visual scene, still falls short of human-level performances, especially in real-world do-

mains. The challenges of common visual relational reasoning tasks, *e.g.* HICO and GQA benchmarks [CWH15b, HM19] are manifested in three aspects: 1) **object-centric learning** to identify objects (including humans) as well as their visual properties; 2) **relational reasoning** to infer all pairwise relationships between the object entities; and 3) **systematic generalization** to reason with visual entities and relations on novel object-relation combinations and extrapolate to longer reasoning hops [BMN18, HDM20]. While existing models have leveraged pre-trained object detectors [RHG15, JMR20] and/or explicit symbolic reasoning methods [YWG18] to tackle these challenges, they leave ample space for improvement.

More recently, **vision transformers** (ViTs) have become the new paradigm for visual recognition and have made great strides in a broad range of visual recognition tasks [DBK20, WXL21b, LLC21]. Several properties of ViT make it a compelling model choice for visual relational reasoning. First, the **self-attention mechanism** in ViT offers a strong relational inductive bias, explicitly modeling the relations between input entities. Second, the design of **image as patches** facilitates the learning of object-centric representations, as evidenced by recent works, *e.g.* DINO and EsViT [CTM21, LYZ21], that demonstrate ViTs trained with self-supervised learning (SSL) capture objects in the image without label annotations.

To investigate the efficacy of the ViT backbone for visual relational reasoning, in particular on systematic generalization, we introduce new systematic splits to canonical benchmarks and compare the ViT backbone with the CNN backbone. Results on GQA show that switching to ViTs in MCAN model [YYC19b] brings an immediate 11% gain in accuracy. However, the performance gap between the original GQA testing split and the new systematic split remains considerable (15% in accuracy) for both backbones. It suggests that generic ViTs still need to be improved to tackle the reasoning task, especially on systematic generalization. Recent works have shown that neural networks can learn representations with better generalization, by learning certain auxiliary tasks of predicting human-specified concepts [HTG20, KNT20]. A natural question emerges: *can we exploit these concepts to*

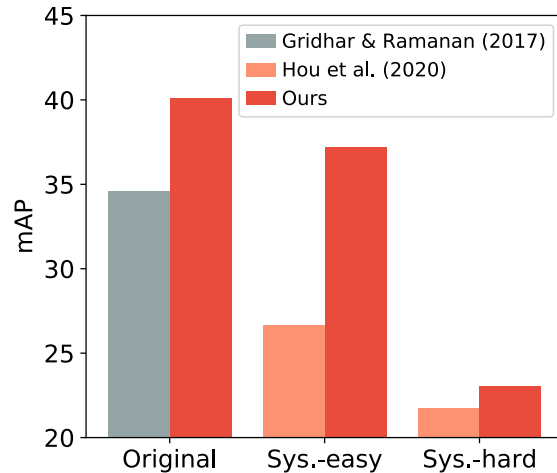


Figure 5.2: **Results on HICO**. Our method improves the best baseline by 16%, 43%, and 7% on the original non-systematic and two new systematic splits. **Sys.:** systematic.

*improve the reasoning ability of ViTs?*

**Our approach** is to make better use of concepts (e.g. the labels in the original training dataset) in the ViT training for better relational reasoning. To this end, we first introduce a novel *concept-feature dictionary*, where each key is a concept and its value is a queue of image features with the same concept, as shown in Figure 5.1. It allows dynamic and flexible training-time image feature retrieval during training. Based on this dictionary, we then augment the canonical ViT training pipeline with two auxiliary tasks: First, to facilitate high-level reasoning about relationships, we design a **global task** that helps cluster images with the same concept together to produce semantically consistent relational representations. Second, to learn better object-centric representations, we develop a **local task** that guides the model to discover object-centric semantic correspondence across images [LYT10]. Thanks to the plug-and-play feature of our concept-feature dictionary, our auxiliary tasks can be easily incorporated into existing ViT training pipelines without additional input pre-processing. We term the resulting model *concept-guided vision transformer* (or RelViT for short).

We evaluate our method on two standard visual relational reasoning benchmarks: HICO

and GQA. Beyond the original independent and identically distributed (I.I.D.) training-testing split, we introduce new systematic splits for each dataset to examine the ability of systematic generalization, *i.e.*, recognizing novel object-relation combinations. Our results show that RelViT significantly outperforms previous approaches. On HICO, it improves the best baseline by 16%, 43%, and 7% on the original non-systematic and two new systematic splits, respectively, as shown in Figure 5.2. On GQA, it further closes the gap of overall accuracy between models using visual backbone feature only and models using additional bounding box features (obtained from pre-trained object detectors) by 13% and 18% on the two splits. We also show that our method is compatible with various ViT variants and robust to hyper-parameters. Finally, our qualitative inspection indicates that RelViT does improve ViTs on learning relational and object-centric representations.

Our main contributions are summarized as follows:

- We propose RelViT, by incorporating visual relational concepts to the ViT training with the newly-introduced concept-guided global and local auxiliary tasks, where a *concept-feature dictionary* is proposed to enable dynamic and flexible image feature retrieval with the concept keys.
- In extensive experiments on the original non-systematic and new systematic split of the HICO and GQA datasets, we demonstrate the advantages of RelViT over various strong baselines for visual relational reasoning.
- We perform ablation studies on RelViT to show the contributions of its key components, its compatibility to various ViT architectures, and its robustness to hyper-parameters. We provide qualitative results to confirm our improved learning of relational and object-centric representations.



## 5.2 Methodology

### 5.2.1 Background

**Vision transformers.** Here we briefly review the architecture of multi-staged ViTs [DBK20]. Given an image  $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ , a ViT model  $g$  first tokenizes the input into  $N$  image tokens (patches) with a resolution of  $(T, T)$ :  $\text{tokenize}(\mathbf{I}) = [t_1, \dots, t_N], t_i \in \mathbb{R}^{T^2 \times C}, N = HW/T^2$ , where  $(H, W)$  and  $C$  denotes the original resolution and number of channel of the image, respectively. Then in each stage, a *patch embedding* and a *multi-head self attention* (MHSA) module is applied to these tokens to produce input for the next stage. The final output of ViT  $g(\mathbf{I})$  is a sequence of tokens  $[z_1, \dots, z_N]$  that correspond to the aforementioned input tokens. For global prediction tasks, *e.g.* image categorization, a summary of the input image can be obtained by either inserting an extra [CLS] token to the input sequence of image tokens or performing an extra pooling operation over the output tokens [ZKH21].

**Self-supervised learning with DINO and EsViT.** Our method is developed upon the recently proposed self-supervised learning (SSL) approach *self-distillation with no labels* (DINO) [CTM21] and its follow-up EsViT [LYZ21]. As shown in Figure 5.1, their main idea is to encourage the output consistency between a teacher  $g_t$  and a student network  $g_s$ , parameterized by  $\theta_t$  and  $\theta_s$ , respectively. Given an input image  $\mathbf{I}$ , both networks map it to a probability distribution  $P_t(\mathbf{I}) = h_t(g_t(\mathbf{I}))$  and  $P_s(\mathbf{I}) = h_s(g_s(\mathbf{I}))$  via an extra projection head  $h(\cdot)$ . The teacher and student network will be updated alternatively by following these two rules: (1) For the student network:  $\theta_s \leftarrow \arg \min_{\theta_s} \mathcal{L}_{\text{Global}}$ , where  $\mathcal{L}_{\text{Global}} = -P_t(\mathbf{I}) \log P_s(\mathbf{I})$ ; (2) For the teacher network,  $\theta_t$  is updated using an exponential moving average (EMA) on  $\theta_s$ :  $\theta_t \leftarrow \lambda \theta_t + (1 - \lambda) \theta_s$ , where  $\lambda$  controls the updating momentum. In practice, multiple views of the input image  $\mathbf{I}$  will be generated via data augmentation and the teacher and student networks will receive different views, preventing the task from being trivialized. EsViT further extends the image-level loss  $\mathcal{L}_{\text{Global}}$  to patch-level by applying dense SSL [WZS21] for learning correspondence between the different views, enhancing the performance on dense

prediction. Readers are encouraged to refer to [CTM21] and [LYZ21] for more details about these two works.

### 5.2.2 ReViT

ReViT is a concept-guided ViT that makes better use of the concepts in the ViT training for the improved relational reasoning. In this section, we first introduce a *concept-feature dictionary* to store and retrieve image features with their concept keys. We then augment the canonical ViT training pipeline with two auxiliary tasks: a global level task and a local level task, both are concept-guided by resorting to the concept-feature dictionary. Intuitively, the global task help cluster images with the same concept together to produce semantically consistent relational features, while the local task guides the model to discover object-centric semantic correspondence across images.

**Concept-feature dictionary.** We assume the total number of concepts is  $M$ , and the set of all concepts  $\mathcal{C} = \{c_1, \dots, c_M\}$ . A *concept-feature dictionary* is denoted by  $D = \{(c_1, Q_1), \dots, (c_M, Q_M)\}$ , where each concept  $c_i$  is associated with a queue  $Q_i$  of image features. During training, each image  $\mathbf{I}$  may come with multiple concepts, which we denote by  $\mathcal{C}_{\mathbf{I}} \subset \mathcal{C}$ . For instance, there may exist several human-object interactions in an image from the HICO dataset, each of which may correspond to a concept. As shown in Figure 5.1, whenever a new image-concept pair  $(\mathbf{I}, \mathcal{C}_{\mathbf{I}})$  comes, we uniformly draw a concept code  $c$  from  $\mathcal{C}_{\mathbf{I}}$ , pick up the queue  $Q$  from the dictionary that corresponds to  $c$ , and then retrieve the image feature  $f$  from  $Q$ . Meanwhile, we pass the input image  $\mathbf{I}$  to the teacher network  $g_t$  to get the new image feature  $f' = g_t(\mathbf{I})$ , and *enqueue* it to  $Q$ . Note that if  $Q$  is full already, we first need to *dequeue* the oldest image feature from  $Q$ . During training, we use the retrieved image feature  $f$  for the two auxiliary tasks below, rather than the input image feature  $f'$ .

Furthermore, the sampling strategy, *i.e.* how to retrieve image feature  $f$  from  $Q$ , plays an important role in the overall performance of our method. We consider the following two sampling strategies:

- *Uniform sampling.* Each image feature is drawn with equal probability from the queue, *i.e.* suppose we have  $N$  features in the queue, then the probability of each feature being sampled is  $1/N$ . This tactic encourages the diversity of the retrieved image features, benefiting the overall performance. However, some older features in the queue may largely fall behind the current model if the teacher network  $g_t$  is updated quickly, eliciting unstable training.
- *“Most-recent” sampling.* The sampling probability mass is allocated based on the freshness of image features, and the most recent feature has the highest chance to be retrieved. Specifically, suppose we have  $N$  features in the queue  $Q$  ( $|Q| \geq N$ ). Then for the  $i$ -th newest feature  $f$ , we define its weight  $w_i = N - i + 1$ . Finally, the probability of the  $i$ -th newest feature being sampled is  $w_i / \sum_{j=1}^N w_j$ . This tactic ensures we retrieve more up-to-date features and thereby stabilizes the learning. But it may hurt the overall performance due to a lack of feature diversity, as the chance of older features being sampled is small.

Note that the feature queue is empty at the beginning of training. In this case, we simply use the input image feature  $f'$  for the auxiliary tasks, and also *enqueue* it to  $Q$  that corresponds to the concept of the input image. As we can show in the next, now our proposed global and local tasks reduce to DINO [CTM21] and EsViT [LYZ21], respectively.

**Concept-guided global task.** Suppose we have two views  $\{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}\}$  of an image  $\mathbf{I}$ , the main idea of our concept-guided global task is to replace  $\mathbf{I}^{(1)}$  in the DINO loss [CTM21] with the image feature  $f$  sampled from the concept-feature dictionary, which becomes

$$\mathcal{L}_{\text{Global}} = -h_t(f) \log h_s(g_s(\mathbf{I}^{(2)})), \quad (5.1)$$

where  $h_t$  and  $h_s$  are the projection head of the teacher and student network, respectively, and  $g_s$  is the student network. Intuitively, minimizing the global loss is equivalent to encouraging the similarity of any two different image features with the same concept. Hence, it can help produce more semantically consistent relational representations, in particular when the concepts stored in the concept-feature dictionary are themselves relational.

Similar inter-class representation learning techniques have been explored before [WHG17, CBJ18]. However, these approaches require a rather complex pre-processing stage, *e.g.* the images have to be split in terms of the concept before training, making them not directly applicable to existing training pipelines. Rather, with our proposed concept-feature dictionary that dynamically saves & retrieves image features from the running storage, our concept-guided global task becomes a plug-n-play task to existing training pipelines.

**Concept-guided local task.** As we mentioned earlier, our concept-guided local task aims at facilitating object-centric learning, by the means of correspondence learning [LYT10, WJE19]. Recent studies have unveiled the possibility of learning correspondence with SSL [WZS21, LYZ21]. However, only low-level correspondence between two augmented (*e.g.* rotated) views of an image can be discovered, while the semantic information of objects is missing. To remedy this, we bring concepts to these methods, endowing them the capability of learning semantic correspondence from images.

Specifically, suppose we have two views  $\{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}\}$  of an image  $\mathbf{I}$ , and we also tokenize the image feature into a sequence of  $N$  local image tokens. Then at the output of ViT, we obtain  $g_t(\mathbf{I}^{(1)}) = [z_1^{(1)}, \dots, z_N^{(1)}]$  and  $g_s(\mathbf{I}^{(2)}) = [z_1^{(2)}, \dots, z_N^{(2)}]$ , where  $z$  denotes the local feature. Prior work, such as EsViT [LYZ21], relies on the local features  $g_t(\mathbf{I}^{(1)})$  and  $g_t(\mathbf{I}^{(2)})$  for the local task. Instead, we replace  $g_t(\mathbf{I}^{(1)})$  with the image feature  $f$  retrieved from the concept-feature dictionary using the concept of the image  $\mathbf{I}$ . We then split  $f$  into multiple local features, *i.e.*  $f = [z_1^{(f)}, \dots, z_N^{(f)}]$  and our concept-guided local loss becomes

$$\mathcal{L}_{\text{Local}} = -\frac{1}{N} \sum_{i=1}^N h_t(z_{j^*}^{(f)}) \log h_s(z_i^{(2)}), \quad j^* = \arg \max_j \text{CosineDistance}(z_j^{(f)}, z_i^{(2)}), \quad (5.2)$$

where  $h_t(\cdot), h_s(\cdot)$  are the projection heads that map local features to probability distributions<sup>1</sup>. Intuitively, it greedily matches the output between two local regions that have

---

<sup>1</sup>Note that the projection head here is different from DINO’s: it works on all output local features. While in DINO, the projection head only works on the summary of input image, *i.e.* the resulting feature after a max-pooling operation or the feature that corresponds to [CLS] in the input.

minimal feature distance – bootstrapping the object-level semantic correspondence among images with the same concept.

**Overall loss.** By combining the global and local tasks, we add an auxiliary task loss  $\mathcal{L}_{\text{aux}}$  to the main loss  $\mathcal{L}_{\text{main}}$  (*e.g.* cross-entropy loss of the reasoning task). The eventual objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \alpha \mathcal{L}_{\text{aux}}, \quad \mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{Global}} + \mathcal{L}_{\text{Local}}, \quad (5.3)$$

where a trade-off weight  $\alpha$  is added for better flexibility. As we mentioned above, our method will reduce to EsViT, a baseline without concept-guided auxiliary tasks, when we use the current input features  $g_t(\mathbf{I}^{(1)})$  instead of  $f$  retrieved from our dictionary for computing  $\mathcal{L}_{\text{Global}}$  and  $\mathcal{L}_{\text{Local}}$ .

## 5.3 Experiments

We conduct experiments on two challenging visual relational reasoning datasets: HICO [CWH15b] and GQA [HM19]. Besides their original non-systematic split, we introduce the systematic splits of each dataset to evaluate the systematic generalization of our method. First, we compare our method against various strong baselines [ML16, GR17, HM18a] on visual relational reasoning, as well as state-of-the-art ViTs. Second, we perform the ablation analysis to examine the key components of our method: ViT backbones, concept-feature dictionaries, and auxiliary tasks. Finally, we provide qualitative results to justify the emerging image clustering in terms of concepts and the learned semantic correspondence. Please see more details of all the evaluated tasks in the supplementary material.

### 5.3.1 Main results I: Human-object Interaction Recognition

**Overview.** HICO [CWH15b] features the human-object interaction (HOI) recognition, *i.e.* predicting all the possible HOI categories of the input image. It contains 600 HOI categories

with 117 unique actions and 80 object classes. The training set includes 38116 images and the test set includes 9658 images. For a fair comparison, we follow the standard practice and mainly focus on those previous methods that do not require extra supervision [FCT18] or data [LXL20, LXH19, JCW21]. By default, we choose PVTv2-b2 [WXL21a] as the ViT backbone. Regarding the concept-feature dictionary, we use the “*most-recent*” *sampling* and a queue length  $|Q|$  of 10. The trade-off weight  $\alpha$  in the overall loss is fixed to 0.1. Other hyper-parameters are inherited from DINO [CTM21].

**Systematic split.** The systematic generalization in HICO has been studied before under the name “zero-shot HOI recognition” [SYH18]. The main idea is to remove some HOI categories from the training set while ensuring all the single actions and objects can still be kept in the remaining HOI categories. We thereby reuse the systematic splits offered by [HPQ20b]. There are two splits: *systematic-easy*, where only the rare HOI classes are removed from the training set; *systematic-hard*, where only the non-rare HOI classes are removed besides the rare ones. The systematic-hard split contains much fewer training instances and thereby is more challenging.

**Concepts.** In HICO, we simply use the 600 HOI categories as our default concepts. We also report results with other concepts (*e.g.* actions, objects) in the ablation study.

**Results.** In Table 5.1, we compare our method with several counterparts. The results read that even a simple model with PVTv2-b2 (**25.4M parameters**) backbone can outperform many previous methods using ResNet-101 (**44.7M parameters**) and lots of bell and whistles. This confirms the great potentials of ViTs in visual relation reasoning. By further adding our global and local tasks, we attain 4-6 mAP gain on original and systematic splits. We also observe that EsViT [LYZ21], a recently proposed SSL approach, also outperforms the ViT-only baseline. Therefore, we combine their SSL task and our concept-guided tasks and reach the peak performance (40.12 mAP) on the original HICO split. Although **we do not utilize any extra supervision**, RelViT+EsViT beats the current state-of-the-art [FCT18] that uses the additional “pose” supervision that does not exist in the HICO dataset. Overall,

Method	Ext. superv.	Backbone	Orig.	Systematic-easy		Systematic-hard	
				Full cls.	Unseen cls.	Full cls.	Unseen cls.
[ML16]*		ResNet-101	33.8	-	-	-	-
[GR17]*	bbox	ResNet-101	34.6	-	-	-	-
[FCT18]*	pose	ResNet-101	39.9	-	-	-	-
[HPQ20b] <sup>†</sup>		ResNet-101	28.57	26.65	11.94	21.76	10.58
ViT-only		PVTv2-b2	35.48	31.06	11.14	19.03	18.85
EsViT ([LYZ21])		PVTv2-b2	38.23	35.15	11.53	22.55	21.84
RelViT (Ours)		PVTv2-b2	39.4	36.99	12.26	22.75	22.66
RelViT + EsViT (Ours)		PVTv2-b2	<b>40.12</b>	<b>37.21</b>	<b>12.51</b>	<b>23.06</b>	<b>22.89</b>

Table 5.1: **Results on HICO dataset.** Some methods requires extra supervision. Bbox/Pose means object-detection or pose estimation is needed. All results are reported in mAP. \*Results reported in the original papers; <sup>†</sup>Introduces the systematic split we use in the experiments. **Full cls.:** results reported on all 600 HOI categories; **Unseen cls.:** results reported on the held-out HOI categories from the training set for testing systematic generalization. **Ext. superv.:** extra supervision.

we raise the results of a fair counterpart [GR17] that only exploits extra bbox supervision (which is included in HICO) by 16% (34.6  $\rightarrow$  40.12) on the original split. For systematic splits, we raise the results of [HPQ20b] by 43% (26.65  $\rightarrow$  37.21) on the systematic-easy split and 7% (21.76  $\rightarrow$  23.06) on the systematic-hard split. Finally, although the gap between systematic and non-systematic split still exists (partly due to the much smaller training set for systematic splits), our method makes significant progress, especially on unseen classes (+12.3 mAP on systematic-hard). This further demonstrates the advantages of our concept-guided ViT in systematic generalization.

