

## 14. Paint by Example: Exemplar-based Image Editing with Diffusion Models (2022.12.12 박희찬)

아카이브 : <https://arxiv.org/pdf/2211.13227v1.pdf>

Github : <https://github.com/Fantasy-Studio/Paint-by-Example>

---

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

Binxin Yang<sup>1\*</sup>      Shuyang Gu<sup>2</sup>      Bo Zhang<sup>2</sup>      Ting Zhang<sup>2</sup>      Xuejin Chen<sup>1</sup>  
Xiaoyan Sun<sup>1</sup>      Dong Chen<sup>2</sup>      Fang Wen<sup>2</sup>

<sup>1</sup> University of Science and Technology of China    <sup>2</sup> Microsoft Research Asia

---

#### REFERENCE

- Stable Diffusion <https://arxiv.org/abs/2112.10752>
- CLIP <https://openai.com/blog/clip/>
- Improved Vector Quantized Diffusion Models <https://arxiv.org/abs/2205.16007>
- classifier-free guidance <https://arxiv.org/abs/2207.12598>
- Diffusion Models Beat GANs on Image Synthesis <https://arxiv.org/abs/2105.05233>

간단한 평

처음엔 단순한 접근으로 시작하고, 문제에 직면하면서  
받아들일 수 있는 실험들을 고민하고, 많은 실험을 통해서 해결하였다.

"a picture is worth a thousand words"



**"a fall landscape with a small cottage next to a lake"**

- Stable Diffusion에서 condition은 기존에는 Text Embedding Vector를 주고 생성하도록 하였음.

language guidance는 이미지 생성을 매우 잘하고 있지만, 세부적인 컨트롤이 힘들다.

사람이 원하는 곳을 원하는 이미지가 자연스럽게 녹아들도록 하는 것을 목표로 함.

\*exemplar guidance

reference image를 source image에 합성을 해야 한다면,

모델은 다음 요소를 학습하고 생성해야 한다.

1. Understanding reference (shape, texture)
2. Transformed view of the object
3. Inpaint area around the object to generate a realistic photo
4. Reference image's resolution considering

색감, 빛을 변환시키는 image harmonization 과 비슷하지만, 완전히 다르다.

---

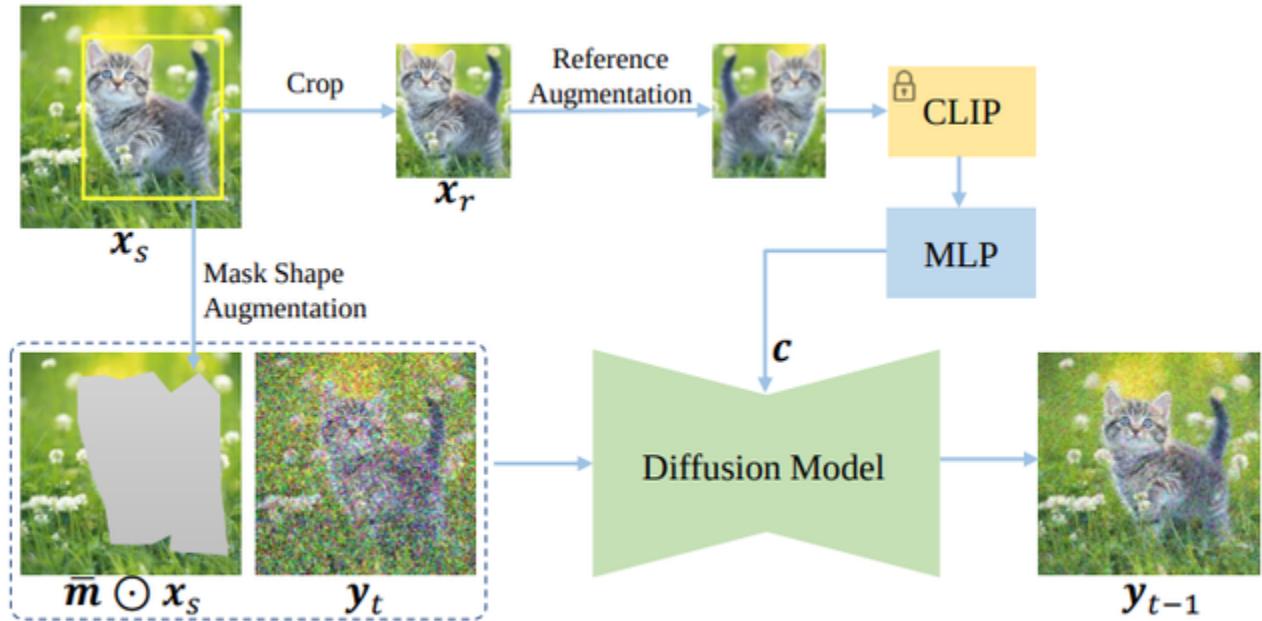


Figure 4. Our training pipeline.

