Introducing Spatial Information and a Novel Evaluation Scheme for Open-Domain Live Commentary Generation

Erica K. Shimomoto[1*], **Edison Marrese-Taylor**[1,3*], **Ichiro Kobayashi**[1,2], **Hiroya Takamura**[1], **Yusuke Miyao**[1,3]
National Institute of Advanced Industrial Science and Technology[1]
Ochanomizu University[2], The University of Tokyo[3]
kidoshimomoto.e@aist.go.jp, edison.marrese@aist.go.jp, koba@is.ocha.ac.jp
takamura.hiroya@aist.go.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

This paper focuses on the task of open-domain live commentary generation. Compared to domain-specific work in this task, this setting proved particularly challenging due to the absence of domain-specific features. Aiming to bridge this gap, we integrate spatial information by proposing an utterance generation model with a novel spatial graph that is flexible to deal with the open-domain characteristics of the commentaries and significantly improves performance. Furthermore, we propose a novel evaluation scheme, more suitable for live commentary generation, that uses LLMs to automatically check whether generated utterances address essential aspects of the video via the answerability of questions extracted directly from the videos using LVLMs. Our results suggest that using a combination of our answerability score and a standard machine translation metric is likely a more reliable way to evaluate the performance in this task.

1 Introduction

Understanding videos has been a central task in developing machine learning, both in computer vision (CV) and natural language processing (NLP). A wide variety of vision-and-language (V&L) tasks has been developed over the years, including generative tasks like video captioning (Venugopalan et al., 2015; Krishna et al., 2017), predictive tasks like action classification (Fabian Caba Heilbron and Niebles, 2015), and localization (Escorcia et al., 2016), and grounding via temporal moment localization (Gao et al., 2017; Hendricks et al., 2017).

This paper focuses on automatically generating live commentary for videos in an open-domain setting, a task proposed recently by Marrese-Taylor et al. (2022). This task is very similar to video captioning, as comments may include descriptions of what is happening in videos, but it differs significantly because the timing and contents of utterances may vary significantly, as a commentator is free to choose what to say, when to say and how to say things (Taniguchi et al., 2019; Kim and Choi, 2020). This setting proved particularly challenging due to the absence of domain-specific features. This fact was clearly shown by the stark contrast of BLEU scores, which amounts to a difference of approximately one order of magnitude (2.38 v/s 24.01) (Ishigaki et al., 2021).

Looking at V&L tasks, we see that approaches often focus on creating multi-modal contextualized embeddings. These embeddings generally rely on activity-centric features derived from image classification models, such as ViT (Dosovitskiy et al., 2020) or activity classification models, such as I3D (Carreira and Zisserman, 2017), and contextualized text embeddings (Devlin et al., 2019). However, previous works also show that solving these tasks often requires reasoning about relationships between objects and subjects present in the video (Hu et al., 2019; Rodriguez-Opazo et al., 2021).

We observe a similar trend in live commentary generation and hypothesize that due to the complexity of the task, which can be seen as a combination of more fundamental ones (Marrese-Taylor et al., 2022), incorporating the spatial dimension is critical to improve performance as it allows models to reason about objects and subjects in the video.

Therefore, we propose a spatial graph to contextualize features derived from an action classification model via a language-conditioned message-passing algorithm utilizing information extracted from objects in the video. Unlike previous work, our graph gives the model freedom to learn relationships between objects, as commentaries might cover aspects not centered on humans. These updated features are then used to generate live commentary via a Transformer encoder-decoder model.

Furthermore, we note that evaluation for live

[*] Authors contributed equally to this work.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10352–10370
November 12-16, 2024 ©2024 Association for Computational Linguistics

commentary generation has so far relied on n-gram overlap metrics such as BLEU (Papineni et al., 2002). However, given the open-ended nature of the task, commentators may choose to talk about a given subject at different times and attend to different points. Thus, we believe evaluation schemes based on text similarity are fundamentally limited, as they cannot correctly capture variations across commentaries for the same video.

Thus, we also propose a novel evaluation approach based on question answerability to complement text similarity-based evaluation. Concretely, we use LLMs to evaluate if commentary utterances address essential aspects of the video by checking if key questions extracted directly from the video using LVLMs can be answered.

Our results highlight the importance of considering spatial information for open-domain live commentary generation, substantially improving performance and allowing this setting to be competitive with in-domain instances of the task. Critically, we show that not constraining the model to human-centered activities leads to the best performance. Finally, our findings suggest that our proposed answerability metric, along with a standard machine translation metric, such as BLEU, can be a more reliable way to evaluate the performance of models in this task, as it helps us identify instances where the generated utterances are not very informative.

2 Related Work

Video Captioning Developing techniques that can automatically describe what happens in a video remains an open challenge. To the best of our knowledge, work by Venugopalan et al. (2015) was the first to tackle the task of describing videos in an open-domain setting. This task was extended by Krishna et al. (2017), who proposed dense video captioning (DVC), where a model is required to detect multiple events that occur in a video and then describe each one using natural language. This task was initially tackled using a pipeline approach, where relevant segments in the input video are first identified. Then, a captioning model generates the natural language descriptions of the identified zones (Wang et al., 2020; Deng et al., 2021). More recently, end-to-end approaches have also been proposed (Zhou et al., 2018; Wang et al., 2021).

Live Commentary Generation To the best of our knowledge, the task of automatically generating live commentaries was first proposed by Ishigaki et al. (2021) in the context of racing car videogame streams, releasing the first dataset annotated for this task, which consisted of gameplay videos aligned with transcribed spoken commentaries. Soon after, Marrese-Taylor et al. (2022) tackled this task in an open-domain setting, detailing the construction of a dataset of transcribed commentary aligned with videos containing human actions in a variety of domains constructed using videos from ActivityNet (Fabian Caba Heilbron and Niebles, 2015). Both works also proposed models to generate such commentaries automatically. While the former model heavily relied on telemetry data extracted from the game API to achieve the best results, the latter did not use domain-specific information, achieving considerably poorer performance. Finally, it is also worth mentioning that other works, including Taniguchi et al. (2019), Qi et al. (2023) and Kim and Choi (2020) have previously worked on the related task of automatically generating commentary for sports matches also relying on domain-specific data.

Role of Spatial Information in V&L tasks Previous work has shown that access to the spatial dimension allows models to perform better in video understanding tasks. For example, Sigurdsson et al. (2017) showed that the performance in action recognition tasks improves significantly if we have a perfect object recognition oracle. Rodriguez-Opazo et al. (2021) also showed that incorporating silver-standard information in the form of automatically detected objects for the temporal video grounding task achieves substantially better performance. Our proposed graph diverges substantially from recent works since we tackle a generative task.

Captioning Evaluation Several methods have been proposed to evaluate the quality of generated descriptions of both images and videos, including BLEU (Papineni et al., 2002), SPICE (Anderson et al., 2016), METEOR (Lavie and Denkowski, 2009), CIDEr (Vedantam et al., 2015) and ML-based metrics such as BERTScore (Zhang et al., 2019). Our work is related to these in that we are also proposing a new evaluation scheme, but our approach is directly tailored to the live commentary generation task.

3 Proposed Approach

Consider a video $V = \{v_n\}_1^N$ which has N video segments. Our goal is to generate a commentary

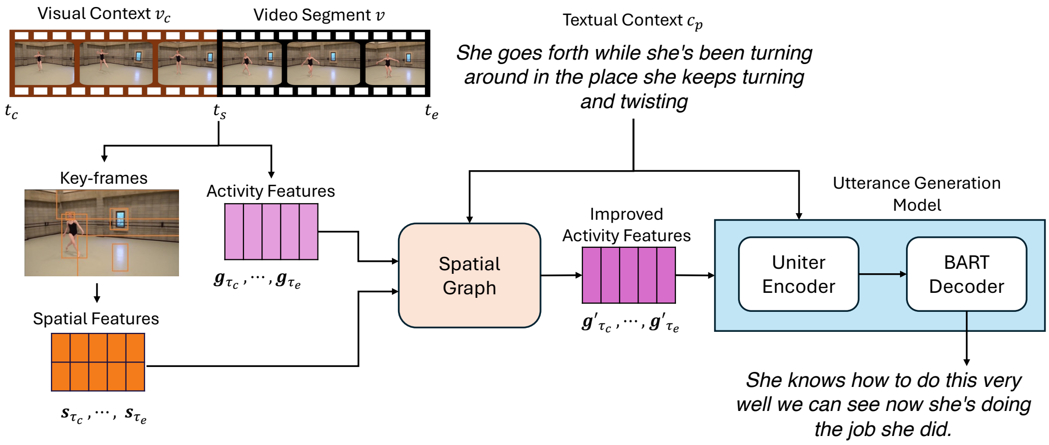

Visual Context v_c
Video Segment v
t_c
t_s
t_e
Textual Context c_p
She goes forth while she's been turning around in the place she keeps turning and twisting
Key-frames
Activity Features
$g_{\tau_c}, \cdots, g_{\tau_e}$
Spatial Features
$s_{\tau_c}, \cdots, s_{\tau_e}$
Spatial Graph
Improved Activity Features
$g'_{\tau_c}, \cdots, g'_{\tau_e}$
Utterance Generation Model
Uniter Encoder
BART Decoder
She knows how to do this very well we can see now she's doing the job she did.

