Paper Reading

Paint by Example: Exemplar-based Image Editing with Diffusion Models

Wenzhen Long

April 27,204

INTRODUCTION

- Creative editing for photos has become a ubiquitous need due to the advances in a plethora of social media platforms. AI-based techniques significantly lower the barrier of fancy image editing that traditionally requires specialized software and labor-intensive manual operations. Recent large-scale language-image (LLI) models, based on either auto-regressive models or diffusion models have shown unprecedented generative power in modeling complex images. However, even the detailed textual description inevitably introduces ambiguity and may not accurately reflect the user-desired effects; indeed, many fine-grained object appearances can hardly be specified by the plain language. In this paper, an example-based image editing method is proposed, which allows accurate semantic manipulation of image content based on user-provided example images or sample images retrieved from databases.

QUESTION

- Different from textguided models, the core challenge is that it is infeasible to collect enough triplet training pairs comprising source image, exemplar and corresponding editing ground truth.
- One workaround is to randomly crop the objects from the input image, which serves as the reference when training the inpainting model. The model trained from such a selfreference setting, however, cannot generalize to real exemplars, since the model simply learns to copy and paste the reference object into the final output

METHOD

- The source image as $x_s \in R^{H \times W \times 3}$, with H and W being the width and height respectively. The edit region could be a rectangular or an irregular shape (at least connected) and is represented as a binary mask m $\in \{0, 1\}^{H \times W}$ where value 1 specifies the editable positions in x_s.
- Given a reference image $x_r \in R^{H \times W \times 3}$containing the desired object, our goal is to synthesize an image y from {x_s, x_r, m}, so that the region where m = 0 remains as same as possible to the source image x_s, while the region where m = 1 depicts the object as similar to the reference image x_r and fits harmoniously.

METHOD

- It is impossible to collect and annotate paired data, i.e. {(xs, xr, m), y}, for the training of exemplar-based image editing. It may take great expense and huge labor to manually paint reasonable output. Thus, we propose to perform self-supervised training.

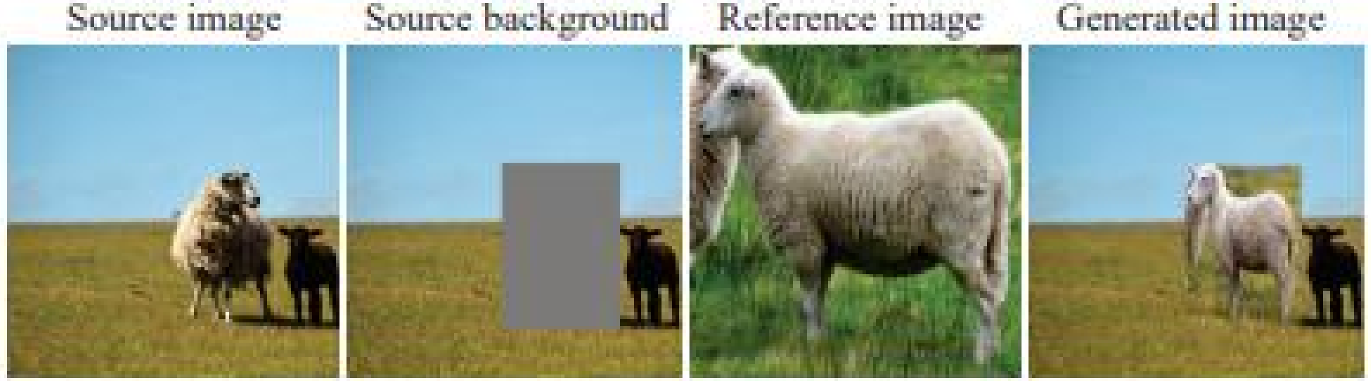


Figure 2. Illustration of the copy-and-paste artifacts of the naive solution. The generated image is extremely unnatural.