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{y}_0, \epsilon} \|\epsilon_\theta(\mathbf{y}_t, \bar{\mathbf{m}} \odot \mathbf{x}_s, \mathbf{c}, t) - \epsilon\|_2^2. \quad (1)$$

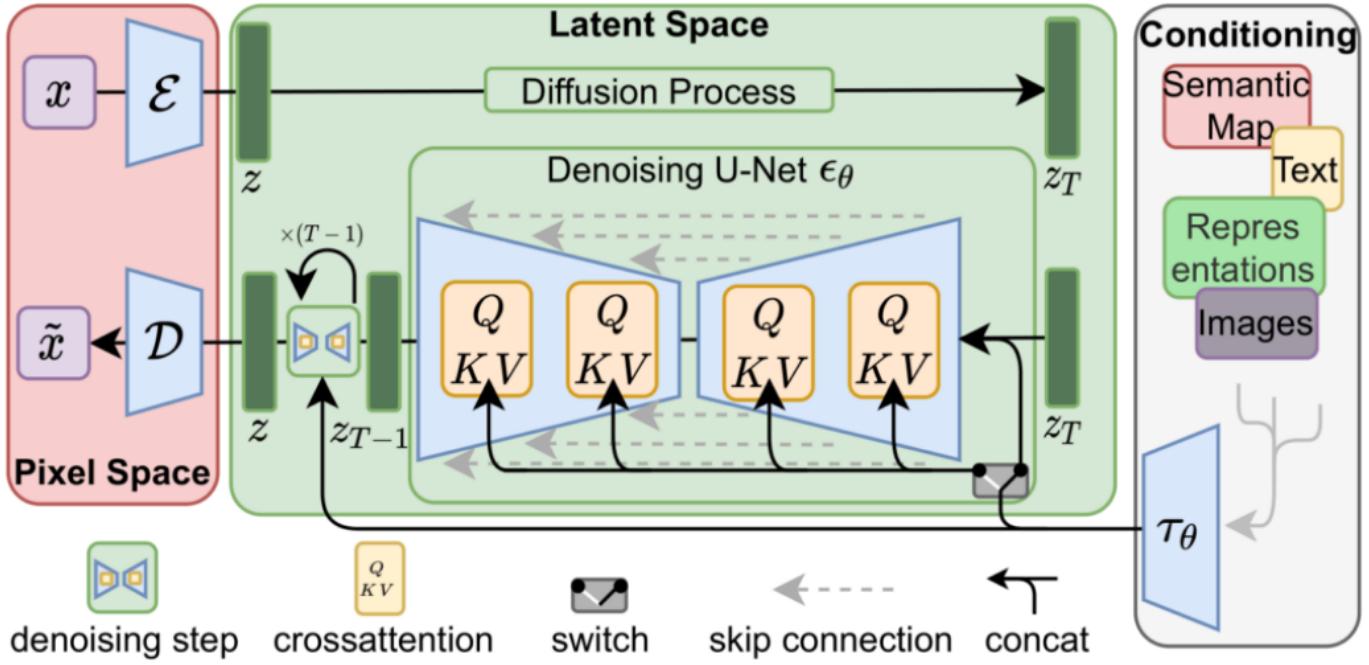
---

### 처음엔 Naive 하게

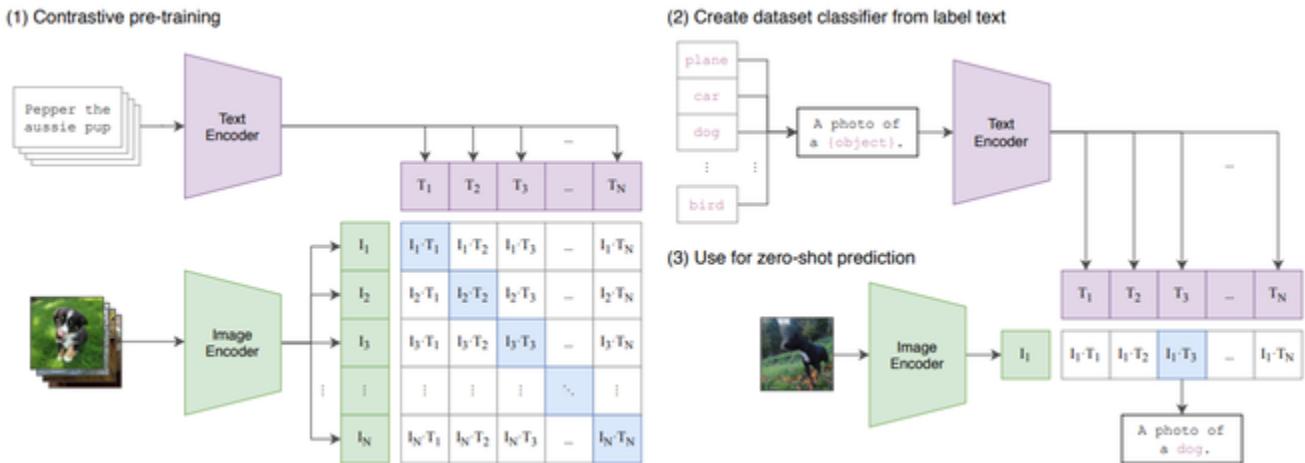
1. Stable Diffusion Model에서 학습 시킬 이미지와 이미지의 Bounding Box를 준비한다.
2. 학습 이미지는 Bounding Box 부분이 마스킹 처리된 이미지이고, Gaussian Noise를 주면서 Diffusion Model에 넣어준다.
3. Diffusion Model에 condition ( $c$ )를 넣어주면서 계속해서 원본 이미지로 복원될 수 있도록 한다.

