

Imagic: Text-Based Real Image Editing with Diffusion Models

Bahjat Kawar*^{1,2}
Huiwen Chang¹
¹Google Research

Shiran Zada*¹
Tali Dekel^{1,3}
²Technion

Oran Lang¹
Inbar Mosseri¹
³Weizmann Institute of Science

Omer Tov¹
Michal Irani^{1,3}

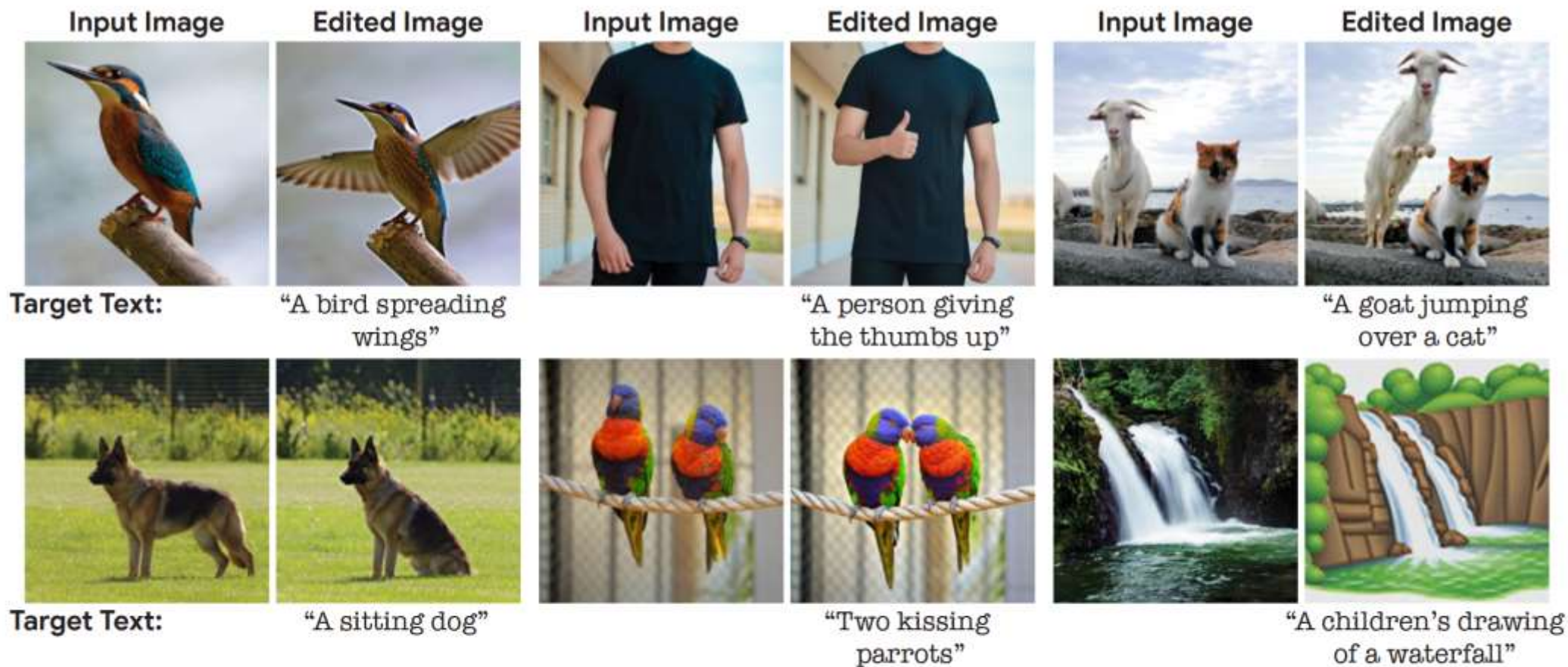
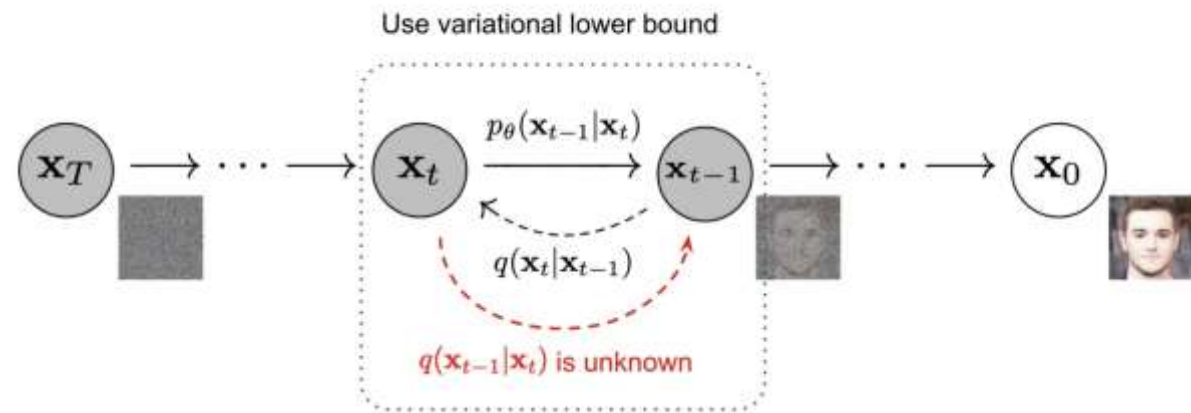
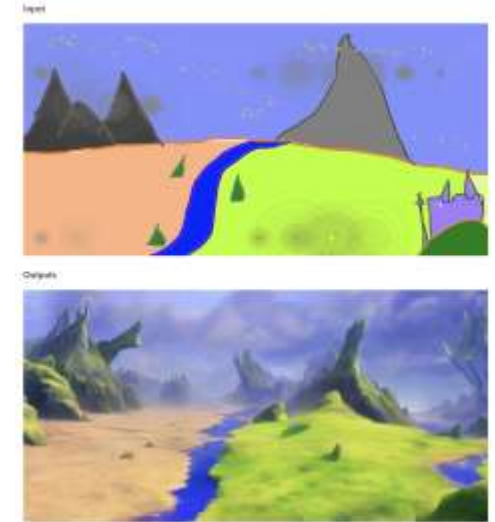
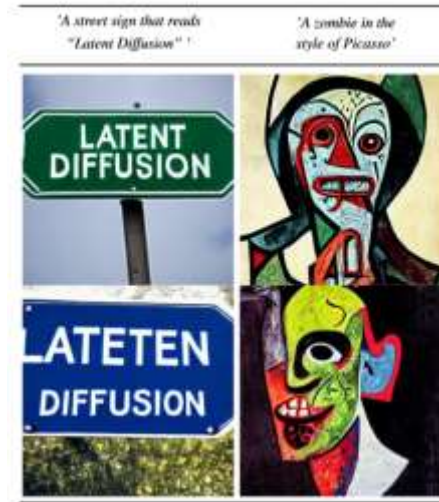


