A. Model Configurations and hyperparameter settings

We summarize the detailed model configurations and hyperparameter settings for ControlVAE in the following three applications: language modeling, disentanglement representation learning and image generation.

A.1. Experimental Details for Language Modeling

For text generation on PTB data, we build the ControlVAE model on the basic VAE model, as in (Bowman et al., 2015). We use one-layer LSTM as the encoder and a three-layer Transformer with eight heads as the decoder and a Multi-Layer Perceptron (MLP) to learn the latent variable $\mathbf{z}$. The maximum sequence length for LSTM and Transformer is set to 100, respectively. And the size of latent variable is set to 64. Then we set the dimension of word embedding to 256 and the batch size to 32. In addition, the dropout is 0.2 for LSTM and Transformer. Adam optimization with the learning rate 0.001 is used during training. Following the tuning guidelines above, we set the coefficients K_p and K_i of P term and I term to 0.01 and 0.0001, respectively. Finally, We adopt the source code on Texar platform to implement experiments (Hu et al., 2019).

For dialog-response generation, we follow the model architecture and hyperparameters of the basic conditional VAE in (Zhao et al., 2017). We use one-layer Bi-directional GRU as the encoder and one-layer GRU as the decoder and two fully-connected layers to learn the latent variable. In the experiment, the size of both latent variable and word embeddings is set to 200. The maximum length of input/output sequence for GRU is set to 40 with batch size 30. In addition, Adam with initial learning rate 0.001 is used. In addition, we set the same K_p and K_i of PI algorithm as text generation above. The model architectures of ControlVAE for these two NLP tasks are illustrated in Table 4, 5.

Table 4. Encoder and decoder architecture for text generation on PTB data.

Encoder	Decoder
Input n words $\times 256$	Input $\in \mathbb{R}^{64}$, $n \times 256$
1-layer LSTM	FC 64×256
FC 64×2	3-layer Transformer 8 heads

A.2. Experimental Details for Disentangling

Following the same model architecture of β-VAE (Higgins et al., 2017), we adopt a convolutional layer and deconvolutional layer for our experiments. We use Adam optimizer with $\beta_1 = 0.90$, $\beta_2 = 0.99$ and a learning rate tuned from 10^{-4}. We set K_p and K_i for PI algorithm to 0.01 and 0.001, respectively. For the step function, we set the step, α, to 0.15 per $K = 5000$ training steps as the information capacity (desired KL- divergence) increases from 0.5 until 18 for 2D Shape data. ControlVAE uses the same encoder and decoder architecture as β-VAE except for plugging in PI control algorithm, illustrated in Table 6.

Table 5. Encoder and decoder architecture for dialog generation on Switchboard (SW) data.

Encoder	Decoder
Input n words $\times 200$	Input $\in \mathbb{R}^{200}$
1-layer bi-GRU	FC 200×400
FC 200×2	1-layer GRU
FC 200×2	

Table 6. Encoder and decoder architecture for disentangled representation learning on 2D Shapes data.

Encoder	Decoder
Input 64×64 binary image	Input $\in \mathbb{R}^{10}$
4×4 conv. 32 ReLU. stride 2	FC. 256 ReLU.
4×4 conv. 32 ReLU. stride 2	4×4 upconv. 256 ReLU.
4×4 conv. 64 ReLU. stride 2	4×4 upconv. 64 ReLU. stride 2.
4×4 conv. 64 ReLU. stride 2	4×4 upconv. 64 ReLU. stride 2
4×4 conv. 256 ReLU.	4×4 upconv. 32 ReLU. stride 2
FC 256. FC. 2×10	4×4 upconv. 32 ReLU. stride 2

A.3. Experimental Details for Image Generation

Similar to the architecture of β-VAE, we use a convolutional layer with batch normalization as the encoder and a deconvolutional layer with batch normalization for our experiments. We use Adam optimizer with $\beta_1 = 0.90$, $\beta_2 = 0.99$ and a learning rate 10^{-4} for CelebA data. The size of latent variable is set to 500, because we find it has a better reconstruction quality than 200 and 400. In addition, we set the desired value of KL-divergence to 170 (same as the original VAE), 180, and 200. For PI control algorithm, we set K_p and K_i to 0.01 and 0.0001, respectively. We also use the same encoder and decoder architecture as β-VAE above except that we add the batch normalization to improve the stability of model training, as shown in Table 7.

B. Examples of Generated Dialog by ControlVAE

In this section, we show an example to compare the diversity and relevance of generated dialog by different methods, as illustrated in Table 8. Alice begins with the open-ended

Table 7. Encoder and decoder architecture for image generation on CelebA data.