Figure 1: Proposed utterance generation model. It contains two main parts: 1) The spatial graph exploits the relationships between spatial information extracted from v and v_c, conditioning them on an attended language representation from previous utterances c_p, improving activity representations; 2) A transformer-based utterance generation model, that receives the improved activity representations and the previous utterances, generating an utterance commenting the video segment v.

utterance for each video segment. We will describe our model, focusing on a single video segment for simplicity. Concretely, given a video segment v with timestamps (t_s, t_e), we feed v to our model along with its visual context v_c with timestamps (t_c, t_s), where $t_c \leq t_s$, resulting in an input video segment v_{in} with timestamps (t_c, t_e). We also feed its textual context c_p, a sequence of tokens from the previous p utterances. Our model then generates an utterance u, commenting on the video segment v. We train our model to minimize the cross entropy between generated and gold-standard utterances.

3.1 Integrating Spatial Information

Figure 1 illustrates our proposed architecture. The model is built on top of the utterance generation model proposed by Marrese-Taylor et al. (2022), a transformer-based model with a UNITER multimodal encoder (Chen et al., 2020) and a BART decoder (Lewis et al., 2020). To integrate spatial information, we introduce a spatial graph inspired by Rodriguez-Opazo et al. (2021) to improve the video representations. These improved video representations and the textual context are fed to the utterance generation model, which then generates the utterances for each video segment.

Spatial Graph The spatial graph exploits the relationships between spatial information in a given scene, conditioning them on an attended language representation from previous utterances. Such information should help the utterance generation model better understand the video's contents, leading to better grounded utterances.

Concretely, we utilize a language-conditioned message-passing algorithm to obtain contextualized video representations inspired by Rodriguez-Opazo et al. (2021). While they designed the graph to account for specific relationships in human-centered activities, we propose a general graph that does not distinguish spatial features, given the open-domain aspect of our task. Figure 2 illustrates our spatial graph. It consists of three semantic nodes, one representing linguistic information and the other two representing visual information.

The **linguistic node** $\mathcal{L}$ is designed to capture essential information in the previous utterances related to the input video segment v_{in}. Using a pre-trained word embedding model, we first encode each word w_j in the textual context c_p as a semantic embedding vector $\boldsymbol{h}_j \in \mathbb{R}^{d_w}$. Then, we initialize an attention module using an aggregated, fixed-length query vector $\boldsymbol{q}$. This vector is constructed using a bi-directional GRU over the word embeddings and mean-pooling. The key component $\boldsymbol{k}$ of the attention module is obtained by projecting the word embeddings $\boldsymbol{h}_j$ using a linear mapping, and the value component $\boldsymbol{v}$ comes from the contextualized word representations from the GRU. The attention module attends to these contextualized representations and returns a re-weighted combination $\boldsymbol{L} = softmax(\boldsymbol{q}\boldsymbol{k}^\top)\boldsymbol{v}$ that initializes the linguistic node.

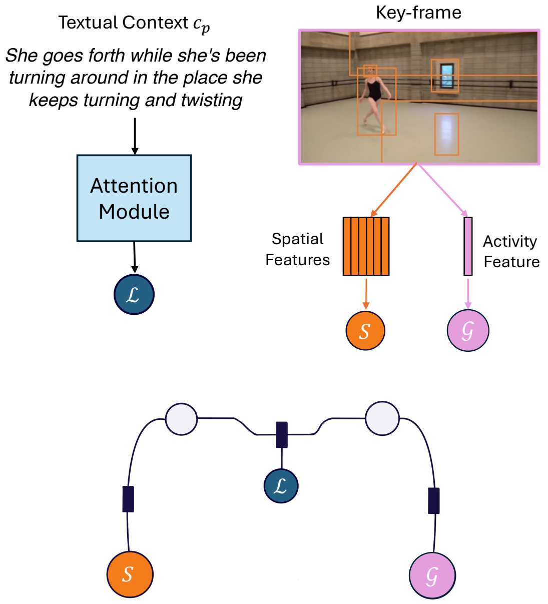


Figure 2: Our proposed OPEN graph. It has one linguistic node $\mathcal{L}$ that captures essential information in the previous utterances related to the input video segment, and two visual nodes: $\mathcal{S}$ that processes spatial information, and $\mathcal{G}$, that processes activity information. A mean-field-like approximation of the message-passing algorithm captures the relationships between visual representations, where messages sent between $\mathcal{S}$ and $\mathcal{G}$ are conditioned on $\mathcal{L}$.

The **visual nodes** consist of an activity node $\mathcal{G}$ and a spatial node $\mathcal{S}$, where each is a latent representation of the corresponding observation. We encode the video V into two different representations to initialize these nodes.

First, we use a function $E : f \mapsto \boldsymbol{g}$, which maps the F video frames into a sequence of activity features $\{\boldsymbol{g}_m \in \mathbb{R}^{d_v}\}$, $m = 1, \dots, M$, accompanied by a mapping function* that allows us to transform timestamps t into feature indices $\tau \in \{1, \dots, M\}$. These activity features are extracted from non-overlapping spans of frames from the whole video V, summarizing spatio-temporal patterns directly from the raw frames. Then, we map the start and end timestamps of v_{in} to feature indices τ_c and τ_e with $1 \leq \tau_c \leq \tau_e \leq M$, and represent v_{in} as a sequence of $T = \tau_e - \tau_c + 1$ features $\boldsymbol{g}_{\tau_c}, \dots, \boldsymbol{g}_{\tau_e}$. These features are observed by the video node $\mathcal{G}$.

Second, we extract spatial information from the key-frame associated with each activity feature $\boldsymbol{g}_\tau$ by using an object detector, resulting in a sequence of spatial features $S = \{\boldsymbol{S}_\tau \in \mathbb{R}^{O \times d_s}\}$, where O is the number of objects detected in each key-frame and d_s is the dimension of the object features. These features are observed by the spatial node $\mathcal{S}$.

Finally, a **mean-field-like approximation of the message-passing algorithm** captures the relationships between spatial and activity representations. Messages are iteratively sent between visual nodes and are conditioned on the linguistic node.

We start by capturing the relationship between the spatial observations of $\mathcal{S}$ and the linguistic node $\mathcal{L}$ using a linear mapping function specific for each node. For simplicity of notation, we will drop the temporal index τ, as the message-passing is done for each activity feature independently.

$$f_{\mathcal{L},\mathcal{S}}(\mathcal{L}, \boldsymbol{s}_{o,i}) = \boldsymbol{W}_{ls}[\mathcal{L}; \boldsymbol{s}_{o,i}] + \boldsymbol{b}_{ls} = \boldsymbol{\Phi}_{o,i}^{\mathcal{L},\mathcal{S}}, \quad (1)$$

$$f_{\mathcal{L},\mathcal{G}}(\mathcal{L}, \boldsymbol{g}_i) = \boldsymbol{W}_{lg}[\mathcal{L}; \boldsymbol{g}_i] + \boldsymbol{b}_{lg} = \boldsymbol{\Phi}_{i}^{\mathcal{L},\mathcal{G}}, \quad (2)$$

where o is the o-th spatial observation in $\boldsymbol{S}_\tau$ and i is the iteration step.

The spatial node $\mathcal{S}$ receives messages from the activity node $\mathcal{G}$. These messages are constructed using a linear mapping function that receives as input the concatenation of the spatial-query relationship and the activity-query relationship:

$$\boldsymbol{\Psi}_{o,i}^{\mathcal{G},\mathcal{L},\mathcal{S}} = f_{\mathcal{G},\mathcal{L},\mathcal{S}}(\boldsymbol{\Phi}_{o,i}^{\mathcal{L},\mathcal{S}}, \boldsymbol{\Phi}_{i}^{\mathcal{L},\mathcal{G}}) \quad (3)$$

Finally, the new representation of the spatial representation is computed as:

$$\boldsymbol{s}_{o,i+1} = \sigma(m_s(\boldsymbol{\Psi}_{o,i}^{\mathcal{G},\mathcal{L},\mathcal{S}}) \odot \boldsymbol{s}_{o,i}), \quad (4)$$

where σ is an activation function, $\odot$ is the Hadamard product, and m_s is a linear function with a bias that constructs the message for the spatial feature $\boldsymbol{s}_o$.