Figure 1. *Imagic* – Editing a single real image. Our method can perform various text-based semantic edits on a single real input image, including highly complex non-rigid changes such as posture changes and editing multiple objects. Here, we show pairs of 1024×1024 input (real) images, and edited outputs with their respective target texts.

Backgrounds

- Diffusion Models
 - Diffusion model for image generation
 - Text-to-Image
 - Image-to-Image
 - Diffusion Process
 - Forward process
 - Used for training
 - Parameters free
 - Reverse process

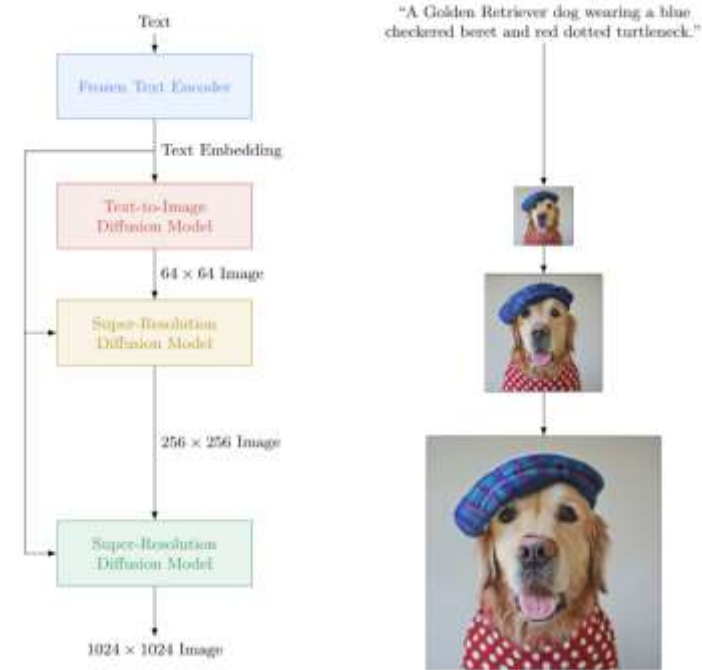


Backgrounds

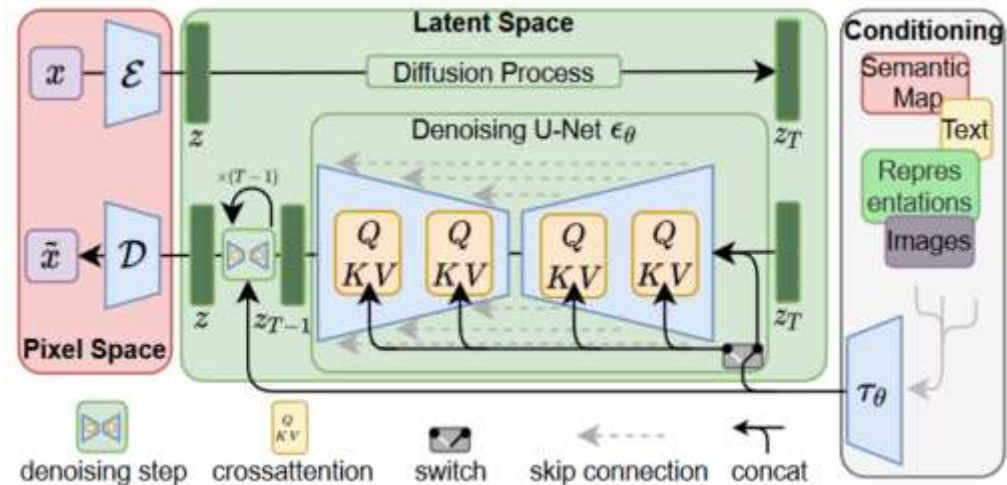
- Diffusion Models
 - Noise prediction
 - U-Net
 - $L_{DM} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_{\theta}(x_t, t)\|_2^2]$