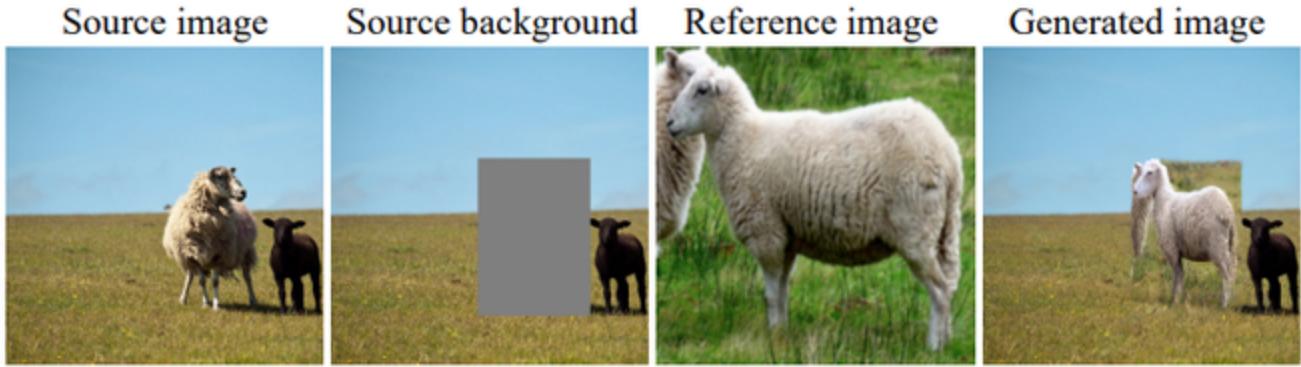
\*기존의 text embedding vector condition을 넣는 것 대신에 CLIP을 통해 image를 embedding vector로 만들어주고 있다.

---



\*CLIP 사용 이유는 Text Embedding Space와 Image Embedding Space는 서로 공유되기 때문





**Figure 3. Illustration of the copy-and-paste artifacts of the naive solution. The generated image is extremely unnatural.**

실험 결과, 배경이 전혀 지워지지 않았음

$$\bar{\mathbf{m}} \odot \mathbf{x}_s + \mathbf{x}_r = \mathbf{x}_s.$$

Diffusion Process를 진행하면서 최종 결과만 보고 생각한다면, 모델은 결국 해당 함수를 만족하기 위해 학습이 진행되기 때문에 모델이 일반화되지 않는 것 뿐 아니라 reference image 자체를 이해하지 못하고, Source Image 와 Reference Image 사이의 연결 점을 못 찾을 것이라고 판단

## 추가 실험

### 1. Stable Diffusion에서 Pretrained Weight를 가져오자. (Image Prior )

처음부터 학습 시키는 것보다, pretrained model를 이용하면,

Embedding Vector에 담긴 Context를 더 잘 활용할 것이라고 생각하고,

CLIP 으로 나온 Image Embedding Vector는 Text Embedding Vector와 공유하므로, 잘 맞는 조합이 될 것이라고 예상

### 2. Embedding Vector 가 문제일까? (Information Bottleneck)

기존 Stable Diffusion 에서 Text Embedding Vector 를 학습한다는 건,

자연어 처리 관점에서 해당 Object 에 대한 Context 정보를 파악하는 과정이라고 생각하면 됨.

하지만, Image Embedding Vector는 Context 파악할 필요 없이 단순히 Copy and Paste 되는 느낌으로 Object 를 기억한다고 주장

Embedding Vector 에 학습 가능한 FC Layer (MLP)를 추가하여 Embedding Vector 내에서 어느 정도 Context를 고려하도록 한다.

### 3. reference에 한번, mask에 한번 (Strong Augmentation)

Augmentation 을 넣어 Self-Supervised Learning 의 문제를 해결한다.

self-supervised learning의 문제로 train - test 간 domain gap의 2가지 mismatch가 발생한다.