### 5.3.2 Main results II: Visual Question Answering

**Overview.** GQA [HM19] is a recent visual question answering (VQA) dataset with a focus on relational reasoning. Each question is also labeled with semantics. By default, it offers both pretrained-CNN grid features and region features obtained through Faster R-CNN [RHG15]. For counterparts, we focus on fair comparisons and therefore exclude those that require massive vision-language pretraining [LYY19]. Notably, **we do not use extra supervision, such as scene graph** [KZG16]. The RelViT configuration is almost the same as in HICO, except that we apply the *uniform sampling* instead as we find it empirically works better. We employ MCAN-Small [YYC19b] as our VQA model and the ImageNet1K-pretrained PVTv2-b2 as our vision backbone. The results are reported on the full validation set of GQA.

**Systematic split.** In GQA, we especially examine the facet of *productivity* in systematic generalization, *i.e.* the ability of reasoning with longer reasoning hops [HDM20]. To this end, we exploit the extra semantics label associated with the GQA questions. We observe that the semantics in GQA break down each question into a sequence of “reasoning hops”, where a distribution of reasoning hops can be found in Figure 3. See the supplementary material for examples. Therefore, our idea is to exclude questions with longer reasoning hops from the training set. We end up only keeping questions with less than 5 reasoning hops in the training set. We refer to this setting as the systematic split (“Sys.”) in the results.

**Concepts.** Inspired by recent research on vision-language pretraining [TB19, LYY19, LYL20], we obtain concepts by parsing the questions into keywords. Specifically, we only keep certain verbs, nouns, and adjectives that contain significant meanings (e.g. actions, objects, characteristics, etc), ending up with 1615 concepts. Due to the space limit, readers may find more details on concept parsing in the supplementary material.

**Results.** We report the comparison results on the original and systematic splits in Table 5.2. The main goal of our experiments on GQA mainly is to verify if our method can help



Method	Bbox feat.*	Backbone	Orig.	Sys.
BottomUp ([AHB18])	✓	ResNet-101	53.21	-
MAC ([HM18b])	✓	ResNet-101	54.06	-
MCAN-Small ([YYC19b])	✓	ResNet-101	58.35	36.21
MCAN-Small ([YYC19b])		ResNet-101	51.1	30.12
ViT-only		PVTv2-b2	56.62	31.39
EsViT ([LYZ21])		PVTv2-b2	56.95	31.76
RelViT (Ours)		PVTv2-b2	57.87	35.48

Table 5.2: **Results on GQA dataset.** All results are reported in overall accuracy. \*With extra Faster R-CNN bbox features.

reduce the gap between models using backbone features only and models using additional bbox features (with dense object detectors). Besides, we also examine to which extent our method can improve systematic generalization. Firstly, we observe that using ViT can largely alleviate the aforementioned gap (51.1  $\rightarrow$  56.62), suggesting that the object-centric representations emerge in ViTs. It implies the potential of using ViTs in eliminating the need for external object detectors. By further adding our proposed auxiliary tasks, we achieve the peak performance and raise the results of MCAN-Small w/o bbox features by 13% (51.1  $\rightarrow$  57.87) on the original split and 18% (30.12  $\rightarrow$  35.48) on the systematic split. **Without any detection pretraining or bbox features**, our method achieves very close results to MCAN-Small w/ bbox features on both two splits. The additional results in appendix demonstrate that the marginal gap could be further eliminated if we apply larger backbone models (PVTv2-b2 has much fewer parameters than ResNet-101).

### 5.3.3 Why do our auxiliary tasks work?

The results in the previous section suggest that RelViT outperforms its counterparts on the challenging relational reasoning tasks. Now we would like to provide more insights into our



Figure 5.3: Histogram of reasoning hops over GQA training questions.

method design by answering the question: why do our auxiliary tasks work? To this end, we perform a diverse set of analyses on accessing the impact of key components in ReViT . We also qualitatively justify the intuitions of two auxiliary tasks. These results are reported on the HICO dataset.

### 5.3.3.1 Ablation study

**Different ViT architectures.** The first aspect we examine is the ViT architecture. Besides the default choice on PVTv2-b2, we test our method with the original ViT-S/16 [DBK20] and another prevalent architecture Swin-Small [LLC21]. The results are presented in Figure 5.4a and Figure 5.4b, respectively. These two architectures can both benefit from our auxiliary tasks and we have similar advantages over counterparts as in the default setting, which confirms our compatibility to various ViT variants. Full quantitative results are provided in the supplementary.

**Implementation of concept-feature dictionary.** We conduct ablations on three facets of concept-feature dictionary: the choice of concepts, sampling tactics, and the size of queue  $|Q|$ . In Figure 5.4c, we compare three different concept choices: actions, objects, and HOIs with our best model. The results suggest that all three choices can bring improvement to the baseline without any feature queue (denoted as “None”) while using HOIs and objects brings larger improvement. We hypothesize that the proposed auxiliary tasks need more “delicate” concepts to guide the ViT training but actions in HICO tend to be vague and even ambiguous [SYH18]. Therefore, albeit the consistent advantages of our method in terms of different concept selections, a careful design of concept space could still be pivotal to achieve the peak performance in relational reasoning.

Furthermore, we show the interplay between sampling strategies and queue size  $|Q|$  in Figure 5.4d. Interestingly,  $|Q|$  has a much smaller impact on the performance with the “*most-recent*” sampling than that with the *uniform sampling*. As we mentioned in Section 5.2.2, the *uniform sampling* could help with more diverse features but could also elicit unstable train-

ing. Larger  $|Q|$  makes the two consequences in the *uniform sampling* more prominent, thus causing worse performance when stable training is the bottleneck (*e.g.* in a small dataset like HICO). Rather, the “*most-recent*” *sampling* can be less sensitive to  $|Q|$  as only the recent features could be sampled even when  $|Q|$  is large.

**Auxiliary tasks.** In Figure 5.4e, we show the results of only adding our global or local task in  $\mathcal{L}_{\text{aux}}$ . Surprisingly, just using the local task is enough to deliver competitive results in the HICO task. This suggests that the real bottleneck in ViTs seems to be better object-centric representations, as our local task is designed for this. Nonetheless, adding our global task can still elicit clear advantages over other counterparts that do not exploit concept-guided learning.

**Robustness to  $\alpha$ .** We sweep the trade-off weight  $\alpha$  from 0.02 to 0.5 and report the results in Figure 5.4f, where **solid** and **dash** represent our method and the baseline, respectively. It is observed that adding the proposed auxiliary tasks always achieves better performances than the baseline, indicating our method is robust to hyper-parameters. Moreover, the improvements become slightly more significant when  $\alpha$  is relatively large (but not too large). The peak performances in different splits all appear around  $\alpha = 0.1$ , which we thus use as our default choice.

### 5.3.3.2 Qualitative inspection

**Features vs. concepts.** To further justify whether our global task can truly facilitate the learned representation to be more relational, we illustrate the learned output features (max-pooling on all the output tokens) by t-SNE visualization in Figure 5.5. Different colors correspond to different HOI categories, *i.e.* the concepts we used in RelViT. The results read that more clusters can be identified over the image features extracted by RelViT; therefore our concept-guided global task can encourage the learned features to be more discriminative regarding the relational concepts than the baselines.

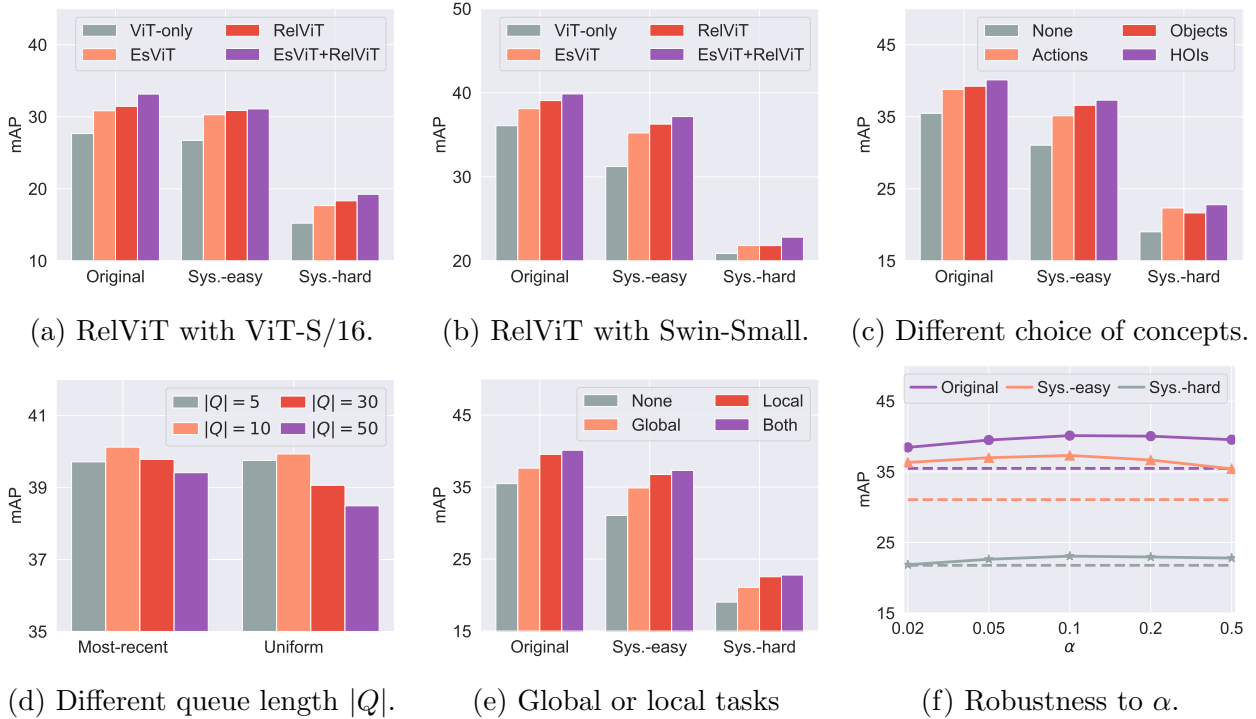


Figure 5.4: **Ablation study on HICO.** We investigate the impact of ViT architectures, implementation of concept-feature dictionary, auxiliary tasks, and the weight  $\alpha$  on the performance of our method. **Sys.:** systematic.

**Semantic correspondence.** We also probe the learned semantic correspondence that could be encouraged by our local task, by intuition. We aim at comparing the correspondence extracted from a model trained with different auxiliary tasks, *i.e.* no auxiliary task, no-concept auxiliary tasks, and our auxiliary tasks. We consider two settings: 1) semantic setting (two images that belong to the same concept, *e.g.* both contains a cat), and 2) non-semantic setting (two views of the same image). Results in Figure 5.6 highlight the high-similarity matches. Although our method and non-concept baseline (EsViT) both work fine in the non-semantic setting, our method can identify the semantic correspondence on more objects thanks to the concept guidance. Not surprisingly, baseline w/o any auxiliary task (ViT-only) performs the worst as it may suffer from over-smoothness [GWL21] and lose all the meaningful spatial information after fine-tuning on the target task.

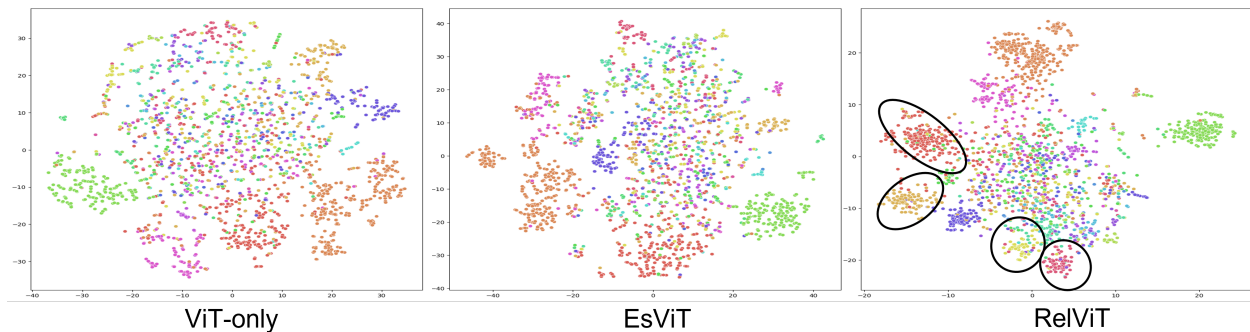


Figure 5.5: Visual illustrations of image features against HOI categories on the HICO test set via t-SNE. We compare the features obtained by ViT without any auxiliary task (ViT-only), ViT with non-concept auxiliary tasks (EsViT), and RelViT. Besides those clusters that are identified with the other two baselines, **clusters that can only be identified with RelViT are highlighted.**



Figure 5.6: Visualization of correspondence. The correspondence is extracted between two views of the same image (upper) and two images that belong to the same concept (lower), using the learned model on HICO. **RelViT can extract correspondence on more objects in the two images (semantic correspondence) setting.** Best viewed on screen.

## 5.4 Related Work

**Systematic generalization in visual reasoning.** Systematic generalization [HDM20, BMN18] characterizes to which extent a learning agent can identify and exploit the underlying entities and relations of the training data, and generalize to novel combinations and longer reasoning hops. There has been extensive research on inspecting and tackling systematic generalization in visual reasoning [JHV17, KM18, HMP16a, KC17]. However, most of them only focus on controlled and synthetic domains [RAB20, ZGJ19b, BHS18b, XMY21b,

[NYM20a], while the open-ended real-world domains are largely neglected with very few exceptions [SYH18, TWC20b, JNY22]. In this paper, we tackle systematic generalization in visual relational reasoning with natural images, thereby filling the gap between synthetic and real domains.

**Object-centric and relational representations.** Many seminal research reveals that ML models can benefit from object-centric and relational representations with better sample efficiency and generalization [FS13, DHS20, MZW18]. However, obtaining such representations from unstructured inputs, *i.e.* raw images, still remains challenging [GKK19b, LWU20b, LBL19, YXM21]. Prevalent approaches adopt a latent variable model to explicitly infer the foreground-background split as well as objects & relations [EHW16a, LWP20b, ZM07], while recent findings suggest that they can be an emerging property of transformers trained with self-supervised objectives [CTM21, LYZ21]. Our goal aligns better with the later regime, as it enables implicit representations and thus could be more versatile and efficient. A key difference is that these methods do not exploit concepts in reasoning benchmarks, making them less capable of learning semantic representations.

**Self-supervised learning for ViTs.** The recent resurgence on self-supervised learning (SSL) of image models has delivered impressive results on many few-shot or zero-shot tasks [OLV18]. From a high-level view, these approaches can be categorized into *contrastive* [HFW20, CKN20] and *non-contrastive* [CH21]. However, not all SSL avenues work well with vision transformers (ViTs) and some delicate design may be needed. [CTM21] found their non-contrastive learning objective (DINO) manifested better quantitative results and emerging properties on ViTs. [CXH21] brought similar results on contrastive SSL. [LYZ21] further introduced patch-level SSL objective to ViTs for dense prediction tasks. In this paper, instead of proposing a new SSL approach, we make better use of concepts for ViT training, which can be directly applied to the existing SSL objectives for the improved visual reasoning.

## 5.5 Conclusion

In this paper, our goal is to seek a better inductive bias for visual relational reasoning, especially on real-world data. We found ViTs to be a promising candidate due to their potential on relational reasoning, object-centric learning, and systematic generalization. We further presented RelViT, a simple yet efficient method for exploiting concepts in the visual relational reasoning tasks to boost the performances of ViTs. In specific, we proposed two auxiliary tasks in RelViT : a global task for semantically consistent relational representation, and a local task for learning object-centric semantic correspondence. These two tasks are made possible through the use of our proposed concept-feature dictionary. RelViT largely outperforms other counterparts on two challenging visual relational reasoning benchmarks. While we mainly focus on extending ViTs to visual reasoning using auxiliary tasks, further exploration of combining our work with architectural modification over ViTs to enable better generalization could be a new direction for future work.

### 5.A A formal description of learning in RelViT

Algorithm 1 formally depicts the execution flow of RelViT.

---

**Algorithm 1** RelViT: Concept-guided Vision Transformer

---

**Input:** A set of training images with concepts  $\{(\mathbf{I}_1, C_1), \dots\}$ , an image augmentation function  $\text{aug}(\cdot)$ , momentum update factor  $\lambda$ , loss weight  $\alpha$ , a concept-feature dictionary  $D$ , teacher and student ViT  $g_t$  and  $g_s$ , parameterized by  $\theta_t$  and  $\theta_s$ , respectively.

- 1: **for**  $(\mathbf{I}_i, C_i)$  in  $\{(\mathbf{I}_1, C_1), \dots\}$  **do**
- 2:    $\mathbf{I}_i^{(1)}, \mathbf{I}_i^{(2)} = \text{aug}(\mathbf{I}_i), \text{aug}(\mathbf{I}_i)$
- 3:   Uniformly draw a concept code  $c \sim C_i$ .
- 4:   Retrieve  $Q$  from  $D$  with  $c$ .
- 5:   **if**  $Q$  is not empty **then**
- 6:     Sample feature  $f \sim Q$ , following some sampling tactics.
- 7:      $\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{Global}}(f, g_s(\mathbf{I}_i^{(2)})) + \mathcal{L}_{\text{Local}}(f, g_s(\mathbf{I}_i^{(2)}))$
- 8:     Insert feature  $g_t(\mathbf{I}_i^{(1)})$  into  $Q$ ; if it is full, remove the oldest feature.
- 9:   **else**
- 10:     $\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{Global}}(g_t(\mathbf{I}_i^{(1)}), g_s(\mathbf{I}_i^{(2)})) + \mathcal{L}_{\text{Local}}(g_t(\mathbf{I}_i^{(1)}), g_s(\mathbf{I}_i^{(2)}))$
- 11:   **end if**
- 12:   Update  $\theta_s$  with the loss function  $\mathcal{L} = \mathcal{L}_{\text{main}} + \alpha\mathcal{L}_{\text{aux}}$ .
- 13:   Update  $\theta_t$  using an EMA:  $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$ .
- 14: **end for**

---

## 5.B Additional details on RelViT

### 5.B.1 Input pipeline

We adopt the following data augmentation pipeline for the generating the additional views for our two auxiliary tasks

1. Randomly crop and resize the image into (224, 224) with scale ratio (0.2, 1.0);
2. Randomly jitter the color of the image on brightness, contrast saturation and hue with probability of (0.4, 0.4, 0.4, 0.1), respectively;



Table 5.3: Hyperparameters for RelViT.

Parameter	Value
Optimizer	AdamW with epsilon $1e^{-1}$ (HICO) / $1e^{-5}$ (GQA)
Gradient clipping norm	No grad clipping (HICO) / 0.5 (GQA)
Base learning rate	$1.5e^{-4}$ (HICO) / $3e^{-5}$ (GQA)
Learning rate schedule	0.1 scale with milestones [15, 25] (HICO) / [8, 10] (GQA)
Batch size	16 (HICO) / 64 (GQA)
Total training epochs	30 (HICO) / 12 (GQA)
Temperature $\tau$ in DINO loss	0.04 for teacher and 0.1 for student, we don't use schedule.
Momentum $m$ for teacher	0.999
Center $m$ for center features	0.9
Sampling method	" <i>most-recent</i> " (HICO) / <i>uniform</i> (GQA)
Queue size $ Q $	10

3. Randomly turn the image into gray scale with probability 0.2;
4. Randomly apply Gaussian blur with kernel size 23 and sigma (0.1, 2.0) and probability 0.5;
5. Randomly flip the image horizontally.

Notably, we apply a random crop operation to ensure that all the input images for our auxiliary tasks contain the same number of patches.

### 5.B.2 Hyper-parameters and baselines

Table 5.3 summarizes the hyper-parameters used by RelViT. We inherit most of the parameters from DINO [CTM21].