METHOD

- In this paper, they propose three principles. 1) Introduce the content bottleneck to force the network to understand and regenerate the content of the reference image instead of just copy. 2) Adopt strong augmentation to mitigate the train-test mismatch issue. This helps the network not only learn the transformation from the exemplar object, but also from the background. 3) Another critical feature for exemplar-based image editing is controllability. We enable control over the shape of the edit region and the similarity degree between the edit region and the reference image.

CONTENT BOTTLENECK

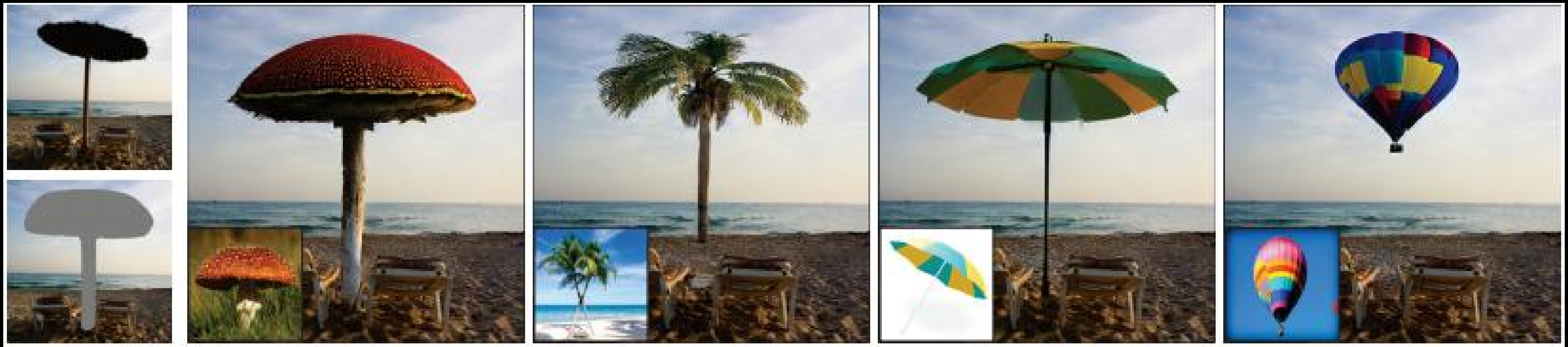
- Increase the difficulty of reconstructing the mask region by compressing the information of the reference image.
- This highly compressed representation tends to ignore the high-frequency details while maintaining the semantic information. It forces the network to understand the reference content and prevents the generator from directly copy-and-paste to reach the optimal results in training.
- To further avoid the trivial solution of directly remembering the reference image, we leverage a well-trained diffusion model for initialization as a strong image prior.

STRONG AUGMENTATION

- Reference image augmentation. The first mismatch is that the reference image x_r is derived from the source image x_s during training, which is barely the case for the testing scenario. To reduce the gap, we adopt several data augmentation techniques (including flip, rotation, blur and elastic transform) on the reference image to break down the connection with the source image. We denote these data augmentation as A. Formally, the condition fed to the diffusion model is denoted as:$c = MLP(CLIP(A(x_r)))$