#### (1) Reference Image Augmentation

train data는 source image 와 reference image가 같은 이미지에서 나오기 때문에, 두 이미지 사이의 연결 점을 약하게 만들기 위해 다양한 Augmentation을 reference image에 극단적으로 적용 시킨다.

#### (2) Mask shape augmentation

mask image는 train 단계에서 reference image에 해당하는 Bounding Box 를 이용하므로, Semantic 한 정보를 잘 추출 할 수 있도록 Augmentation을 적용 해준다.

#### (3) Control the mask shape

mask augmentation 을 통하여, Inference 단계에서 어떤 mask가 들어와도 잘 컨트롤 할 수 있도록 해준다.

#### (4) Control the similarity degree

edited area 와 reference image의 similarity degree를 조절하기 위해, classifier-free guidance 전략을 사용.

$$\begin{aligned} & \log p(\mathbf{y}_t | \mathbf{c}) + (s - 1) \log p(\mathbf{c} | \mathbf{y}_t) \\ & \propto \log p(\mathbf{y}_t) + s(\log p(\mathbf{y}_t | \mathbf{c}) - \log p(\mathbf{y}_t)), \end{aligned} \quad (4)$$

$$\tilde{\epsilon}_{\theta}(\mathbf{y}_t, \mathbf{c}) = \epsilon_{\theta}(\mathbf{y}_t, \mathbf{v}) + s(\epsilon_{\theta}(\mathbf{y}_t, \mathbf{c}) - \epsilon_{\theta}(\mathbf{y}_t, \mathbf{v})).$$

기존 classifier-free guidance 는 v 대신에 null 값을 이용해 unconditional score function을 구했는데,

본 논문에서는 learnable parameter로 대체하여 사용.

\*conditional diffusion model

diffusion model 을 내가 원하는 것을 생성 할 수 있도록, condition을 reverse process에 주는 diffusion model

classifier를 이용한 sampling 전략을 채용하여 이미지 생성 시 높은 퀄리티의 이미지를 얻고 있음

$$q(\mathbf{x}_{1:T} | \mathbf{x}_0) := \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

$$\mathbf{x}_t \approx \mathbf{x}_{t-1} - \frac{\beta(t)\Delta t}{2} \mathbf{x}_{t-1} + \sqrt{\beta(t)\Delta t} \mathcal{N}(\mathbf{0}, \mathbf{I})$$



$$d\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)} d\omega_t$$

Anderson 1982  
**Reverse Generative Diffusion SDE:**

$$d\mathbf{x}_t = \underbrace{\left[ -\frac{1}{2}\beta(t)\mathbf{x}_t - \beta(t)\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) \right]}_{\text{"Score Function"}} dt + \sqrt{\beta(t)} d\bar{\omega}_t$$

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x} | \mathbf{y})] dt + g(t) d\mathbf{w}$$

Conditional Diffusion Model 에서는 Diffusion SDE 식에서 Score Function에 condition을 추가로 입력

gradient log p (x|y)  $\rightarrow$  gradient log p(x) - gradient log p (y | x) 로 변환

gradient log p(x)로 condition 없이 Diffusion model 로 학습하면서,

gradient  $\log p(y|x)$ 에서 학습된 condition을 이용

#### \*classifier guidance

GAN Model 들은 Generator가 Discriminator를 속이도록 학습만 하면, IS / FID 같은 평가 지표에서 좋은 성능을 보여주는데(이미지 퀄리티 상승).. 이를 위해, Diversity 와 Fidelity의 Trade-off에서 Fidelity를 중점으로 학습하였음. Conditional Diffusion Model에서 GAN처럼 Fidelity에 더욱 집중

prior  $p(x|y)$  가 있다면, posterior  $p(y|x)$  도 있다고 가정.

prior probability  $p(x|y)$  를 maximize 하는게, conditional diffusion model의 목표라 할 때,

대부분의 논문에서 posterior probability  $p(y|x)$  가 있다는 가정을 찾길 실패하거나, 무시하고 있음

(posterior issue)

---

**Algorithm 1** Classifier guided diffusion sampling, given a diffusion model  $(\mu_\theta(x_t), \Sigma_\theta(x_t))$ , classifier  $p_\phi(y|x_t)$ , and gradient scale  $s$ .