 - Modeling conditional distribution
 - U-Net: Res-block + Attention module
 - Condition encoder(Text encoder)

Backgrounds

- Diffusion Models
 - Image-level diffusion
 - Imagen



- Latent space diffusion
 - Stable Diffusion



Introduction

- Settings
 - Input a source image and a prompt which describe the wanted changes on source image
 - Output target images



"A zebra."



"A horse with a saddle"



"A horse with its head down"



"A brown horse in a grass field"



"A cartoon of a horse"



"A pistachio cake"



"A chocolate cake"



"A strawberry cake"



"A wedding cake"



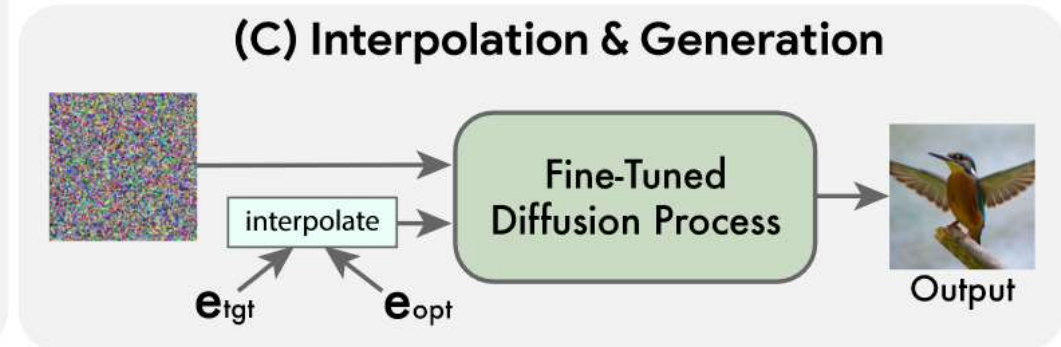
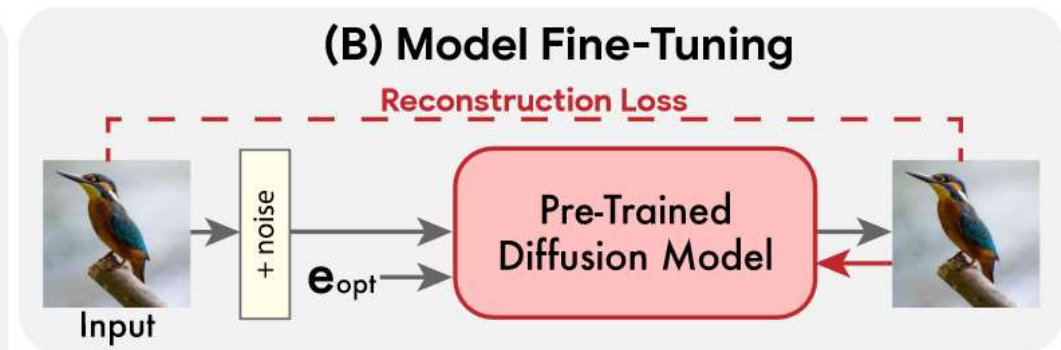
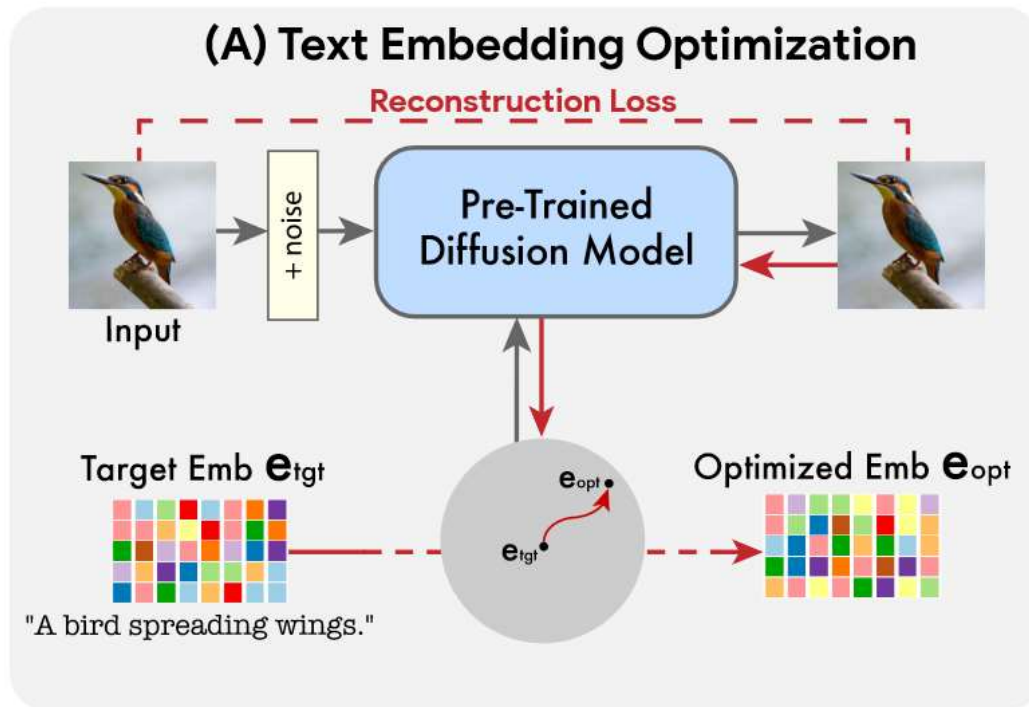
"A slice of cake"