Table 5.4 summarizes the key details about the loss implementation of different baselines

Table 5.4: Key details about the loss implementation in baselines and RelViT .

	$\mathcal{L}_{\text{Global}}$	$\mathcal{L}_{\text{Local}}$	Compare <code>student(aug(img))</code> with
DINO	x		<code>teacher(aug(img))</code>
EsViT	x	x	<code>teacher(aug(img))</code>
RelViT	x	x	<code>queues[concept(img)].pop()</code>
RelViT + EsViT	x	x	<code>teacher(aug(img))</code> and <code>queues[concept(img)].pop()</code>

and RelViT.

## 5.C Additional details on the datasets

### 5.C.1 HICO

#### 5.C.1.1 Original and systematic splits

Besides the official training/testing split, we adopt the splits for systematic generalization presented in [HPQ20b]. It offers two splits that follow different strategies to select held-out HOI categories. **Systematic-easy** only select *rare* HOI categories (with less than 10 training samples), while **Systematic-hard** select *non-rare* categories instead. Therefore, the training set of **Systematic-hard** will contain much fewer samples and become more challenging. Some basic statistics of these training/testing splits can be found in Table 5.5.

Splits	#Training samples	#Training HOIs	#Testing samples	#Testing HOIs
Original	38118	600	9658	600
Systematic-easy	37820	480	9658	600
Systematic-hard	9903	480	9658	600

Table 5.5: Statistics of the splits of HICO dataset.

### 5.C.1.2 Implementation of $\mathcal{L}_{\text{main}}$

In HICO, there might be multiple HOIs for a single image. We, therefore, formulate the HOI prediction task as a multi-class classification problem. Specifically, the model makes 600 binary classifications and  $\mathcal{L}_{\text{main}}$  in (5.3) is a binary cross-entropy loss.

## 5.C.2 GQA

### 5.C.2.1 Original and systematic splits

We introduce a systematic split for the GQA dataset that is based on reasoning hops. Specifically, we remove those questions that have more than 4 reasoning hops from the training set. Some basic statistics of these training/testing splits can be found in Table 5.6.

Splits	#Training samples	#Testing samples
Original	943000	132062
Systematic	711945	32509

Table 5.6: Statistics of the splits of GQA dataset.

### 5.C.2.2 Reasoning hops

Since all the questions and answers in the GQA dataset are synthetic, it additionally provides “semantics” that characterizes the reasoning procedure that generates the answer from a question and a visual scene. These semantics are composed of multiple “reasoning primitives” that act like functions: receiving arguments and generating output for the next reasoning step. It is believed they can reflect whether a question will require complex multi-hop reasoning – a pivotal angle of systematic generalization. Therefore we develop our systematic split with it. Table 5.7 provides a few examples on semantics.

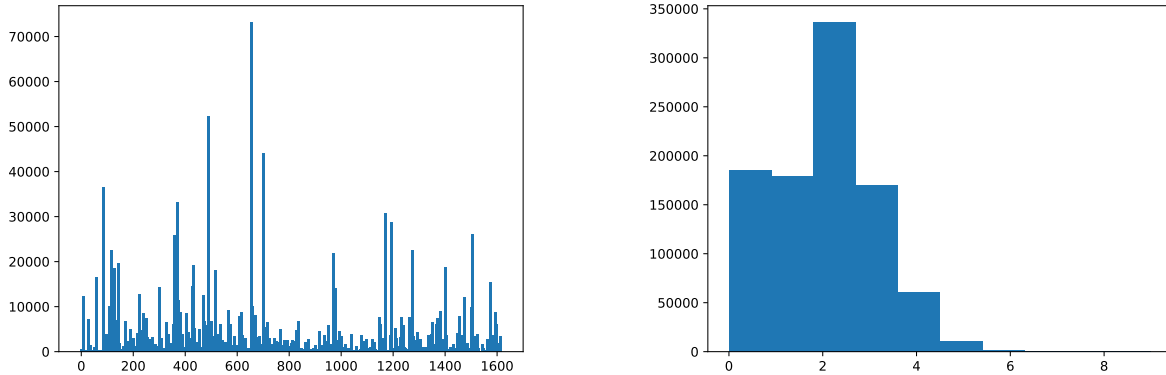
Question	Semantics (Reasoning hops)
Is the pizza with the pepper small and covered?	<pre> relate([0], pizza, with, s(1130674)); filter([1], pizza); verify([2], covered); verify_size([2], small); and([3,4]); </pre>
Do you see any tablecloths or dressers?	<pre> select([], dresser); exist([0], ?); select([], tablecloth); exist([2], ?); or([1, 3], ?); </pre>
Are there microwave ovens to the right of the appliance near the window?	<pre> select([], window); relate([0], appliance, near, s(1297947)) relate([1], microwave, right, s(1297947)); exist([2], ?); </pre>

Table 5.7: Examples of semantics (reasoning hops) in GQA dataset.

### 5.C.2.3 Concept parsing

We obtain the concepts in the GQA dataset by parsing the questions into word tokens. Specifically, we construct a set of concepts that contain nouns, verbs, and adjectives that are with significant meaning. We also manually filter some ambiguous words from this set. The resulting concept set contains 1615 concepts.

We use the python nltk package to process the question. The parsing procedure starts with part-of-speech tagging, where we only keep nouns (NN), verbs (VB) and adjectives (JJ). Then we lemmatize the remaining words to obtain the minimal form of them. Finally, we remove those that do not present in the pre-selected concept list. Additionally, we skip



(a) Histogram of number of questions per concept.

(b) Histogram of number of concepts per question.

Figure 5.7: Histograms of concepts in GQA training set.

questions with “No” as the answer as the question may be unrelated to the image. We provide the statistics of the concepts in GQA in Table 5.8. The number of associated questions of all the 1615 concepts and a histogram on the number of concepts for each question is presented in Figure 5.7a and Figure 5.7b, respectively.

### 5.C.2.4 Implementation of $\mathcal{L}_{\text{main}}$

GQA is formulated as a classification problem, *i.e.* the learner needs to select an answer from the pre-defined answer set; thus  $\mathcal{L}_{\text{main}}$  in (5.3) is a cross-entropy loss.

## 5.D Additional results

### 5.D.1 RelViT with larger backbone models

As we mentioned in Section 5.3.1, the ViT backbone we use (PVTv2-b2) only has **25.4M** parameters, even less than the commonly-used ResNet-101 (**44.7M** parameters). Therefore, we evaluate RelViT with larger ViT backbones: PVTv2-b3 (**45.2M** parameters) and Swin-base (**88M** parameters) [LLC21] and provide the results on HICO and GQA below:

Item	Value
Questions without concept	166217 out of 943000 (17.6%)
Concepts without any question	14
Concepts with < 10 questions	209
Averaged #questions per concept	1030.9
Median #questions per concept	106
Top 20 concepts and their #associated questions	man 52295 animal 44070 furniture 36523 white 33141 front 30779 person 28751 vehicle 26133 woman 25769 bottom 22624 black 22517 device 21962 food 19683 fence 19172 chair 18872 table 18649 hold 18090 shirt 16483 blue 15434 car 14838

Table 5.8: Statistics of concepts in GQA training set.

Table 5.9: Results with larger ViT models on HICO.

HICO mAP	[FCT18]	RelViT + EsViT (PVTv2-b2)	RelViT + EsViT (PVTv2-b3)	RelViT + EsViT (Swin-base)
Original	39.9	40.12	42.61	<b>43.98</b>
Systematic-easy	-	37.21	39.92	<b>42.04</b>
Systematic-hard	-	23.06	25.56	<b>28.36</b>

Table 5.10: Results with larger ViT models on GQA.

GQA overall accuracy	MCAN-Small (w/ bbox)	RelViT (PVTv2-b2)	RelViT (PVTv2-b3)	RelViT (Swin-base)
original	58.35	57.87	61.41	<b>65.54</b>
systematic	36.21	35.48	36.25	<b>37.51</b>

## CHAPTER 6

# Learning Hybrid Latent Representations with LDEBM

### 6.1 Introduction

Text modeling has achieved impressive progress with the fast development of neural generative models [SSB16, LLB17, ZZE17, GAS18, ZKZ18]. It allows near human-level text generation quality and also leads to a wide range of real-world applications such as dialog system [YGT13] and machine translation [BDD93]. Although the quality of generation (*e.g.*, fluency and diversity) is the primary concern of most work, interpretability of the generation process has drawn much attention recently. Among the existing frameworks, the Deep Latent Variable Model (DLVM) is especially suitable for the task, as the learned latent space could capture high-level structures with semantic meanings like topics [WGX19] and dialog actions [ZLE18]; such latent space could further enable more interpretable text modeling, featuring unsupervised text attributes discovery [WMB17], conditional and controllable text generation [FLG19, SZM20], and semi-supervised text classification [PW21].

In essence, DLVM summarizes the observed sample (*e.g.*, a piece of text) into inferred latent variables. Earlier text-modeling methods with DLVM mostly follow the formulation of Variational Auto-Encoder (VAE) [KW13, RMW14, BVV16], which assumes a continuous latent space. More recently, [ZLE18] explore the possibility of using a discrete latent space to capture dialog actions; [SZM20] propose to use VAE with the mixture of Gaussians as the prior, demonstrating promising interpretability of dialog utterance generation. To further improve the expressivity of the latent space, [PW21] leverage the flexibility of



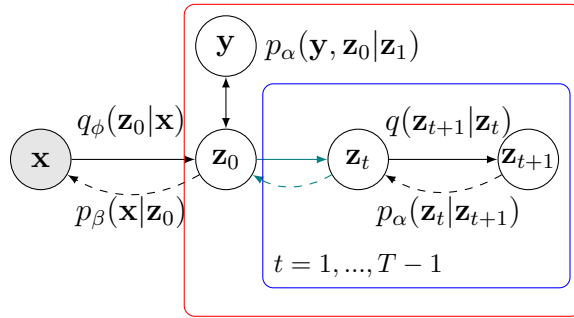


Figure 6.1: **Graphical illustration of the latent diffusion process.** We construct the forward and reverse diffusion processes in the latent space. The symbolic one-hot vector is coupled with the initial latent vector  $\mathbf{z}_0$ . The latent and diffused latent variables are highlighted by the red and blue plates, respectively. The cyan arrows indicate that  $\mathbf{z}_0$  is connected with only  $\mathbf{z}_1$ . We learn a sequence of EBMs to model the reverse diffusion process  $p_\alpha(\mathbf{z}_t|\mathbf{z}_{t+1})$ .

*energy-based prior* [PHN20] and learn a structured latent space for interpretable text generation and classification. Specifically, they propose a symbol-vector coupling prior model. The continuous latent variables are coupled with discrete one-hot symbol variables, allowing better discrete structure induction without sacrificing the generation quality offered by the continuous latent space. However, similar to learning an EBM in data space, the learning of energy-based prior requires Markov Chain Monte Carlo (MCMC) sampling, whose quality can degenerate in practice [GWJ19, NHZ19, NHH20, GSP20], especially on data with complex latent structures; it often leads to instability during training. As we demonstrate empirically in section 6.4.1, this phenomenon is particularly concerning when adopting the variational learning scheme to update model parameters.

To remedy this MCMC sampling issue, we may take a look at the endeavor of EBM learning in general. Among the recent efforts, methods drawn inspiration from the diffusion probabilistic models [SWM15, HJA20, SE20, SSK20] have demonstrated superior results. In particular, [GSP20] propose a diffusion recovery likelihood method to learn and sample from a sequence of EBMs defined on increasingly noisy versions of a dataset; the models are trained by optimizing conditional likelihoods, which are more tractable than the marginal likelihood. It greatly mitigates the burden of sampling during training. A natural question

thus emerges: *Can we leverage the methodology of diffusion models to address the learning issue of energy-based prior?*

In this work, we make the first attempt to address the learning issue of energy-based prior through leveraging diffusion models in the latent space, with a focus on interpretable text modeling. We first unveil the non-trivial symbiosis between latent-space EBMs and diffusion models. Specifically, we focus on the symbol-vector coupling prior; we construct a flexible process that restores the hidden structure in text data by noise-level-aware sampling from a learned sequence of conditional EBMs in the latent space. A variational learning framework is then derived from it. We further employ a geometric clustering-based regularization that complements the symbol-inducing information bottleneck to improve the quality of learned latent space. We term the resulting model LDEBM. Compared to [GSP20], which deals with EBMs in the data space, LDEBM is directly applicable to text data with or without labels; it extracts interpretable latent structures that benefit potential downstream tasks such as semi-supervised classification. Although there are methods using diffusion models in the latent space, some of which have achieved very impressive image generation results, *e.g.*, [VKK21], few of them to our knowledge have explored (unsupervised) symbol induction in the latent space especially on text data. In addition, our method can be trained from scratch and form a well-structured latent space without pretraining, as required by concurrent works on image modeling such as [VKK21] and [NVA21]. In our experiments on generative modeling and interpretable text modeling, LDEBM largely outperforms strong counterparts in terms of both generation quality and interpretability of the learned latent space.

### **Contributions.**

(1) We introduce a novel symbiosis of the latent space EBM and diffusion model in a variational learning framework; the model can be trained from scratch, is directly applicable to text data with or without labels, and shows superior sampling quality. (2) We develop a geometric clustering-based regularization jointly with the information bottleneck that tackles the mode-collapse problem in variational learning of the latent space EBM. (3) Our

experiments demonstrate that the proposed model learns a well-structured latent space and delivers strong results on interpretable text modeling.

## 6.2 Preliminaries: Symbol-Vector Coupling EBM

We assume that for an observed high-dimensional sample  $\mathbf{x} \in \mathbb{R}^D$ , there exists  $\mathbf{z} \in \mathbb{R}^d$  as its compact continuous latent variables. We assume that  $\mathbf{y}$  is the symbolic one-hot vector indicating one of  $K$  categories that  $\mathbf{z}$  belongs to. The complete-data distribution is  $p_\theta(\mathbf{y}, \mathbf{z}, \mathbf{x}) = p_\alpha(\mathbf{y}, \mathbf{z})p_\beta(\mathbf{x}|\mathbf{z})$ , where  $p_\alpha(\mathbf{y}, \mathbf{z})$  is the joint prior model with parameters  $\alpha$ , and  $p_\beta(\mathbf{x}|\mathbf{z})$  is the top-down generation model with parameters  $\beta$ ; henceforth, we use  $\theta = (\alpha, \beta)$  to summarize the parameters. Given  $\mathbf{z}$ ,  $\mathbf{y}$  and  $\mathbf{x}$  are independent; *i.e.*,  $\mathbf{z}$  is sufficient for  $\mathbf{y}$  in this model.

[PW21] propose to formulate the joint prior model,  $p_\alpha(\mathbf{y}, \mathbf{z})$ , as an EBM,

$$p_\alpha(\mathbf{y}, \mathbf{z}) = \frac{1}{Z_\alpha} \exp(\langle \mathbf{y}, f_\alpha(\mathbf{z}) \rangle) p_0(\mathbf{z}), \quad (6.1)$$

where  $p_0(\mathbf{z})$  is a reference distribution, assumed to be the non-informative prior (*e.g.*, isotropic Gaussian or uniform) of the conventional generation model,  $f_\alpha(\mathbf{z}) \in \mathbb{R}^K$  is parameterized by a small multi-layer perceptron, and  $Z_\alpha$  is the normalizing constant or partition function. The energy term  $\langle \mathbf{y}, f_\alpha(\mathbf{z}) \rangle$  in Eq. (6.1) forms an associative memory that couples the symbol  $\mathbf{y}$  and the dense vector  $\mathbf{z}$ . Given  $\mathbf{z}$ ,

$$p_\alpha(\mathbf{y}|\mathbf{z}) \propto \exp(\langle \mathbf{y}, f_\alpha(\mathbf{z}) \rangle) \quad (6.2)$$

becomes a softmax classifier, where  $f_\alpha(\mathbf{z})$  provides the logit scores for the  $K$  categories. Marginally, we have

$$p_\alpha(\mathbf{z}) = \frac{1}{Z_\alpha} \exp(F_\alpha(\mathbf{z})) p_0(\mathbf{z}), \quad (6.3)$$

where the marginal energy term is in a log-sum-exponential form,  $F_\alpha(\mathbf{z}) = \log \sum_{\mathbf{y}} \exp(\langle \mathbf{y}, f_\alpha(\mathbf{z}) \rangle)$ .

It is shown that the coupling between  $\mathbf{z}$  and  $\mathbf{y}$  enables a symbol-aware continuous vector

computation during prior and posterior sampling, which helps to induce a structural latent space [PW21]. Finally, the prior model  $p_\alpha(\mathbf{y}, \mathbf{z})$  stands on a generation model  $p_\beta(\mathbf{x}|\mathbf{z})$ . In text modeling, let  $\mathbf{x} = (\mathbf{x}^{(t)}, t = 1, \dots, T)$  be a sentence, where  $\mathbf{x}^{(t)}$  is the  $t$ -th token.  $p_\beta(\mathbf{x}|\mathbf{z})$  can be defined as a conditional autoregressive model,  $p_\beta(\mathbf{x}|\mathbf{z}) = \prod_{t=1}^T p_\beta(\mathbf{x}^{(t)}|\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t-1)}, \mathbf{z})$ . The complete model  $p_\theta(\mathbf{y}, \mathbf{z}, \mathbf{x})$  with the energy-based prior  $p_\alpha(\mathbf{y}, \mathbf{z})$  and the generation model  $p_\beta(\mathbf{x}|\mathbf{z})$  is termed as Symbol-Vector Coupling Energy-Based Model (SVEBM).

In principle, a SVEBM can be learned through maximizing the log-likelihood function, where the learning gradient is  $\nabla_\theta \log p_\theta(\mathbf{x}) = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})}[\nabla_\theta(\log p_\alpha(\mathbf{z}) + \log p_\beta(\mathbf{x}|\mathbf{z}))]$ . To estimate the expectation, one may sample from the prior  $p_\alpha(\mathbf{z})$  and the posterior  $p_\theta(\mathbf{z}|\mathbf{x})$  with Langevin dynamics [WT11]. Since  $f_\alpha$  is a small network, prior sampling is particularly affordable. In comparison, the posterior sampling can be more expensive as it requires back-propagating through the generation network. One promising solution is to follow the variational learning scheme [KW13] that amortizes the posterior sampling from  $p_\theta(\mathbf{z}|\mathbf{x})$  by an inference network  $q_\phi(\mathbf{z}|\mathbf{x})$ ; MCMC-based sampling can be used for prior samples.