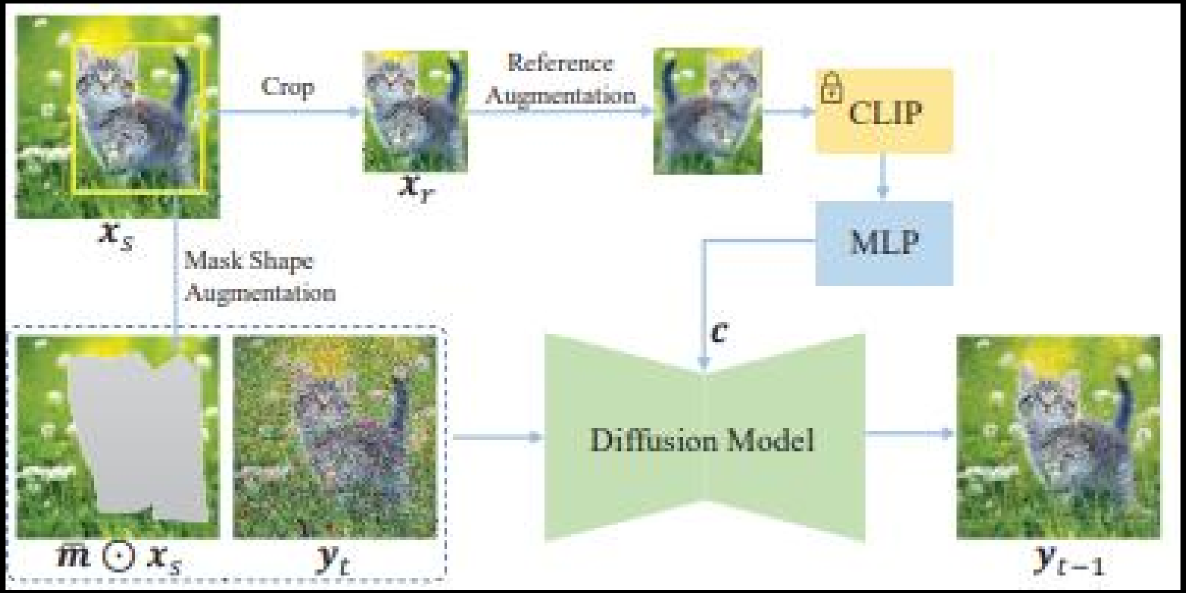
CONTROL THE MASK SHAPE

- Another benefit of mask shape augmentation is that it increases the control over mask shape in the inference stage. In practical application scenarios, a rectangle mask usually can not represent the mask area precisely. e.g. the sun umbrella in Figure. In some cases people would like to edit a specific region while preserving the other area as much as possible, this leads to the demand for handling irregular mask shapes. By involving these irregular masks into training, our model is able to generate photo-realistic results given various shape masks.



