

Supplementary Material

A. Experimental Details

A.1. Implementation of Similarity Metrics

This subsection lists approaches for computing image-to-image similarities and text-to-text similarities. In the image-text-image and text-image-text tasks, the evaluation metrics and distance metrics are closely related, i.e., the evaluation metrics for one task serve as distance metrics for the other task. Hence, we present them jointly in this section.

CLIP: The CLIP ViT-L/14 model³ is pre-trained on a 400M text-image pair dataset. Specifically, we employ the outputs from the visual projection layer with an embedding size of 768 for the CLIP image encoder. For the CLIP text encoder, the outputs from the textual projection layer with an embedding size of 512 are used.

DINO: The pretrained DINO ViT-B/8 model⁴ are used to obtain the image embeddings. DINO has been trained on the images dataset in a self-supervised way and has shown superior performance on representation learning tasks. The input image size is set to 384 and the output embedding size to 768.

FID, IS: Following [46], the torch-fidelity library⁵ is used to compute the fidelity scores of the generated images.

SBERT: We use Sentence-BERT⁶ with embedding size of 384 and take the [CLS] embeddings from the last transformer layer as the representation for the input text.

WMD: We use the open implementation of NLTK library⁷ to compute WMD between a source sentence and a target sentence. Specifically, we use the *glove-wiki-gigaword* vector embeddings with 200 dimensions as the choice for the WMD. Since WMD is a distance metric rather than a score metric, we plot the y -axis in the reversed order to represent that the higher the y , the better the text, as shown in the second subfigure in Figure 3.

A.2. Details of Generation Models and Benchmarks

We use the NoCaps [1] validation set of 4500 images and a subset of 2000 images from the COCO Karpathy test split [27] to support the evaluation. For BLIP⁸, we use the ViT-L model finetuned for the image captioning task. For SD⁹, we use the weights sd-v1-4.ckpt and DDPM scheduler

³adapted from the public library <https://huggingface.co/docs/transformers/modeldoc/clip>

⁴adapted from <https://github.com/facebookresearch/dino>

⁵public implementation available from <https://github.com/toshas/torch-fidelity>

⁶public implementation from <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁷<https://www.nltk.org/>

⁸<https://github.com/salesforce/BLIP>

⁹<https://github.com/huggingface/diffusers>

with sampling steps of 50 and a guidance scale of 7.5. The output image size is 512. The number of minimal words is set to 5 and the maximum number of words is set to 30. We set $p=0.9$ in the Top- p strategy.

B. Impact of Parameter of Top- p Sampling

p	0.1	0.2	0.3	0.4	0.5
CIDEr	111.8	111.2	110.0	107.2	103.4
p	0.6	0.7	0.8	0.9	0.95
CIDEr	96.2	89.7	83.1	78.7	69.2

Table 6. Impact of p on the Top- p sampling performance of BLIP ViT-L.

Top- p sampling, which is also referred to as nucleus sampling, is a text generation method that samples words from a set of candidates whose cumulative probability exceeds a specified threshold p . By varying p , we achieve a trade-off between the diversity and accuracy of the generated text. Broadly speaking, larger values of p result in more diverse captions, whereas smaller values of p lead to less variable yet more accurate captions for an input image.

Note that in Table 2 of the BLIP paper [35], the Top- p sampling method is used to generate a diverse set of captions which are then utilized for bootstrapping the BLIP model. However, the evaluation result in that row is obtained by the beam search method. To examine the effect of p , we report the performance of BLIP ViT-L on the NoCaps dataset for image captioning, under different choices of p , in Table 6.

C. Influence of Sampling Methods for Image Captioning

Several sampling methods exist for generating text in the image captioning model. Apart from the Top- p sampling approach presented in the main text, we conduct a quantitative evaluation of two different sampling strategies, including Top- k [16] with $k=10$, and Tempered sampling [9, 39] with $T=0.7$. Each sampling method serves as a baseline. Table 7 verifies that our conclusion is solid across different sampling methods.

D. Impact of Number of Candidates N

We investigate the effect of the number of candidates N on our findings. We conduct experiments using the BLIP ViT-L model on the NoCaps dataset for image captioning, utilizing the Top- p sampling method. For image generation, we use the SD model and the DDPM scheduler. We sample N candidates for each input image or text in each experiment. We compare our approach to the baseline method, where a candidate is selected randomly.