## 6.3 Latent Diffusion Energy-Based Model

### 6.3.1 A Symbiosis between SVEBM and Diffusion Model

Contrasting to the vanilla sampling process of the latent variables in SVEBM, LDEBM follows the philosophy of diffusion probabilistic models [SWM15]; it assumes a sequence of perturbed samples,  $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_T$ , to construct a flexible process that restores the structure in data. First, we define the forward diffusion process that systematically and gradually destroys structure in a data distribution:  $\mathbf{z}_0 \sim q_\phi(\mathbf{z}_0|\mathbf{x}); \mathbf{z}_{t+1} = \sqrt{1 - \sigma_{t+1}^2} \mathbf{z}_t + \sigma_{t+1} \boldsymbol{\epsilon}_{t+1}$ , where  $t = 0, 1, \dots, T - 1$  and  $\boldsymbol{\epsilon}_t$  is the zero-mean standard Gaussian noise. The scaling factor  $\sqrt{1 - \sigma_{t+1}^2}$  ensures that the sequence is a spherical interpolation between the posterior sample and the Gaussian white noise. The forward trajectory and the Markov transition between

each perturbed samples  $\mathbf{z}_1, \dots, \mathbf{z}_T$  are thus

$$\begin{aligned}
 q_\phi(\mathbf{z}_{0:T}|\mathbf{x}) &= q_\phi(\mathbf{z}_0|\mathbf{x}) \prod_{t=0}^{T-1} q(\mathbf{z}_{t+1}|\mathbf{z}_t); \\
 q(\mathbf{z}_{t+1}|\mathbf{z}_t) &= \mathcal{N}(\mathbf{z}_{t+1}; \sqrt{1 - \sigma_{t+1}^2} \mathbf{z}_t, \sigma_{t+1}^2 \mathbf{I}).
 \end{aligned}
 \tag{6.4}$$

Our goal is to learn the generative distribution that describes the same trajectory but in reverse. Inspired by [GSP20], we start by constructing a sequence of *marginal* EBMs at each diffusion step in the latent space. The *conditional* EBMs aiming at recovering  $\mathbf{z}_0$  from noisy inputs then follow as:

$$\begin{aligned}
 p_\alpha(\tilde{\mathbf{z}}_t|\mathbf{z}_{t+1}) &= \\
 &= \frac{1}{\tilde{Z}_{\alpha,t}(\mathbf{z}_{t+1})} \exp\left(F_\alpha(\tilde{\mathbf{z}}_t, t) - \frac{1}{2\sigma_{t+1}^2} \|\tilde{\mathbf{z}}_t - \mathbf{z}_{t+1}\|^2\right),
 \end{aligned}
 \tag{6.5}$$

where  $t = 0, 1, \dots, T - 2$ . We denote  $\tilde{\mathbf{z}}_t = \sqrt{1 - \sigma_{t+1}^2} \mathbf{z}_t$  for brevity.  $F_\alpha(\tilde{\mathbf{z}}_t, t)$  is the neural network that parameterizes the energy function at each diffusion step, and  $\tilde{Z}_{\alpha,t}(\mathbf{z}_{t+1}) = \int \exp(F_\alpha(\tilde{\mathbf{z}}_t, t) - \frac{1}{2\sigma_{t+1}^2} \|\tilde{\mathbf{z}}_t - \mathbf{z}_{t+1}\|^2) d\tilde{\mathbf{z}}_t$  is the partition function of each conditional EBM. For  $t = T - 1$ ,  $p_\alpha(\tilde{\mathbf{z}}_t|\mathbf{z}_{t+1}) = \frac{1}{\tilde{Z}_{\alpha,t}} \exp(F_\alpha(\tilde{\mathbf{z}}_t, t) - \frac{1}{2\sigma_{t+1}^2} \|\tilde{\mathbf{z}}_t\|^2)$  since the diffused samples at time step  $T$  should be close to Gaussian white noise; the distribution of  $\tilde{\mathbf{z}}_{T-1}$  can thus be exponentially tilting of a zero-mean Gaussian distribution.

Eq. (6.5) shares the idea of denoising generative modeling [BYA13], where a denoising autoencoder is trained by maximizing the conditional probabilities of the observed samples given their noisy versions. Compared to the vanilla definition (see Eq. (6.3)), the noise-level-aware quadratic term constrains the energy landscape to be localized around the noisy sample; this makes the latent space much less multi-modal and easier to sample from. To be specific, [GSP20] show that  $p_\alpha(\tilde{\mathbf{z}}_t|\mathbf{z}_{t+1})$  is approximately a single-mode Gaussian distribution when  $\sigma$  is sufficiently small; it greatly reduces the burden of MCMC sampling. After sampling  $\tilde{\mathbf{z}}_t$  from the model, we can easily obtain  $\mathbf{z}_t = \tilde{\mathbf{z}}_t / \sqrt{1 - \sigma_{t+1}^2}$ .

Next, we show that the forward and reverse process in the latent space can be naturally integrated into the variational learning scheme to amortize the time-consuming posterior

sampling. Similar to VAE, the ELBO in SVEBM is

$$\begin{aligned} \text{ELBO}_{\theta,\phi} &= \log p_{\theta}(\mathbf{x}) - \mathbb{D}_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x})) \\ &= \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log p_{\beta}(\mathbf{x}|\mathbf{z})] - \mathbb{D}_{\text{KL}}(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\alpha}(\mathbf{z})), \end{aligned} \tag{6.6}$$

where  $\mathbb{D}_{\text{KL}}$  denotes the Kullback-Leibler divergence. Since we now consider the full trajectory of the perturbed samples, in LDEBM we may optimize

$$\begin{aligned} \text{ELBO}_{\text{Diff},\theta,\phi} &= \mathbb{E}_{q_{\phi}(\mathbf{z}_0|\mathbf{x})} [\log p_{\beta}(\mathbf{x}|\mathbf{z}_0) - \log q_{\phi}(\mathbf{z}_0|\mathbf{x})] \\ &\quad + \mathbb{E}_{q_{\phi}(\mathbf{z}_0|\mathbf{x}),q(\mathbf{z}_{1:T}|\mathbf{z}_0)} \left[ \log \frac{p_{\alpha}(\mathbf{z}_{0:T})}{q(\mathbf{z}_{1:T}|\mathbf{z}_0)} \right], \end{aligned} \tag{6.7}$$

which is a valid ELBO by applying Jensen’s inequality to Eq. (6.6). The joint training of inference, prior and generation models can be largely reduced to finding the agreement of the forward and reverse Markov transitions defined by  $q_{\phi}$  and  $p_{\theta}$ , respectively.

Finally, we show how to introduce the symbolic one-hot vector  $\mathbf{y}$  into our formulation. We assume a complete data distribution that considers the full trajectory of the perturbed latent variables,  $p_{\theta}(\mathbf{y}, \mathbf{z}_{0:T}, \mathbf{x})$ . Among several possibilities for coupling the symbolic vector  $y$  with the latent variables, two major options arise: We can couple the symbol with the whole trajectory, *i.e.*,  $p_{\theta}(\mathbf{y}, \mathbf{z}_{0:T}, \mathbf{x}) = p_{\alpha}(\mathbf{y}, \mathbf{z}_{0:T})p_{\beta}(\mathbf{x}|\mathbf{z}_{0:T})$ ; or we can couple the symbol with only the clean posterior sample  $\mathbf{z}_0$ , *i.e.*,  $p_{\theta}(\mathbf{y}, \mathbf{z}_{0:T}, \mathbf{x}) = p(\mathbf{z}_T)p_{\alpha}(\mathbf{y}, \mathbf{z}_0|\mathbf{z}_1) \prod_{t=1}^{T-1} p_{\alpha}(\mathbf{z}_t|\mathbf{z}_{t+1})p_{\beta}(\mathbf{x}|\mathbf{z}_0)$ . We prefer the latter one, since it is sufficient to model the reverse Markovian transition, while enabling a simpler and more efficient training scheme following [HJA20] (see section 6.3.4). Of note, coupling only  $\mathbf{z}_0$  to  $\mathbf{y}$  means that we condition only the final reverse diffusion step  $[\mathbf{z}_0|\mathbf{z}_1]$  on  $\mathbf{y}$  when performing controllable generation. This could be a bit counter-intuitive as no label information is injected in previous reverse steps. Theoretically,  $\mathbf{y}$  and  $\mathbf{z}_{1:T}$  are independent given  $\mathbf{z}_0$  in our formulation; however, we empirically observe that  $\mathbf{y}$  and  $\mathbf{z}_t$  for  $t > 0$  are nearly independent even marginally, after we integrating out  $\mathbf{z}_{0:t-1}$  in our model. In other words,  $p_{\alpha}(\mathbf{y}|\mathbf{z}_t)$ ,  $t > 0$  are in general non-informative since adding noise in the latent space could be much more corrupting than adding noise in the data space. The model learns to enjoy the less multi-modal energy landscape in previous reverse steps; it then seeks

the given mode only in the most informative final reverse step. Specifically, we achieve this coupling by similarly defining  $p_\alpha(\mathbf{y}, \mathbf{z}_0|\mathbf{z}_1)$  as in Eq. (6.1) and using the log-sum-exponential form for learning as in Eq. (6.3). Please refer to figure 6.1 for a graphical illustration of our model.

### 6.3.2 Information Bottleneck

To learn the symbolic vector  $\mathbf{y}$ , we may consider adopting the Information Bottleneck (IB) principle [TPB00], an appealing approach for inducing symbolic representations. In this section, we re-interpret the above ELBO as a cooperative learning objective, defined as the divergence between two joint distributions; we then show how this formulation helps to incorporate the IB-based regularization into LDEBM in a principled manner.

As shown in [HNF19], the variational learning scheme can be regarded as performing alternating projection between two joint distributions,  $Q_\phi$  and  $P_\theta$ . In our modeling, we have:  $Q_\phi(\mathbf{x}, \mathbf{z}_{0:T}) = q_{data}(\mathbf{x})q_\phi(\mathbf{z}_{0:T}|\mathbf{x})$ , and  $P_\theta(\mathbf{x}, \mathbf{z}_{0:T}) = p(\mathbf{z}_T) \prod_{t=0}^{T-1} p_\alpha(\mathbf{z}_t|\mathbf{z}_{t+1})p_\beta(\mathbf{x}|\mathbf{z}_0)$ ; we use  $q_{data}(\mathbf{x})$  to denote the data distribution of  $\mathbf{x}$  for notation consistency. Maximizing  $\mathbb{E}_{q_{data}(\mathbf{x})}[\text{ELBO}_{Diff, \theta, \phi}(\mathbf{x})]$  over  $(\theta, \phi)$  is equivalent to minimizing the following divergence:

$$\begin{aligned} \mathbb{D}_{\text{KL}}(Q_\phi \| P_\theta) &= \mathbb{D}_{\text{KL}}(q_{data}(\mathbf{x}) \| p_\theta(\mathbf{x})) \\ &+ \mathbb{E}_{q_{data}(\mathbf{x})}[\mathbb{D}_{\text{KL}}(q_\phi(\mathbf{z}_{0:T}|\mathbf{x}) \| p_\theta(\mathbf{z}_{0:T}|\mathbf{x}))], \end{aligned} \tag{6.8}$$

since  $\mathcal{H}(\mathbf{x}) = -\mathbb{E}_{q_{data}(\mathbf{x})}[\log q_{data}(\mathbf{x})]$ , *i.e.*, the entropy of data distribution is fixed. Minimizing the KL-divergence  $\min_\theta \min_\phi \mathbb{D}_{\text{KL}}(Q_\phi \| P_\theta)$  defines a cooperative game, with the dynamics that  $q_\phi$  and  $p_\theta$  run towards each other.

Since the initial posterior sample  $\mathbf{z}_0$  is coupled with the symbolic vector  $\mathbf{y}$ , it should be the most informative latent variable for inducing the discrete symbol. We can therefore plug in Eq. (6.8) with a mutual information term between  $\mathbf{z}_0$  and  $\mathbf{y}$ :  $\mathcal{I}(\mathbf{z}_0, \mathbf{y}) = \mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y}|\mathbf{z}_0)$ , which essentially incorporates the IB as we show below. Given the distribution  $Q_\phi(\mathbf{x}, \mathbf{z}_{0:T})$ , we can first define the marginal distribution of  $\mathbf{z}_0$  as the aggregated posterior by integrating

out  $\mathbf{z}_{1:T}$ :  $q_\phi(\mathbf{z}_0) = \mathbb{E}_{q_{data}(\mathbf{x})}[q_\phi(\mathbf{z}_0|\mathbf{x})]$ . The entropy of  $\mathbf{z}_0$  and conditional entropy of  $\mathbf{z}_0$  on  $\mathbf{x}$  then follow as  $\mathcal{H}(\mathbf{z}_0)$  and  $\mathcal{H}(\mathbf{z}_0|\mathbf{x})$ , respectively. Taken together, the KL-Divergence with  $\lambda\mathcal{I}(\mathbf{z}_0, \mathbf{y})$  can therefore be parsed as

$$\begin{aligned}\mathcal{L} &= \mathbb{D}_{\text{KL}}(Q_\phi\|P_\theta) - \lambda\mathcal{I}(\mathbf{z}_0, \mathbf{y}) \\ &= \mathcal{C} + \mathcal{L}_{\text{RC}} + \mathcal{L}_{\text{EBM}} + \mathcal{L}_{\text{IB}},\end{aligned}\tag{6.9}$$

where  $\mathcal{C} = -\mathcal{H}(\mathbf{x}) + \sum_{t=0}^{T-1} \mathcal{H}(\mathbf{z}_{t+1}|\mathbf{z}_t)$  does not involve learnable parameters,  $\mathcal{L}_{\text{RC}} = -\mathbb{E}_{Q_\phi}[\log p_\beta(\mathbf{x}|\mathbf{z}_0)]$  is the reconstruction loss,  $\mathcal{L}_{\text{EBM}} = \mathbb{D}_{\text{KL}}(q_\phi(\mathbf{z}_0)\|p_\alpha(\mathbf{z}_{0:T}))$  corresponds with learning latent space models, and  $\mathcal{L}_{\text{IB}} = \mathcal{I}(\mathbf{x}, \mathbf{z}_0) - \lambda\mathcal{I}(\mathbf{z}_0, \mathbf{y})$  is the IB, where  $\mathcal{I}(\mathbf{x}, \mathbf{z}_0) = \mathcal{H}(\mathbf{z}_0) - \mathcal{H}(\mathbf{z}_0|\mathbf{x})$  is the mutual information between  $\mathbf{x}$  and  $\mathbf{z}_0$  under  $Q_\phi$ ;  $\lambda \geq 0$  controls the expressivity of  $\mathbf{z}_0$  to  $\mathbf{y}$ .

### 6.3.3 Geometric Clustering Anchors the Modes

As shown in the previous section, IB provides an elegant solution for inducing the symbolic vector  $\mathbf{y}$ . In this section, we further introduce an approach that facilitates the emergence of  $\mathbf{y}$  from a geometric perspective. To induce a latent space with interpretable structures, ideally, the location of data points in the latent space encodes their semantic meaning, *i.e.*, it indicates the semantic class; semantically similar points should be placed closer and produce the same symbolic vector  $\mathbf{y}$ . This resembles geometric clustering algorithms: Labels of data points are assigned based on their geometric (typically Euclidean) distance from each other. Below, we show how to realize this intuition in LDEBM.

Let us consider the joint distribution  $p_\theta(\mathbf{x}, \mathbf{y})$ . We can decompose its log-likelihood into  $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{x}) + \log p_\theta(\mathbf{y}|\mathbf{x})$  as in [GWJ19], where  $\log p_\theta(\mathbf{x})$  is substituted with the ELBO derived in section 6.3.1.  $p_\theta(\mathbf{y}|\mathbf{x})$  is the classification model in the latent space:  $p_\theta(\mathbf{y}|\mathbf{x}) \approx \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{x})}[p_\alpha(\mathbf{y}|\mathbf{z}_0)]$ .  $p_\alpha(\mathbf{y}|\mathbf{z}_0)$  is the softmax classifier of  $\mathbf{y}$  based on  $\mathbf{z}_0$  similarly as in Eq. (6.2). Therefore, we can encode the semantic information from the label  $\mathbf{y}$  into  $\mathbf{z}_0$  through learning the classifier  $p_\alpha(\mathbf{y}|\mathbf{z}_0)$ . In case there is full or partial access to the



ground-truth semantic class labels, we could directly utilize these labels to supervise the classifier, jointly with the existing ELBO objective. When no label is provided, we generate pseudo label  $\hat{\mathbf{y}}$  by clustering  $\mathbf{z}_0$ , which optimizes  $\mathbb{E}_{\mathbf{y}} \log p_{\theta}(\mathbf{x}, \mathbf{y})$  instead;  $\mathbb{E}_{\mathbf{y}}$  is defined by the clustering algorithm. It is akin to the EM algorithm, where geometric clustering serves as a hard-decision E-step to help induce  $\mathbf{y}$ . In practice, we employ K-means to cluster  $\mathbf{z}_0$ . In section 6.4.1, we empirically show that this strategy learns a better latent space and significantly alleviates the mode-collapse problem.

### 6.3.4 Algorithms and Implementation

#### Training and sampling algorithms.

For learning the prior model, we have for each  $t = 0, 1, \dots, T - 1$ :

$$\begin{aligned} \nabla_{\alpha} \text{ELBO}_t &= \mathbb{E}_{q_{\phi}(\tilde{\mathbf{z}}_t, \mathbf{z}_0 | \mathbf{x})} [\nabla_{\alpha} F_{\alpha}(\tilde{\mathbf{z}}_t, t)] \\ &\quad - \mathbb{E}_{q_{\phi}(\mathbf{z}_{t+1}, \mathbf{z}_0 | \mathbf{x}), p_{\alpha}(\tilde{\mathbf{z}}_t | \mathbf{z}_{t+1})} [\nabla_{\alpha} F_{\alpha}(\tilde{\mathbf{z}}_t, t)]. \end{aligned} \tag{6.10}$$

Let  $\psi = \{\beta, \phi\}$  collect the parameters of the inference (encoder) and generation (decoder) models.

$$\begin{aligned} \nabla_{\psi} \text{ELBO} &= \nabla_{\psi} \mathbb{E}_{q_{\phi}(\mathbf{z}_0 | \mathbf{x})} [\log p_{\beta}(\mathbf{x} | \mathbf{z}_0) - \log q_{\phi}(\mathbf{z}_0 | \mathbf{x})] \\ &\quad - \nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{z}_{0:T} | \mathbf{x})} \left[ \log p(\mathbf{z}_T) + \sum_{t=0}^{T-1} \log p_{\alpha}(\mathbf{z}_t | \mathbf{z}_{t+1}) \right]. \end{aligned} \tag{6.11}$$

Recall that we denote  $\tilde{\mathbf{z}}_t = \sqrt{1 - \sigma_{t+1}^2} \mathbf{z}_t$ .  $\mathbb{E}_{p_{\alpha}(\tilde{\mathbf{z}}_t | \mathbf{z}_{t+1})}$  is approximated by MCMC samples from the prior.  $\mathbb{E}_{q_{\phi}(\mathbf{z}_0 | \mathbf{x})}$  is approximated by samples from the inference network. We also add the gradient from  $\mathcal{I}(\mathbf{z}_0, \mathbf{y})$ , denoted as  $\nabla \mathcal{I}$ , to Eqs. (6.10) and (6.11) during training to incorporate IB.

Note that the expectation in Eq. (6.10) requires MCMC sampling (*e.g.*, Langevin dynamics [WT11]) of the prior model. For a target distribution  $\pi(\tilde{\mathbf{z}})$ , the dynamics iterates  $\tilde{\mathbf{z}}^{k+1} = \tilde{\mathbf{z}}^k + \frac{s^2}{2} \nabla_{\tilde{\mathbf{z}}} \log \pi(\tilde{\mathbf{z}}^k) + s \boldsymbol{\epsilon}^k$ , where  $k$  indexes the iteration of the dynamics,  $s$  is a small step size, and  $\boldsymbol{\epsilon}^k \sim \mathcal{N}(0, \mathbf{I})$  is the Gaussian noise. In this work, we follow the heuristics

---

**Algorithm 2 Learning algorithm.**

---

**series input:** initial parameters  $(\alpha, \beta, \phi)$ , learning rate  $\eta$ , observed unlabeled examples  $\{\mathbf{x}^{(i)}\}_{i=1}^M$ , observed labeled examples  $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=M+1}^{M+N}$  (alternative, needed in controllable generation or semi-supervised learning).

**repeat**

**series posterior sampling:** For each  $\mathbf{x}^{(i)}$ , sample  $\mathbf{z}_0^{(i)} \sim q_\phi(\mathbf{z}_0|\mathbf{x}^{(i)})$  using inference network.

**series prior sampling:** For each  $\mathbf{z}_0^{(i)}$ , sample diffusion step  $t$  from  $\text{Unif}(\{0, \dots, T-1\})$ , and the perturbed pair  $(\tilde{\mathbf{z}}_t^{(i)}, \mathbf{z}_{t+1}^{(i)})$  following Eq. (6.4). Set  $\tilde{\mathbf{z}}_t^{(i)}$  as the positive sample  $\tilde{\mathbf{z}}_t^{(i)+}$ . Initialize the MCMC using  $\mathbf{z}_{t+1}^{(i)}$  and update by Eq. (6.12) for  $K$  steps to obtain  $\tilde{\mathbf{z}}_t^{(i)-}$ .

**series learning prior model:** Update  $\alpha$  with

$$\eta(\sum_i [\nabla_\alpha F_\alpha(\tilde{\mathbf{z}}_t^{(i)+}, t) - \nabla_\alpha F_\alpha(\tilde{\mathbf{z}}_t^{(i)-}, t)] - \nabla_\alpha \mathcal{I}).$$

**series learning inference and generation models:**

Update  $\beta$  and  $\phi$  with Eq. (6.11) and  $\nabla_\phi \mathcal{I}$ .