A similar process is applied for the activity node $\mathcal{G}$. However, as the messages need to consider all spatial observations, they are constructed using an aggregation of all spatial-query relationships:

$$\boldsymbol{\Psi}_{i}^{\mathcal{S},\mathcal{L},\mathcal{G}} = f_{\mathcal{S},\mathcal{L},\mathcal{G}}(\boldsymbol{\Phi}_{i}^{\mathcal{L},\mathcal{G}}, \sum_o \boldsymbol{\Phi}_{o,i}^{\mathcal{L},\mathcal{S}}) \quad (5)$$

$$\boldsymbol{g}_{i+1} = \sigma(m_g(\boldsymbol{\Psi}_{i}^{\mathcal{S},\mathcal{L},\mathcal{G}}) \odot \boldsymbol{g}_i) \quad (6)$$

The messages are iteratively sent a total of I times, and the resulting representation of the visual node $\mathcal{G}$ is fed to the utterance generation model.

*We apply the mapping $\tau = (t \cdot \text{fps} \cdot M)/F$ to transform time t to feature index τ. F is the total number of frames in the video and M is total number of features extracted from it.

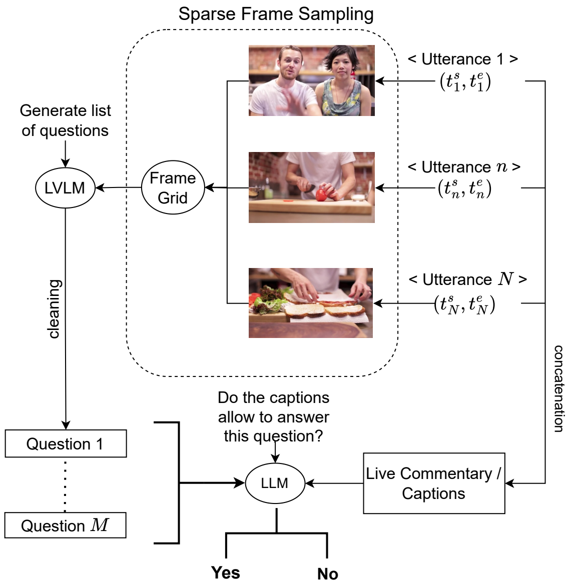


Figure 3: Summary of our approach for question answerability-based evaluation: (1) Generate questions via an LVLM, which receives a visual summary of the video shown on the upper part of the diagram; (2) Evaluate question answerability via an LLM by counting how many questions can be answered from the generated commentary, shown on the lower part of the diagram.

3.2 Evaluation via question-answerability

The task of live commentary generation has been previously motivated by suggesting commentaries can help make spectators more excited, more immersed, and better informed about the content they are viewing (Schaffrath, 2003; Ishigaki et al., 2021). However, what kind of information is provided to the spectators remains unclear.

The work of Marrese-Taylor et al. (2022) showed that inter-annotator agreement in terms of commentary content measured via BLEU-4 and SPICE is significantly lower compared to other video-to-text datasets such as ActivityNet Captions (Krishna et al., 2017) (2.59 and 0.034 v/s 4.86 and 0.12, respectively). They attribute this difference to the nature of the task, where each annotator has to decide when and what to speak independently. While it means that each commentator may produce significantly different output utterances, we think it is reasonable to assume there will be a set of common-ground facts shared across annotators, which depend only on the visual contents of the video.

To that end, we propose to rely solely on the visual modality to generate a set of questions for each video, which one should expect to be answerable by looking at the transcript of a live commentary. These questions can later be compared with the output of a given model to measure how many key aspects of the video are covered. Figure 3 provides an overview of our proposal.

First, to generate relevant questions, we propose to rely on a Large Vision-and-Language Model (LVLM). We prompt it with a visual summary of the video, requesting it to generate a list of relevant questions for someone providing live commentary. Our prompt asks the model to generate questions whose answers describe aspects of the actions in the video to listeners who cannot see it for themselves. Critically, we request only to include questions with definite answers, following Liu et al. (2023).

To construct the visual summary, we sample frames from each temporal segment defined by the gold standard utterances and construct a grid-like input image. Noting that many frames in a segment can be blurry, we select the sharpest frame that appears closest to the middle of the segment. Please see §B in our supplementary material for more details, including our full prompt in §C.

Once the questions have been collected, they are processed to remove noise. First, we remove questions that ask for details of specific frames. As the questions generated by LVLM can be similar or related, we rely on a pre-trained NLI model to identify groups of generated questions that are entailed to each other. After identifying these groups, we randomly sample one question from each group.

During early experiments, we observed that the LVLM could not generate relevant questions in some cases. This issue happened when the video contained multiple cuts-scenes or dark/blurry imagery. To obtain questions for these videos, we propose to sample from the most common questions across all videos for which we have questions. Please see §D for examples of these questions. We empirically found that this set contains critical information-seeking, general questions that apply well to our dataset.

Then, we evaluate question answerability via an LLM, which we prompt to answer if a given question can be answered given the generated commentary (please see §C for the entire prompt).

The final step in our evaluation scheme is to compute a score that adequately summarizes the number of key video aspects covered. As a base score, we propose to use the ratio between the total number of questions available for a video V, which we denote as Q, and the number of questions that are answerable via the judge LLM, denoted as q, such that $\rho = q/Q$. We further length-correct this score to discourage the utterances from being too short or too long via a parameter L such that our final **answerability score** is defined as $\rho^* = (q + \log(l/L))/Q$.

4 Experimental Setting

Data We work with the dataset by Marrese-Taylor et al. (2022), which, to the best of our knowledge, is the only one to tackle live commentary generation in the open-domain scenario. This dataset was built on top of the videos in the ActivityNet Dataset (Caba Heilbron et al., 2015), where human annotators were asked to record commentary-like narrations of the videos in English under two settings: (1) without prior knowledge and (2) after watching them once. These audios were then transcribed into text. It consists of 25k commentaries, covering a total of 6,771 videos. In our experiments, we considered only the annotations in setting (1), resulting in 6,854 commentaries for training and 6142 commentaries for validation.

After an initial inspection, we noticed some quality problems in the data that prompted us to conduct an in-depth human assessment, which we detail in §A. This study indicated that a significant portion of the problems we identified were partially due to issues in the transcripts. Therefore, we relied on WHISPER (Radford et al., 2023) (*openai/whisper-large*), a speech-to-text model based on a Transformer encoder-decoder with 1,550 M parameters, to once again transcribe the recordings, which were kindly provided to us by the authors[†]. In Section 5.1, we detail the impact of this step via several experiments and directly use our answerability metric to assess data quality.

Model Our Transformer-based model uses $d_m = 512$, with a total of 257M parameters, and is trained with a maximum learning rate of 10^{-4} with Adam and a linear annealing for 5% of the epochs, with a batch size of 4 using 4 NVIDIA V100 GPUs. During inference, we utilize beam search with a beam size of 5. It takes around 20 minutes to train one epoch and 1 hour to evaluate once.

For the initial representation of the video, we utilize the same offline video encoding function of the model proposed by Marrese-Taylor et al. (2022), relying on the I3D features released by Rodriguez-Opazo et al. (2021). Furthermore, we obtain the initial representation for the textual context for the spatial graph using a pre-trained GLoVe (Pennington et al., 2014) embeddings (glove.840B.300d).

The spatial data is based on an object detector, which is applied to a set of key-frames associated with each activity representation to extract. Following Rodriguez-Opazo et al. (2021), we select these key-frames using the Laplace variance algorithm (Pech-Pacheco et al., 2000). Our approach is agnostic to the choice of object detector, but in this paper, we use Faster RCNN (Ren et al., 2015; Anderson et al., 2018), which was trained on the Visual Genome (Krishna et al., 2016) dataset. The dataset contains 1,600 object categories, manually assigned to either human or object labels by Rodriguez-Opazo et al. (2021).

Evaluation via question answerability We use LLaVA v1.6 (Liu et al., 2023) as our LVLM to generate the questions. We empirically found that this version could adequately handle context from up to (at least) 8 frames sampled from the video. We use the same technique detailed above to identify sharp frames (for more details, please refer to §B). For the NLI-based data cleaning procedure, we relied on a multi-task finetuned DeBERTa-v3 (He et al., 2023) (*sileod/deberta-v3-small-tasksource-nli*), which can be used for zero-shot NLI. Finally, as our judge LLM, we used Llama3 (AI@Meta, 2024), specifically the 8B version (*Meta-Llama-3-8B-Instruct*), which we quantize to 4 bits via QLoRA (Dettmers et al., 2023).