Method	Nocaps								COCO			
	In-domain		Near-domain		Out-domain		Overall		Karpathy Test			
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	B@3	B@4	CIDEr	SPICE
Top- <i>k</i> Sampling	79.5	13.2	78.4	12.4	83.4	11.7	79.6	12.4	39.7	28.1	108.4	20.9
Ours	84.3	13.5	81.6	12.8	91.6	12.7	84.0	12.9	40.1	28.7	111.6	21.5
Gain (%)	+6.0	+2.3	+4.1	+3.2	+9.8	+8.5	+5.5	+4.0	+1.1	+2.3	+2.9	+2.9
Tempered Sampling	83.9	13.2	82.7	12.7	89.8	12.0	84.3	12.6	33.0	22.2	92.4	19.5
Ours	87.8	13.4	87.0	13.1	98.1	12.8	89.3	13.1	33.6	22.8	95.4	20.4
Gain (%)	+4.6	+1.5	+5.2	+3.1	+9.2	+6.7	+5.9	+4.0	+1.8	+2.6	+3.2	+4.7

Table 7. Comparison of different sampling methods and our proposed method on Nocaps and COCO datasets. Our method outperforms every sampling method on all metrics. The relative gain of our method compared to each sampling method is given in the last row in each block. B@*k*: BLEU@*k*.

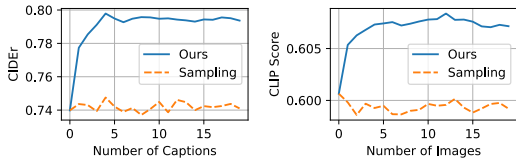


Figure 9. Evaluation of the choice of the number of sample candidates.

Figure 9 depicts the impact of the number of candidates on image-text-image (left) and text-image-text (right) tasks. The figure (left) shows the image captioning score (*y*-axis) for our approach and the baseline method, with the *x*-axis representing the number of candidate captions for each input image. As shown in the figure, the performance of the baseline method remains consistent as it randomly samples captions with varying qualities. Conversely, our approach improves significantly after the number of captions has reached around five. When the number of candidates is limited, there may not be enough high-quality captions to choose from, resulting in lower performance. Nevertheless, our approach remains effective even in that stage. After acquiring a reasonable number of candidates, our method consistently outperforms the baseline method by a significant margin. A similar conclusion can be inferred for the text-image-text task, as demonstrated on the right side of the figure.

E. Qualitative Results for Image and Text Generation

To reinforce the findings in Section 3, we provide additional qualitative examples on the NoCaps dataset. The annotations and explanations for each figure are included in their respective captions. Figure 10 shows both positive examples and negative examples for the image-text-image task, whereas Figure 11 presents visualizations for the text-image-text task. All results are obtained from BLIP ViT-L and SD models.

F. Analysis of Different Image-to-Text and Text-to-Image Generative Models

We conduct further experiments with different image captioning and text-to-image models. Table 8 shows the result of BLIP with a VAE-based image generative model LAFITE for the image-text-image task, as well as SD with BLIP-2 for the text-image-text task. In general, our finding, that better reconstruction leads to better generation performance, still holds. We find that SD performs better than LAFITE, very likely due to its larger training data. When coupled with BLIP-2, SD improved the performance upon the baseline, but it performed worse than that in the case of BLIP.

Text Generation					
I2T	T2I	I-C	N-C	O-C	E-C
Baseline		75.1	72.4	78.7	74.1
BLIP [35]	SD [46]	77.3	78.3	88.8	80.3
BLIP [35]	LAFITE [64]	75.4	76.2	82.9	77.4
Image Generation					
T2I	I2T	CLIP↓	FID↓	IS↑	
Baseline		40.54	32.37	41.19 ± 2.91	
SD [46]	BLIP [35]	33.47	29.59	45.64 ± 2.40	
SD [46]	BLIP-2 [34]	39.41	31.34	42.34 ± 2.23	

Table 8. Comparison of different combinations of generation models. We evaluate two image captioning models and two image generation models on the NoCaps dataset. I2T: Image-to-text model. T2I: Text-to-image model. I-C/N-C/O-C/E-C: In-/Near-/Out-/Entire-domain CIDEr.