---

```
Input: class label  $y$ , gradient scale  $s$ 
 $x_T \leftarrow$  sample from  $\mathcal{N}(0, \mathbf{I})$ 
for all  $t$  from  $T$  to 1 do
     $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$ 
     $x_{t-1} \leftarrow$  sample from  $\mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$ 
end for
return  $x_0$ 
```

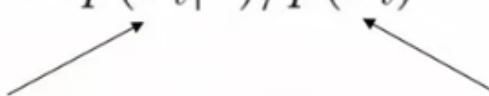
---

해당 논문에서도 지적한 내용으로, classifier는 따로 추가 학습이 필요하며 classifier guidance 를 넣는 건 디퓨전 프로세스 (p)에서 adversarial sample이 생성될 가능성성이 높아서 pre-trained DDPM을 가져오는 것을 권장

#### \*classifier-free guidance

classifier guidance 는 샘플의 퀄리티는 높지만 (Fidelity), 다양성 (Diversity) 가 부족한 문제가 있음.

$$p(\mathbf{c}|\mathbf{x}_t) \propto p(\mathbf{x}_t|\mathbf{c})/p(\mathbf{x}_t)$$

  
Conditional diffusion model      Unconditional diffusion model

$$\nabla_x \log p_\gamma(x | y) = \nabla_x \log p(x) + \gamma \nabla_x \log p(y | x)$$

$$\tilde{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1+w)\epsilon_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\epsilon_\theta(\mathbf{z}_\lambda)$$

classifier-free guidance 기법은 classifier 없어도, conditional image sampling을 잘 수행한다.

\*paint by example 과 다른 점은,  $e(z, \text{null})$  로 입력하여 unconditional score를 구하게 함



Figure 1: Classifier-free guidance on the malamute class for a 64x64 ImageNet diffusion model. Left to right: increasing amounts of classifier-free guidance, starting from non-guided samples on the left.

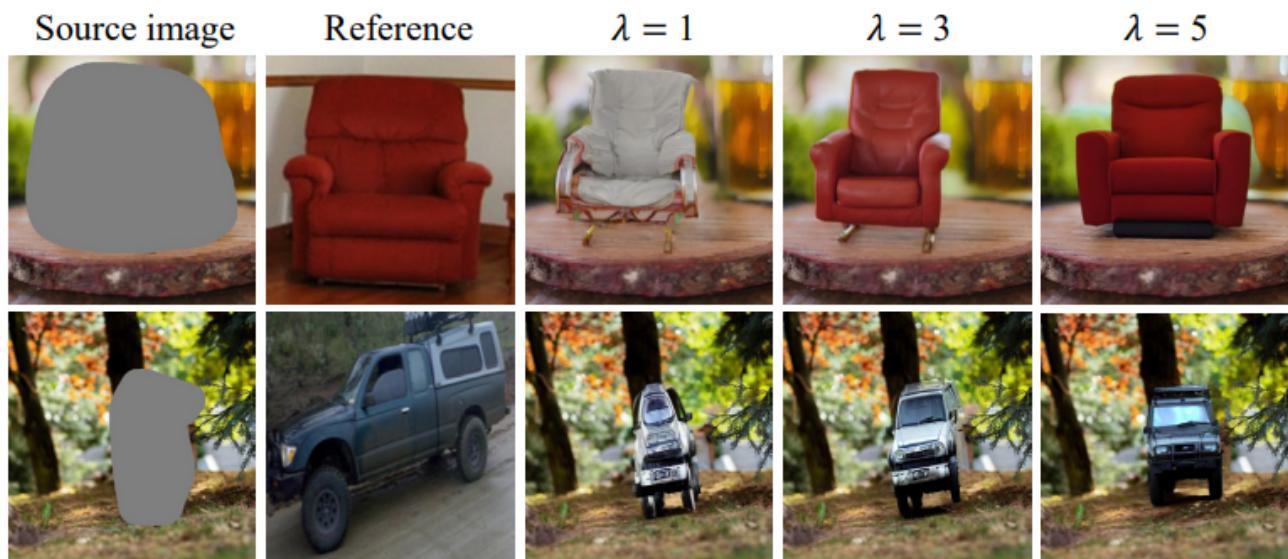


Figure 7. Effect of classifier-free guidance scale  $\lambda$ . A larger  $\lambda$  makes the generated region more similar to the reference.

---

#### 실험 스펙

OpenImages 데이터셋 (총 16 million 이미지중, 1.9 million 이미지 사용 600 classes)