Encoder	Decoder
Input 128 × 128 × 3 RGB image	Input $\in \mathbb{R}^{500}$
4 × 4 conv. 32 ReLU. stride 2	FC. 256 ReLU.
4 × 4 conv. 32 ReLU. stride 2	4 × 4 upconv. 256 ReLU. stride 2
4 × 4 conv. 64 ReLU. stride 2	4 × 4 upconv. 64 ReLU. stride 2.
4 × 4 conv. 64 ReLU. stride 2	4 × 4 upconv. 64 ReLU. stride 2
4 × 4 conv. 256 ReLU. stride 2	4 × 4 upconv. 32 ReLU. stride 2
FC 4096. FC.2 × 500	4 × 4 upconv. 32 ReLU. stride 2

conversation on choosing a college. Our model tries to predict the response from Bob. The ground truth response is "um - hum". We can observe from Table 8 that ControlVAE (KL=25, 35) can generate diverse and relevant response compared with the ground truth. In addition, while cyclical annealing can generate diverse text, some of them are not very relevant to the ground-truth response.

Table 8. Examples of generated dialog for different methods. Our model tries to predict the response from Bob. The response generated by ControlVAE (KL=25,35) are relevant and diverse compared with the ground truth. However, some of reponse generated by cost annealing and cyclical annealing are not very relevant to the ground-truth data

<table>
<tr><td colspan="2"><b>Context</b>: (Alice) and a lot of the students in that home town sometimes ⟨ unk ⟩ the idea of staying and going to school across the street so to speak
<b>Topic</b>: Choosing a college <b>Target</b>: (Bob) um - hum</td></tr>
<tr><td><b>ControlVAE-KL-25</b></td><td><b>ControlVAE-KL-35</b></td></tr>
<tr><td>yeah</td><td>uh - huh</td></tr>
<tr><td>um - hum</td><td>yeah</td></tr>
<tr><td>oh that's right um - hum</td><td>oh yeah oh absolutely</td></tr>
<tr><td>yes</td><td>right</td></tr>
<tr><td>right</td><td>um - hum</td></tr>
<tr><td><b>Cost annealing (KL=17)</b></td><td><b>Cyclical anneal (KL=21.5)</b></td></tr>
<tr><td>oh yeah</td><td>yeah that's true do you do you do it</td></tr>
<tr><td>uh - huh</td><td>yeah</td></tr>
<tr><td>right</td><td>um - hum</td></tr>
<tr><td>uh - huh and i think we have to be together</td><td>yeah that's a good idea</td></tr>
<tr><td>oh well that's neat yeah well</td><td>yeah i see it too,it's a neat place</td></tr>
</table>

C. $\beta(t)$ of ControlVAE for Image Generation on CelebA data

Fig. 7 illustrates the comparison of $\beta(t)$ for different methods during model training. We can observe that $\beta(t)$ finally converges to 1 when the desired value of KL-divergence is set to 170, same as the original VAE. At this point, ControlVAE becomes the original VAE. Thus, ControlVAE can be customized by users based on different applications.