**if** labeled data  $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$  is available **then**

series update  $\gamma = (\alpha, \phi)$  using  $\mathbf{y}^{(i)}$ :

Learning gradient  $\eta \sum_i \nabla_\gamma \log p_{\alpha_t}(\mathbf{y}^{(i)}|\mathbf{z}_0^{(i)})$  is provided by ground-truth label.

**else if** only unlabeled data is available **then**

series update  $\gamma = (\alpha, \phi)$  using pseudo-label  $\hat{\mathbf{y}}^{(i)}$ :

Geometric clustering generates  $\hat{\mathbf{y}}^{(i)}$  for each  $\mathbf{x}^{(i)}$ .  $\eta \sum_i \nabla_\gamma \log p_{\alpha_t}(\hat{\mathbf{y}}^{(i)}|\mathbf{z}_0^{(i)})$ , *i.e.*, the gradient comes from pseudo-label generated by geometric clustering.

**end if**

**until** converged.

---

in [GSP20] and set the step size  $s_t = b\sigma_t c_t$ , where  $b < 1$  is a tuned hyperparameter, and  $c_t = \sqrt{\prod_{i=1}^t \sigma_i/\sigma_1}$  is a scaling factor. Let  $t$  indexes the diffusion step;  $K$  steps of Langevin dynamics thus iterates

$$\begin{aligned} \tilde{\mathbf{z}}_t^{k+1} &= \tilde{\mathbf{z}}_t^k + \frac{b^2 \sigma_t^2 c_t^2}{2} \left( \nabla_{\tilde{\mathbf{z}}} F_\alpha(\tilde{\mathbf{z}}_t^k, t) - \frac{1}{\sigma_t^2} (\tilde{\mathbf{z}}_t^k - \mathbf{z}_{t+1}) \right) \\ &\quad + b\sigma_t c_t \boldsymbol{\epsilon}^k. \end{aligned} \tag{6.12}$$

---

**Algorithm 3** Synthesizing algorithm.

---

series input:  $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$   
series output:  $\mathbf{z}_0$   
**for**  $t = T - 1$  series to  $t = 0$  **do**  
    Initialize  $\tilde{\mathbf{z}}_t = \mathbf{z}_{t+1}$ .  
    **for**  $k = 1$  series to  $K$  **do**  
        Update  $\tilde{\mathbf{z}}_t$  using Eq. (6.12).  
    **end for**  
     $\mathbf{z}_t = \tilde{\mathbf{z}}_t / \sqrt{1 - \sigma_{t+1}^2}$   
**end for**

---

In principle, training the model amounts to minimizing the ELBO in Eq. (6.7), which requires a summation over all the diffusion steps; it involves sampling a full forward trajectory. To make the training simpler and more efficient, following [HJA20], we randomly choose one diffusion step from the summation to optimize at each training iteration. After training, we initialize the reverse trajectory from Gaussian white noise. The synthesized sample at each step serves to initialize an MCMC that samples from the model of the previous step. The learning and synthesizing algorithms are summarized in Algorithms 2 and 3, respectively.

**Implementation.**

For the K-means algorithm, we use the implementation of [JDJ19], which explicitly deals with the empty clusters and trivial parameterization problems. To emphasize that the proposed model shows better capability of modeling latent space, we use the same encoder and decoder as [PW21] for all the experiments. We use a shared network  $F_\alpha(\tilde{\mathbf{z}}_t, t)$  for each  $t = 0, 1, \dots, T - 1$ ;  $T = 6$ ;  $t$  is encoded by sinusoidal position embedding as in [HJA20], and we set  $\sigma_t^2$  to increase linearly. For Langevin dynamics, we use  $K = 50$  and  $b^2 = 0.002$  throughout the experiments.

## 6.4 Experiments

Through a series of experiments, we empirically examine the capability of our model for generative modeling and interpretability on text modeling tasks.

### 6.4.1 Generative Modeling

#### 2D synthetic data.

We first perform experiments of our model on 2D synthetic datasets as a sanity check to validate our assumptions; results are displayed in figure 6.2. The gap between LDEBM and SVEBM is very clear. As mentioned in section 6.1, for more complex datasets (*e.g.*, datasets with more modes or more complex data structure), SVEBM struggles to capture regularities in the data; the model is prone to collapse, which features an exploding KL-term and poor performance on generation. In contrast, LDEBM without geometric clustering already overcomes this problem, performing relatively well in terms of modeling both *posterior*  $\mathbf{x}$  and *prior*  $\mathbf{x}$ . Although LDEBM without geometric clustering faithfully reconstructs the data and shows significant improvement on generation quality, the generated distribution can be slightly distorted, and some modes are missing. The problem is clearer in the latent space: Mode-collapse occurs in the *prior*  $\mathbf{z}$  distribution, where the latent structure is broken. LDEBM with geometric clustering maintains the number of modes as in the data distribution and induces a highly-structural latent space, echoing our intuition in section 6.3.3. figure 6.3 shows the structural similarity between data distribution and the learned latent distribution.

#### Language generation.

Following previous state-of-the-art competitors [ZLE18, SZM20, PW21], we evaluate the quality of generation on a real-world text dataset, Penn Treebanks (PTB) [MMS93] as pre-processed by [MKB10]. We report four metrics of the generation performance: Reverse Perplexity (rPPL) [ZKZ18], BELU [PRW02], Word-Level KL Divergence (wKL), and Negative Log-Likelihood (NLL); table 6.1 summarizes results.

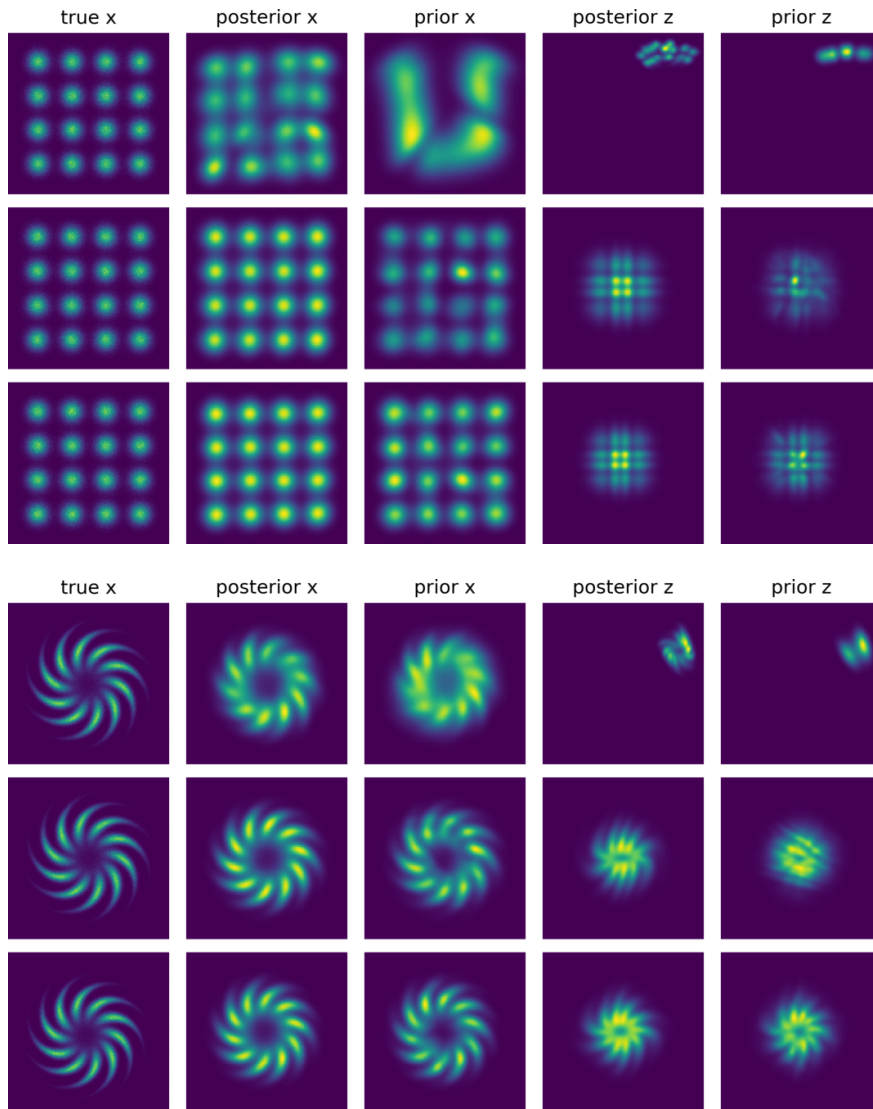


Figure 6.2: **Evaluation on 2D synthetic data:** a mixture of 16 Gaussians (upper panel) and a 10-arm pinwheel-shaped distribution (lower panel). In each panel, the top, middle, and bottom row display densities learned by SVEBM-IB, our model w/o geometric clustering, and our full model, respectively. In each row, from left to right, it displays the data distribution and the Kernel Density Estimations (KDEs) of:  $\mathbf{x}$  generated by amortized posterior  $\mathbf{z}$  samples,  $\mathbf{x}$  by MCMC sampled prior  $\mathbf{z}$  samples, posterior  $\mathbf{z}$  samples, and prior  $\mathbf{z}$  samples.

The proposed model, either w/ or w/o geometric clustering, demonstrates the best performance on reconstruction (highest BLEU) and fitting capacity (lowest NLL) than all baseline models. Moreover, the higher expressivity of our models enables the generation of high-

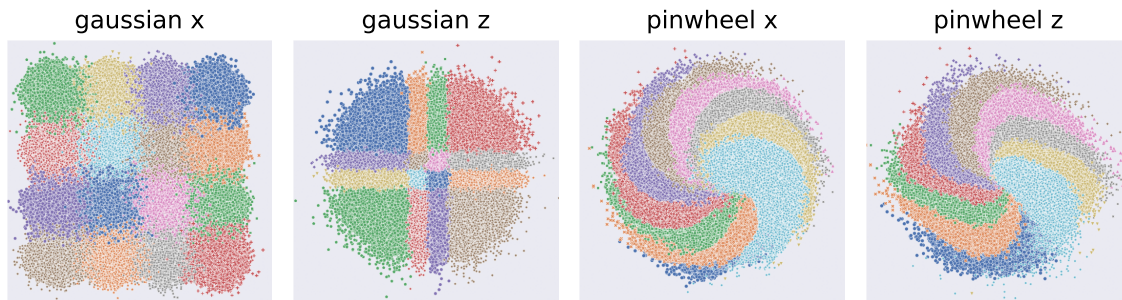


Figure 6.3: **Visualization of color-coded data points.** We visualize data points and the corresponding inferred latent variables of two 2D synthetic datasets (*gaussian* and *pinwheel*). Data points with different labels are assigned with different colors.

quality sentences. The lowest rPPL indicates that our models can further improve over these strong baselines on fluency and diversity of generated text; the lowest wKL indicates that the word distribution of the generated sentences is the most consistent with that of the original data.

### Sentence completion.

Further, we show that the trained model enables text completion on a masked JerichoWorld dataset [AR21]. We perform conditional sampling in the latent space to complete the masked sentences.

## 6.4.2 Interpretable Text Modeling

In this section, we move on to evaluate our model on the interpretability of text modeling.

### Unsupervised text attributes discovery.

First, we examine the efficacy of our model on the unsupervised text attributes discovery task. We assess the model on the DD dataset [LSS17], a chat-oriented dataset of 13,118 daily conversations with human-annotated dialog action and emotion labels for the utterances. The interpretability is evaluated through the ability to unsupervisedly capture the utterance attributes of DD. We flatten the dialogues for text modeling and use  $p_{\theta}(\mathbf{y}|\mathbf{x})$  to infer the utterance label. In particular, we take the *argmax* of the classification head as the

Table 6.1: **Results of language generation on PTB dataset.** We highlight our model results in gray color. The best and second-best performances are marked in bold numbers and underlines, respectively; tables henceforth follows this format.

Model	rPPL $\downarrow$	BLEU $\uparrow$	wKL $\downarrow$	NLL $\downarrow$
Test Set	-	100.0	0.14	-
RNN-LM	-	-	-	101.21
AE	730.81	10.88	0.58	-
VAE	686.18	3.12	0.50	100.85
DAE	797.17	3.93	0.58	-
DVAE	744.07	1.56	0.55	101.07
DI-VAE	310.29	4.53	0.24	108.90
semi-VAE	494.52	2.71	0.43	100.67
semi-VAE + $\mathcal{I}$	260.28	5.08	0.20	107.30
GM-VAE	983.50	2.34	0.72	99.44
GM-VAE + $\mathcal{I}$	287.07	6.26	0.25	103.16
DGM-VAE	257.68	8.17	0.19	104.26
DGM-VAE + $\mathcal{I}$	247.37	8.67	0.18	105.73
SVEBM	180.71	9.54	0.17	95.02
SVEBM-IB	177.59	9.47	0.16	94.68
Ours w/o GC	<u>168.32</u>	<u>11.12</u>	<u>0.07</u>	<b>79.84</b>
Ours	<b>164.57</b>	<b>11.16</b>	<b>0.06</b>	<u>82.38</u>

inferred label. Following [ZLE18], we recruit homogeneity to evaluate the consistency between ground-truth action and emotion labels and those inferred from our model. table 6.3 displays the results of our model and baselines. It shows that the proposed model outperform other baselines in reconstruction by a large margin and give a much better homogeneity on both the dialog action and emotion. The superior performance of LDEBM equipped with

Table 6.2: **Sentence completion on JerichoWorld dataset.** The gray words in the input sentences are masked with <unk> token.

Input	<p>... A bathroom lies to the south, while a door to the east leads to the living room. On the bed are a driver’s license, some keys and a wallet</p> <p>On the end table is a telephone.</p>
Pred.	<p>... A bathroom lies to the south, while a door to the east leads to the living room. On the bed is a wallet. On the end table are a telephone and some keys.</p>
Input	<p>... All around you the crowd is in a state of pandemonium. The paths of least resistance are up, down and west.</p>
Pred.	<p>... All around you the crowd is in a state of pandemonium. The paths of least resistance are down and east.</p>

latent space geometric clustering again verifies our intuition in section 6.3.3.

### Conditional response generation.

Next, we evaluate our model on dialog generation with SMD [EKC17] and DD datasets. We evaluate the quality of generated responses using BELU and three word-embedding-based topic similarity metrics [SZM20]: embedding average [ML08], embedding extrema [FPL14], and embedding greedy [RL12]. table 6.4 shows that LDEBM has competitive performance compared with SVEBM-IB on SMD and outperforms the strong baselines on all metrics on DD; see qualitative examples in tables 6.5 and 6.6.

### Sentence sentiment control.

Finally, we inspect the capability of our model for controllable generation on Yelp reviews,



Table 6.3: **Results of interpretable text modeling on DD.** We use mutual information (MI), BLEU, and homogeneity with actions and emotions for evaluation.

Model	MI <sup>↑</sup>	BLEU <sup>↑</sup>	Act. <sup>↑</sup>	Emo. <sup>↑</sup>
DI-VAE	1.20	3.05	0.18	0.09
semi-VAE	0.03	4.06	0.02	0.08
semi-VAE + $\mathcal{I}$	1.21	3.69	0.21	0.14
GM-VAE	0.00	2.03	0.08	0.02
GM-VAE + $\mathcal{I}$	1.41	2.96	0.19	0.09
DGM-VAE	0.53	7.63	0.11	0.09
DGM-VAE + $\mathcal{I}$	1.32	7.39	0.23	0.16
SVEBM	0.01	11.16	0.03	0.01
SVEBM-IB	2.42	10.04	0.59	0.56
Ours w/o GC	<u>2.44</u>	<u>16.72</u>	<u>0.65</u>	<u>0.63</u>
Ours	<b>3.94</b>	<b>28.75</b>	<b>0.74</b>	<b>0.74</b>

Table 6.4: **Dialog evaluation results on SMD and DD.** Models are assessed using four metrics: BLEU, average, extrema, and greedy word embedding based similarity.

Data	Model	BLEU <sup>↑</sup>	Avg. <sup>↑</sup>	Extr. <sup>↑</sup>	Grdy. <sup>↑</sup>
SMD	DI-VAE	7.06	76.17	43.98	60.92
	DGM + $\mathcal{I}$	10.16	78.93	48.14	64.87
	SVE-IB	<b>12.01</b>	<b>80.88</b>	<u>51.35</u>	<u>67.12</u>
	w/o GC	11.44	<u>80.16</u>	51.26	66.51
	Ours	<u>11.51</u>	<b>80.88</b>	<b>51.57</b>	<b>67.13</b>
DD	DGM + $\mathcal{I}$	2.19	74.73	<u>45.85</u>	<u>64.28</u>
	SVE-IB	<u>2.23</u>	<u>77.37</u>	43.32	63.99
	Ours	<b>3.72</b>	<b>78.89</b>	<b>46.19</b>	<b>65.87</b>

Table 6.5: **Samples of unsupervisedly discovered action categories and corresponding utterances on SMD.**

Action	Request-weather
Utterance	I need to know if it is going to be foggy in Fresno today and tomorrow car.
	Manhattan, please.
	Will it be cloudy on Monday?
	I need current weather data about New York, specifically information about the temperature.
Action	Request-city
Utterance	In what city are you interested?
	What city would you like to know the weather about?
	Okay, what city should I look in?

pre-processed by [LJH18]. The Yelp dataset is of larger scale, containing 180,000 negative reviews and 270,000 positive ones. For a controllable generation process, the symbolic vector  $\mathbf{y}$  is provided to guide the sampling in latent space. Following [PW21], we train the model with sentiment supervision and use the same pre-trained classifier to determine the sentiment of the generated sentence. The pre-trained classifier has an accuracy of 98.5% on the testing data and thus can accurately evaluate the sentiment of given sentences. The quantitative and qualitative results are summarized in tables 6.7 and 6.8, respectively. LDEBM generates positive and negative reviews with a nearly saturate accuracy, significantly outperforming all the baselines.

Table 6.6: **Dialog cases generated by LDEBM given the context.** On SMD, we provide the same context but with different  $\mathbf{y}$  values to generate each response; actions indicated by  $\mathbf{y}$  are listed in parentheses. On DD, LDEBM can well capture the dialog topic; we provide the ground-truth response in each case for reference.

SMD	
Ctx.	<i>User:</i> What gas stations are here? <i>Sys:</i> There is a Chevron.
Ref.	That’s good! Please pick the quickest route to get there and avoid all heavy traffic!
Pred.	(Req.-address) What is the address? (Req.-route) Please set the quickest route to go.
DD	
Ctx.	<i>A:</i> Hi. Have you got a personal computer? <i>B:</i> Certainly. What ’ s the matter? <i>A:</i> I wonder if you often trade with others on the internet.
Ref.	Sure. I often buy things or do business through it without going out to the physical stores.
Pred.	Yes, but I think it is a little different way.

### 6.4.3 Semi-supervised Classification

In this experiment, we switch from neural sequence models used in previous experiments to neural document models [MYB16, CTS18]; we show our model can be similarly extended to semi-supervised settings as in [PW21] and benefit from the better learned latent space. Our model is evaluated on AGNews [ZZL15], a popular benchmark for text classification with 127,600 documents from 4 classes. table 6.9 shows that LDEBM performs the best when

Table 6.7: **Accuracy of sentence attribute control on Yelp.**

Model	Overall <sup>†</sup>	Positive <sup>†</sup>	Negative <sup>†</sup>
DGM-VAE + $\mathcal{I}$	64.7%	95.3%	34.0%
CGAN	76.8%	94.9%	58.6%
SVEBM-IB	<u>90.1%</u>	<u>95.1%</u>	<u>85.2%</u>
Ours	<b>99.0%</b>	<b>98.8%</b>	<b>99.1%</b>

Table 6.8: **Generated positive and negative reviews on Yelp.**

	The food here was very tasty and our server was very attentive.
Positive	I was very satisfied for my birthday party! Definitely the best Philly Cheesesteaks I've ever been. They are the best customer service ever!
Negative	Ugh the staff is so incompetent and rude. It just can't make it worse. Avoid this company at all costs. Just ruined the experience with a horrible attitude on it.

having only partial access to ground-truth data labels; it further validates the proposed formation for learning a well-structured latent space.