64 x V100 GPU

512x512 이미지로 40 epochs, 7일동안

---

#### 실험 결과

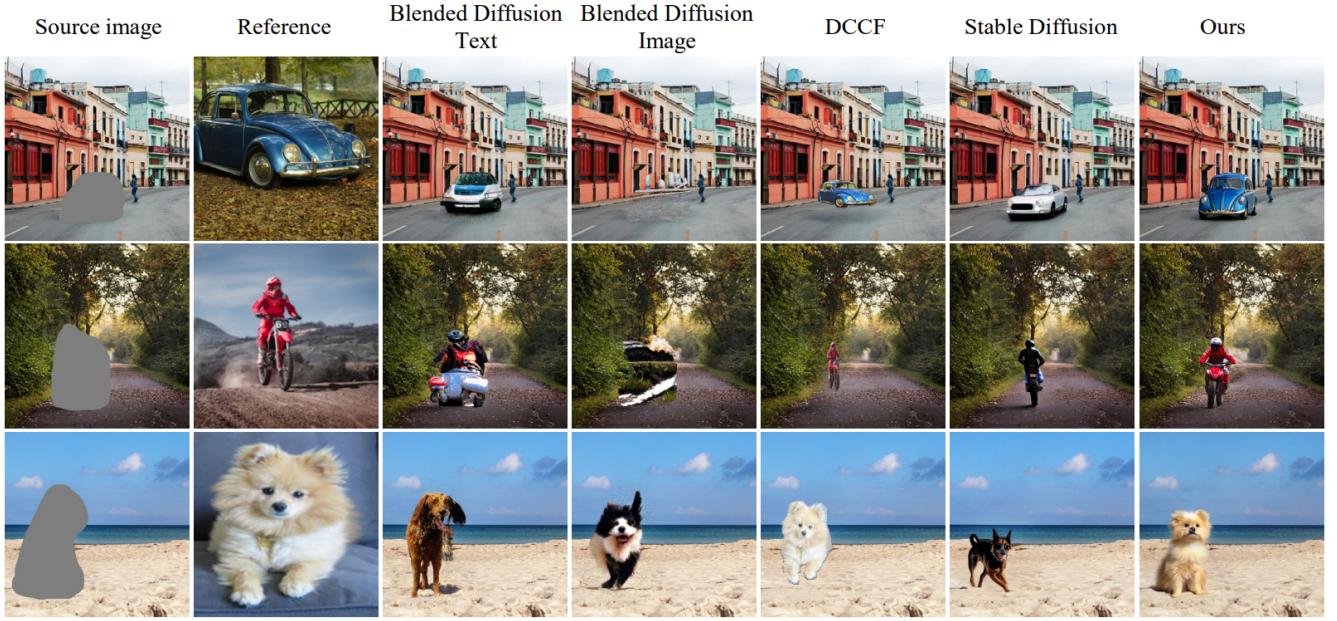


Figure 5. Qualitative comparison with other approaches. Our method can generate results that are semantically consistent with the input reference images in high perceptual quality.

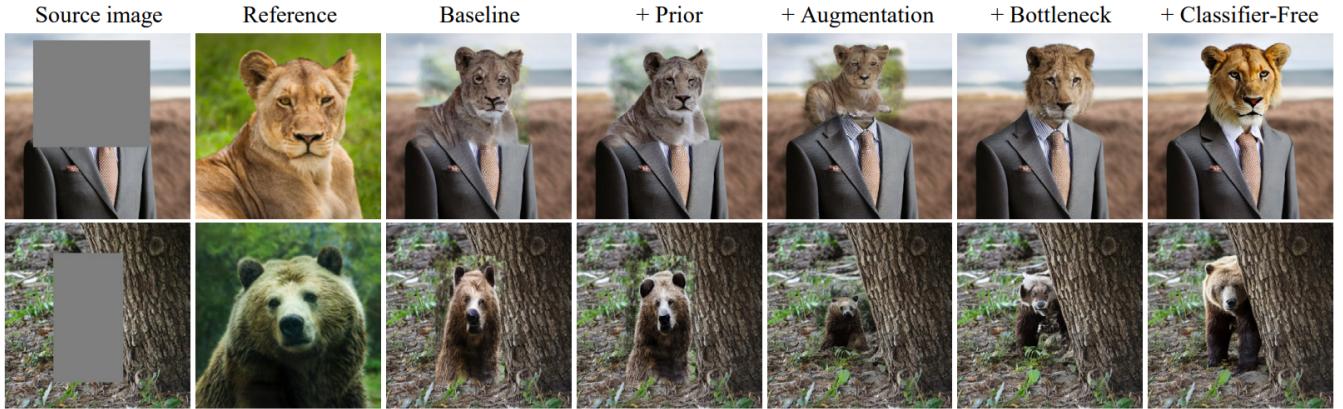


Figure 6. Visual ablation studies of individual components in our approach. We gradually eliminate the boundary artifacts through these techniques and finally achieve plausible generated results.

Table 3. Quantitative comparison of different variants of our method. We achieve the best performance by leveraging all these techniques.