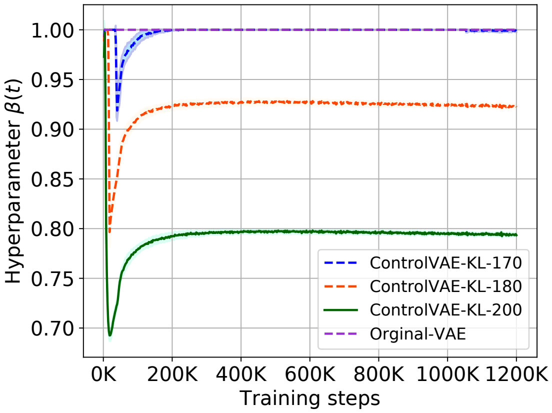
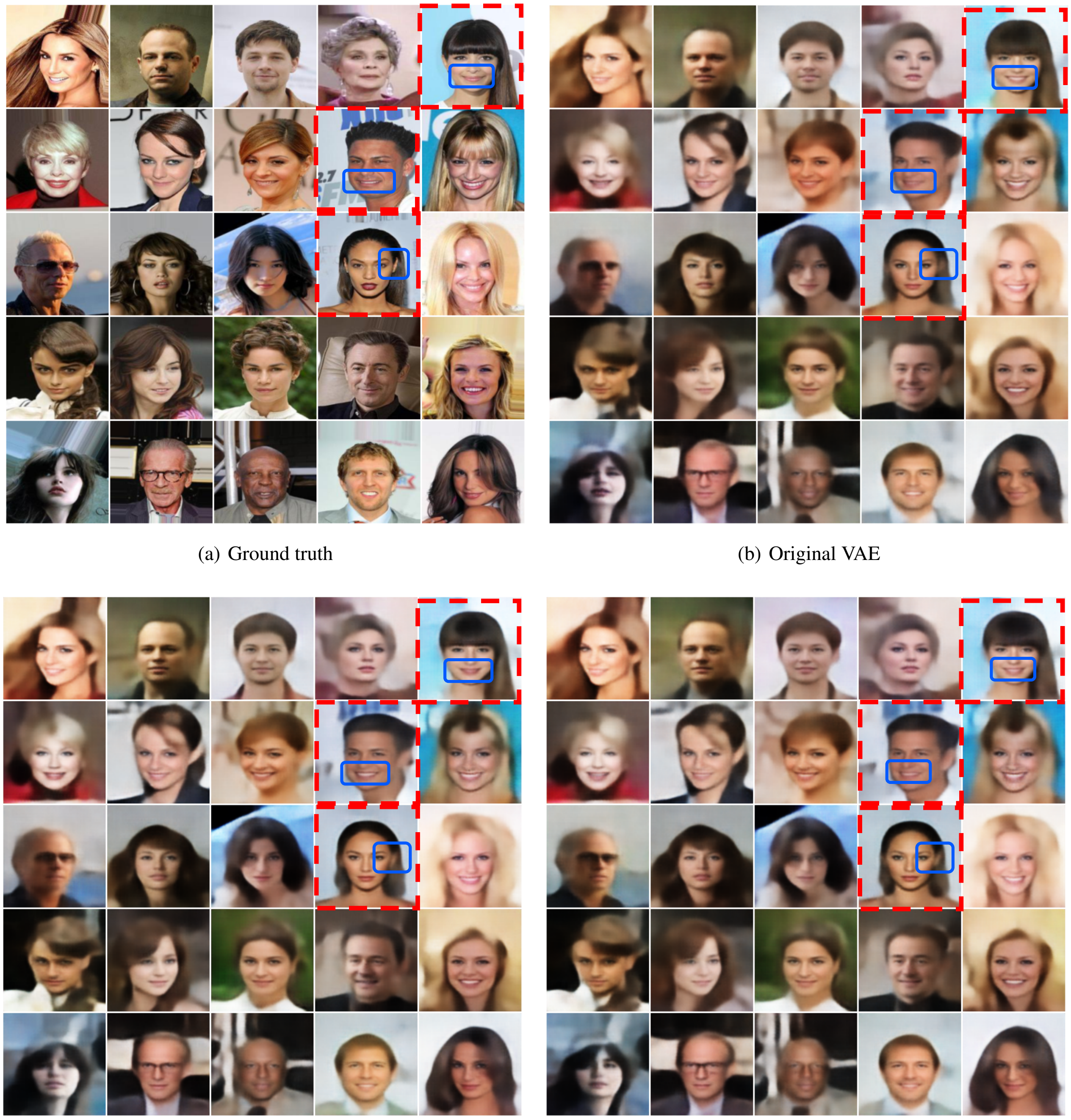


Figure 7. Hyperparameter $\beta(t)$ of ControlVAE for image generation on CelebA data for 3 random seeds. If we set the desired value of KL-divergence to 170, the hyperparameter, $\beta(t)$, gradually approaches 1. It means the ControlVAE becomes the original VAE.

D. Examples of Reconstruction Images by VAE and ControlVAE

We also show some reconstruction images by ControlVAE and the original VAE in Fig. 8. It can be observed that images reconstructed by ControlVAE-KL-200 (KL = 200) has the best reconstruction quality compared to the original VAE. Take the woman in the first row last column as an example. The woman does not show her teeth in the groundtruth image. However, we can see the woman reconstructed by the original VAE smiles with mouth opening. In contrast, the woman reconstructed by ControlVAE-KL-200 hardly show her teeth when smiling. In addition, we also discover from the other two examples marked with blue rectangles that ControlVAE-KL-200 can better reconstruct the "smile" from the man and the "ear" from the woman compared to the original VAE. Therefore, we can conclude that our ControlVAE can improve the reconstruction quality via slightly increasing (control) KL-divergence compared to the original VAE. Note that, if we want to improve the generation quality by sampling the latent code, we can reduce the KL divergence, which will be explored in the future.



(a) Ground truth

(b) Original VAE

(c) ControlVAE-KL-200

(d) ControlVAE-KL-170

Figure 8. Examples of images recontructed by different methods and ground truth. From the images marked with blue rectangles, we can see that ControlVAE-KL-200 (KL=200) can better reconstruct woman's month opening (first row last column), man's smiling with teeth (second row fourth column), and woman'ear (third row fourth column) than the original VAE based on the ground-truth data in (a).