## 6.5 Discussions and Related Work

### Text modeling.

VAE has been one of the most prominent latent variable models for generative modeling

Table 6.9: **Accuracy on AGNews.** We report semi-supervised classification accuracy with varied number of labeled data.

Model	200	500	2500	10000
Glove-ID	70.4	78.0	84.1	87.1
Glove-OD	68.8	78.8	85.3	88.0
VAMPIRE	82.9	84.5	85.8	87.7
Hard EM	83.9	84.6	85.1	86.9
CatVAE	84.6	85.7	86.3	87.5
SVEBM	84.5	84.7	86.0	88.1
SVEBM-IB	<u>86.4</u>	<u>87.4</u>	<u>87.9</u>	<u>88.6</u>
Ours	<b>87.4</b>	88.1	<b>89.2</b>	<b>90.1</b>

[KW13, RMW14]. It is first applied to text modeling in [BVV16], followed by a wide range of work attacking challenging text generation problems using the shared framework of VAE. These include dialog generation [SSB16, SSL17, WMB17, ZZE17, ZLE18, FLG19], machine translation [ZXS16], text summarization [LLB17], and paraphrase generation [GAS18]. In parallel, extensive efforts have been made to address issues like posterior collapse [BVV16, HMP16b, ZZE17, ZKZ18, HSN18, LHN19, FLL19] and mode-collapse [SZM20] in training VAE to further improve the language modeling performance and text generation quality.

The interpretability of the generation process is naturally brought up as the generation quality achieves impressive progress. Recently, [ZLE18], [SZM20], and [PW21] have explored interpretable text generation with deliberately designed latent spaces. [ZLE18] use a discrete latent space to capture dialog actions; [SZM20] adopt a mixture of Gaussians as the VAE prior. To further improve the expressivity of latent space, [PW21] propose a symbol-vector coupling energy-based prior to learn a structured latent space. The coupling formulation provides a natural interface to induce the symbolic representation, which eliminates the need of training extra auxiliary inference networks for symbol induction. Our formulation inherits the advantages from [PW21] by choosing an appropriate symbol-vector coupling

scheme and principally incorporating the IB. We further develop a geometric clustering-based regularization that complements the IB; it alleviates the mode-collapse problem in variational learning of the latent space model.

### **Energy-based model.**

EBMs [XLZ16, NHZ19, NHH20, HNZ20] have drawn growing interest in generative modeling. As an interesting branch, [PHN20] learn an EBM in the latent space as a prior model for continuous latent variables; it greatly improves the expressivity over non-informative priors and demonstrates strong performance on downstream tasks, *e.g.*, image segmentation, molecule generation, and trajectory prediction [YXM21, PHW20, PZX21, JMH19, JMS18]. However, both EBM and latent space EBM require MCMC sampling to learn the model. The degenerate sampling quality in practice can lead to poor generation quality and instability in training [GWJ19, DLT21]. We leverage diffusion models as a cure for the vanilla latent space EBM in this work; the proposed model shows reliable sampling quality in practice.

### **Diffusion model.**

Diffusion models [SWM15, HJA20, GSP20], originating from [SWM15], learn from a sequence of noise-perturbed versions of the data. From such perturbed data, one can learn the conditional model to invert the diffusion process and generate high-quality samples given noisy inputs. On another front, [SE19, SE20, SSK20] extend the denoising score matching method [Vin11], modeling the diffusion process with continuous time step. Our formulation moves the model to the latent space in a variational framework with two benefits: (a) learning in a lower-dimensional space enables faster sampling and better convergence, and (b) learning the diffusion model in a continuous latent space avoids the discreteness of text data, which hinders the direct application of vanilla diffusion models to text modeling [AJH21].

Similar to our work, [WL21], [SSM21], [NVA21], and [VKK21] have proposed to learn a diffusion model in the latent space. Specifically, [WL21] empirically demonstrate that a diffusion prior can perform better than the non-informative Gaussian prior when jointly

trained with a VAE. [SSM21] combine contrastive learning with diffusion models in the latent space of VAEs for controllable generation. [NVA21] and [VKK21] extend the idea of [SSK20] in the latent space: [NVA21] perform controllable image generation by training a latent energy-based attribute classifier on a pre-trained generator; [VKK21] train score-based denoising diffusion models in the latent space of a powerful VAE [VK20]. Both methods have achieved very impressive image generation results. However, the listed methods are generally limited to image generation with tailored or pre-trained encoders and decoders. In contrast, our method is a general improvement for the sampling quality of latent space EBM; it is not restricted to a certain data type. Moreover, the proposed model can be trained from scratch to form a well-structured latent space, in contrast to [VKK21] and [NVA21] which require a pre-learned latent space.

## 6.6 Conclusion and Future Works

We presented LDEBM, a novel symbiosis between symbol-vector coupling EBM and diffusion model that offers the best of both worlds. The proposed model shows reliable sampling quality, learns a well-structured and meaningful latent space from scratch, and can be flexibly extended to scenarios where data labels are available. It demonstrates superior performance over strong baselines on interpretable text modeling. We hope our work inspires future research along this challenging but promising research direction. A potential follow-up research problem is to reuse powerful pre-trained language models. One could consider integrating pre-trained models with our method to realize high-quality controllable generation at low computational cost.

## 6.A Extra Experiment Details and Discussion

### 6.A.1 Network Architecture and Hyperparameters

We provide detailed network architecture for the latent space model of this work in table 6.10 and table 6.11; we adopt the same architecture throughout the experiments. Spectral normalization [MKK18] is used to regularize parameters in linear layers. The encoder and decoder in all models are the same as in [PW21], implemented with a single-layer GRU with a hidden size of 512. The key hyperparameters of LDEBM for each dataset are listed in table 6.12. Of note, we use the same dimension of the latent space as in [PW21] for a fair comparison.

$\lambda_1$  generally controls how fast  $q_\phi$  and  $p_\theta$  run towards each other.  $\lambda_2$  refers to the hyperparameter in Eq. (6.9); it controls the trade-off between the compressivity of  $\mathbf{z}_0$  about  $\mathbf{x}$  and its expressivity to  $\mathbf{y}$ .  $\lambda_3$  controls the weight of classification loss mentioned in section 6.3.3; recall that we use pseudo-label  $\hat{\mathbf{y}}$  inferred by the geometric clustering algorithm or the ground-truth label  $\mathbf{y}$  to supervise  $p_\alpha(\mathbf{y}|\mathbf{z}_0)$  in our modeling. For controllable generation and semi-supervised classification, we find it important to have a larger weight on the classification loss so that the model is forced to capture the major modes of the data.

For optimization, we use Adam optimizer [KB14] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  for all the experiments. On all the datasets but 2D synthetic datasets and AGNews dataset, we use a batch size of 128 and a constant learning rate of  $1e - 3$  for encoder and decoder without weight decay. For LDEBM, we use a constant learning rate of  $1e - 4$ . We use a larger batch size of 1000 on 2D synthetic datasets. On the AGNews dataset, we use the same set of hyperparameters as in [PW21] for optimization. The batch size is set to 200; the initial learning rate is  $1e - 4$  for encoder and decoder, and  $1e - 5$  for LDEBM. Learning rates are exponentially decayed with a decay rate of 0.998 for each model. Encoder and LDEBM have a weight decay rate of  $2e - 3$  and  $1e - 3$ , respectively.



## 6.A.2 Experiment Settings and Baselines

### Experiment settings.

For generative modeling, following previous methods [SZM20, PW21], the NLL term is computed with importance sampling [BGS16] using 500 importance samples. To compute rPPL, we set the generated sample size as 40,000, which is the same size as PTB training set. We recruit ASGD Weight-Dropped LSTM [MKS18] to compute rPPL as in previous works.

In terms of conditional response generation, for word-embedding-based evaluation on SMD and DD, we use the publicly available GloVe [PSM14] word embeddings of 300 dimension trained on 840B tokens, and report the score from 1 response per context. We use a context window size of 5 during training and evaluation.

The maximum length of each sentence is set to 40 words for most datasets and 70 words for the JerichoWorld dataset. On JerichoWorld dataset, we extract the description of each state as the text data.

### Baselines.

On PTB, DD and SMD, our model is compared with the following baselines: (1) RNNLM [MKB10], the language model implemented with GRU [CMG14]; (2) AE [VLL10], the deterministic auto-encoder which has no regularization to the latent space; (3) DAE, the AE with a discrete latent space; (4) VAE [KW13], the vanilla VAE with a continuous latent space and a non-informative Gaussian prior; (5) DVAE, the VAE with a discrete latent space; (6) DI-VAE [ZLE18], a DVAE variant with a mutual information term between the observed piece of text  $\mathbf{x}$  and its inferred latent variable  $\mathbf{z}$ ; (7) semi-VAE [KMR14], the semi-supervised VAE model with independent discrete and continuous latent variables; (8) GM-VAE, the VAE with a Gaussian mixture prior; (9) DGM-VAE [SZM20], the GM-VAE with a dispersion term that avoids the mode-collapse of Gaussian mixture prior; (10) semi-VAE +  $\mathcal{I}(\mathbf{x}, \mathbf{y})$ , GM-VAE +  $\mathcal{I}(\mathbf{x}, \mathbf{y})$ , DGM-VAE +  $\mathcal{I}(\mathbf{x}, \mathbf{y})$ , are the same models as (7),

(8), and (9) respectively, but with a mutual information term between  $\mathbf{x}$  and  $\mathbf{y}$  computed using separate inference networks for  $\mathbf{y}$  and  $\mathbf{z}$ . We compare with the close competitors (11) SVEBM, the symbol-vector coupling prior model and (12) SVEBM-IB, SVEBM with regularization based on information-bottleneck.

On Yelp dataset, we additionally include text conditional GAN [SRS18] as a baseline for controllable generation. On AGNews dataset, we further compare our model to VAMPIRE [GDC19], a VAE-based semi-supervised text learning model. Other baselines include its supervised learning variants: (1) the model trained with Glove embedding pre-trained on 840 billion words (Glove-OD); (2) the model trained with Glove embedding on in-domain unlabeled data (Glove-ID). We also include more recent baselines such as Hard EM and CatVAE [JWS20] that improve over VAMPIRE.

Table 6.10: **Network architecture for the LDEBM prior.**  $N$  is set to 12 for all the experiments.

Layers	Output size	Note
Time Embedding		
Input: $t$	1	Index of diffusion step
Sin. embedding	200	
Linear, LReLU	200	negative_slope 0.2
Linear	200	
Input Embedding		
Input: $\mathbf{z}$	$d_{\text{lat}}$	
Linear, LReLU	200	negative_slope 0.2
Linear	200	
Context Embedding (for response generation only)		
Input: $\mathbf{z}_{\text{ctx}}$	512	ctx. embedding
Linear, LReLU	200	negative_slope 0.2
Linear	200	

Table 6.11: **Network architecture for the LDEBM prior (Cont'd)**.  $N$  is set to 12 for all the experiments.

LDEBM Prior		
Input:	$\mathbf{z}, t$ $*\mathbf{z}_{\text{ctx}}$	$1, d_{\text{lat}}$ 512 optional $\mathbf{z}_{\text{ctx}}$
Embedding	200	Embedding of each input
Concatenate	400	w/o ctx.
	600	w/ ctx.
LReLU, Linear	200	negative_slope 0.2
N ResBlocks	200	LReLU, Linear + Input
LReLU, Linear	$K$	$K$ class logits
Log-Sum-Exp	1	energy score

Table 6.12: **Hyperparameters of LDEBM.** DD-CLS presents the set of hyperparameters used in unsupervised clustering on DD dataset. DD-GEN presents the set of hyperparameters used in conditional response generation on DD dataset.

DATASET	$d_{\text{lat}}$	$K$	$\lambda_1$	$\lambda_2$	$\lambda_3$
2D GAUSSIAN	2	16	1	0.05	0.05
2D PINWHEEL	2	10	1	0.05	0.05
PTB	40	20	0.1	0.05	0.05
JERICO	40	20	0.1	0.05	0.05
DD-CLS	32	125	0.01	0.05	0.5
DD-GEN	32	125	1	0.05	0.05
SMD	32	125	10	10	5
YELP	40	2	50	50	200
AGNEWS	20	4	1e-3	5	200

# CHAPTER 7

## Conclusion

In this dissertation, we introduce our contributions to the task of visual and relational reasoning, aiming at shrinking the human-machine gap in terms of learning and reasoning with real-world sensory input, zero-shot and few-shot generalization, and adaptation to novel modalities. We propose to study this problem from two angles: establishing benchmarks, where we focus on deepening our understanding of the limitations of existing AI reasoning systems in the aforementioned challenges (Chapter 2 and Chapter 3); and framework developing, where we propose a unified framework for visual and relational reasoning by drawing inspiration from human language system, and demonstrate some promising results on getting machine closer to human-level performances in these tasks (Chapter 4, Chapter 5 and Chapter 6). Here, we would like to pinpoint some insights as follows:

- Albeit the seemingly rich tasks and benchmarks we have, still, it is always a good idea to establish new benchmarks if you have identified a significant drawback of the existing models and systems. In our case, both the Bongard-HOI and SQA3D benchmarks are introduced to expose the limitations of current few-shot and multimodal reasoning systems and they demonstrate their uniqueness. Rather, sticking to the existing benchmarks leads to a false sense of progress and may hinder how we perceive the challenge ahead.
- Compared to the explicit reasoning systems as in many neural-symbolic efforts [LHH20], we argue that learning human-like representations could be more crucial to the success of closing the human-machine gap. We’ve demonstrated that in an end-to-end learned model, good representations could eliminate the need for additional explicit or implicit

but parameter-hungry reasoning modules. We believe human-like representations emerge from experiences and therefore could use some help from scalable learning pipelines.

- In the canonical perception-action loop of intelligent agents, we find reasoning bridges both parties, as not only it provides a “system-2” within the agent but it also serves as the abstraction layer that is easier to cope with rather than working with the low-level control directly [ABB22]. Therefore, many challenges in acting can effectively be reduced to reasoning problems, *e.g.* generalization, few-shot learning, *etc.* We might need to solve reasoning before solving acting agents.

What’s next? This is never an easy question to address. But we will try to have a glance here. First of all, there are still tons of work that needs to be done to tackle the challenges we posed in this dissertation. What seems to be a key to few-shot generalization? Will learning from massive experiences help? When it comes to more challenging modalities like embodied 3D scenes, what are the ingredients that might be missing in our framework, as it calls for a sense of agency – which is clearly beyond the scope of the current framework for reasoning from a third-person perspective. Extending the work of this dissertation in these two directions (few-shot and embodied reasoning) could further facilitate what we just suggested: from a unified reasoning model to a unified agent that can ultimately act in the real world.

Another exciting direction to explore is the elephant in the room of the reasoning community – the large models. We acknowledge the breakthrough in human-level reasoning brought by these models. But as many have pointed out, its success has a shadow, which includes the drawbacks of the human-level capabilities pinpointed at the beginning of this dissertation. Either using these large models as better foundations of reasoning or reconciling the principles introduced in this dissertation with their methodologies could light up the path to the next generation of thinking machines.

## REFERENCES

- [AAL15a] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. “VQA: Visual Question Answering.” In *International Conference on Computer Vision (ICCV)*, 2015. 3
- [AAL15b] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. “Vqa: Visual question answering.” In *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015. 51, 55, 56
- [AAX20] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. “Referit3D: Neural listeners for fine-grained 3D object identification in real-world scenes.” In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 422–440, 2020. 40, 42, 49, 54
- [AB94] Rolf Adams and Leanne Bischof. “Seeded region growing.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **16**(6):641–647, 1994. 71
- [ABB22] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. “Do as i can, not as i say: Grounding language in robotic affordances.” *arXiv preprint arXiv:2204.01691*, 2022. 39, 151
- [ADL22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. “Flamingo: a visual language model for few-shot learning.” *arXiv preprint arXiv:2204.14198*, 2022. 1, 46
- [AGP15] Aurore Avarguès-Weber, Martin Giurfa, Joshua Plotnik, Nicola S Clayton, Robert Seyfarth, Dorothy L Cheney, Brad Mahon, H Clark Barrett, Pascal Boyer, Jerry A Fodor, et al. *The conceptual mind: New directions in the study of concepts*. MIT Press, 2015. 1, 4
- [AGR00] John R Anderson, James G Greeno, Lynne M Reder, and Herbert A Simon. “Perspectives on learning, thinking, and activity.” *Educational Researcher*, **29**(4):11–13, 2000. 39
- [AHB18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. “Bottom-up and top-down attention for image captioning and visual question answering.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018. 105



- [AJH21] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. “Structured denoising diffusion models in discrete state-spaces.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 142
- [AMF10] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. “Contour detection and hierarchical image segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **33**(5):898–916, 2010. 71
- [AMK22] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. “ScanQA: 3D Question Answering for Spatial Scene Understanding.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19129–19139, 2022. 40, 42, 45, 46, 47, 53, 54, 55, 60, 63
- [And18] Jacob Daniel Andreas. *Learning from Language*. University of California, Berkeley, 2018. iii
- [AR21] Prithviraj Ammanabrolu and Mark O Riedl. “Modeling Worlds in Text.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 134
- [ASS12] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. “SLIC superpixels compared to state-of-the-art superpixel methods.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **34**(11):2274–2282, 2012. 71
- [AWT18] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018. 38, 42, 54
- [BDD93] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. “The mathematics of statistical machine translation: Parameter estimation.” *Computational Linguistics*, 1993. 120
- [BGJ11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011. 76
- [BGS16] Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. “Importance Weighted Autoencoders.” In *International Conference on Learning Representations (ICLR)*, 2016. 145
- [BHS18a] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. “Measuring abstract reasoning in neural networks.” In *ICML*, pp. 511–520, 2018. 9, 12, 15, 17, 19, 24, 28

- [BHS18b] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. “Measuring abstract reasoning in neural networks.” In *International conference on machine learning*, pp. 511–520. PMLR, 2018. 110
- [BJ96] David Braddon-Mitchell and Frank Jackson. *Philosophy of Mind and Cognition*. Blackwell, 1996. 4
- [BJ01] Yuri Y Boykov and M-P Jolly. “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2001. 70
- [BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **39**(12):2481–2495, 2017. 70
- [BLR16] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. “Neural photo editing with introspective adversarial networks.” *arXiv preprint arXiv:1609.07093*, 2016. 82
- [BMJ19] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. “Phyre: A new benchmark for physical reasoning.” In *Advances in Neural Information Processing Systems*, pp. 5083–5094, 2019. 28
- [BMN18] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. “Systematic generalization: what is required and can it be learned?” *arXiv preprint arXiv:1811.12889*, 2018. 94, 109
- [BMR20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “Language models are few-shot learners.” *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, **33**:1877–1901, 2020. 1, 4, 41, 47, 48
- [BMW19] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. “Monet: Unsupervised scene decomposition and representation.” *arXiv preprint arXiv:1901.11390*, 2019. 71, 73, 74
- [Bon68] Mikhail Moiseevich Bongard. “The recognition problem.” Technical report, Foreign Technology Div Wright-Patterson AFB Ohio, 1968. 4, 9, 11, 28
- [Bro90] Rodney A Brooks. “Elephants don’t play chess.” *Robotics and autonomous systems*, **6**(1-2):3–15, 1990. 54

- [BVV16] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. “Generating sentences from a continuous space.” In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2016. 120, 141
- [BW20] Yaniv Benny and Lior Wolf. “OneGAN: Simultaneous Unsupervised Learning of Conditional Image Generation, Foreground Segmentation, and Fine-Grained Clustering.” In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 71, 73, 83, 84, 85
- [BYA13] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. “Generalized Denoising Auto-Encoders as Generative Models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 125
- [CAD19] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. “Unsupervised Object Segmentation by Redrawing.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 71, 73, 83, 88
- [Car00] Susan Carey. “The origin of concepts.” *Journal of Cognition and Development*, 1(1):37–41, 2000. 1
- [CBJ18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. “Deep clustering for unsupervised learning of visual features.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018. 100
- [CCN20] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. “Scanrefer: 3D object localization in rgb-d scans using natural language.” In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 202–221, 2020. 40, 42, 49, 54, 56
- [CDF17] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. “Matterport3D: Learning from rgb-d data in indoor environments.” *arXiv preprint arXiv:1709.06158*, 2017. 54
- [CDH16] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. “InfoGAN: interpretable representation learning by information maximizing Generative Adversarial Nets.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 82
- [CFG20] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. “Improved Baselines with Momentum Contrastive Learning.” *CoRR*, **abs/2003.04297**, 2020. 23, 24, 26