Introduction

- Fast, cheap(relatively) **single image-text pair finetuning** methodology for Generative Image Editing
 - Two versions: Imagen \ Stable Diffusion (v1)
 - Time and resources for finetuning a single image
 - Imagen: 8 minutes on two TPU V4
 - Stable Diffusion: 7 minutes on one A100
- Balance text alignment with fidelity
 - Maintaining foreground and background contexts
 - Achieve purpose modification

Methodology

- Three procedures for image editing
 - Maintain fidelity
 - (A) Text Embedding Optimization
 - (B) Model Fine-Tuning
 - Balance text alignment with fidelity
 - (C) Interpolation and Generation

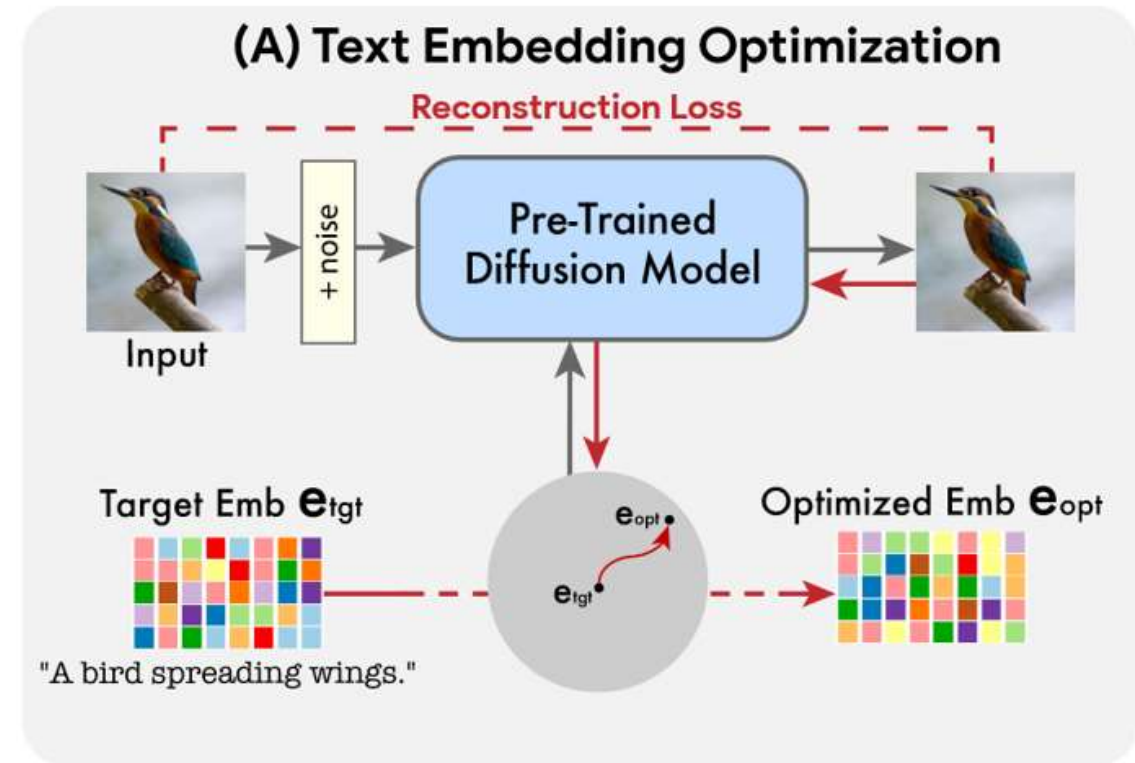


Methodology

- Text Embedding Optimization
 - Introduced T5[1] text encoder to get **target** text embedding $\mathbf{e}_{tgt} \in \mathbb{R}^{T \times d}$
 - Diffusion model is frozen
 - Using reconstruction objective to optimize text embedding