G. Implementations of Tokenizer Transformation

We elaborate on the gradient backpropagation process between the output of BLIP and the input of SD, shown on the left side of Figure 6. Our framework includes three types of text tokenizers: BLIP utilizes the word-piece tokenizing method; BLIP-2 uses the byte-pair-encoding tokenizing method; SD employs the CLIP tokenizer, which uses the word-level tokenizing strategy. One of the challenges we faced is to align the output token distributions from BLIP with those from SD, allowing SD to interpret

the output of BLIP. To address this issue, we use these tokenization strategies to tokenize each sentence in the COCO training set and learn a one-to-one hard-coded mapping from a source token to a target token. Despite using different strategies, we found that these tokenizing methods have a high ratio of overlapping between tokens. Specifically, when applied to the COCO training set, more than 60% captions can be tokenized into the same set of tokens for BLIP and SD. For unmatched tokens, we map them to the most similar ones or to the [UNK] token. We visually examined this method by generating images conditioned on token distributions and found it to be practical. Additionally, we remove the prefix tokens of BLIP, *a photo of*, and add [BOS] and [EOS] tokens for SD.

H. Discussion on the Loss Function

Parameter Update While optimizing \mathcal{L}_{IR} for image captioning, both SD and BLIP are trained. \mathcal{L}_{TG} only updates the parameters of BLIP. Likewise, both SD and BLIP are trained when optimizing \mathcal{L}_{TR} , and only SD is updated by \mathcal{L}_{IG} . The reasons for training SD in \mathcal{L}_{IR} are twofold. First, SD needs to adapt to new distributions coming from BLIP. As in the standard training, the input of BLIP is discrete tokens. Whereas in our approach, the input is token distributions. Second, SD could be improved because of additional training data sampled from BLIP. In addition, since they are refreshed at each iteration, one model is able to provide better samples to train the other model throughout the training.

Connection to CycleGAN CycleGAN is a generative model that learns bidirectional mappings from domain X to domain Y , where X and Y are images. Its cyclic loss ensures that the mapping between the input and output domains is consistent, i.e., if we take an image from domain X , pass it through the generator network to obtain an image in domain Y , and then pass that image through the generator network again to obtain an image in domain X , we should obtain an image that is similar to the original image in domain X . Likewise, we aim to enforce consistent mapping between the input and output domains, but we deal with two distinct domains, i.e., image and language, which may require much more complex mapping. In addition, we also optimize a single objective, similar to CycleGAN which is trained on the weighted cyclic loss and adversarial loss. CycleGAN employs a hyperparameter λ to control the relative importance of the cyclic loss to the adversarial loss, we found that a similar weighting did not yield substantial differences in performance in our work. Therefore we do not adjust this hyperparameter in our approach.

Pseudo-Code The pseudo-code of our train framework is presented in Algorithm 1.

Algorithm 1 Training Framework

Model UNet ϵ_ψ , SD’s Text Encoder π , BLIP b_θ
Input an image-text pair $(\mathbf{x}_0, \mathbf{y})$

- 1: **repeat**
- # Image-Text-Image (BLIP \rightarrow SD)
- 2: $\mathbf{x}_0, \mathbf{y} \sim q(\mathbf{x}_0, \mathbf{y})$ \triangleright Sample an image-text pair from the dataset
- 3: $\hat{\mathbf{y}} = b_\theta(\mathbf{x}_0, \tilde{\mathbf{y}})$ $\triangleright \hat{\mathbf{y}} \in \mathbb{R}^{L \times V}$: Output token distribution.
 $\tilde{\mathbf{y}}$: (causal) masked text input
- 4: $\mathcal{L}_1 = \text{CE}(\mathbf{y}, \hat{\mathbf{y}})$ \triangleright CE: cross entropy loss
- 5: $\mathbf{c} = \pi(\hat{\mathbf{y}})$ \triangleright Text encoder encodes BLIP’s output into embeddings
- 6: $t \sim \text{Uniform}(\{1, \dots, T\})$ \triangleright Sample a timestep for the diffusion process
- 7: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ \triangleright Sample noise for timestep t
- 8: $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$ \triangleright Add noise to image
- 9: $\hat{\epsilon} = \epsilon_\psi(\mathbf{x}_t, \mathbf{c})$ \triangleright UNet predicts noise $\hat{\epsilon}$ from the noisy image \mathbf{x}_t
- 10: $\mathcal{L}_2 = \|\epsilon - \hat{\epsilon}\|^2$
- # Text-Image-Text (SD \rightarrow BLIP)
- 11: $\mathbf{x}_0, \mathbf{y} \sim q(\mathbf{x}_0, \mathbf{y})$ \triangleright Sample an image-text pair from the dataset
- 12: $\mathbf{c} = \pi(\mathbf{y})$ \triangleright Text encoder encodes input text into embeddings
- 13: $t \sim \text{Uniform}(\{1, \dots, T\})$ \triangleright Sample a timestep for the diffusion process
- 14: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ \triangleright Sample noise for timestep t
- 15: $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$ \triangleright Add noise to image
- 16: $\hat{\epsilon} = \epsilon_\psi(\mathbf{x}_t, \mathbf{c})$ \triangleright UNet predicts noise $\hat{\epsilon}$ from the noisy image \mathbf{x}_t
- 17: $\mathcal{L}_3 = \|\epsilon - \hat{\epsilon}\|^2$
- 18: $\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{1 - \alpha_t} \hat{\epsilon})$ \triangleright 1-step approximation of \mathbf{x}_0
- 19: $\mathcal{L}_4 = \text{CE}(\mathbf{y}, b_\theta(\hat{\mathbf{x}}_0, \tilde{\mathbf{y}}))$ $\triangleright \tilde{\mathbf{y}}$: (causal) masked text input
- 20: Take gradient descent step on
 BLIP: $\nabla_\theta(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4) = \nabla_\theta(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_4)$
 SD: $\nabla_\psi(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4) = \nabla_\psi(\mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4)$
- 21: **until** converged
- 22: **return** ψ, θ