Method	FID ( $\downarrow$ )	QS ( $\uparrow$ )	CLIP Score ( $\uparrow$ )
Baseline	3.61	76.71	85.90
+ Prior	3.40	77.63	<b>88.79</b>
+ Augmentation	3.44	76.86	81.68
+ Bottleneck	3.26	76.62	81.41
+ Classifier-Free	<b>3.18</b>	<b>77.80</b>	84.97

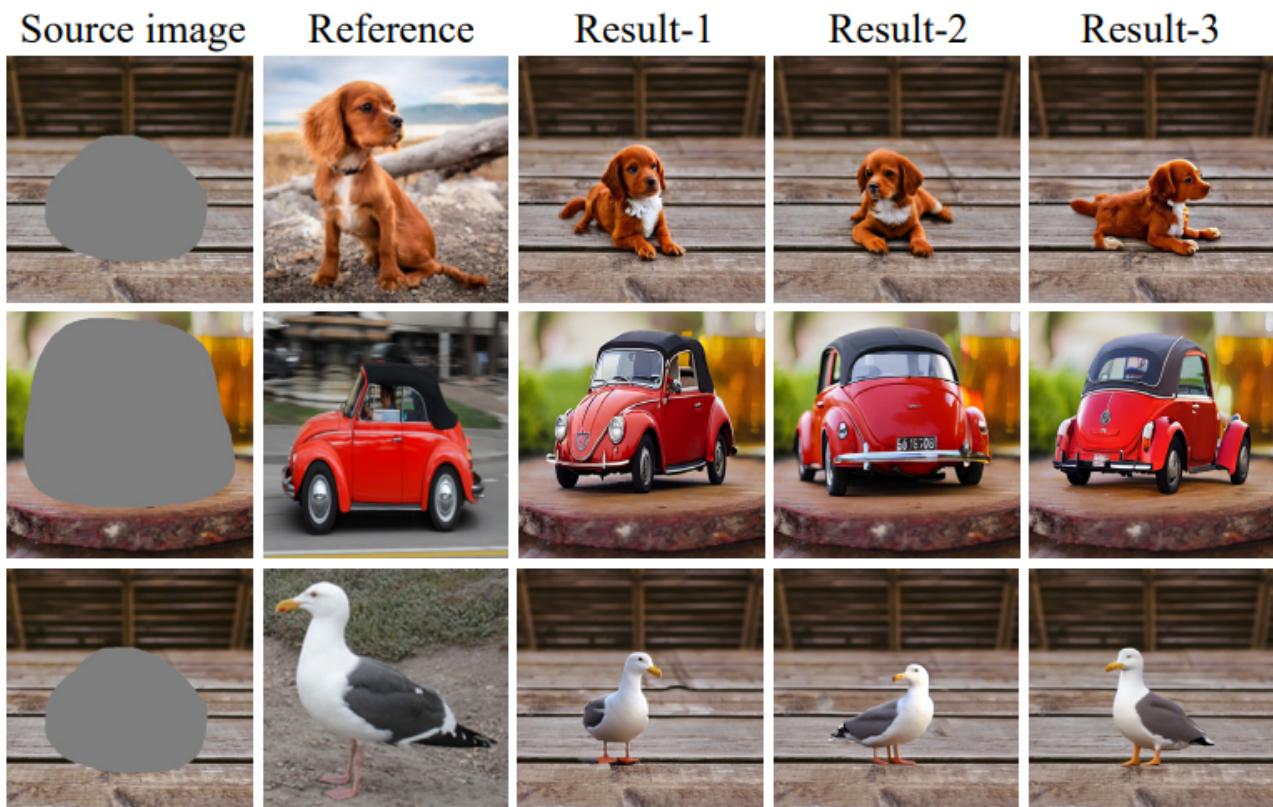


Figure 10. Our framework can synthesize realistic and diverse results from the same source image and exemplar image.



Figure 8. Comparison between progressively precise textual description and image as guidance. Using image as condition can maintain more fine-grained details.

