SINE: SINgle Image Editing with Text-to-Image Diffusion Models – Supplementary Materials

Zhixing Zhang[1] Ligong Han[1] Arnab Ghosh[2] Dimitris Metaxas[1] Jian Ren[2]
[1]Rutgers University [2]Snap Inc.

1. Overview

In the supplementary material, we provide the following:

- More comparisons with exiting works on editing one single real image (Sec. 2).
- More results for applying SINE on editing a single image and the novel image manipulation tasks that our approach can enable (Sec. 3).
- Results for applying Model-base Classifier-free Guidance on multiple image editing tasks (Sec. 4).
- More ablation analysis (Sec. 5).
- Discussion about the limitation of our method and possible future work (Sec. 6).

2. More Comparisons

2.1. Test-time Efficiency Comparisons

We evaluate the test-time efficiency of Dreambooth [9], Textual-Inversion [5] and our work with one RTX8000 GPU. The run time for Dreambooth [9], Textual-Inversion [5], and our work is $12.20s$, $23.20s$, and $19.69s$, respectively. We use the same resolution and number of denoising steps for testing.

2.2. Quantitative Comparisons

We conduct quantitative experiments with the following setting: 1) We prepare the images for 7 objects; 2) For each object, we apply 8 different prompts and resolution pairs on the diffusion models to sample 4 editing results; 3) We calculate LPIPS for the image alignment and CLIP-score for text alignment. Fig. 1 shows the comparison between Dreambooth [9] and Textual Inversion [5], demonstrating the advantages of our approach.

2.3. Qualitative Comparisons

Besides the comparison with existing works shown in the main paper, we provide more results by comparing our approach with Prompt-to-Promt [6]. In addition, we compare our methods with training-free single-image editing approaches, including SDEdit [7] and ILVR [4].

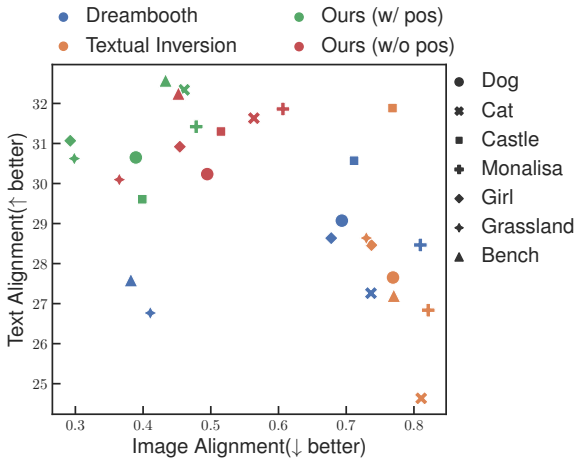


Figure 1. **Quantitative Comparisons.** We compare our method with Dreambooth [9] and Textual Inversion [5] over multiple objects.

We first show the technical differences between our works and training-free methods in Tab. 1. SDEdit [7] applies the diffusion process on an image or a user-created semantic map to conduct the denoising procedure, conditioned with the desired output. ILVR [4] guides the denoising process by replacing the low-frequency part of the sample with that of the target reference image.

The visual comparisons are illustrated in Fig. 2. As can be seen, our approach significantly outperforms other methods for generating high-fidelity images with the maximal keeping of the details in the source image.

We also compare our patch-based method with Anyres-GAN [3]. The main difference is that Anyres-GAN uses high-resolution data to train the model for high-resolution generation, requiring an image *dataset*. Also, all training patches from Anyres-GAN are cropped from higher/equal-resolution images and keep original spatial configuration. In our method, we use a single image for training, and the pre-trained auto-encoder of the stable diffusion model provides high-resolution generation. We further apply the method from Anyres-GAN to our model and show the re-

Table 1. The differences between our approach and other training-free methods for single image editing.