5 Results

5.1 Improved data quality

We begin by showing the impact of using WHISPER transcriptions by employing our answerability score ρ^*. For this study, we feed the original transcripts (o) and our versions (w) to the question-answering-based evaluation. We repeat this process for the training and evaluation portions of the data. For the original transcripts, we obtained answerability scores of 0.447 and 0.439 for the training and evaluation portions, respectively. These scores improve respectively to 0.457 and 0.446 with WHISPER transcripts, where differences are statistically significant at $\alpha = 0.05$, via Welch's paired t-test. These results suggest that WHISPER can better recover key information from the recordings.

To check the impact of using WHISPER, we evaluate the performance of utterance generation via the model proposed by Marrese-Taylor et al. (2022)

[†] Some of the recordings were missing, so we only transcribed 92.85% of recordings in the ORIGINAL dataset.

Model		**BLEU**	ρ^*
No Spatial	Data		
BASELINE	(o)	2.29 ± 0.13	0.423 ± 0.031
	(w)	2.74 ± 0.04	0.436 ± 0.003
With Spatial	AH		
SPECIFIC	3	24.50 ± 2.55	0.545 ± 0.018
GENERAL	3	23.43 ± 1.20	0.555 ± 0.009
	2	21.26 ± 1.99	0.558 ± 0.018
	1	24.25 ± 3.03	0.550 ± 0.016
OPEN (Ours)	1	$\mathbf{28.49} \pm \mathbf{1.52}$	$\mathbf{0.564} \pm \mathbf{0.009}$

Table 1: Utterance generation performance when introducing spatial information through a spatial graph. BASELINE refers to the utterance generation model proposed by Marrese-Taylor et al. (2022). SPECIFIC refers to the spatial graph proposed by Rodriguez-Opazo et al. (2021). GENERAL refers to a modified version of SPECIFIC with only one type of spatial node while keeping the multi-head attention. OPEN refers to our proposed graph. AH refers to the number of attention heads in the attention module in the spatial graphs.

(BASELINE) using the new transcripts. We choose the visual context to span from 6 seconds prior to the beginning of the video segment and feed the previous 3 gold standard utterances as textual context, ($t_c = t_s - 6, p = 3$), as this setting led to the best results with the original transcripts. To allow direct comparison with previous results, we also use the BLEU score between the gold standard and generated utterances (BLEU). Finally, we trained them three times to better estimate each model's performance. Table 1 reports the average and standard deviation.

Using WHISPER data led to an improvement in all considered metrics. Although only the improvement in BLEU is statistically significant at $\alpha = 0.05$, given these results and the answerability score of the training and evaluation portions of the dataset, WHISPER transcripts are likely superior to the original transcripts.

5.2 Spatial Information Integration

In this section, we conduct a series of experiments to investigate the effects of spatial information integration. For all experiments, unless stated differently, we train the models three times using different seeds and report the average and standard deviation of all considered metrics.

Spatial Features We evaluate and compare our proposed model OPEN with the BASELINE model, which does not use spatial information. Given the results in Section 5.1, we used only WHISPER data and set the context window to ($t_c = ts - 6, p = 3$). Furthermore, we set the number of iterations of the message-passing algorithm to $I = 2$.

Table 1 shows the results. Compared to the BASELINE, OPEN performed significantly better regarding all metrics considered. Such results agree with the human assessment of the data (§A), which revealed that most of the dataset's utterances comment on things explicitly shown in the videos. Thus, having more information regarding the video in the form of spatial features can help the model generate better utterances.

Ablation Study Given that the videos in the dataset come from the ActivityNet, where most of them show human-centered activities, we were curious to see how the specific spatial graph proposed by Rodriguez-Opazo et al. (2021) (SPECIFIC) would perform. This graph assumes that the activities shown in the videos are in the form of an action (*verb*) being performed by a *subject*, involving *objects*. Therefore, it contains two different nodes to process the spatial features.

Furthermore, we also explore a variation with only one visual node, giving more freedom to the model to establish its connections with the activity representation while keeping the multi-attention head that processes the textual context (GENERAL). We tested this model with 1 to 3 attention heads to understand their impact on performance. We refer the readers to §E for details on these variations.

The results are shown in Table 1. All the variations performed much better than the BASELINE, where our proposed graph performed best. These results reinforce the role of spatial information in this task. GENERAL results indicate that multiple attention heads cannot explore different aspects of the textual context without the guidance of the different spatial nodes, as adding attention heads hurts the performance. Even in its best performance, with only one attention head, GENERAL achieved roughly the same performance as SPECIFIC.

These results reveal that despite using the same videos, the task of live commentary generation is very different from the temporal video grounding. Commentary utterances may comment on things not directly related to the human performing the activity, such as describing the environment depicted, which makes SPECIFIC unsuitable for the task. On the other hand, the flexibility of our proposed graph

Model	Visual	BLEU	ρ^*
OPEN (Ours)	$\mathcal{G}$	20.66 ± 0.09	0.564 ± 0.010
	$\mathcal{S}$	25.45 ± 2.68	$\mathbf{0.570 \pm 0.010}$
	$\mathcal{G}+\mathcal{S}$	$\mathbf{28.49 \pm 1.52}$	0.564 ± 0.009

Table 2: Impact on the performance of our proposed OPEN graph when using only one type of visual feature.

Model	P	R	F1
Llama-3	0.87	0.49	0.60
gemma-2	0.88	0.44	0.56
gpt-3.5	0.87	0.47	0.58

Table 3: LLMs' performance in assessing the answerability of generated question based on generated commentary utterances. P, R, and F1 refer to precision, recall, and f1-score, respectively.

can better accommodate these open commentaries, achieving the best performance.

Finally, we also investigate the contribution of each visual feature in our proposed OPEN graph by removing one of the visual nodes from OPEN. Results are shown in Table 2. We can see that using only the activity node $\mathcal{G}$ or only the spatial node $\mathcal{S}$ led to great performance improvement compared to the baseline model, where spatial features contribute more to the performance. While all three variations presented similar ρ^*, the combination of both types of features led to the best BLEU score.

The role of context To better understand the impact of the context window on both visual and textual sides, we experimented with different windows. Given our computing restrictions, we only trained the model with each window once. Results are summarized in Figure 4a.

BLEU scores show that feeding no visual/textual context $(t_c = t_s, p = 0)$ leads to very poor performance. Adding only visual or only textual context results in performance improvement where textual context seems to be more critical. We believe that feeding previous utterances helps the model generate utterances that are more cohesive between one and another. Nevertheless, a combination of both textual and visual contexts led to the best results.

Finally, it is important to note that our results do not necessarily imply that longer context windows will always lead to better performance. We must consider the technical limitations, such as the maximum input length of our utterance generation model. Additionally, looking too far back in a video may not be beneficial, as it may not be relevant to the current events. Moreover, longer context windows, especially on the visual side, can significantly increase the computational cost of training and evaluation. Therefore, the choice of context window should be made with these factors in mind.

5.3 Question Answerability Score

Generated Questions We start by checking the capabilities of LVLMs to generate relevant questions, as well as LLMs' capabilities in assessing if questions can be answered given a commentary.

First, we evaluate the questions generated by LLaVA via crowdsourcing using 200 videos. For each video and its generated questions, three workers are asked to watch the video and 1) State if the question is relevant; 2) Rate the question's answerability on a scale from 0 to 4. We found that $87\% \pm 18\%$ of the questions were deemed relevant, with an average answerability score of 2.95 ± 0.37. Therefore, we assume that LLaVA can generate relevant and answerable questions from the videos.

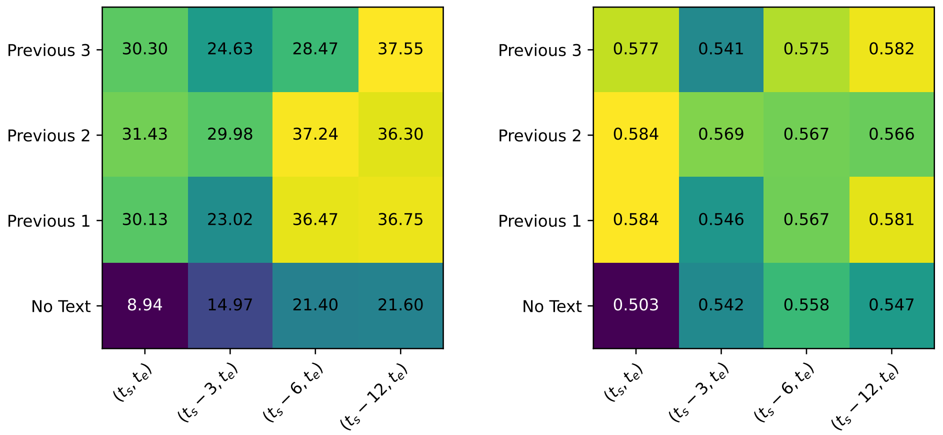
Next, we use crowdsourcing to gather ground-truth human answerability labels in a similar setting from above, where instead of the video, we show transcripts of the gold-standard commentary. We deem a question *answerable* if any worker gives it a score ≥ 3. Then, we compare these labels with the assessment of the following LLMs: Meta-Llama-3-8B-Instruct (used in the other section's results), gpt-3.5-turbo-0125, and gemma-2-9b-it.