- [CGN21] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. “Scan2cap: Context-aware dense captioning in rgb-d scans.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3203, 2021. 40, 48, 54
- [CH21] Xinlei Chen and Kaiming He. “Exploring simple siamese representation learning.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. 110
- [Che11] Anthony Chemero. *Radical embodied cognitive science*. MIT press, 2011. 1
- [Cho06] Noam Chomsky et al. *Language and mind*. Cambridge University Press, 2006. iii
- [Cho19] François Chollet. “On the measure of intelligence.” *arXiv preprint arXiv:1911.01547*, 2019. 28
- [CKN20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations.” In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 110
- [CKS22] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. “All You May Need for VQA are Image Captions.” *arXiv preprint arXiv:2205.01883*, 2022. 48
- [CMG14] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 145
- [CMH14] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. “Global contrast based salient region detection.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **37**(3):569–582, 2014. 70
- [Coh91] Laurent D Cohen. “On active contour models and balloons.” *CVGIP: Image understanding*, **53**(2):211–218, 1991. 71
- [CPK17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **40**(4):834–848, 2017. 70
- [CSM19] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. “Touchdown: Natural language navigation and spatial reasoning in visual street environments.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12538–12547, 2019. 54

- [CTM21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. “Emerging properties in self-supervised vision transformers.” *arXiv preprint arXiv:2104.14294*, 2021. 93, 94, 97, 98, 99, 102, 110, 113
- [CTS18] Dallas Card, Chenhao Tan, and Noah A Smith. “Neural Models for Documents with Metadata.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 139
- [CTY06] Nick Chater, Joshua B Tenenbaum, and Alan Yuille. “Probabilistic models of cognition: Conceptual foundations.” *Trends in Cognitive Sciences*, **10**(7):287–291, 2006. 70
- [CWH15a] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. “HICO: A Benchmark for Recognizing Human-Object Interactions in Images.” In *ICCV*, 2015. 15, 21, 26
- [CWH15b] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. “HICO: A Benchmark for Recognizing Human-Object Interactions in Images.” In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 94, 101
- [CWL20] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. “A New Meta-Baseline for Few-Shot Learning.” *arXiv preprint arXiv:2003.04390*, 2020. 13, 19, 24
- [CXH21] Xinlei Chen, Saining Xie, and Kaiming He. “An empirical study of training self-supervised vision transformers.” *arXiv preprint arXiv:2104.02057*, 2021. 110
- [DA05] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005. 4
- [DBK20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” *arXiv preprint arXiv:2010.11929*, 2020. 6, 94, 97, 106
- [DCS17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. “ScanNet: Richly-annotated 3D reconstructions of indoor scenes.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. 40, 42, 43, 49, 54, 58

- [DDC22] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. “Episodic Memory Question Answering.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19119–19128, 2022. 55
- [DDG18] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. “Embodied question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2018. 40, 54
- [DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database.” In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009. 9, 23
- [DHS15] Jifeng Dai, Kaiming He, and Jian Sun. “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 75
- [DHS20] David Ding, Felix Hill, Adam Santoro, and Matt Botvinick. “Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures.” *arXiv preprint arXiv:2012.08508*, 2020. 110
- [DLT21] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. “Improved contrastive divergence training of energy based models.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 142
- [DVH22] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. “ProcTHOR: Large-Scale Embodied AI Using Procedural Generation.” *arXiv preprint arXiv:2206.06994*, 2022. 54
- [EHR20] S. M. Ali Eslami, Irina Higgins, and Danilo J. Rezende. “Representation Learning Without Labels.” <https://icml.cc/Conferences/2020/Schedule?showEvent=5751>, 2020. 4, 5
- [EHW16a] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. “Attend, infer, repeat: Fast scene understanding with generative models.” *Advances in Neural Information Processing Systems*, **29**:3225–3233, 2016. 110
- [EHW16b] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. “Attend, Infer, Repeat: Fast Scene Understanding with Generative Models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 71, 73

- [EKC17] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. “Key-Value Retrieval Networks for Task-Oriented Dialogue.” 2017. 136
- [EKJ20] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. “GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations.” In *International Conference on Learning Representations (ICLR)*, 2020. 71, 73, 74
- [EVW10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. “The pascal visual object classes (voc) challenge.” *International Journal of Computer Vision (IJCV)*, **88**(2):303–338, 2010. 70
- [FCT18] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. “Pairwise body-part attention for recognizing human-object interactions.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 51–67, 2018. 102, 103, 119
- [Fe 03] Li Fe-Fei et al. “A Bayesian approach to unsupervised one-shot learning of object categories.” In *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1134–1141. IEEE, 2003. 27
- [FHC18] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. “Speaker-follower models for vision-and-language navigation.” *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, **31**, 2018. 54
- [FLG19] Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. “Implicit Deep Latent Variable Models for Text Generation.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 120, 141
- [FLL19] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Çelikyilmaz, and Lawrence Carin. “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing.” In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 141
- [Fod75] Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975. iii, 4, 5
- [FPL14] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. “Bootstrapping dialog systems with word embeddings.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 136

- [FS13] Ali Farhadi and Mohammad Amin Sadeghi. “Phrasal Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(12):2854–2865, 2013. 110
- [GAM13] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. “Perceptual organization and recognition of indoor scenes from RGB-D images.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 564–571, 2013. 54
- [GAS18] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. “A deep generative framework for paraphrase generation.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 120, 141
- [GDC19] Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. “Variational Pretraining for Semi-supervised Text Classification.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 146
- [GDG15] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. “Draw: A recurrent neural network for image generation.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2015. 73
- [GKA21] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. “AGQA: A benchmark for compositional spatio-temporal reasoning.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021. 55
- [GKK19a] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. “Multi-object representation learning with iterative variational inference.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019. 71, 73, 74, 80, 86
- [GKK19b] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. “Multi-object representation learning with iterative variational inference.” In *International Conference on Machine Learning*, pp. 2424–2433. PMLR, 2019. 110
- [GKS17a] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3



- [GKS17b] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. “Making the v in vqa matter: Elevating the role of image understanding in visual question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. 45
- [GM15] Saurabh Gupta and Jitendra Malik. “Visual Semantic Role Labeling.” *arXiv preprint arXiv:1505.04474*, 2015. 15, 21, 26
- [GPM14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. “Generative Adversarial Nets.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 71, 73
- [GPT22] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. “Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5067–5077, 2022. 48
- [GR17] Rohit Girdhar and Deva Ramanan. “Attentional pooling for action recognition.” *arXiv preprint arXiv:1711.01467*, 2017. 101, 103
- [GRB16] Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hotloo Hao, Jürgen Schmidhuber, and Harri Valpola. “Tagger: Deep unsupervised perceptual grouping.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 71, 73, 86
- [Gre98] James G Greeno. “The situativity of knowing, learning, and research.” *American psychologist*, **53**(1):5, 1998. 39
- [GSP20] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. “Learning Energy-Based Models by Diffusion Recovery Likelihood.” In *International Conference on Learning Representations (ICLR)*, 2020. 121, 122, 125, 130, 142
- [GSS17] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. “Neural expectation maximization.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 71, 73
- [GWJ19] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. “Your classifier is secretly an energy based model and you should treat it like one.” In *International Conference on Learning Representations (ICLR)*, 2019. 121, 128, 142

- [GWL21] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. “Improve vision transformers training by suppressing over-smoothing.” *arXiv preprint arXiv:2104.12753*, 2021. 108
- [GZW07] Cheng-en Guo, Song-Chun Zhu, and Ying Nian Wu. “Primal sketch: Integrating structure and texture.” *Computer Vision and Image Understanding (CVIU)*, **106**(1):5–19, 2007. 71, 77
- [HBQ21] Zhi Hou, Yu Baosheng, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. “Detecting Human-Object Interaction via Fabricated Compositional Learning.” In *CVPR*, 2021. 26
- [HDM20] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. “Compositionality decomposed: how do neural networks generalise?” *Journal of Artificial Intelligence Research*, **67**:757–795, 2020. 94, 104, 109
- [HFW20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. “Momentum contrast for unsupervised visual representation learning.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. 110
- [HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 70, 83
- [HGD20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. “Mask R-CNN.” *IEEE TPAMI*, **42**(2):386–397, 2020. 31
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7, 121, 126, 131, 142
- [HLZ21] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. “Vlgrammar: Grounded grammar induction of vision and language.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 40
- [HM18a] Drew A Hudson and Christopher D Manning. “Compositional attention networks for machine reasoning.” *arXiv preprint arXiv:1803.03067*, 2018. 101
- [HM18b] Drew A Hudson and Christopher D Manning. “Compositional attention networks for machine reasoning.” *arXiv preprint arXiv:1803.03067*, 2018. 105
- [HM19] Drew A Hudson and Christopher D Manning. “Gqa: A new dataset for real-world visual reasoning and compositional question answering.” In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019. 3, 94, 101, 104

- [HMP16a] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. “beta-VAE: Learning basic visual concepts with a constrained variational framework.” 2016. 109
- [HMP16b] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.” In *International Conference on Learning Representations (ICLR)*, 2016. 141
- [HNF19] Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. “Divergence triangle for joint training of generator model, energy-based model, and inferential model.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 127
- [HNZ20] Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. “Joint training of variational auto-encoder and latent energy-based model.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 142
- [HPQ20a] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. “Visual compositional learning for human-object interaction detection.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 584–600. Springer, 2020. 3, 27
- [HPQ20b] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. “Visual compositional learning for human-object interaction detection.” In *European Conference on Computer Vision*, pp. 584–600. Springer, 2020. 102, 103, 114
- [HSN18] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. “Lagging Inference Networks and Posterior Collapse in Variational Autoencoders.” In *International Conference on Learning Representations (ICLR)*, 2018. 141
- [HTG20] Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. “Grounded language learning fast and slow.” *arXiv preprint arXiv:2009.01719*, 2020. 94
- [HWQ21] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. “Vln bert: A recurrent vision-and-language bert for navigation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1643–1653, 2021. 54
- [HWW18] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. “Weakly-supervised semantic segmentation network with deep seeded region growing.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 75

- [HYC01] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. “Learning to learn using gradient descent.” In *International Conference on Artificial Neural Networks*, pp. 87–94. Springer, 2001. 27
- [HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In *CVPR*, pp. 770–778, 2016. 20, 23, 83
- [IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2015. 82
- [JCH20] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. “Lemma: A multi-view dataset for learning multi-agent multi-task activities.” In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 55
- [JCW21] Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Zicheng Liu, and Jenq-Neng Hwang. “Is Object Detection Necessary for Human-Object Interaction Recognition?” *arXiv preprint arXiv:2107.13083*, 2021. 102
- [JDJ19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. “Billion-scale similarity search with gpus.” *IEEE Transactions on Big Data*, 2019. 131
- [JHM17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. “CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning.” In *CVPR*, 2017. 20, 28, 83, 86
- [JHV17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017. 3, 109
- [JHV19] Xu Ji, João F Henriques, and Andrea Vedaldi. “Invariant information clustering for unsupervised image classification and segmentation.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 70, 74
- [JJ94] Michael I Jordan and Robert A Jacobs. “Hierarchical mixtures of experts and the EM algorithm.” *Neural Computation*, **6**(2):181–214, 1994. 72, 76
- [JJN91] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. “Adaptive mixtures of local experts.” *Neural Computation*, **3**(1):79–87, 1991. 72, 76

- [JLZ22] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. “EgoTaskQA: Understanding Human Tasks in Egocentric Videos.” In *The 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*, 2022. 55
- [JMH19] Mingxuan Jing, Xiaojian Ma, Wenbing Huang, Fuchun Sun, and Huaping Liu. “Task transfer by preference-based cost learning.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 142
- [JMR20] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. “In defense of grid features for visual question answering.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10267–10276, 2020. 94
- [JMS18] Mingxuan Jing, Xiaojian Ma, Fuchun Sun, and Huaping Liu. “Learning and inferring movement with deep generative model.” *arXiv preprint arXiv:1805.07252*, 2018. 142
- [JNY22] Xiaojian Ma Huaizu Jiang, Weili Nie, Zhiding Yu, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. “Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 110
- [Jul81] Bela Julesz. “Textons, the elements of texture perception, and their interactions.” *Nature*, **290**(5802):91–97, 1981. 77
- [JWS20] Shuning Jin, Sam Wiseman, Karl Stratos, and Karen Livescu. “Discrete Latent Variable Representations for Low-Resource Text Classification.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 146
- [JWY13] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. “Salient object detection: A discriminative regional feature integration approach.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 70
- [Kan18] Asako Kanezaki. “Unsupervised image segmentation by backpropagation.” In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 70, 74
- [Kar94] By A Karmiloff-Smith. “Beyond modularity: A developmental perspective on cognitive science.” *European journal of disorders of communication*, **29**(1):95–105, 1994. 1, 4
- [KB14] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” *arXiv preprint arXiv:1412.6980*, 2014. 92, 144

- [KBH17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. “Simple does it: Weakly supervised instance and semantic segmentation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 75
- [KC17] Alexander Kuhnle and Ann Copestake. “Shapeworld-a new test methodology for multimodal language understanding.” *arXiv preprint arXiv:1704.04517*, 2017. 109
- [KJY11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. “Novel dataset for fine-grained image categorization: Stanford dogs.” In *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011. 83, 88
- [KLG18] Keizo Kato, Yin Li, and Abhinav Gupta. “Compositional learning for human object interaction.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 234–251. Springer, 2018. 3, 27
- [KM18] Hyunjik Kim and Andriy Mnih. “Disentangling by factorising.” In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018. 109
- [KMH17] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. “Ai2-thor: An interactive 3D environment for visual ai.” *arXiv preprint arXiv:1712.05474*, 2017. 54
- [KMK20] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. “Unifiedqa: Crossing format boundaries with a single qa system.” *arXiv preprint arXiv:2005.00700*, 2020. 1, 41, 47, 48, 50
- [KMR14] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. “Semi-supervised learning with deep generative models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 145
- [KNT20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. “Concept bottleneck models.” In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020. 94
- [KOJ17] Ranjay Krishna, Yuke Zhu andd Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.” *IJCV*, **123**(1):32–73, 2017. 15, 26

- [KRA20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale.” *IJCV*, 2020. 15, 26
- [KSD13] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. “3d object representations for fine-grained categorization.” In *ICCV workshops*, 2013. 83, 88
- [KW13] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes.” *arXiv preprint arXiv:1312.6114*, 2013. 87, 120, 124, 141, 145
- [KWT88] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. “Snakes: Active contour models.” *International Journal of Computer Vision (IJCV)*, **1**(4):321–331, 1988. 71
- [KZG16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.” 2016. 104
- [KZS15] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese neural networks for one-shot image recognition.” In *ICML deep learning workshop*, volume 2. Lille, 2015. 27
- [LBL19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. “Challenging common assumptions in the unsupervised learning of disentangled representations.” In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019. 110
- [LBM19] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. “Unsupervised part-based disentangling of object shape and appearance.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 74
- [LBP19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.” *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, **32**, 2019. 46
- [Lec89] Yvan G Leclerc. “Constructing simple stable descriptions for image partitioning.” *International Journal of Computer Vision (IJCV)*, **3**(1):73–102, 1989. 71

- [LHH20] Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Ying Nian Wu, and Song-Chun Zhu. “Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning.” In *International Conference on Machine Learning*, pp. 5884–5894. PMLR, 2020. 150
- [LHN19] Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. “A Surprisingly Effective Fix for Deep Latent Variable Modeling of Text.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 141
- [LJ93] Barbara Landau and Ray Jackendoff. ““What” and “where” in spatial language and spatial cognition.” *Behavioral and brain sciences*, **16**(2):217–238, 1993. 52
- [LJH18] Juncen Li, Robin Jia, He He, and Percy Liang. “Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer.” In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. 138
- [LKB16] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. “Visual Relationship Detection with Language Priors.” In *ECCV*, 2016. 26
- [LLB17] Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. “Deep Recurrent Generative Decoder for Abstractive Text Summarization.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 120, 141
- [LLC21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. “Swin transformer: Hierarchical vision transformer using shifted windows.” *arXiv preprint arXiv:2103.14030*, 2021. 94, 106, 117
- [LLL20] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. “Detailed 2D-3D Joint Representation for Human-Object Interaction.” In *CVPR*, 2020. 12, 15, 17, 27
- [LLW15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Face Attributes in the Wild.” In *ICCV*, December 2015. 15
- [LLW20] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. “PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection.” In *CVPR*, 2020. 16
- [LLZ21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. “Less is more: Clipbert for video-and-language learning via sparse sampling.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7331–7341, 2021. 41, 48, 49, 59, 60, 64



- [LMB14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context.” In *European conference on computer vision*, pp. 740–755. Springer, 2014. 16, 26, 70, 83
- [LML19] Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. “Pointnetgpd: Detecting grasp configurations from point sets.” In *Proceedings of International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635. IEEE, 2019. 39
- [LMR19] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. “Meta-learning with differentiable convex optimization.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019. 19, 24, 26, 27
- [LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 70
- [LSG11] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. “One shot learning of simple visual concepts.” In Laura A. Carlson, Christoph Hölscher, and Thomas F. Shipley, editors, *CogSci*, 2011. 15
- [LSS17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 134
- [LST15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. “Human-level concept learning through probabilistic program induction.” *Science*, **350**(6266):1332–1338, 2015. 9
- [LWP20a] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. “SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition.” In *International Conference on Learning Representations (ICLR)*, 2020. 71, 73
- [LWP20b] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. “Space: Unsupervised object-oriented scene representation via spatial attention and decomposition.” *arXiv preprint arXiv:2001.02407*, 2020. 110
- [LWU20a] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. “Object-centric learning with slot attention.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 71, 73, 74, 80, 86

- [LWU20b] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. “Object-centric learning with slot attention.” *arXiv preprint arXiv:2006.15055*, 2020. 110
- [LXH19] Yonglu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Haoshu Fang, and Cewu Lu. “HAKE: Human Activity Knowledge Engine.” *CoRR*, **abs/1904.06539**, 2019. 12, 15, 17, 23, 25, 27, 28, 29, 102
- [LXL20] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Haoshu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. “Pastanet: Toward human activity knowledge engine.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 382–391, 2020. 102
- [LYB18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. “Tvqa: Localized, compositional video question answering.” *arXiv preprint arXiv:1809.01696*, 2018. 50, 55
- [LYL20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. “Oscar: Object-semantics aligned pre-training for vision-language tasks.” In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 121–137. Springer, 2020. 46, 104
- [LYT10] Ce Liu, Jenny Yuen, and Antonio Torralba. “Sift flow: Dense correspondence across scenes and its applications.” *IEEE transactions on pattern analysis and machine intelligence*, **33**(5):978–994, 2010. 95, 100
- [LYY19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. “Visualbert: A simple and performant baseline for vision and language.” *arXiv preprint arXiv:1908.03557*, 2019. 104
- [LYZ21] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. “Efficient Self-supervised Vision Transformers for Representation Learning.” *arXiv preprint arXiv:2106.09785*, 2021. 93, 94, 97, 98, 99, 100, 102, 103, 105, 110
- [MHH17] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. “dSprites: Disentanglement testing Sprites dataset.” <https://github.com/deepmind/dsprites-dataset/>, 2017. 83, 86
- [MIB23] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. “Dissociating language and thought in large language models: a cognitive perspective.” *arXiv preprint arXiv:2301.06627*, 2023. 1