Guidance	Finetune	Compatible w/ LDM [8]	Position Control	Admits Multiple Inputs
SINE (Ours)	Required	✓	✓	✓
ILVR [4]	Not Required	✗	✗	✗
SDEdit [7]	Not Required	✓	✗	✗

Figure 2. **Comparison results.** We compare our method with ILVR [4], SDEdit [7], and Prompt-to-Prompt [6] on editing single real image. Note that when the hyper-parameter N is set to 1, the process of ILVR is equivalent to stochastic SDEdit. We adopt the official implementation of ILVR for conducting experiments on SDEdit. For ILVR, we set downsample ratio of $N = 8$. For SDEdit, we use the stochastic q sample. In both cases, we set $K = 400$.

sults in Fig. 3. Anyres-GAN *fails to disentangle* the correlation between content and spatial location and generates repeated patterns or blurry images.

We further show the difference between our method and the one provided in Blended Diffusion(BD) [2]. There are four major differences: 1) BD needs an editing mask while we do not; 2) BD requires test-time optimization while we do not; 3) BD only performs region-replacement editing while we can do global editing, *e.g.,* style transfer; 4) The identity of the object is changed when using BD for editing (a similar issue for the later work Blended Latent Diffusion [1]), while we can keep the identity of the object. We also provide comparisons in Fig. 4

3. More Editing Results and Tasks

3.1. More Editing Results

We provide more editing results in Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, and Fig. 10. All results are obtained by fine-tuning the large-scale text-to-image model [8] using our proposed patch-based method at the resolution of 512×512 and sampling with our introduced model-based classifier-free guidance at a higher resolution, *e.g.*, 768×1024. Images on the top-left corner of these results are the real images utilized for fine-tuning. We specify the hyper-parameters used during sampling in the caption of each image.

Our method can be applied to non-rigid editing tasks. We have shown non-rigid editing in Figs. 1&3 of the main paper, *i.e.*, coffee machine dog and jumping cat. In Fig. 11, we provide more examples, *i.e.*, changing dog pose, and closing human eyes, as follows.

3.2. More Editing Tasks

Face manipulation. Our method demonstrates promising editing ability for *in-the-wild human faces*. As shown in Fig. 12, our approach can edit locally and globally on human faces for various facial manipulation tasks, *e.g.*, image stylization, adding accessories, and age changing.

Content removal. In Fig. 13a, we show the content re-

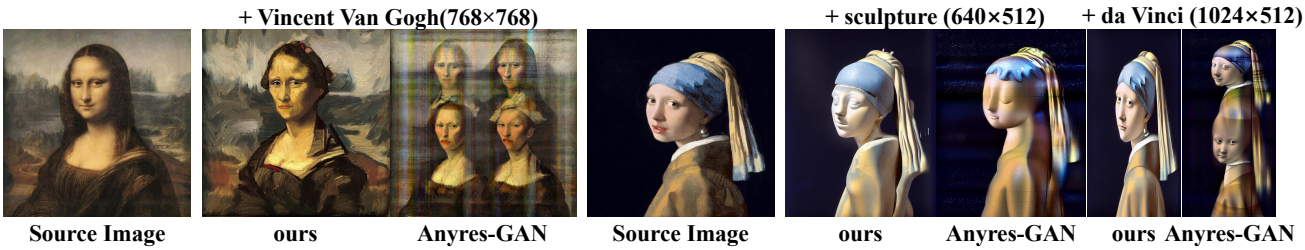


Figure 3. **Comparison results with Anyres-GAN [3].** We apply the training method proposed in Anyres-GAN in our problem and show the editing results.

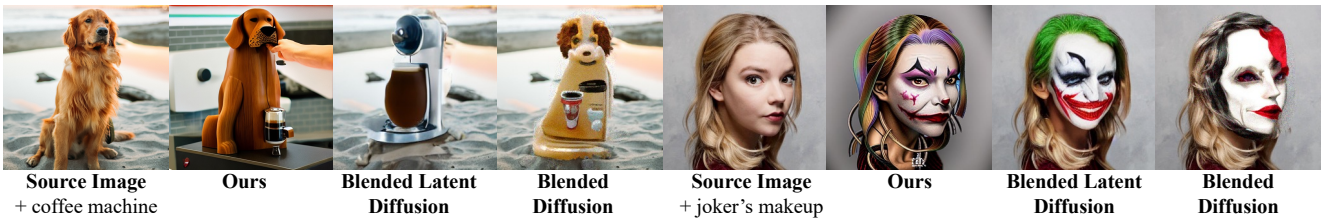


Figure 4. **Comparison results with Blended Diffusion [2] and Blended Latent Diffusion [1].**