CONTROL THE SIMILARITY DEGREE

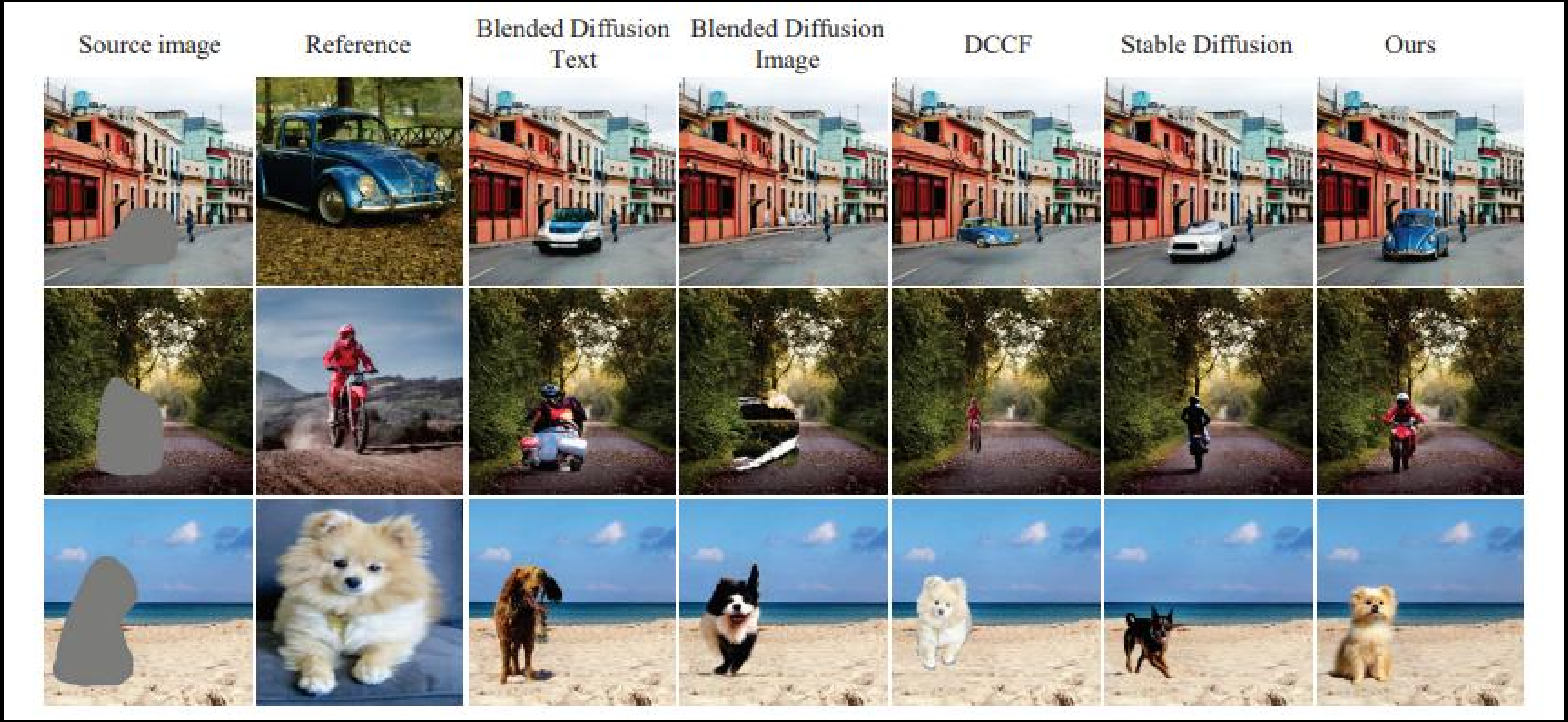
- To control the similarity degree between the edited area and the reference image, we find that classifier-free sampling strategy is a powerful tool.
- Above all, the overall framework of our method is illustrated in Figure.



EXPERIMENTS

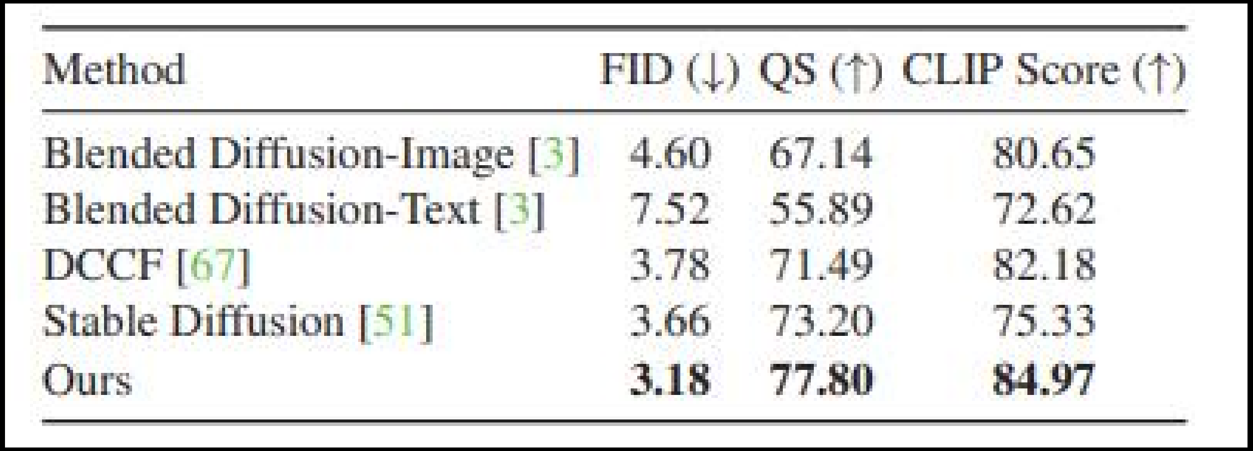
- Use the following three metrics to evaluate the generated images. 1) FID score, which is widely used to evaluate generated results. We follow and use CLIP model to extract the feature, calculating the FID score between 3, 500 generated images and all images from COCO testing set. 2) Quality Score(QS) , which aims to evaluate the authenticity of each single image. We take average of it to measure the overall quality of generated images. 3) CLIP score , evaluating the similarity between the edited region and the reference image. Specifically, we resize these two images to 224 × 224, extract the features via CLIP image encoder and calculate their cosine similarity. Higher CLIP score indicates the edited region is more similar to reference image.