I. Ablation Study

Table 9 summarizes the different losses and weight-freezing strategies, highlighting that improvement comes from the proposed reconstruction loss. For simplicity, we only list the effect of different settings on BLIP. Our default setting jointly optimizes both pipelines with both models trainable $p_1 : \mathbf{I} \xrightarrow{\text{BLIP}} \mathbf{T}(L_{TG}) \xrightarrow{\text{SD}} \mathbf{I}(L_{IR})$, and $p_2 : \mathbf{T} \xrightarrow{\text{SD}} \mathbf{I}(L_{IG}) \xrightarrow{\text{BLIP}} \mathbf{T}(L_{TR})$. The third column in the table shows the training signals that BLIP can receive from the two pipelines under the specific setting. ”-” means that loss has no effect as the model is frozen.

Weights Update				Loss Function				Effect on BLIP	Experiment	CIDEr
BLIP ^{p1}	SD ^{p1}	SD ^{p2}	BLIP ^{p2}	L_{TG}	L_{IR}	L_{IG}	L_{TR}			
✓	✓	✓	✓	✓	✓	✓	✓	p_1 : ground truth + reconstruction p_2 : augmentation	Ours, Tab. 4	111.8
✓	×	×	×	✓	×	-	-	p_1 : ground truth	Tab. 4	109.7
✓	×	✓	×	✓	✓	✓	✓	p_1 : ground truth + reconstruction	Ablation	110.9
×	✓	×	✓	-	✓	-	✓	p_2 : augmentation	Ablation	110.2
✓	×	×	×	×	✓	-	-	p_1 : reconstruction	Tab. 6	102.3

Table 9. Analysis of different training paradigms of loss terms and model frozen strategies.

J. Qualitative Results for Training Framework

We provide additional qualitative examples of image captioning and image generation by our trained framework in Figure 12 and Figure 13, respectively. Detailed annotations and explanations can be found in the corresponding figure captions.