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t, \epsilon} \left[\|\epsilon - f_{\theta}(\mathbf{x}_t, t, \mathbf{e})\|_2^2 \right]$$

- Produces optimized embedding \mathbf{e}_{opt}

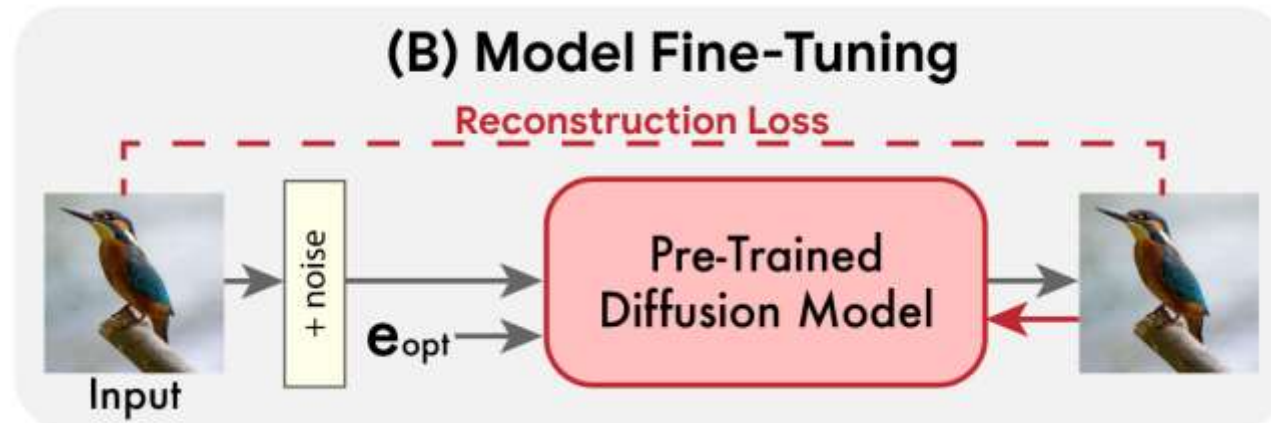


Methodology

- Model Fine-Tuning

- Further help the model generate high fidelity editing results
- Using \mathbf{e}_{opt} as text embedding to finetune Diffusion Model with the objective of:

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t, \epsilon} \left[\|\epsilon - f_{\theta}(\mathbf{x}_t, t, \mathbf{e})\|_2^2 \right]$$



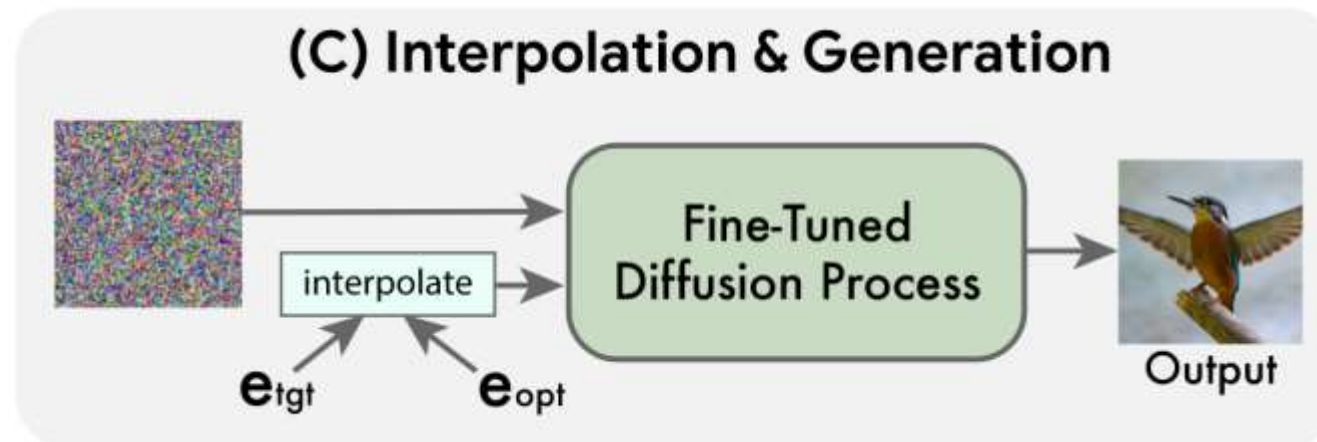
Methodology

- Interpolation and Generation

- Simple linear interpolation between a and b can achieve a result that takes into account fidelity and text alignment

$$\bar{\mathbf{e}} = \eta \cdot \mathbf{e}_{tgt} + (1 - \eta) \cdot \mathbf{e}_{opt}$$

- Generation process follow normal txt2img pipeline, using fine-tuned diffusion model.