moval using our approach. We fine-tune the pre-trained large-scale text-to-image model with the language descriptor as "a [$*$] dog with a flower in mouth". At sampling time, we use text prompts such as "a dog" and "a [$*$] dog" for the pre-trained and fine-tuned models. The pre-trained model can successfully remove the flower held in the mouth of the dog.

Style generation. Our method can also be employed to learn the underlying style of an image. As shown in Fig. 13b, the model is fine-tuned with the text, "a painting in the [$*$] style". When sampling results, we feed the pre-trained model a prompt as "painting of a forest". The model can successfully synthesize images with the specified content in the style of the given real image.

Style transfer. Our model-based classifier-free guidance can be leveraged to combine multiple models for providing the guidance. We show the result in Fig. 13c by doing a style transfer task with dual-model guidance. We fine-tune two models using prompts: "picture of a [$*$] dog" and "painting in [$*$] style". During inference, we give the pre-trained model the prompt "painting of a dog" and fine-tuned models with prompts the same as training. With guidance from two separate models, our method can generate images with the content from one and style from the other and achieve stylized generation.

4. Results on Multiple Images

We train our model on the number of images from 1 to 6 and show the editing results in Fig. 14. We can see the following: 1) Our method can be trained on multiple images; 2) With more images, the generated object contains features from multiple source images; 3) The model has a better geometry understanding of the object as it can synthesize the object from a different view direction.

5. More Ablations

Analysis on guidance step K and guidance weight v. We conduct experiments by varying the guidance step K and guidance weight v in Fig. 15, Fig. 16, and Fig. 17. We use the same random seed and generate results with specific text prompts at a fixed resolution by varying the parameters. These experiments show the same behavior of our approach as mentioned in Sec. 4.3 of our main paper. By adjusting these two parameters, we can find an optimal combination specifically for the image and the target language guidance. In most cases, we adopt the parameters setting of $K = 400$ and $v = 0.7$. However, we want our model to maintain more fidelity or apply a stronger edit in some instances. For example, the "optimal" setting we decide for experiments in Fig. 15 is $v = 0.5$ and $K = 400$.

Analysis on regularization loss. Dreambooth [9] proposes to leverage Prior-Preservation Loss(PPL) to address the issues of overfitting and language drift. They propose to generate 200 samples with the pre-trained model using the prompt "a [class noun]". Then, during fine-tuning, they use these samples to regulate the model with the Prior-Preservation Loss to maintain the generalization ability of the model. However, in our experiments, as shown in Fig. 18, this loss does not improve the final results due to the uniqueness of certain pictures/paintings. On the con-

Figure 5. **A children's painting of a castle.** The generation resolution is set to $H = 768$ and $W = 1024$. We use $K = 400$ and $v = 0.7$ in this sample.

trary, more artifacts are introduced to the results, and the fidelity of the editing results decreases. Therefore, given the motivation of editing unique images, we forfeit the generalization ability provided by regularizing the model with the samples generated by the pre-trained model. We encourage our model to overfit a single image for the fidelity of the editing results.

6. Limitations

We present some failure cases in Fig. 19. As mentioned in the main paper, when confusing guidance is given to the model or drastic change is to be applied, our method produces unsatisfying results. The language comprehension limitation of the pre-trained model and the over-fitting issue of our fine-tuned model can cause this. It would be an interesting future direction to explore how to over-fit on one single image without "forgetting" prior knowledge.