Table 3 shows the results. LLMs miss some questions humans regard as answerable but are remarkable at identifying which ones are. Please see §F for more details on this evaluation.

Live Commentary Evaluation Finally, we evaluate the performance of our utterance generation model using the proposed answerability score ρ^*.

Table 1 shows a clear improvement from BASELINE to the models with spatial information. Also, spatial models present a higher answerability score than the gold standard utterances. This result underscores the importance of leveraging spatial information when generating live commentary.

Figure 4b summarizes ρ^* when using different context windows. We can see that while some contexts display an agreement between the BLEU scores and the answerability metric, some contexts with high BLEU scores received low answerability scores. Such discrepancies come from the fact that while BLEU evaluates the generated utterances



(a) BLEU scores.

(b) ρ^* scores.

Figure 4: Results for our proposed utterance generation model according to different visual and textual context window sizes. The y-axis denotes the amount of textual content fed, while values on the x-axis denote the time interval of video features fed.

against the gold standard utterances, the answerability score evaluates them against the video. Therefore, even a perfect BLEU score might not lead to a high answerability score if the gold standard utterance is not very informative. By manually inspecting the generation results from settings with very high BLEU scores, we notice several examples of such behavior. Please see §G for examples.

Furthermore, we also see context windows with high answerability but low BLEU scores. In such cases, we believe there are two main reasons: First, the generated utterances may include the necessary information to answer the questions but present poor grammar. Second, given that there are many different suitable commentaries for the same video segment, which often depends on how knowledgeable the commentator is regarding the topics in the videos, how engaging the commentator wants to be with the audience, among other factors, the generated utterances might lead to a very low BLEU score because it did not cover the same aspects covered by the standard gold utterances.

Therefore, a combination of our answerability score and a standard machine translation metric, such as BLEU, is likely a more reliable way to evaluate the performance in this task.

6 Conclusions

In this paper, we tackled the open-domain live commentary task and proposed improvements to two critical aspects of the task: (1) We integrated spatial information by proposing a novel spatial graph that is flexible in dealing with the open-domain characteristics of the commentaries; (2) We proposed a novel evaluation scheme for live commentary that automatically checks whether utterances addressed essential aspects of the video via answerability of questions extracted directly from the videos.

Our OPEN graph significantly improved performance, making it competitive with in-domain instances of the task. Our results highlight the importance of considering spatial information for open-domain live commentary generation, where not constraining the model to human-centered activities led to the best performance.

Moreover, results show that our proposed answerability score helps us identify instances where the generated utterances are not very informative. However, it is not very robust in terms of grammar issues. Therefore, combining our answerability score and a standard machine translation metric might be more suitable for evaluating this task's performance.

Finally, our findings highlight potential future research. We express reservations about the current dataset, particularly when the answerability score indicates a likely poor commentary gold standard. It opens up the possibility of a new dataset iteration, where we could engage other annotators to enhance the quality of the commentary. Additionally, we recognize the need to refine our proposed answerability metric, as it only measures the number of questions answered without differentiating the types of questions answered.

Limitations

In this work, we utilize a dataset containing YouTube videos with annotations in English, and we focus solely on this language. Results show that our proposed model can significantly outperform previous work by incorporating information extracted from the spatial dimension. While we believe this shows the overall effectiveness of our approach in the task of commentary generation in general, we have no evidence to suggest how well these capabilities could generalize to other languages. This point may prove especially important in low-resource cases, where access to pre-trained models is limited.

The availability of pre-trained models in languages other than English poses a significant constraint on our research. As our work primarily relies on English-specific models, the ability to conduct our answerability-based evaluation in other languages is severely limited.

Finally, access to computational resources was crucial in enabling the experiments performed for this work. We train models that contain hundreds of millions of parameters using pre-trained models like BART and UNITER while also relying on an LVLM (LLaVa v1.6) and an LLM (Llama3), which have several billion parameters for our answerability-based evaluation.

Ethics Statement

This work does not present any direct ethical issues. The dataset used to evaluate our proposed approach was shared with us directly from the original authors, and data characteristics relevant to our task were described in the experimental evaluation section. The paper includes references for further information.

Moreover, the annotators used in our human study were volunteers. They were made aware that collected data would be anonymous and used only for the purpose of this research, upon which they agreed to share their evaluations.

Finally, our proposed model was developed solely for generating live video commentary, showing a significant step in our research. However, it's important to note that relying on LLVMs and LLMs to evaluate the generated utterances carries a risk. We urge caution and thorough evaluation when dealing with the generated questions.

Acknowledgements

This paper is based on results obtained from the project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) was used for experiments. Finally, we are also grateful to the NVIDIA Corporation, which donated one of the GPUs used in this research.

References

AI@Meta. 2024. Llama 3 model card.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 382–398, Cham. Springer International Publishing.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*, pages 6077–6086.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 104–120, Cham. Springer International Publishing.

Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. 2021. Sketch, Ground, and Refine: Top-Down Dense Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. DAPs: Deep Action Proposals for Action Understanding. *ECCV*.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. 2017. TURN TAP: temporal unit regression network for temporal action proposals. *ICCV*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-Conditioned Graph Networks for Relational Reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10294–10303.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating Racing Game Commentary from Vision, Language, and Structured Data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1056–1065. Association for Computing Machinery, New York, NY, USA.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715.

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. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Preprint*, arxiv:2304.08485.

Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. 2022. Open-domain Video Commentary Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7326–7339, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. 2000. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings*

15th International Conference on Pattern Recognition. ICPR-2000, volume 3, pages 314–317. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5391–5395.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*.

Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. Dori: Discovering object relationships for moment localization of a natural language query in a video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1079–1088.

Michael Schaffrath. 2003. Mehr als 1:0! Bedeutung des Live-Kommentars bei Fußballübertragungen– eine explorative Fallstudie [more than 1:0! the importance of live commentary on football matches – an exploratory case study]. *Medien und Kommunikationswissenschaft*, 51.

Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. 2017. What actions are needed for understanding human actions in videos? In *Proceedings of the IEEE international conference on computer vision*, pages 2137–2146.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating Live Soccer-Match Commentary from Play Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7096–7103.

Yufei Tian, Tuhin Chakrabarty, Fred Morstatter, and Nanyun Peng. 2021. Identifying distributional perspectives from colingual groups. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 178–190, Online. Association for Computational Linguistics.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado. Association for Computational Linguistics.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-End Dense Video Captioning With Parallel Decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857.

Teng Wang, Huicheng Zheng, and Mingjing Yu. 2020. Dense-Captioning Events in Videos: SYSU Submission to ActivityNet Challenge 2020. *arXiv:2006.11693 [cs]*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. End-to-End Dense Video Captioning With Masked Transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

A Dataset Analysis

For our human study of the data, we randomly sample approximately 1% of the videos from the training set and recruit two volunteers (a female and a male, both around 30 years old) who are asked to watch the videos and evaluate the nature and quality of the transcripts by answering the following questions: (1) Is it related to the contents of the video? (2) Does it comment on additional information not shown in the video? (3) Does it talk about something already passed in the video? (4) Does the transcript for this utterance likely contain problems due to automatic speech-to-text?

We display each commentary separated by utterance, each annotated with its corresponding timestamps, allowing the annotators to scroll through the video to make a proper judgment.

Aspect	Ratio	Cohen's kappa
(1) Content related	90.34%	0.43
(2) Add. information	19.15%	0.65
(3) Delayed	63.26%	0.07
(4) STT issue	9.73%	0.30

Table 4: Summary of our human evaluation results of the original transcriptions, where ratio indicates the percentage of utterances for which at least one of the annotators' answers was "yes" to the corresponding aspect. Cohen's kappa indicates the degree of inter-annotator agreement.

Metric	Original	Whisper
Avg. N. of Utt. / Video	17.77	26.40
Avg. Utterance Length	14.86	11.05
Avg. Transcript Length	264.11	291.87

Table 5: Comparison between the Original transcripts and the new ones we obtained using Whisper. Utterance and transcript length are shown in terms of number of words. Whisper led to transcripts with shorter utterances, increasing the granularity of described actions and events.

Original
'He talks to the camera , still with the iron in hand , gesturing to it , he takes the iron and now starts ironing the ski in front of him . Going a circular motion , really melting the wax onto the skis, finishing he puts it back on the little shelf in front of him and continue speaking to the camera while gesturing to the ski .'
Whisper
' He talks to the camera, still with the iron in hand.', ' He takes the iron and starts ironing the ski in front of him.' ' Going in a circular motion.' ' Really melting the wax onto the skis.' ' Finishing, he puts it back on the little shelf in front of him.' ' And continues speaking to the camera while gesturing to the ski.'