Input	Sample #1	Sample #2	Sample #3	Sample #4	Sample #5
	<p>the pink golf cart is parked outside a trailer</p> <p>CIDEr: 348.8 CLIP Text: 0.849</p>	<p>the pink golf cart is parked outside a trailer</p> <p>CIDEr: 343.2 CLIP Text: 0.799</p>	<p>the small golf cart has pink paint and has a name written on it</p> <p>CIDEr: 152.4 CLIP Text: 0.821</p>	<p>a pink golf cart with a girly sign on the side</p> <p>CIDEr: 343.4 CLIP Text: 0.864</p>	<p>a pink cart that is sitting outside</p> <p>CIDEr: 141.7 CLIP Text: 0.771</p>
	<p>a black and yellow bird is perched on a tree branch</p> <p>CIDEr: 343.2 CLIP Text: 0.873</p>	<p>a black and yellow bird sitting in a tree</p> <p>CIDEr: 307.1 CLIP Text: 0.805</p>	<p>a yellow and black bird perched in a tree</p> <p>CIDEr: 211.1 CLIP Text: 0.869</p>	<p>two birds perched on a tree branch in the forest</p> <p>CIDEr: 85.8 CLIP Text: 0.639</p>	<p>the two birds are sitting on the branch</p> <p>CIDEr: 62.4 CLIP Text: 0.581</p>
	<p>there are a lot of chinese lanterns hanging from the ceiling</p> <p>CIDEr: 200.2 CLIP Text: 0.687</p>	<p>colorful oriental lanterns hung from the ceiling and light the room</p> <p>CIDEr: 139.5 CLIP Text: 0.735</p>	<p>many different colorful hanging lanterns on the ceiling</p> <p>CIDEr: 178.5 CLIP Text: 0.784</p>	<p>brightly lit colorful lanterns on the ceiling of a shop</p> <p>CIDEr: 156.8 CLIP Text: 0.725</p>	<p>several colorful lamps with picture frames hanging</p> <p>CIDEr: 31.7 CLIP Text: 0.590</p>
	<p>bottles of milk with name tags for sale</p> <p>CIDEr: 125.5 CLIP Text: 0.655</p>	<p>three bottles of milk on a shelf in a store</p> <p>CIDEr: 260.4 CLIP Text: 0.764</p>	<p>a few bottles of cow milk sitting in a row</p> <p>CIDEr: 116.6 CLIP Text: 0.647</p>	<p>milk products are on a shelf in a supermarket</p> <p>CIDEr: 86.3 CLIP Text: 0.652</p>	<p>a display in a grocery store with several different drinks</p> <p>CIDEr: 5.4 CLIP Text: 0.534</p>
	<p>a white plate topped with a piece of french toast and topped with strawberries</p> <p>CIDEr: 221.3 CLIP Text: 0.818</p>	<p>a close up of french toast with strawberries</p> <p>CIDEr: 333.6 CLIP Text: 0.825</p>	<p>plate of french toast topped with powdered sugar and sliced strawberries</p> <p>CIDEr: 297.1 CLIP Text: 0.858</p>	<p>french toast with strawberries and powdered sugar on a white plate</p> <p>CIDEr: 366.6 CLIP Text: 0.862</p>	<p>a white plate with a white dish of food and some strawberries</p> <p>CIDEr: 69.6 CLIP Text: 0.629</p>
	<p>a statue of a gnome and potted flower</p> <p>CIDEr: 186.4 CLIP Text: 0.772</p>	<p>a close up of a gnome on a porch</p> <p>CIDEr: 150.3 CLIP Text: 0.590</p>	<p>a gnome statue standing next to a flower pot</p> <p>CIDEr: 190.5 CLIP Text: 0.796</p>	<p>the garden gnome is standing on the pavement with his arms out</p> <p>CIDEr: 156.6 CLIP Text: 0.568</p>	<p>a statue and potted plant next to steps</p> <p>CIDEr: 42.8 CLIP Text: 0.596</p>