QUALITATIVE ANALYSIS



QUANTITATIVE ANALYSIS

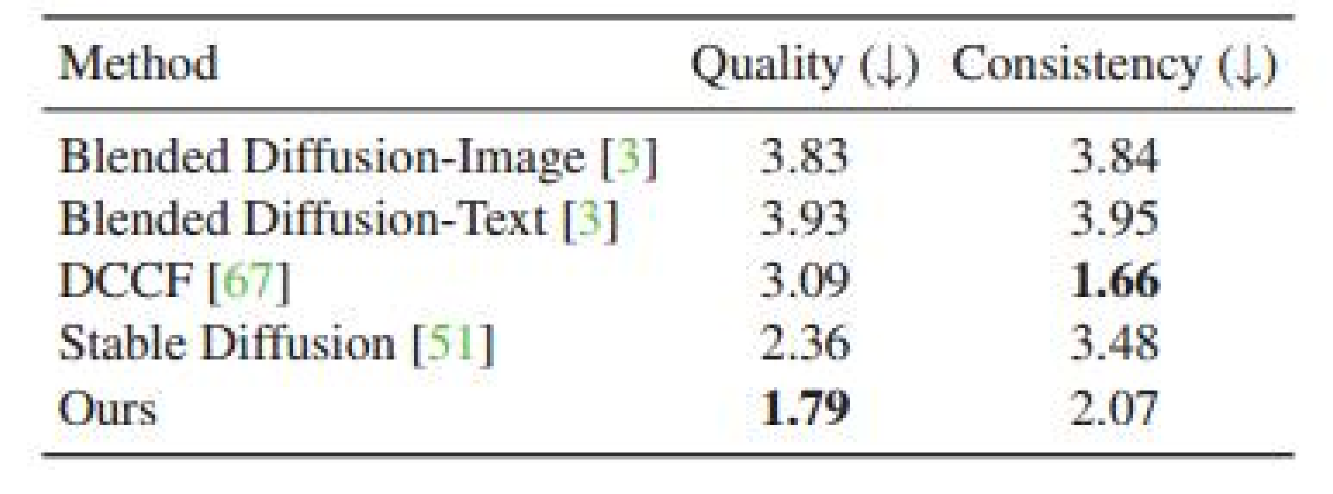
- The image-based editing method (including Blended Diffusion (image) and DCCF) reaches a high CLIP score, demonstrating that they are able to preserve the information from condition image, while the resulting image is of poor quality. The generated result from Stable Diffusion is much more plausible according to the FID and QS. However, it can hardly incorporate the conditional information of the image. Our approach achieves the best performance on all of these three metrics, verifying that it can not only generate high-quality images but also maintain the conditional information.



Method	FID (↓)	QS (↑)	CLIP Score (↑)
Blended Diffusion-Image [3]	4.60	67.14	80.65
Blended Diffusion-Text [3]	7.52	55.89	72.62
DCCF [67]	3.78	71.49	82.18
Stable Diffusion [51]	3.66	73.20	75.33
Ours	**3.18**	**77.80**	**84.97**

USER STUDY

Participants are given unlimited time to rank the score from 1 to 5 (1 is the best, 5 is the worst) on two perspectives independently: the image quality and the similarity to the reference image.