Table 6: Example of the difference in granularity given by Whisper compared to the Original transcripts, where we show how the latter contained only a single utterance, which Whisper was able adequately process.

A total of 1,044 utterances were evaluated. The summary of the results can be found in Table 4. The Cohen's kappa scores show moderate agreement between our annotators, except for the delayed aspect. We noticed that defining when some action or event started is subjective, likely leading to very different perceptions of when something has already passed in the video.

We can see that most utterances talk about things explicitly shown in the videos, with a small portion adding extra information. This point also suggests that we could benefit from exploring more information from the frames, such as using spatial information. Furthermore, over half of the evaluated utterances described events already passed in the video. This observation is good evidence that using visual and textual context is essential. Finally, we note that almost 10% of the annotated utterances were considered to have speech-to-text issues. While small, this number could affect the model's performance.

Furthermore, annotators pointed out another issue: Some of the utterances were very long, spanning a large portion of the commented video and describing different actions and activities in the video all at once. While such utterances could help the utterance generation model gain a more general understanding of the video, it might miss critical granular actions worth commenting on.

Table 5 briefly describes and summarizes the nature of the new transcripts we obtained via Whisper, compared against the originals. As can be seen, Whisper led to more relatively short utterances per video, increasing the granularity of the utterances. Table 6 illustrates the difference between the transcripts.

We also utilize our answerability score to compare the live commentary annotations with the dense video captioning annotations. We compute the scores for two validation portions of the ActivityNet Captions dataset (Krishna et al., 2017). Averaging these values, we obtained an answerability score of 0.177, much lower than scores obtained for the live commentary validation set reported in §5.1. These results provide further empirical evidence that the live commentary generation task differs from traditional video captioning.

B LLaVA's Visual Summary

To construct our visual summary of the video, we build a grid where the sampled frames are sorted temporally, and each one is labeled with an index from 1 to n, where n is the total number of frames to sample. Figure 5 shows an example of how this grid-like input looks when feeding 8 frames.

While there are a few commercial alternatives to use as LVLMs, we choose to work with open-weight models only in this work. To ensure that the chosen model could handle the amount of visual context we intended to feed into it via our grid-like

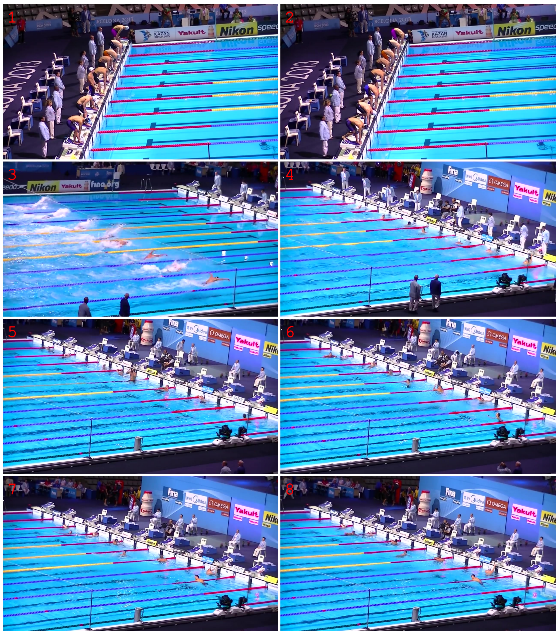


Figure 5: Example of the input fed to the LVLM for video *v_r3dM-5cZ7e8*‘, which led to the following questions being generated: ’How many swimmers are visible in the video?’, ’Are there any visible spectators or audience members in the frames?’, ’Can you identify any specific brands or sponsors in the frames?’, ’What type of swimming event is it, based on the number of lanes and the presence of lane dividers?’ .

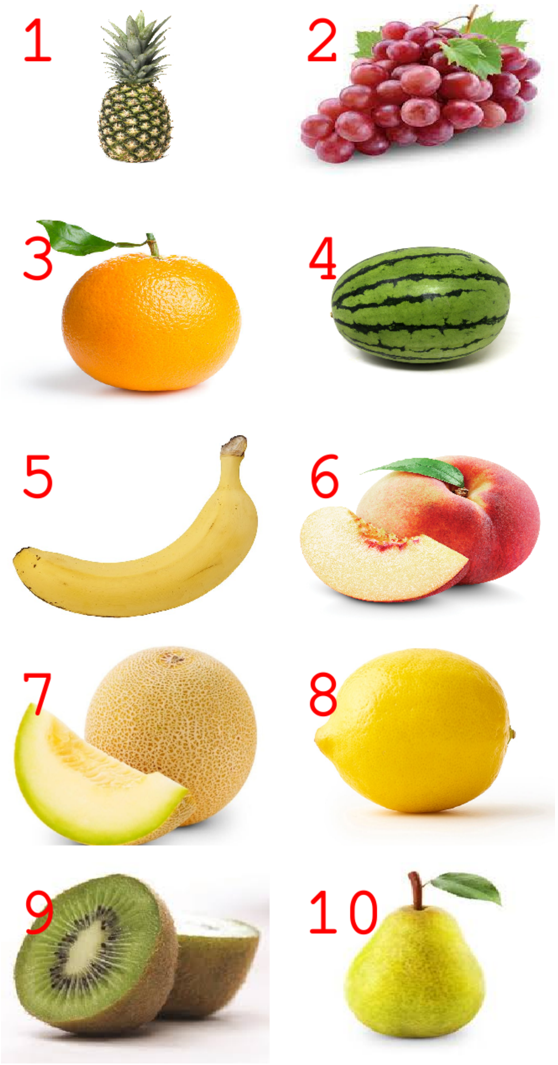


Figure 6: Fruit images utilized in our experiments with LLaVA.

video summary, we designed a simple experiment to test the limits of the three versions of LLaVA available at the time of writing this paper.

Concretely, we construct grid-like inputs with images of up to 10 fruits, shown in 6 and ask the model to name the fruits for each position in the grid, and overall; which we measure via accuracy and recall, respectively.

Figure 7 summarizes our results, suggesting that LLaVA v1.6 34b is the best candidate for our proposed evaluation scheme.

C Prompts to LVLM and LLMs

Figure 8 shows our prompt to LVLM to generate the evaluation questions used in our answerability scheme. Figure 9 shows our prompt to the LLM, to judge if the generated questions can be answered from the generated commentary utterances.

D General questions

Here, we present some examples of general questions sampled from the set of most common questions across all videos for the videos that the LVLM could not generate relevant questions:

- How many people are visible in the video?
- Can you describe the attire of the individuals in the video?
- What is the primary activity taking place in the video?
- Can you describe the overall mood or atmosphere of the video?
- Are there any indications of the time of day or lighting conditions in the video?

E Spatial Graph variations

Figure 10 illustrates the spatial graph’s variations tested during our experiments.

The SPECIFIC graph (Figure 10a) was proposed by Rodriguez-Opazo et al. (2021) for the task

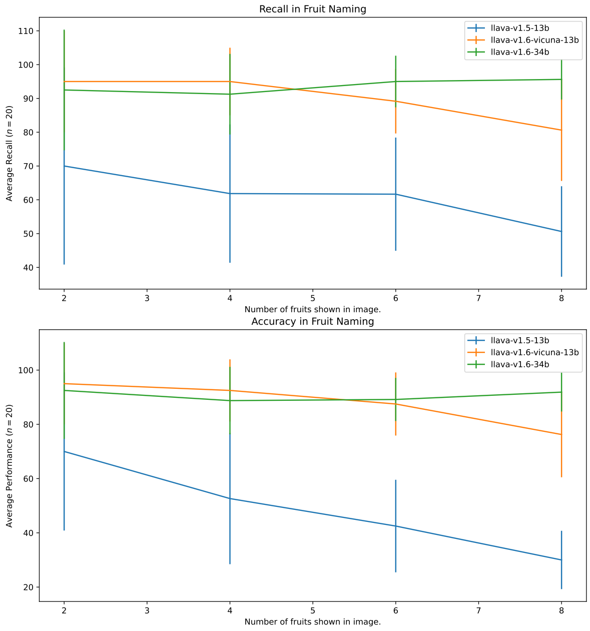


Figure 7: Performance on simple fruit naming tasks to probe the amount of context that the LLaVA models can handle, in terms of Recall (top) and Accuracy (bottom), which we surmise should correlate with their ability to perform our tasks for the answerability-based evaluation.