- [MKB10] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. “Recurrent neural network based language model.” In *Interspeech*, 2010. 132, 145
- [MKK18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. “Spectral Normalization for Generative Adversarial Networks.” In *International Conference on Learning Representations (ICLR)*, 2018. 144
- [MKS18] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. “Regularizing and Optimizing LSTM Language Models.” In *International Conference on Learning Representations (ICLR)*, 2018. 145
- [ML08] Jeff Mitchell and Mirella Lapata. “Vector-based models of semantic composition.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008. 136
- [ML16] Arun Mallya and Svetlana Lazebnik. “Learning models for actions and person-object interactions with transfer to question answering.” In *European Conference on Computer Vision*, pp. 414–428. Springer, 2016. 101, 103
- [MMS93] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a large annotated corpus of English: the penn treebank.” *Computational Linguistics*, 1993. 132
- [MNY22] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. “Relvit: Concept-guided vision transformer for visual relational reasoning.” *arXiv preprint arXiv:2204.11167*, 2022. 6, 28, 47, 64
- [MRC18] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. “A Simple Neural Attentive Meta-Learner.” *ICLR*, 2018. 27
- [MRF19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. “Ok-vqa: A visual question answering benchmark requiring external knowledge.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019. 3, 45, 55, 56
- [MYB16] Yishu Miao, Lei Yu, and Phil Blunsom. “Neural variational inference for text processing.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2016. 139
- [MYM18] Joe Marino, Yisong Yue, and Stephan Mandt. “Iterative amortized inference.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 73

- [MYZ22] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. “SQA3D: Situated Question Answering in 3D Scenes.” *arXiv preprint arXiv:2210.07474*, 2022. 4
- [MZB21] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Ed Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. “ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition.” In *ICCV*, 2021. 15
- [MZW18] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Joshua B Tenenbaum, and Daniel LK Yamins. “Flexible neural representation for physics prediction.” *arXiv preprint arXiv:1806.08047*, 2018. 110
- [NHH20] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. “On the anatomy of mcmc-based maximum likelihood learning of energy-based models.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 121, 142
- [NHZ19] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. “Learning non-convergent non-persistent short-run MCMC toward energy-based model.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 87, 121, 142
- [NVA21] Weili Nie, Arash Vahdat, and Anima Anandkumar. “Controllable and compositional generation with latent-space energy-based models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 122, 142, 143
- [NYM20a] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. “Bongard-logo: A new benchmark for human-level concept learning and reasoning.” *Advances in Neural Information Processing Systems*, **33**, 2020. 3, 110
- [NYM20b] Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, and Anima Anandkumar. “Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning.” In *NeurIPS*, 2020. 9, 12, 15, 19, 23, 24, 26, 28
- [OAC20] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzshanskyi. “GECToR – Grammatical Error Correction: Tag, Not Rewrite.” In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170, Seattle, WA, USA â†’ Online, July 2020. Association for Computational Linguistics. 56
- [OBK17] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. “Exploiting saliency for object segmentation from image

- level labels.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 75
- [OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding.” *arXiv preprint arXiv:1807.03748*, 2018. 110
- [OSL18] Pavel Ostyakov, Roman Suvorov, Elizaveta Logacheva, Oleg Khomenko, and Sergey I Nikolenko. “Seigan: Towards compositional image generation by simultaneously learning to segment, enhance, and inpaint.” *arXiv preprint arXiv:1811.07630*, 2018. 71, 73
- [PCM15] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. “Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 75
- [PDS18] Trung T Pham, Thanh-Toan Do, Niko Sünderhauf, and Ian Reid. “Scenecut: Joint geometric and object segmentation for indoor scenes.” In *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2018. 74
- [PGM19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 81
- [PH16] Genevieve Patterson and James Hays. “COCO Attributes: Attributes for People, Animals, and Objects.” *European Conference on Computer Vision*, 2016. 9, 23
- [PHN20] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. “Learning Latent Space Energy-Based Prior Model.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7, 72, 75, 76, 87, 121, 142
- [PHW20] Bo Pang, Tian Han, and Ying Nian Wu. “Learning latent space energy-based prior model for molecule generation.” *arXiv preprint arXiv:2010.09351*, 2020. 142
- [PKD15] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. “Constrained convolutional neural networks for weakly supervised segmentation.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 75
- [PRB18] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. “Virtualhome: Simulating household activities via

- programs.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8494–8502, 2018. 54
- [PRW02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002. 132
- [PSD18] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. “Film: Visual reasoning with a general conditioning layer.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. 54
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 145
- [PSS21] Alexander Pashevich, Cordelia Schmid, and Chen Sun. “Episodic transformer for vision-and-language navigation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15942–15952, 2021. 54
- [PW21] Bo Pang and Ying Nian Wu. “Latent space energy-based model of symbol-vector coupling for text generation and classification.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 76, 120, 123, 124, 131, 132, 138, 139, 141, 144, 145
- [PZX21] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. “Trajectory Prediction with Latent Belief Energy-Based Model.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 142
- [QLH19] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. “Deep hough voting for 3D object detection in point clouds.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9277–9286, 2019. 47, 52
- [QWA20] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. “Reverie: Remote embodied visual referring expression in real indoor environments.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9982–9991, 2020. 54
- [RAB20] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. “A benchmark for systematic generalization in grounded language understanding.” *arXiv preprint arXiv:2003.05161*, 2020. 110

- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 70
- [RFL21] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. “The MECCANO Dataset: Understanding Human-Object Interactions From Egocentric Videos in an Industrial-Like Domain.” In *WACV*, pp. 1569–1578, 2021. 26
- [RHG15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks.” *Advances in neural information processing systems*, **28**:91–99, 2015. 94, 104
- [RHG17] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” *IEEE TPAMI*, **39**(6):1137–1149, 2017. 20, 23
- [RJL18] Pranav Rajpurkar, Robin Jia, and Percy Liang. “Know what you don’t know: Unanswerable questions for SQuAD.” *arXiv preprint arXiv:1806.03822*, 2018. 48
- [RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ““ GrabCut” interactive foreground extraction using iterated graph cuts.” *ACM Transactions on Graphics (TOG)*, **23**(3):309–314, 2004. 70, 88
- [RL12] Vasile Rus and Mihai Lintean. “An optimal assessment of natural language student input using word-to-word similarity metrics.” In *International Conference on Intelligent Tutoring Systems*, 2012. 136
- [RMC15] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” *arXiv preprint arXiv:1511.06434*, 2015. 82
- [RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic backpropagation and approximate inference in deep generative models.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2014. 120, 141
- [ROF92] Leonid I Rudin, Stanley Osher, and Emad Fatemi. “Nonlinear total variation based noise removal algorithms.” *Physica D: nonlinear phenomena*, **60**(1-4):259–268, 1992. 82
- [RP15] Matteo Ruggero Ronchi and Pietro Perona. “Describing Common Human Visual Actions in Images.” In *BMVC*, 2015. 26

- [RRB20] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. “Rapid learning or feature reuse? towards understanding the effectiveness of maml.” *ICLR*, 2020. 19, 24, 27
- [RTR18] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. “Meta-learning for semi-supervised few-shot classification.” *ICLR*, 2018. 27
- [SBB16] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. “Meta-learning with memory-augmented neural networks.” In *International conference on machine learning*, pp. 1842–1850, 2016. 27
- [SCG18] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. “Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions.” In *International Conference on Learning Representations (ICLR)*, 2018. 71, 73
- [SE19] Yang Song and Stefano Ermon. “Generative Modeling by Estimating Gradients of the Data Distribution.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 142
- [SE20] Yang Song and Stefano Ermon. “Improved Techniques for Training Score-Based Generative Models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 121, 142
- [SGH19] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. “Analysing mathematical reasoning abilities of neural models.” *arXiv preprint arXiv:1904.01557*, 2019. 28
- [SGT21] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. “Embodied bert: A transformer model for embodied, language-guided visual task completion.” *arXiv preprint arXiv:2108.04927*, 2021. 54
- [SHK12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor segmentation and support inference from rgb-d images.” In *Proceedings of European Conference on Computer Vision (ECCV)*, 2012. 74
- [Sim69] Herbert A Simon. *The Sciences of the Artificial*. The MIT Press, 1969. 4
- [SK01] Thomas F Shipley and Philip J Kellman. *From fragments to objects: Segmentation and grouping in vision*. Elsevier, 2001. 70
- [SKM19] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. “Habitat: A platform for embodied ai research.” In *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019. 39, 54



- [SLY17] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. “A corpus of natural language for visual reasoning.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, 2017. 3
- [SM00] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **22**(8):888–905, 2000. 70
- [SMF22] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. “Cliport: What and where pathways for robotic manipulation.” In *Conference on Robot Learning (CoRL)*, pp. 894–906. PMLR, 2022. 39, 54
- [SMG14] Andrew M Saxe, James L McClelland, and Surya Ganguli. “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.” In *International Conference on Learning Representations (ICLR)*, 2014. 82, 90
- [SOL19] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. “Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 71, 73
- [SRB17] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. “A simple neural network module for relational reasoning.” In *NeurIPS*, 2017. 18, 20
- [SRS18] Sandeep Subramanian, Sai Rajeswar, Alessandro Sordoni, Adam Trischler, Aaron Courville, and Christopher Pal. “Towards text generation with adversarially learned neural outlines.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 146
- [SSB16] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. “Building end-to-end dialogue systems using generative hierarchical neural network models.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2016. 120, 141
- [SSK20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. “Score-Based Generative Modeling through Stochastic Differential Equations.” In *International Conference on Learning Representations (ICLR)*, 2020. 121, 142, 143
- [SSL17] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. “A hierarchical latent variable encoder-decoder model for generating dialogues.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 141

- [SSM21] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. “D2C: Diffusion-Denoising Models for Few-shot Conditional Generation.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 142, 143
- [SSZ17] Jake Snell, Kevin Swersky, and Richard S. Zemel. “Prototypical Networks for Few-shot Learning.” *Advances in Neural Information Processing Systems*, 2017. 19, 24, 26, 27
- [STG20] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. “Alfred: A benchmark for interpreting grounded instructions for everyday tasks.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020. 54
- [SWM15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2015. 121, 124, 142
- [SX14] Shuran Song and Jianxiong Xiao. “Sliding shapes for 3D object detection in depth images.” In *European conference on computer vision*, pp. 634–651. Springer, 2014. 54
- [SYC20] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. “Alfworld: Aligning text and embodied environments for interactive learning.” *arXiv preprint arXiv:2010.03768*, 2020. 54
- [SYH18] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. “Scaling human-object interaction recognition through zero-shot learning.” In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1568–1576. IEEE, 2018. 102, 106, 110
- [SZM20] Wenxian Shi, Hao Zhou, Ning Miao, and Lei Li. “Dispersed Exponential Family Mixture VAEs for Interpretable Text Generation.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2020. 120, 132, 136, 141, 145
- [SZZ18] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. “A corpus for reasoning about natural language grounded in photographs.” *arXiv preprint arXiv:1811.00491*, 2018. 3
- [TB19] Hao Tan and Mohit Bansal. “Lxmert: Learning cross-modality encoder representations from transformers.” *arXiv preprint arXiv:1908.07490*, 2019. 104

- [TF96] Joshua Tenenbaum and William Freeman. “Separating style and content.” *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, **9**, 1996. 54
- [TPB00] Naftali Tishby, Fernando C Pereira, and William Bialek. “The information bottleneck method.” *arXiv preprint physics/0004057*, 2000. 127
- [TUR50] A. M. TURING. “I.—COMPUTING MACHINERY AND INTELLIGENCE.” *Mind*, **LIX**(236):433–460, 10 1950. 1
- [TWC20a] Damien Teney, Peng Wang, Jiewei Cao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. “V-PROM: A benchmark for visual reasoning using visual progressive matrices.” In *AAAI*, 2020. 9, 10, 12, 15, 28
- [TWC20b] Damien Teney, Peng Wang, Jiewei Cao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. “V-prom: A benchmark for visual reasoning using visual progressive matrices.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12071–12078, 2020. 110
- [TZ02] Zhuowen Tu and Song-Chun Zhu. “Image segmentation by data-driven Markov chain Monte Carlo.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **24**(5):657–673, 2002. 70
- [TZD20] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. “Meta-dataset: A dataset of datasets for learning to learn from few examples.” In *ICLR*, 2020. 3, 9, 15, 19
- [UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance normalization: The missing ingredient for fast stylization.” *arXiv preprint arXiv:1607.08022*, 2016. 82
- [UVL20] D Ulyanov, A Vedaldi, and V Lempitsky. “Deep image prior.” *International Journal of Computer Vision (IJCV)*, **128**(7), 2020. 74
- [VBL16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. “Matching networks for one shot learning.” In *NeurIPS*, pp. 3630–3638, 2016. 3, 9, 15, 27
- [Vin11] Pascal Vincent. “A connection between score matching and denoising autoencoders.” *Neural Computation*, 2011. 142
- [VK20] Arash Vahdat and Jan Kautz. “NVAE: A Deep Hierarchical Variational Autoencoder.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 143

- [VKK21] Arash Vahdat, Karsten Kreis, and Jan Kautz. “Score-based Generative Modeling in Latent Space.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 122, 142, 143
- [VLL10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.” 2010. 145
- [VMB21] Andrey Voynov, Stanislav Morozov, and Artem Babenko. “Object segmentation without labels with large-scale generative models.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 75
- [VSP17a] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In *Advances in neural information processing systems*, pp. 5998–6008, 2017. 1
- [VSP17b] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, **30**, 2017. 21, 46, 47
- [WBM10] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. “Caltech-UCSD Birds 200.” Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 83
- [WBW11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. “The Caltech-UCSD Birds-200-2011 Dataset.” Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 15
- [WCL22] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. “HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 40
- [WDM19] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. “Embodied question answering in photorealistic environments with point cloud perception.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6659–6668, 2019. 40, 42
- [WGX19] Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. “Topic-Guided Variational Auto-Encoder for Text Generation.” In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 120

- [WHC19] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. “Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6629–6638, 2019. 54
- [WHG17] Xiaolong Wang, Kaiming He, and Abhinav Gupta. “Transitive invariance for self-supervised visual representation learning.” In *Proceedings of the IEEE international conference on computer vision*, pp. 1329–1338, 2017. 100
- [Win71] Terry Winograd. “Procedures as a representation for data in a computer program for understanding natural language.” Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971. 3
- [WJE19] Xiaolong Wang, Allan Jabri, and Alexei A Efros. “Learning correspondence from the cycle-consistency of time.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2566–2576, 2019. 100
- [WKM19a] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. “Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames.” *arXiv preprint arXiv:1911.00357*, 2019. 54
- [WKM19b] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. “Detectron2.” <https://github.com/facebookresearch/detectron2>, 2019. 83
- [WL21] Antoine Wehenkel and Gilles Louppe. “Diffusion Priors In Variational Autoencoders.” *arXiv preprint arXiv:2106.15671*, 2021. 142
- [WMB17] Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. “Latent intention dialogue models.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2017. 120, 141
- [WR12] Erik Weitnauer and Helge Ritter. “Physical bongard problems.” In *Ifip international conference on artificial intelligence applications and innovations*, pp. 157–163. Springer, 2012. 28
- [WT11] Max Welling and Yee W Teh. “Bayesian learning via stochastic gradient Langevin dynamics.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2011. 76, 80, 90, 124, 129
- [WTB22] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. “Emergent abilities of large language models.” *arXiv preprint arXiv:2206.07682*, 2022. 52

- [WWS22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. “Chain of thought prompting elicits reasoning in large language models.” *arXiv preprint arXiv:2201.11903*, 2022. 52
- [WXL21a] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. “Pvtv2: Improved baselines with pyramid vision transformer.” *arXiv preprint arXiv:2106.13797*, 2021. 102
- [WXL21b] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.” *arXiv preprint arXiv:2102.12122*, 2021. 94
- [WXT21] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. “Debiased Visual Question Answering from Feature and Sample Perspectives.” **34**:3784–3796, 2021. 51
- [WXW18] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. “Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation.” In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 37–53, 2018. 54
- [WYC21] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. “STAR: A benchmark for situated reasoning in real-world videos.” In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 55
- [WZS21] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. “Dense contrastive learning for self-supervised visual pre-training.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021. 97, 100
- [XK17] Xide Xia and Brian Kulis. “W-net: A deep model for fully unsupervised image segmentation.” *arXiv preprint arXiv:1711.08506*, 2017. 70, 74
- [XLZ16] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. “A theory of generative convnet.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2016. 142
- [XLZ19] Xu Xie, Hangxin Liu, Zhenliang Zhang, Yuxing Qiu, Feng Gao, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. “Vrgym: A virtual testbed for physical and interactive ai.” In *Proceedings of the ACM Turing Celebration Conference-China*, pp. 1–6, 2019. 54

- [XMY16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. “Msr-vtt: A large video description dataset for bridging video and language.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016. 64
- [XMY21a] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. “Halma: Humanlike abstraction learning meets affordance in rapid problem solving.” *arXiv preprint arXiv:2102.11344*, 2021. 28
- [XMY21b] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. “Halma: Humanlike abstraction learning meets affordance in rapid problem solving.” *arXiv preprint arXiv:2102.11344*, 2021. 110
- [YCH21] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. “3D Question Answering.” *arXiv preprint arXiv:2112.08359*, 2021. 42, 54, 55
- [YGL20] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. “CLEVRER: Collision Events for Video Representation and Reasoning.” In *ICLR*, 2020. 28
- [YGT13] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. “Pomdp-based statistical spoken dialog systems: A review.” *Proceedings of the IEEE*, 2013. 120
- [YKB17] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. “Lr-gan: Layered recursive generative adversarial networks for image generation.” In *International Conference on Learning Representations (ICLR)*, 2017. 71, 73
- [YLS19] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. “Unsupervised moving object detection via contextual information separation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 71, 74
- [YLS21] Yanchao Yang, Brian Lai, and Stefano Soatto. “DyStaB: Unsupervised Object Segmentation via Dynamic-Static Bootstrapping.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 71, 74
- [YWG18] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. “Neural-symbolic vqa: Disentangling reasoning from vision and language understanding.” *arXiv preprint arXiv:1810.02338*, 2018. 94
- [YXM21] Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. “Unsupervised Foreground Extraction via Deep Region Competition.” *Advances in Neural Information Processing Systems*, **34**, 2021. 6, 110, 142

- [YYC19a] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. “Deep modular co-attention networks for visual question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6281–6290, 2019. 41, 48, 59, 64
- [YYC19b] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. “Deep modular co-attention networks for visual question answering.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6281–6290, 2019. 94, 104, 105
- [YYD21] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. “CLEVR3D: Compositional Language and Elementary Visual Reasoning for Question Answering in 3D Real-World Scenes.” *arXiv preprint arXiv:2112.11691*, 2021. 42, 54, 55
- [ZBF19] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. “From recognition to cognition: Visual commonsense reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6720–6731, 2019. 55
- [ZGB16] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. “Visual7w: Grounded question answering in images.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016. 55
- [ZGJ19a] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. “Raven: A dataset for relational and analogical visual reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5317–5327, 2019. 3, 28
- [ZGJ19b] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. “Raven: A dataset for relational and analogical visual reasoning.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5317–5327, 2019. 110
- [ZKH21] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. “Scaling vision transformers.” *arXiv preprint arXiv:2106.04560*, 2021. 97
- [ZKZ18] Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. “Adversarially regularized autoencoders.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 120, 132, 141
- [ZLE18] Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. “Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 120, 132, 135, 141, 145



- [ZLW14] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. “Saliency optimization from robust background detection.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 70
- [ZM07] Song-Chun Zhu and David Mumford. *A stochastic grammar of images*. Now Publishers Inc, 2007. 110
- [ZSQ17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. “Pyramid scene parsing network.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 70
- [ZWH21a] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. “End-to-end human object interaction detection with hoi transformer.” In *CVPR*, 2021. 37
- [ZWH21b] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. “End-to-End Human Object Interaction Detection with HOI Transformer.” In *CVPR*, 2021. 13, 21, 22, 24, 27
- [ZWM98] Song-Chun Zhu, Yingnian Wu, and David Mumford. “Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling.” *International Journal of Computer Vision (IJCV)*, **27**(2):107–126, 1998. 71
- [ZWS18] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. “HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection.” In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018. 15
- [ZXS16] Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. “Variational Neural Machine Translation.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 141
- [ZY96] Song-Chun Zhu and Alan Yuille. “Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **18**(9):884–900, 1996. 70, 71, 72, 76, 80
- [ZZE17] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. “Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 120, 141
- [ZZL15] Xiang Zhang, Junbo Zhao, and Yann LeCun. “Character-level convolutional networks for text classification.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 139

- [ZZL19] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. “Joint learning of saliency detection and weakly supervised semantic segmentation.” In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 75