Method	Quality (↓)	Consistency (↓)
Blended Diffusion-Image [3]	3.83	3.84
Blended Diffusion-Text [3]	3.93	3.95
DCCF [67]	3.09	**1.66**
Stable Diffusion [51]	2.36	3.48
Ours	**1.79**	2.07

ABLATION STUDY

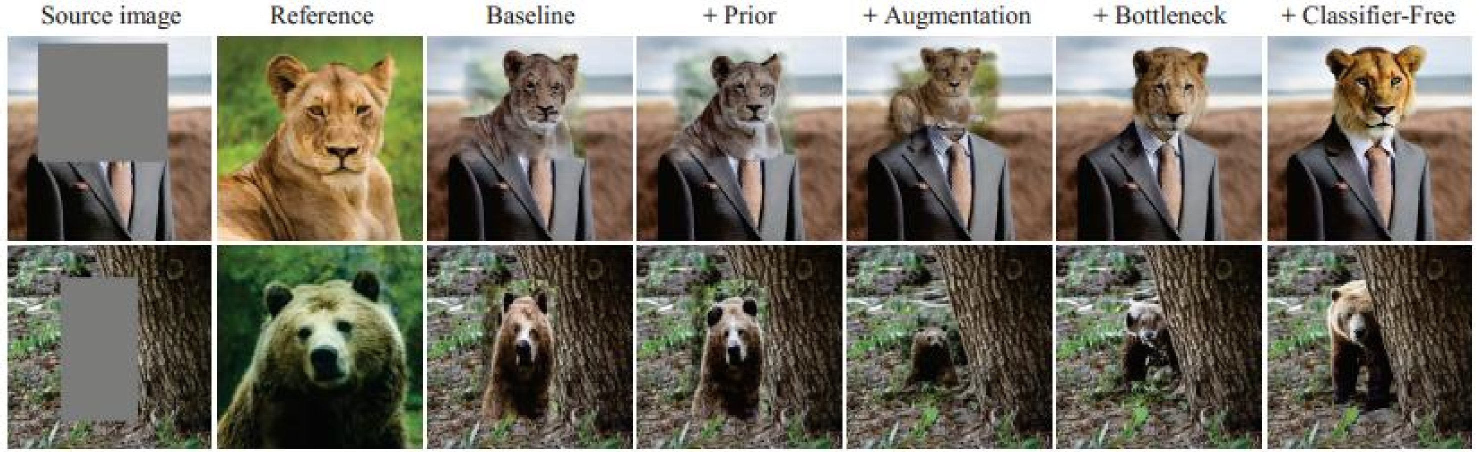


Figure 5. Visual ablation studies of individual components in our approach. We gradually eliminate the boundary artifacts through these techniques and finally achieve plausible generated results.