Prompt

The image presented contains a set of frames sampled from a video. The frames are sorted temporally in a grid and each one is labeled with an index from 1 up to 8 in red. Write a list of questions asking about the content of the video which would be relevant for someone providing live commentary of such video, describing aspects of the actions to listeners who cannot see it for themselves, counting the number of participants, and making educated guesses about the overall context of the video, such as the location where actions are taking place. Only include questions that have definite answers:
(1) one can see the content in the frames that the question asks about and can answer confidently;
(2) one can determine confidently from the frames that it is not in the frames.
Do not ask any question that cannot be answered confidently. Do not ask questions that involve a specific frame.

Figure 8: Prompt fed to the LVLM to obtain relevant questions for a video in our dataset, when provided with a visual summary.

Prompt

Below there is a piece of text describing the contents of a video over time.
{video_description}
With the information provided in the above text, indicate if the following question be answered.
{question}
Reply by only saying 'YES' or 'NO', and nothing else.

Figure 9: Prompt fed to the judge LLM to assess question answerability, where {video_description} and {question} are placeholders for a given transcript and questions, respectively.

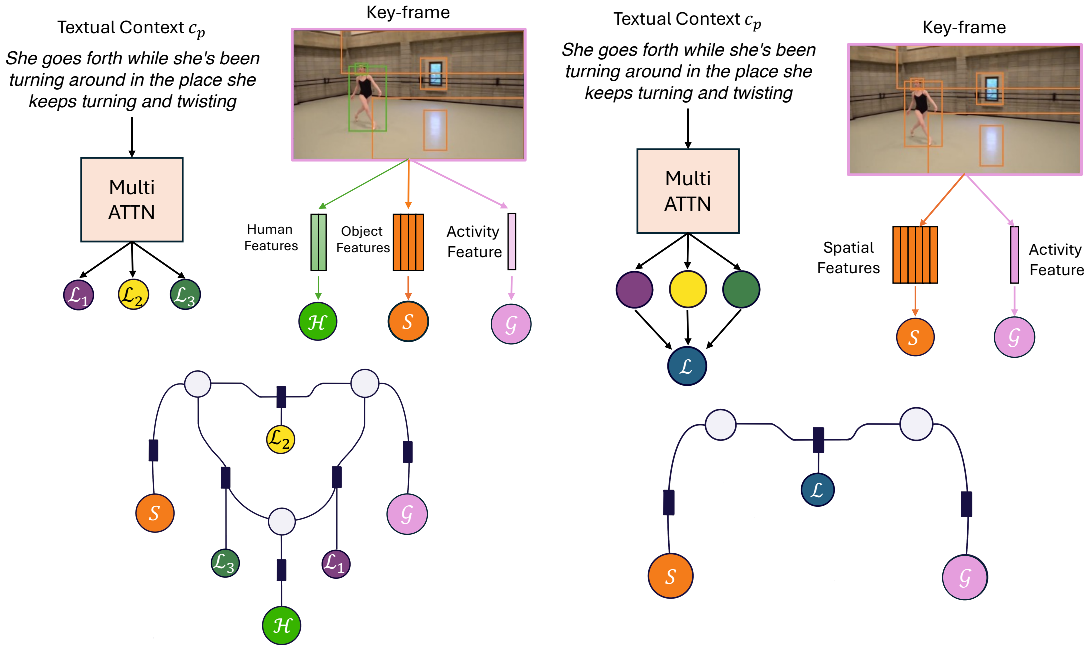
of temporal video grounding. For details on the message-passing algorithm, please refer to the original paper.

The GENERAL graph (Figure 10b) is a variation between SPECIFIC and our proposed graph OPEN. It keeps the multi-attention head that processes the textual context c_p but only considers one visual node for the spatial features. The attended textual context representations are concatenated and then go through a linear mapping function with bias, $\boldsymbol{L} = m_l([\boldsymbol{L}_1; \boldsymbol{L}_2; \boldsymbol{L}_3])$, which initializes the linguistic node $\mathcal{L}$.

F Human Evaluation via Crowdsourcing

We relied on Amazon Mechanical Turk to perform our crowdsourcing experiments. In order to make sure our annotators are closely following our request, we implement *attention questions*, which have been used by previous work to ensure annotation quality (Tian et al., 2023, 2021; Jo et al., 2021; Mittal et al., 2022). Concretely, our attention question states "*[This is a test, please check this checkbox and reply 'Neither Agree nor Disagree' to the question below.]*", which we add a random position on 30% of our tasks. We simply discard any work done by annotators that fail to reply correctly to our attention questions. Figure 11 shows an example our annotation interface.

In order to compensate our workers adequately, we first ran test tasks to measure how long the workers take to solve each task, finding that on average 2 minutes are required. Since each worker first needs to watch the video before answering our questions (which on average last 2 minutes), we decided to compensate them at a rate of 0.25 USD per task, which we estimate should roughly take 4 mins on average. This leads to a pay rate of 7



(a) SPECIFIC.

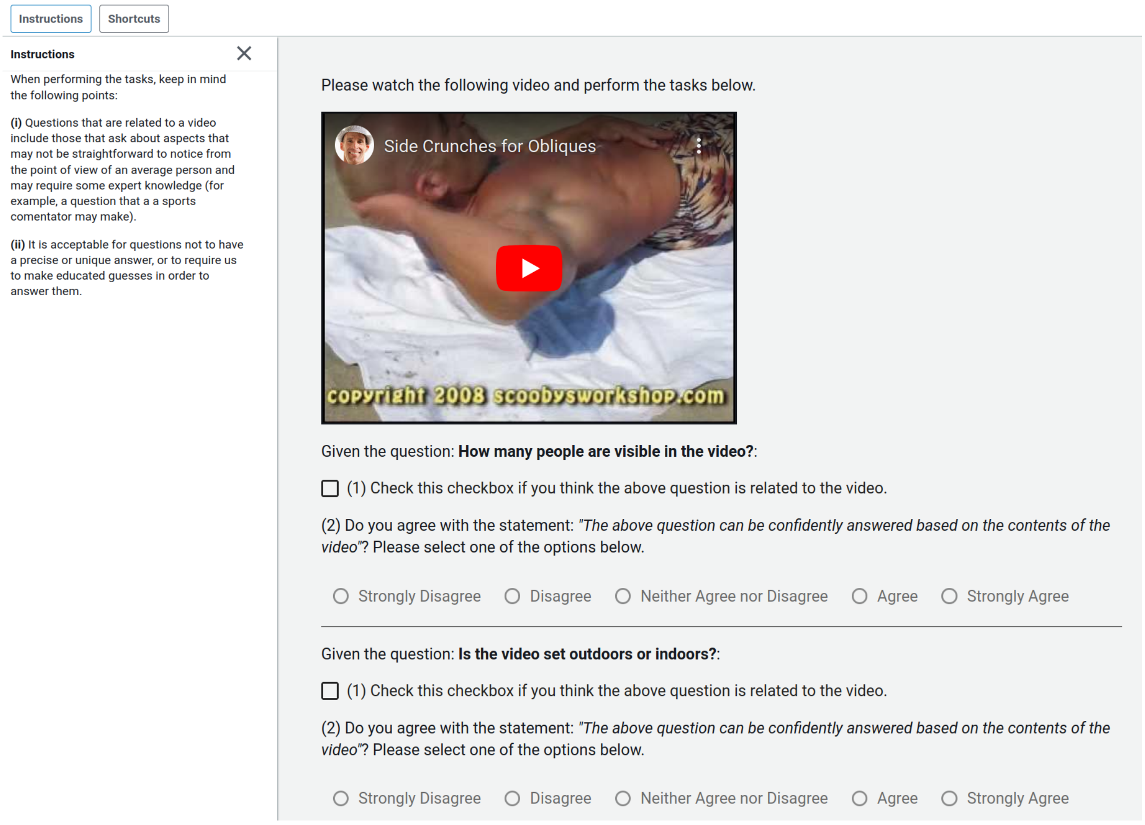
(b) GENERAL.

Figure 10: Illustration of the variations of the spatial graph used in the experiments.

USD/hour, roughly above the local minimum wage of 1,080 JPY/hour.

G Video Commentary Examples

Figures 12 and 13 show examples of commentaries that received contrasting BLEU and answerability scores.



Instructions
Shortcuts
Side Crunches for Obliques
copyright 2008 scoobysworkshop.com

Instructions

When performing the tasks, keep in mind the following points:

(i) Questions that are related to a video include those that ask about aspects that may not be straightforward to notice from the point of view of an average person and may require some expert knowledge (for example, a question that a a sports comentator may make).

(ii) It is acceptable for questions not to have a precise or unique answer, or to require us to make educated guesses in order to answer them.

Please watch the following video and perform the tasks below.

Given the question: **How many people are visible in the video?**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the video"*? Please select one of the options below.