	<p>an otter in the water looking forward at something</p> <p>CIDEr: 173.6 CLIP Text: 0.834</p>	<p>a close up of an otter's face in water</p> <p>CIDEr: 152.1 CLIP Text: 0.717</p>	<p>an otter that is sitting in the water</p> <p>CIDEr: 207.7 CLIP Text: 0.822</p>	<p>a river otter in the water looking to his left</p> <p>CIDEr: 144.2 CLIP Text: 0.819</p>	<p>a close up shot of an animal in the water</p> <p>CIDEr: 31.2 CLIP Text: 0.508</p>
	<p>a glass of beer on the bar has an orange wedge on it</p> <p>CIDEr: 136.3 CLIP Text: 0.644</p>	<p>a glass of beer sitting on a table</p> <p>CIDEr: 290.2 CLIP Text: 0.763</p>	<p>a beverage sitting in a glass on top of a table</p> <p>CIDEr: 46.0 CLIP Text: 0.708</p>	<p>a close up of a glass of alcohol on a table</p> <p>CIDEr: 42.1 CLIP Text: 0.623</p>	<p>a small orange slice in a glass sitting on top of a wooden table</p> <p>CIDEr: 56.5 CLIP Text: 0.571</p>
	<p>a couple of people on a snowmobile outside a hotel</p> <p>CIDEr: 262.8 CLIP Text: 0.783</p>	<p>people are sitting on a snowmobile and waiting for something</p> <p>CIDEr: 160.0 CLIP Text: 0.702</p>	<p>people standing on a snow bike in front of a hotel</p> <p>CIDEr: 124.6 CLIP Text: 0.618</p>	<p>a man and two women are riding on a snow bike</p> <p>CIDEr: 38.1 CLIP Text: 0.707</p>	<p>two men and a woman sit on the front of a ski machine outside of a building</p> <p>CIDEr: 11.1 CLIP Text: 0.630</p>
	<p>an eggplant, carrots and a knife sitting on the table</p> <p>CIDEr: 229.1 CLIP Text: 0.780</p>	<p>a knife and two fresh carrots next to each other</p> <p>CIDEr: 116.1 CLIP Text: 0.734</p>	<p>four different types of carrots on a cutting board</p> <p>CIDEr: 69.8 CLIP Text: 0.633</p>	<p>three different colored vegetables laying next to a knife</p> <p>CIDEr: 93.2 CLIP Text: 0.675</p>	<p>the vegetables have been cut into two large ones</p> <p>CIDEr: 11.9 CLIP Text: 0.605</p>
	<p>the small lemurs are perched on the tree branch</p> <p>CIDEr: 28.7 CLIP Text: 0.642</p>	<p>a baby lemur up in a tree</p> <p>CIDEr: 35.3 CLIP Text: 0.629</p>	<p>a black and white lemur on top of a tree</p> <p>CIDEr: 29.9 CLIP Text: 0.594</p>	<p>a monkey up in a tree looking at the camera</p> <p>CIDEr: 176.4 CLIP Text: 0.763</p>	<p>a large gray and white monkey standing on top of a tree</p> <p>CIDEr: 156.0 CLIP Text: 0.624</p>
	<p>the silhouette of a person walks on a bridge</p> <p>CIDEr: 82.7 CLIP Text: 0.596</p>	<p>a person carrying an umbrella walking along a wooden bridge</p> <p>CIDEr: 152.7 CLIP Text: 0.733</p>	<p>woman holding umbrella walking on wooden bridge in the sun</p> <p>CIDEr: 158.3 CLIP Text: 0.749</p>	<p>a man is walking across the bridge with an umbrella</p> <p>CIDEr: 195.5 CLIP Text: 0.723</p>	<p>a woman with an umbrella walking across a bridge</p> <p>CIDEr: 217.9 CLIP Text: 0.800</p>