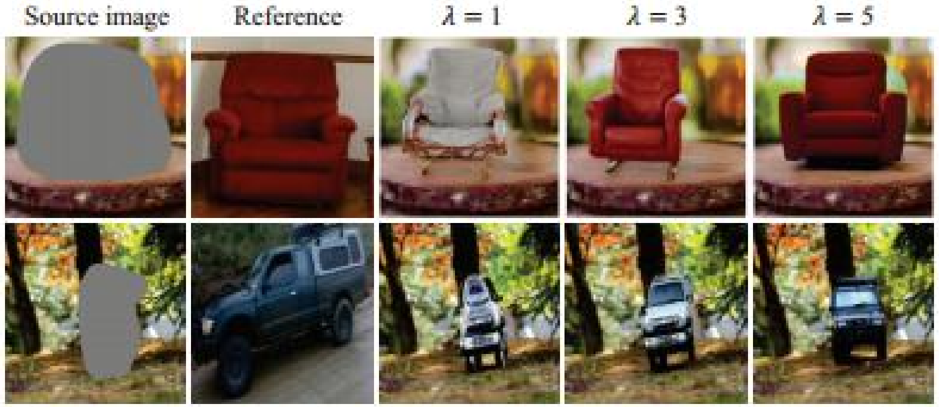
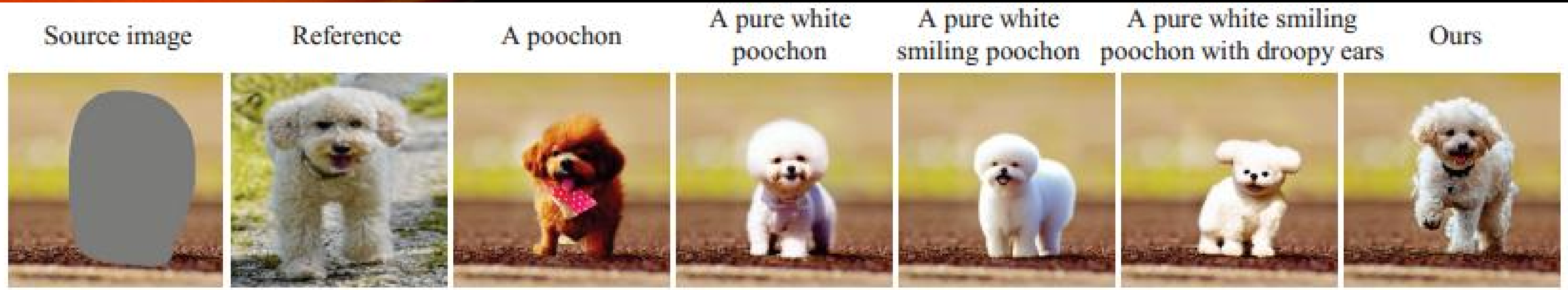


Figure 6. Effect of classifier-free guidance scale λ. A larger λ makes the generated region more similar to the reference.



Source image
Reference
A poochon
A pure white poochon
A pure white smiling poochon
A pure white smiling poochon with droopy ears
Ours

Figure 7. Comparison between progressively precise textual description and image as guidance. Using image as condition can maintain more fine-grained details.

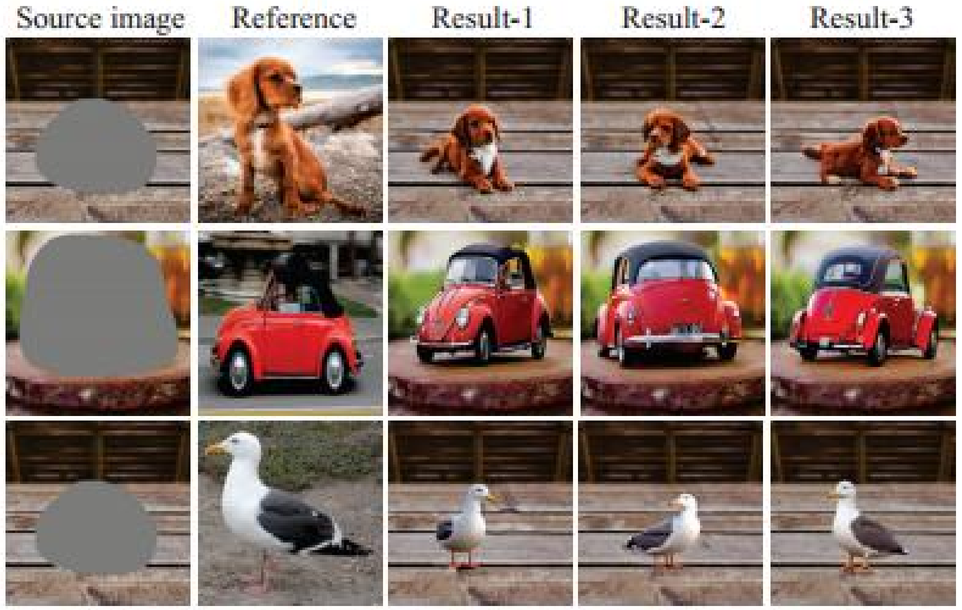


Figure 9. Our framework can synthesize realistic and diverse results from the same source image and exemplar image.