Also, as can be noticed in the second row of Fig. 12, the color of the sweater is changed in most cases. Also, the background letters are twisted after editing. The identity and the color are also changed for the non-rigid editing result on the right of Fig. 11. Even though our method can perform editing with maximal protection of the details in the source image, editing strictly on a specific part of an image is also worth further exploration.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2, 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 3

[3] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution

Figure 6. **A painting of a castle in the style of Claude Monet.** The output resolution is set to $H = 768$ and $W = 1024$. We use $K = 400$ and $v = 0.65$ in this example.

image synthesis. *ECCV*, 2022. 1, 3

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 1, 2

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2

[7] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1, 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1, 3

Figure 7. **A photo of a lake with many sailboats.** The output resolution is set to $H = 768$ and $W = 1024$. We use $K = 400$ and $v = 0.7$ in this case.

Figure 8. **A desert.** The output resolution is set to $H = 768$ and $W = 1024$. We use $K = 500$ and $v = 0.8$ in this case.

Figure 9. **A desert.** The output resolution is set to $H = 768$ and $W = 1024$. We use $K = 500$ and $v = 0.8$ in this case.

Figure 10. **A watercolor painting of a girl.** The output resolution is set to $H = 1024$ and $W = 768$. We use $K = 400$ and $v = 0.6$ in this case.

Figure 11. **Non-rigid Editing Results.**

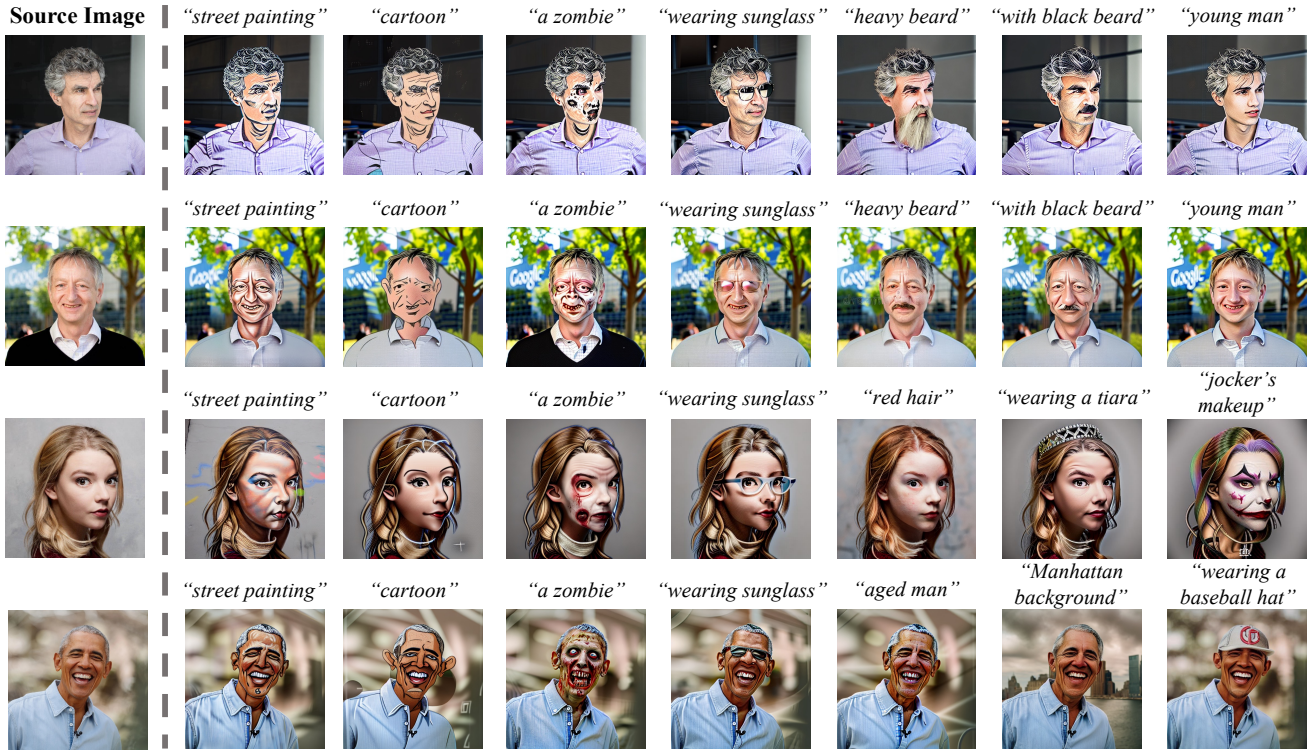
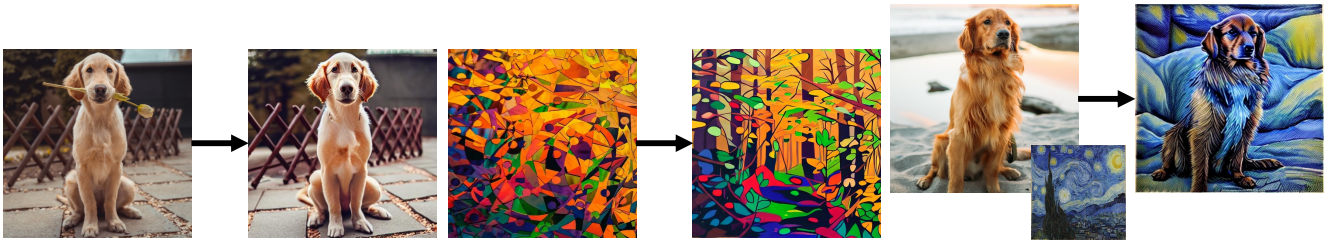