Methodology

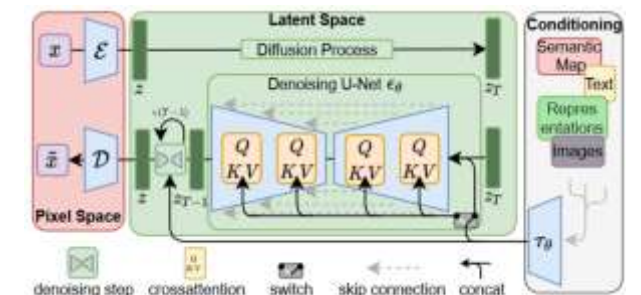
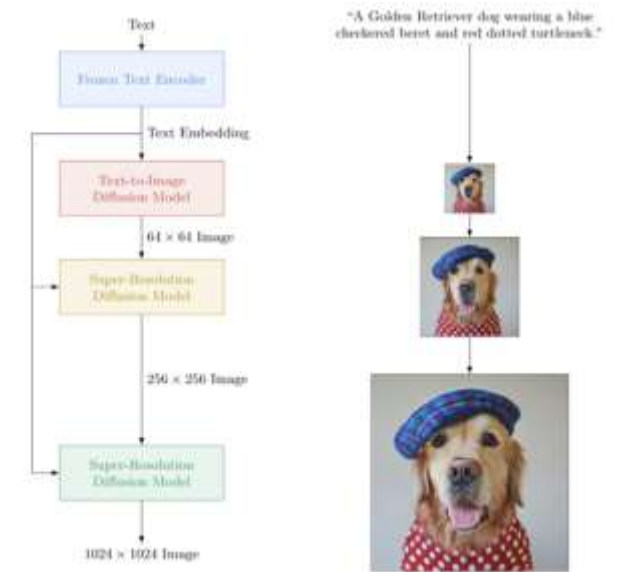
- Implementation Details

- Imagen

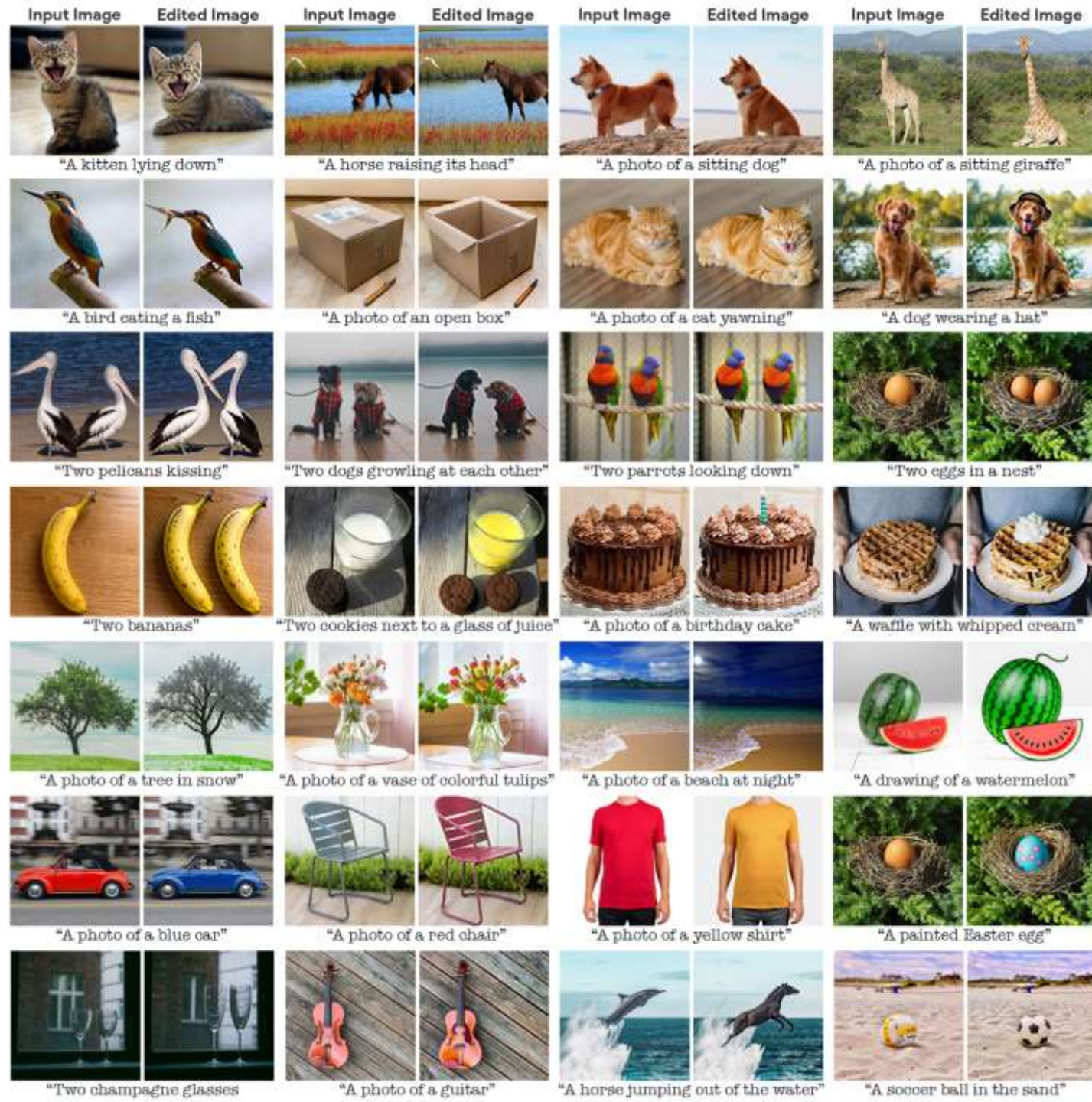
- Optimize text embedding on 64x64 for 100 steps (Adam 1e-3)
 - Finetuning diffusion model on 64x64 for 1500 steps
 - Finetuning on 64x64 → 256x256 SR diffusion for 1500 steps
 - Finetuning on 64x64 → 256x256 adds little effect, thus directly use pretrained model