Figure 10. Examples for the image-text-image task using Top- p sampling. The first column displays the input images, followed by the generated caption and its corresponding generated image. We rank the generated text-image pairs based on the similarity of the images and show the score of the caption below each text. In the first row, *golf car* in the first sample is a more accurate description than *cart* in the fifth sample so that the first generated image is closer to the input image. Additionally, we show a few failed examples in the last two rows.

Input	Sample #1	Sample #2	Sample #3	Sample #4	Sample #5
People are standing around a black luxury vehicle.	 Visual Score: 0.559 FID: 179.6	 Visual Score: 0.489 FID: 209.5	 Visual Score: 0.549 FID: 213.4	 Visual Score: 0.456 FID: 217.2	 Visual Score: 0.452 FID: 152.1
A woman dressed in a black outfit is playing the harp on a stage	 Visual Score: 0.680 FID: 55.2	 Visual Score: 0.642 FID: 97.8	 Visual Score: 0.609 FID: 71.2	 Visual Score: 0.356 FID: 575.0	 Visual Score: 0.482 FID: 453.2
A three layer white cake with blue, pink, red and green figures on top.	 Visual Score: 0.610 FID: 269.6	 Visual Score: 0.668 FID: 280.7	 Visual Score: 0.644 FID: 333.1	 Visual Score: 0.544 FID: 333.9	 Visual Score: 0.533 FID: 382.9
A crowd stands near a memorialized motorcycle that has a passenger car attached to it.	 Visual Score: 0.722 FID: 126.0	 Visual Score: 0.583 FID: 153.9	 Visual Score: 0.676 FID: 111.2	 Visual Score: 0.611 FID: 146.8	 Visual Score: 0.428 FID: 216.0
A saxophone lays on a glass table in front of a window, and we can see its reflection in the table.	 Visual Score: 0.614 FID: 106.9	 Visual Score: 0.592 FID: 160.4	 Visual Score: 0.772 FID: 71.5	 Visual Score: 0.410 FID: 389.2	 Visual Score: 0.437 FID: 303.5
A black classic car with a black license plate.	 Visual Score: 0.799 FID: 99.1	 Visual Score: 0.610 FID: 207.0	 Visual Score: 0.719 FID: 213.4	 Visual Score: 0.596 FID: 138.9	 Visual Score: 0.541 FID: 136.7
A curled up pretzel on a desk next to a computer mouse and wires.	 Visual Score: 0.668 FID: 140.9	 Visual Score: 0.754 FID: 93.2	 Visual Score: 0.592 FID: 174.8	 Visual Score: 0.652 FID: 100.7	 Visual Score: 0.452 FID: 348.8
A red bag of buttered popcorn sitting in a chair.	 Visual Score: 0.825 FID: 314.5	 Visual Score: 0.832 FID: 290.5	 Visual Score: 0.819 FID: 250.3	 Visual Score: 0.751 FID: 261.3	 Visual Score: 0.487 FID: 365.3
A hockey player skating with a hockey puck.	 Visual Score: 0.643 FID: 45.8	 Visual Score: 0.739 FID: 155.5	 Visual Score: 0.520 FID: 144.9	 Visual Score: 0.624 FID: 67.0	 Visual Score: 0.574 FID: 88.7
Three young men are in a room playing the bass, the guitar and a trombone.	 Visual Score: 0.473 FID: 130.8	 Visual Score: 0.568 FID: 130.8	 Visual Score: 0.559 FID: 236.8	 Visual Score: 0.399 FID: 186.3	 Visual Score: 0.488 FID: 153.2
A person is sitting in a chair in front of audio equipment.	 Visual Score: 0.531 FID: 302.9	 Visual Score: 0.512 FID: 194.6	 Visual Score: 0.496 FID: 235.4	 Visual Score: 0.602 FID: 354.8	 Visual Score: 0.721 FID: 305.4
A black bottle of perfume with some floral details	 Visual Score: 0.512 FID: 311.1	 Visual Score: 0.463 FID: 415.3	 Visual Score: 0.552 FID: 299.0	 Visual Score: 0.533 FID: 301.0	 Visual Score: 0.643 FID: 206.5