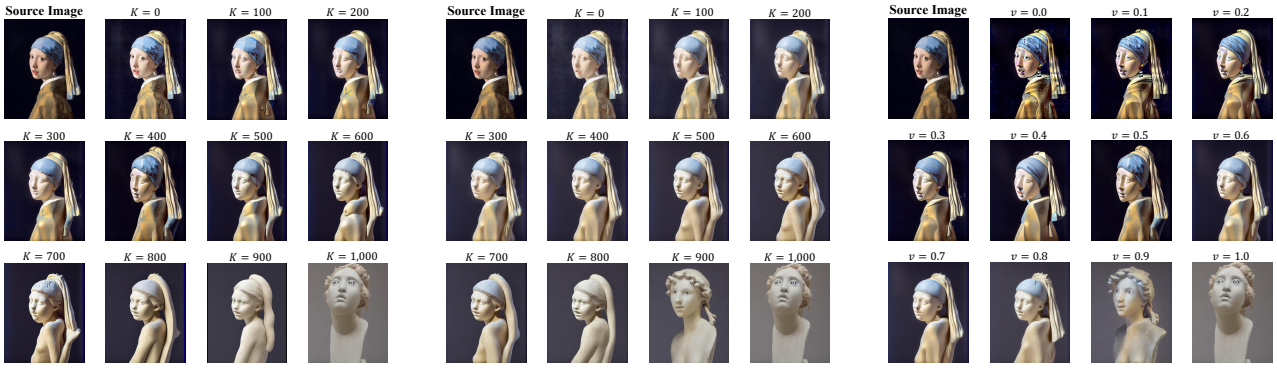
Figure 12. **In-the-wild human face manipulation.** We conduct various editing on human face photos, locally or globally. The models are trained and edited at a resolution of 512×512.



(a) Content Removal (b) Style Generation (c) Style Transfer

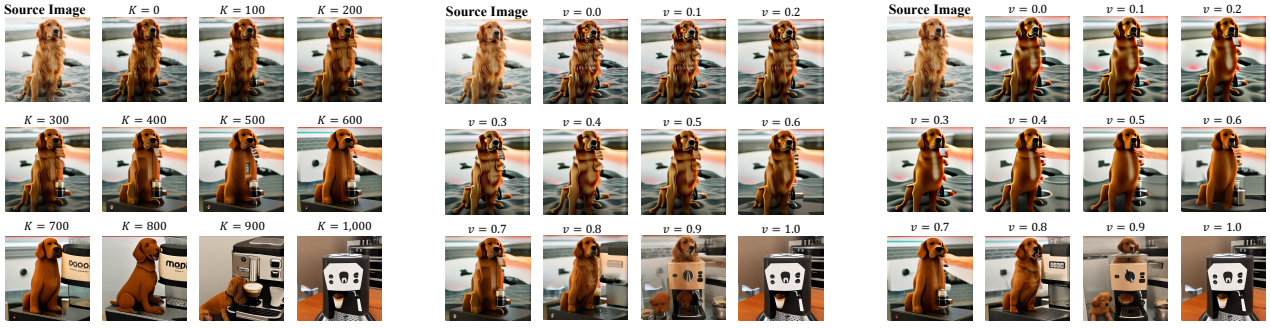
Figure 13. **More applications.** We show how our approach can be applied to various tasks in image editing, such as content removal (a), style generation (b), and style transfer (c).