- Stable Diffusion

- Optimize text embedding on latent space diffusion for 1000 steps (Adam 2e-3)
 - Finetuning for 1500 steps (Adam 5e-7)



Experiments



Experiments

- User study
 - TEdBench
 - a novel collection of 100 pairs of input images and target texts describing a desired complex non-rigid edit.

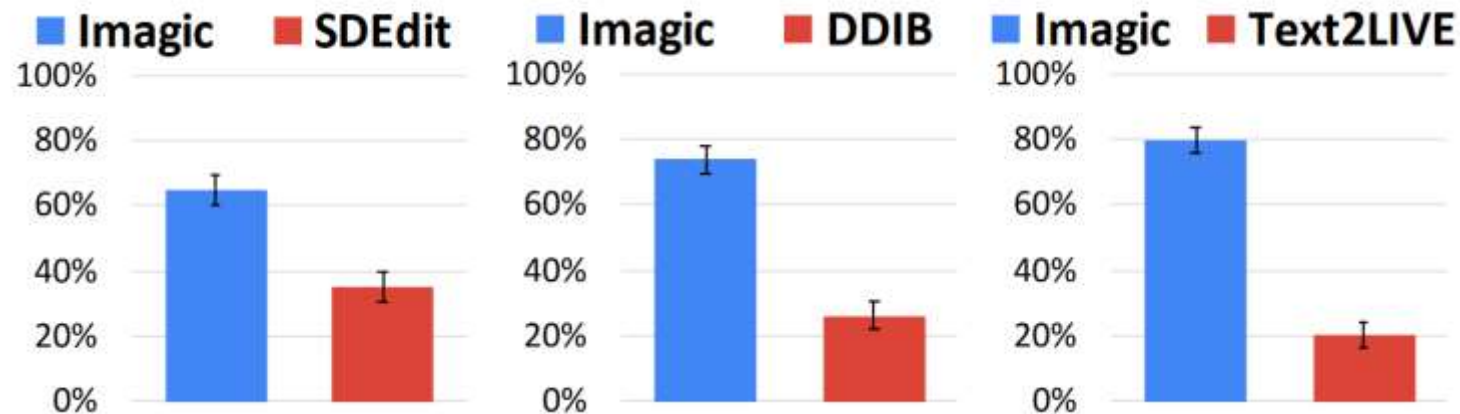


Figure 8. **User study results.** *Preference rates (with 95% confidence intervals) for image editing quality of Imagic over SDEdit [35], DDIB [59], and Text2LIVE [7].*