Figure 11. Examples for the text-image-text task. The first column shows the input text, followed by generated image and its corresponding generated text. We rank the generated image-text pairs by the similarity of the text and show the score of the image in the box. As shown in the first line, the image in the first sample represents the input text better than the image in the fifth sample, and the similarity of the reconstructed text reflected this comparison. Further, we show some failed examples in the last two rows.


	<p>GT: A little girl with long straight reddish brown hair tilts her head to the right and smiles at us.</p> <p>BLIP: a close up of a child with a toothbrush</p> <p>Ours: a little girl with a smile on her face</p>		<p>GT: Three men in sports uniforms smiling next to each other</p> <p>BLIP: a group of men standing next to each other on a podium</p> <p>Ours: three men standing on a podium with medals around their necks</p>		<p>GT: A man with a ponytail that is playing the saxophone.</p> <p>BLIP: a man playing a trumpet in a black and white photo</p> <p>Ours: a man playing a saxophone in a black and white photo</p>
	<p>GT: People on motorcycles in the street with helmets.</p> <p>BLIP: a group of people riding scooters down a street</p> <p>Ours: a group of people riding motorcycles down a street</p>		<p>GT: A little girl in a blue shirt preparing to shoot a basket.</p> <p>BLIP: a group of young people playing a game of basketball</p> <p>Ours: a young girl holding a basketball in a gym</p>		<p>GT: A woman standing beside an vintage two door car.</p> <p>BLIP: a woman standing in front of an old car</p> <p>Ours: a woman in a dress and hat standing next to an old car</p>
	<p>GT: A man wearing a jacket and doing tricks on a bike.</p> <p>BLIP: a man riding a bike down the middle of a street</p> <p>Ours: a man is doing a trick on a bicycle</p>		<p>GT: An empty cup of coffee on a saucer by a phone.</p> <p>BLIP: a cup of coffee sitting on top of a white saucer</p> <p>Ours: a cup of coffee on a saucer next to a cell phone</p>		<p>GT: Two zebras standing close together on dirt surface.</p> <p>BLIP: a couple of zebra standing next to each other</p> <p>Ours: two zebras standing next to each other on a dirt road</p>
	<p>GT: Pilot and smiling copilot sitting in cockpit of airplane.</p> <p>BLIP: a couple of people that are sitting in a plane</p> <p>Ours: a man and a woman sitting in the cockpit of an airplane</p>		<p>GT: Woman with red hair holding silver digital camera in front of her face.</p> <p>BLIP: a woman taking a picture of herself with a camera</p> <p>Ours: a woman holding a camera in front of her face</p>		<p>GT: A person on a motorized vehicle in a red sweater.</p> <p>BLIP: a man riding an electric scooter in a parking lot</p> <p>Ours: a man on a segway at a car show</p>
	<p>GT: A small plant grows out of some dirt in a small pot.</p> <p>BLIP: a small green tree sitting on top of a wooden table</p> <p>Ours: a bonsai tree in a pot on a table</p>		<p>GT: Three women are dancing while wearing a black and white dress.</p> <p>BLIP: a group of people that are standing on a stage</p> <p>Ours: a group of women in black and white dresses dancing</p>		<p>GT: A man drinking a beer on a golf cart.</p> <p>BLIP: a man sitting in a golf cart on a golf course</p> <p>Ours: a man sitting in a golf cart drinking a beverage</p>
	<p>GT: Glass curtained windows, rest above new-looking brick siding, fronted by a pair of healthy sunflowers.</p> <p>BLIP: a large yellow sunflower in front of a brick wall</p> <p>Ours: two sunflowers in front of a brick wall</p>		<p>GT: A man skis downhill on icy snow, one winter.</p> <p>BLIP: a man riding skis down a snow covered slope</p> <p>Ours: a person in an orange and white ski suit skiing down a hill</p>		<p>GT: A woman in black rides a brightly painted cruiser motorcycle with a small dog on her lap.</p> <p>BLIP: a man riding a motorcycle down a street</p> <p>Ours: a person riding a motorcycle with a dog on the back</p>
	<p>GT: A closeup of a large lobster in water</p> <p>BLIP: a close up of a red and white crab under water</p> <p>Ours: a close up of a red and white lobster</p>		<p>GT: Grey dolphin swimming with half body out of water</p> <p>BLIP: a dolphin swimming in the ocean near a boat</p> <p>Ours: a dolphin is swimming in the blue water</p>		<p>GT: Empty living room with black sofa and yellow striped curtains</p> <p>BLIP: a living room filled with furniture and a flat screen tv</p> <p>Ours: a living room with a black couch and a television</p>
	<p>GT: A chef stands by an appliance making something good.</p> <p>BLIP: a man standing in front of a pizza oven</p> <p>Ours: a man in a red apron standing in front of a machine</p>		<p>GT: A white bicycle chained to a lamp post.</p> <p>BLIP: a bicycle parked next to a pole on a city street</p> <p>Ours: a white bicycle is chained to a pole</p>		<p>GT: A blonde woman wearing a blue and white striped shirt and pink scarf.</p> <p>BLIP: a woman with a cell phone in her hand</p> <p>Ours: a woman wearing a blue and white striped shirt</p>
	<p>GT: A white limousine driving down a road.</p> <p>BLIP: a white limousine parked on the side of a road</p> <p>Ours: a white car parked on the side of a road</p>		<p>GT: A person sitting on one of the four stools standing in a line</p> <p>BLIP: a woman sitting on a stool next to two stools</p> <p>Ours: a woman sitting at a bar with her legs crossed</p>		<p>GT: Large piece of brown cake on top of a tray.</p> <p>BLIP: a loaf of bread sitting on top of a metal rack</p> <p>Ours: a close up of a piece of food on a rack</p>
	<p>GT: A red grandfather clock is sitting by a white wall.</p> <p>BLIP: a red grandfather clock sitting on top of a wooden floor</p> <p>Ours: a tall red clock sitting on top of a wooden floor</p>		<p>GT: Cookies are on a tray being cooked under heat.</p> <p>BLIP: a close up of chocolate chip cookies on a baking sheet</p> <p>Ours: a bunch of cookies that are on a table</p>		<p>GT: A bunch of colorful balloons stand outside of tall buildings.</p> <p>BLIP: a bunch of green and yellow balloons in front of a building</p> <p>Ours: a bunch of balloons that are in the air</p>

Figure 12. Qualitative results for image captioning. We compare the performance of our approach and BLIP ViT-B baseline. A ground truth (GT) caption is given for each image. On average, our method provides more accurate descriptions for input images. The last two rows show examples of our model performing worse than the baseline method.

Input	Ours				SD			
A green car parked outside of a house.								
A kitchen with a table with chairs and a ceiling fan.								
White translucent jellyfish light up the darkness deep below the water.								
Man is enjoying playing while wearing their helmet.								
A dresser has some books on the top of it.								
A clock stands next to a white wall on wood floor.								
The lifeguard is sitting in a high chair with chairs and umbrellas behind him.								
Having a picnic outside your car is fun.								
Kitchen utensils alongside food in a large table								
A grey and white two story house between a grey fence with green bushes around the fence								

Figure 13. Qualitative results for image generation. We compare the original SD model with our finetuned model. Each row displays firstly the input text, followed by four images generated by our approach, and four images generated by the SD baseline. We can see that compared with the baseline method, the image generated by our method better reflects the semantics of the input text.