Figure 14. **Results on Multiple Images.**



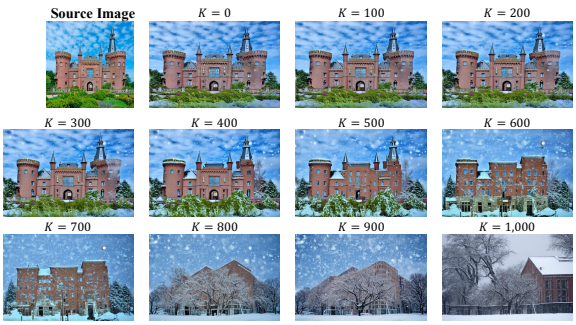
(a) We set $v = 0.5$ and change K (b) We set $v = 0.7$ and change K. (c) We set $K = 400$ and change v

Figure 15. **"A sculpture of a girl"** with the resolution of $H = 640$ and $W = 512$.

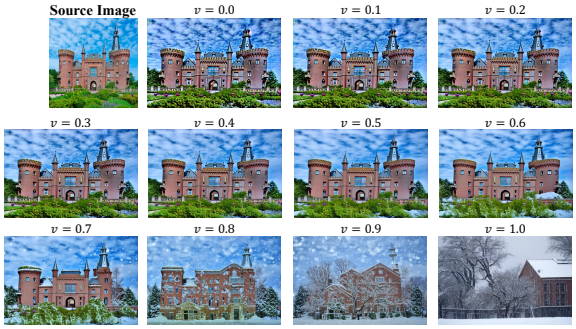


(a) We set $v = 0.7$ and change K (b) We set $K = 400$ and change v (c) We set $K = 600$ and change v

Figure 16. **"A coffee machine in the shape of a dog"** with the resolution of $H = 512$ and $W = 512$.



(a) Set $v = 0.7$ varying K



(b) Set $K = 400$ varying v

Figure 17. **"A castle covered by snow"** with the resolution of $H = 512$ and $W = 768$.

Figure 18. Analysis of Prior-Preservation Loss (PPL).

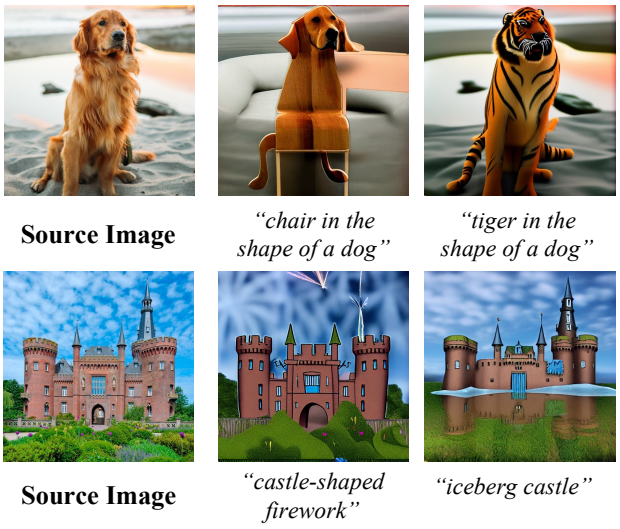


Figure 19. **Failure cases.** We showcase where our method fails to generate results with high fidelity and text alignment.