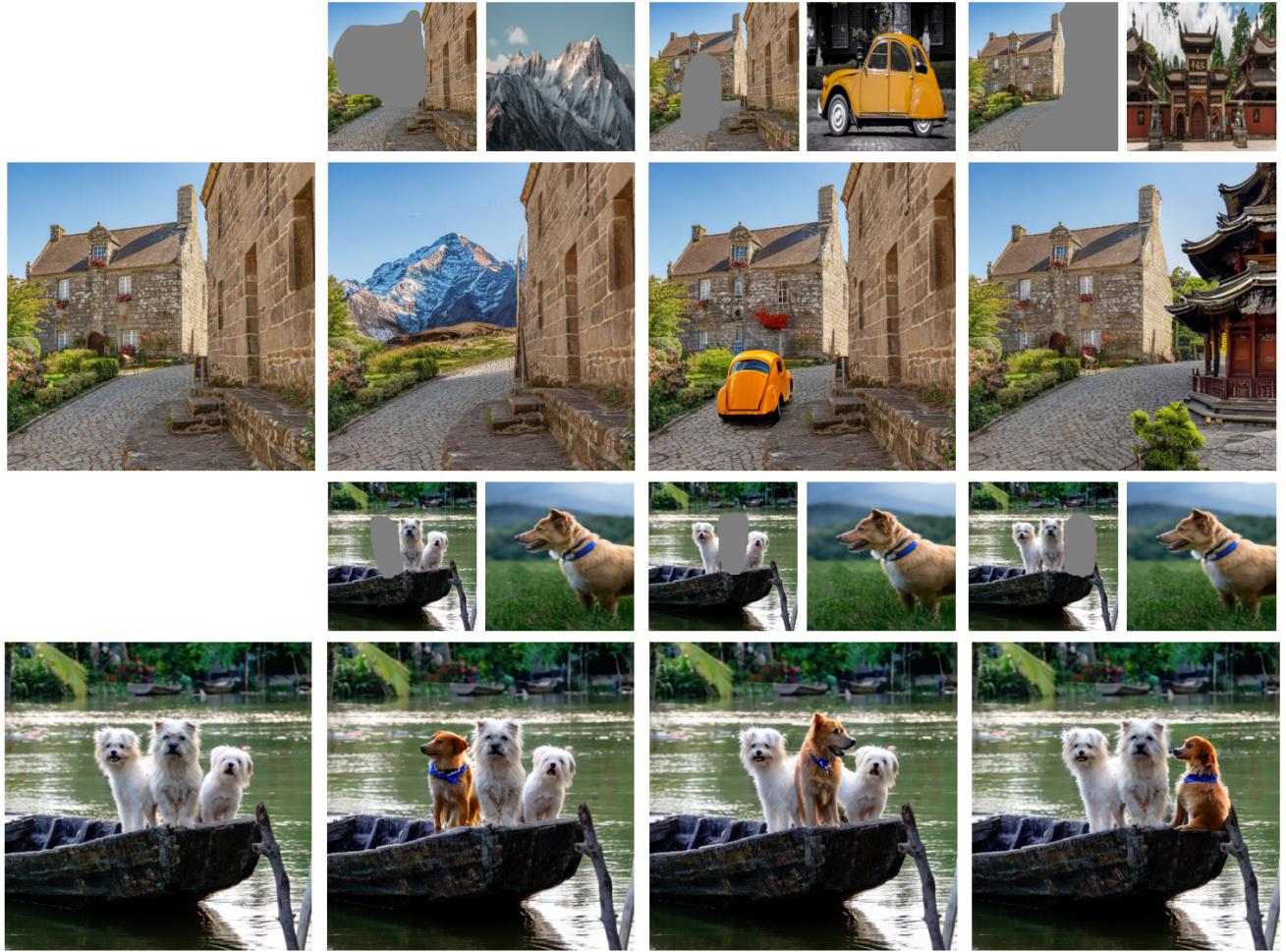


Figure 11. Our method enables the user to edit different regions in the same source images.



Figure 12. Results of the same object with different source images. Our method is robust for different objects or different source images, even for some complicated objects, like 'Big Ben'.

## Appendix C. Limitation

Because the majority of the training data is natural photos, our method does not perform well with some artificial images, such as oil paintings. Furthermore, for some rarer objects like dinosaur, our method can hardly understand them well. We present some failure cases in Fig. 13.

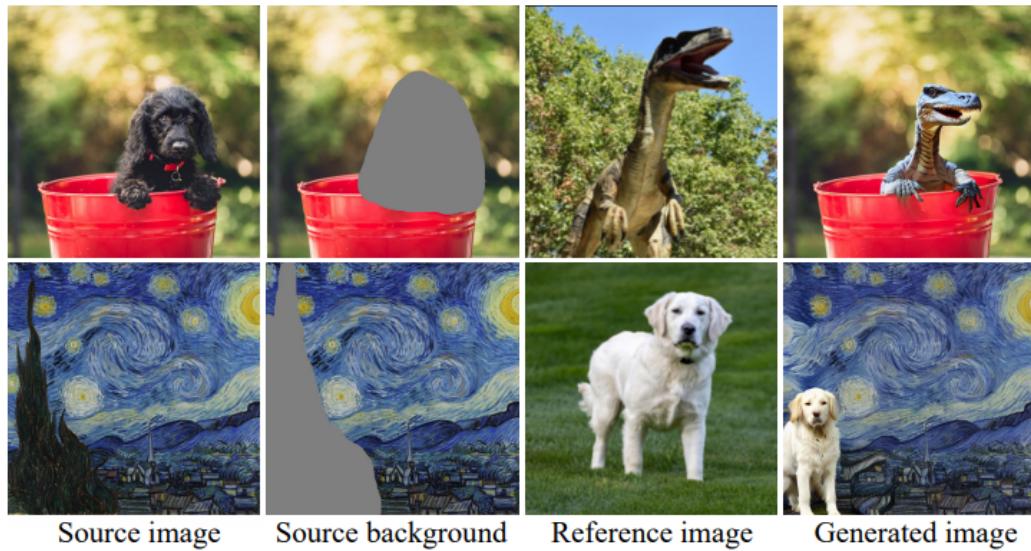


Figure 13. Some failure cases.