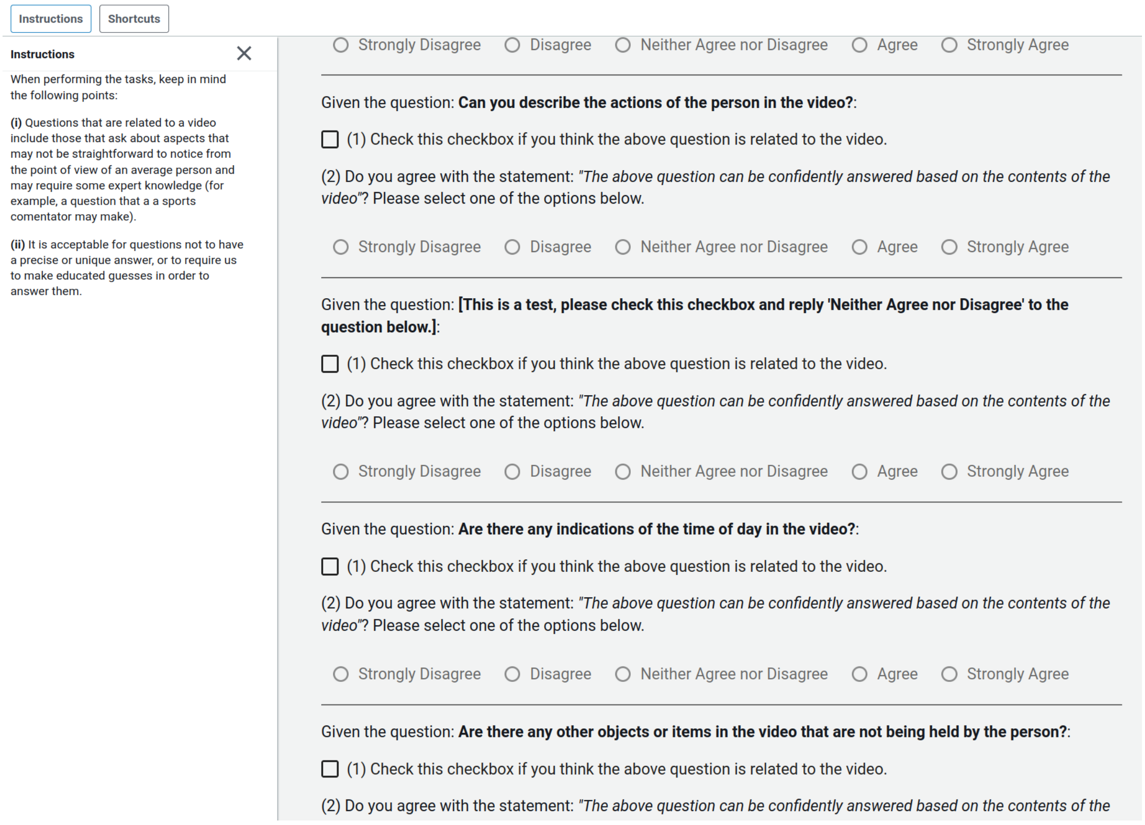
○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Given the question: **Is the video set outdoors or indoors?**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the video"*? Please select one of the options below.

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree



Instructions

When performing the tasks, keep in mind the following points:

(i) Questions that are related to a video include those that ask about aspects that may not be straightforward to notice from the point of view of an average person and may require some expert knowledge (for example, a question that a a sports comentator may make).

(ii) It is acceptable for questions not to have a precise or unique answer, or to require us to make educated guesses in order to answer them.

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Given the question: **Can you describe the actions of the person in the video?**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the video"*? Please select one of the options below.

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Given the question: **[This is a test, please check this checkbox and reply 'Neither Agree nor Disagree' to the question below.]**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the video"*? Please select one of the options below.

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Given the question: **Are there any indications of the time of day in the video?**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the video"*? Please select one of the options below.

○ Strongly Disagree ○ Disagree ○ Neither Agree nor Disagree ○ Agree ○ Strongly Agree

Given the question: **Are there any other objects or items in the video that are not being held by the person?**:

□ (1) Check this checkbox if you think the above question is related to the video.

(2) Do you agree with the statement: *"The above question can be confidently answered based on the contents of the*

Figure 11: Screenshot of our annotation interface, showing examples of regular (top) and attention questions (bottom), while also displaying how our workers see our quick instructions/tips.

Context Window: $(t_c = t_s - 6, p = 2)$
Video: v_jWPr92KwXeY

Questions:
0 How many people are visible in the video?
1 What is the primary activity taking place in the video?
2 What is the weather condition in the video?
3 Can you describe the terrain in the video?
4 Are there any trees or vegetation visible in the video?
5 Can you determine the time of day from the video?
6 Are there any visible landmarks or signs that indicate the location of the video?
7 Are there any safety measures or protective gear visible in the video?
8 Can you estimate the speed of the participants from the video?
9 Are there any other people or objects in the background of the video?
10 Can you describe the clothing or attire of the participants in the video?
11 Are there any visible signs of motion or movement in the video?
12 Can you identify any specific techniques or maneuvers being performed by the participants?
13 Are there any indications of a competition or event in the video?
14 Can you describe the snow conditions in the video?
15 Are there any indications of the altitude or elevation in the video?
16 Are there any visible changes in the environment or conditions over the course of the video?
17 Can you identify any potential hazards or obstacles in the video?
18 Are there any indications of the participants' skill level or experience from the video?

BLEU: 74.18
ρ^*: 0.073
Questions answered: {16, 1, 3, 17}

Gold standard utterances:
"He jumps down what looks like a cliff."
"This is Jeff Hambleton."
"He continues to move down the hill.

Generated utterances:
"He jumps down what looks like a hill."
"This is insane insane."
"He continues to move down the hill."

Figure 12: Generation example where the BLEU score is very high, but the answerability score ρ^* is very low. Although our model generating utterances almost equal to the gold standard, the answerability score is low because the gold standard utterances themselves are not very informative.

Context Window: $(t_c = t_s, p = 1)$
Video: v_5n8wY8hwy3Y

Questions:
0 What is the main subject of the video?
1 Can you describe the toy cars in the video?
2 How many toy cars are visible in the video?
3 Are the toy cars designed to resemble characters from a popular animated movie?
4 What is the setting of the video?
5 Is there any interaction between the toy cars and a person in the video?
6 Can you describe the actions of the toy cars in the video?
7 Are there any obstacles or structures in the video that the toy cars interact with?
8 Can you identify any specific scenes or moments in the video that stand out?

BLEU: 18.57
ρ^*: 0.815
Questions answered: $\{0, 2, 4, 5, 6, 7, 8\}$

Gold standard utterances:
And now it looks like he’s pretending that the blue car is putting wheels on the back of the toy car with no wheels.
And now he’s directing the yellow car. And now he puts the blue and the yellow car together and they have to move off the screen.
And now they go to what looks like another car. And it is some sort of a Play-Doh toy.
It’s something where he has a knob on top that you twist.
on the thing to shape these little
black pieces of dough. Alright, now
he’s shaped about 1, 2, 3, 4, 5, 6 pieces of
dough. Now he’s sticking all
...
Okay, so it’s gonna give it to the other car.
Which is odd.
I don’t know. The other car really don’t need it.
But we don’t know if that car need it or not, you know.
There’s a vehicle up underneath it. It’s creating something to make. It’s putting something in a little tiny dump truck.
It’s grinding up something.
It’s grinding up something to put in this kid’s little dump truck. I think they’re made out of clay or something.
It’s grinding up these pieces. This little dump truck is filled full of something in the holding part of the dump truck.
They pull it out and put it on the side.
They put it on the front of this all over this little car or dump truck.
They lay the pieces on it and move it to the side.
Now the white car is being put up there.

Generated utterances:
And now it looks like he’s just laying the bumper car in the blue bucket with the safety stick that goes.
And now they’re turning the red and then the man in the yellow, and the white, they both turn off the screen.
And now we go to what looks like another fire and it’s just a sort of a yellow container.
It appears that it has a screwdriver on where you kind of put the sandwich around.
on the place to make some beautiful thing.
black slices of bread she’s now cutting
she’s got two pieces of sandwich sandwich that’s about in sequence.
bits, now he’s sticking all everything.
...
Okay, so it’s easier to grab the name.
Which is sort.
I don’t know the other cars, I think it’s really dirty.
But if I’m not sure that you can see anything in this car.
Maybe it’s a tire. Putting something into the chimney. Pulls it up. Some kind of material.
It’s tightening up something.
It’s some sort of stuff. They’re putting stuff out in a bucket to get used to decorate this grass.
It’s cutting some sort of powder that’s all the bonfire. One of the cars is coming up inside a tire.
They pull it out and put it on the side.
They put it on the top of this whole boat, all over this little bumper car,
They lay the wood on it and move it to the side.
Now the white car is being put up there.

Figure 13: Generation example where the BLEU score is low, but the answerability score ρ^* is very high. Despite the generated utterances containing information necessary to answer many of the questions, generated utterances have really poor grammar.