Ablation Study

- Text Embedding Optimization



Ablation Study

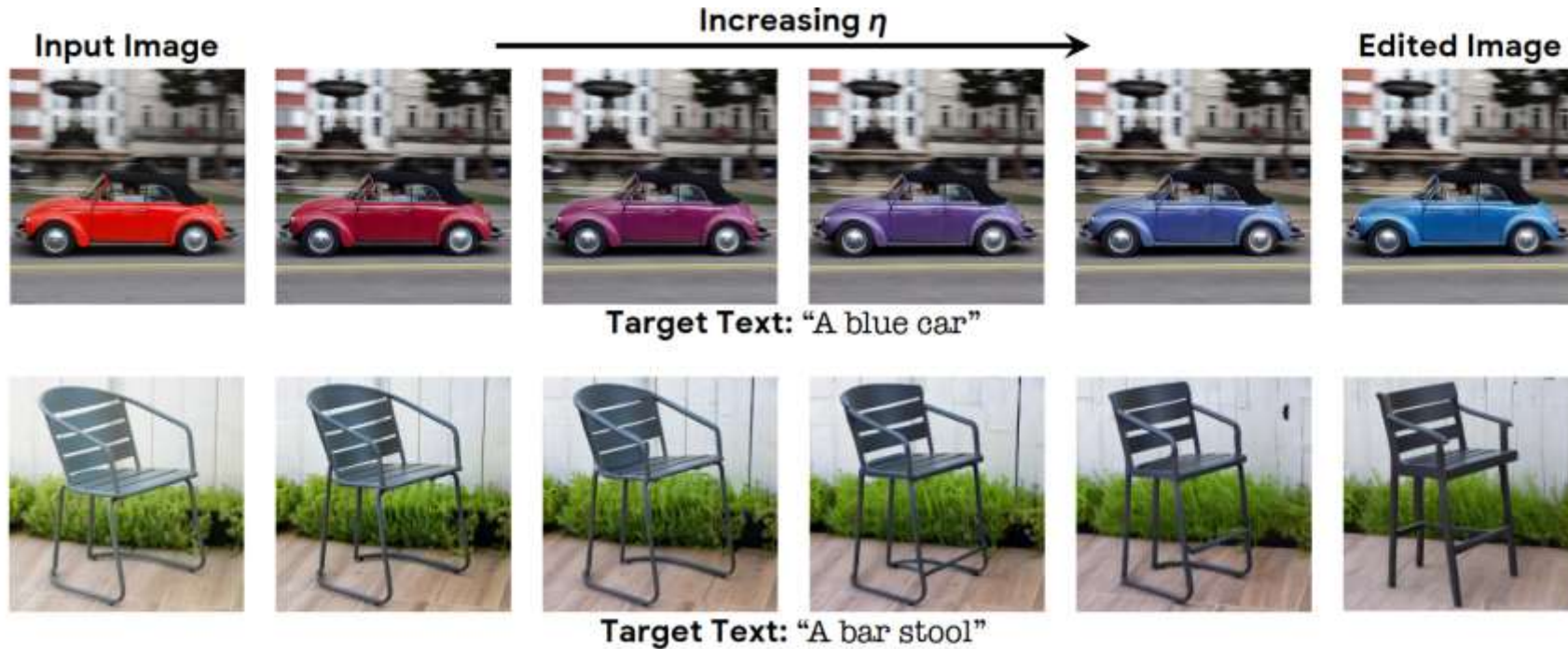
- Model Fine-Tuning



Figure 7. **Embedding interpolation.** Varying η with the same seed, using the pre-trained (top) and fine-tuned (bottom) models.

Ablation Study

- Interpolation and Generation



Ablation Study

- Interpolation and Generation
 - When using different random seeds and different samples, users need to make small adjustments to the value of η to get the best results
 - Select the value range of η by two evaluation indicators

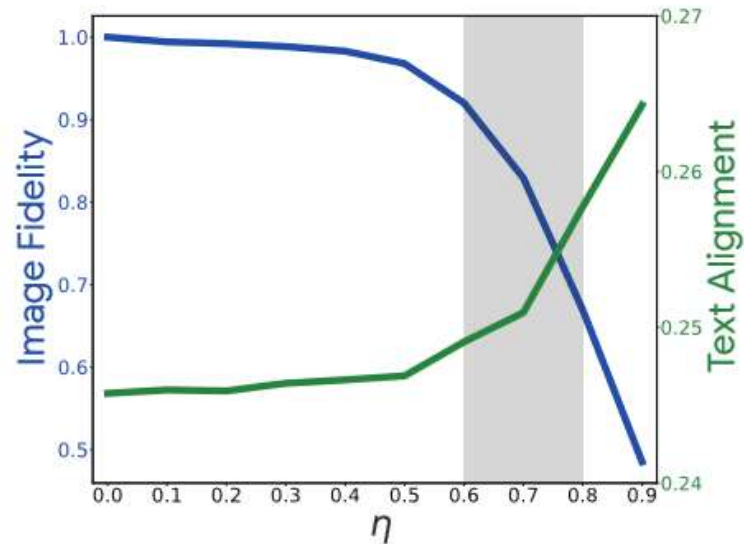


Figure 9. **Editability–fidelity tradeoff.** *CLIP score (target text alignment) and $1-LPIPS$ (input image fidelity) as functions of η , averaged over 150 inputs. Edited images tend to match both the input image and text in the highlighted area.*

- η in $[0.6, 0.8]$

Limitations

- Subtle cases
- Affects extrinsic image details

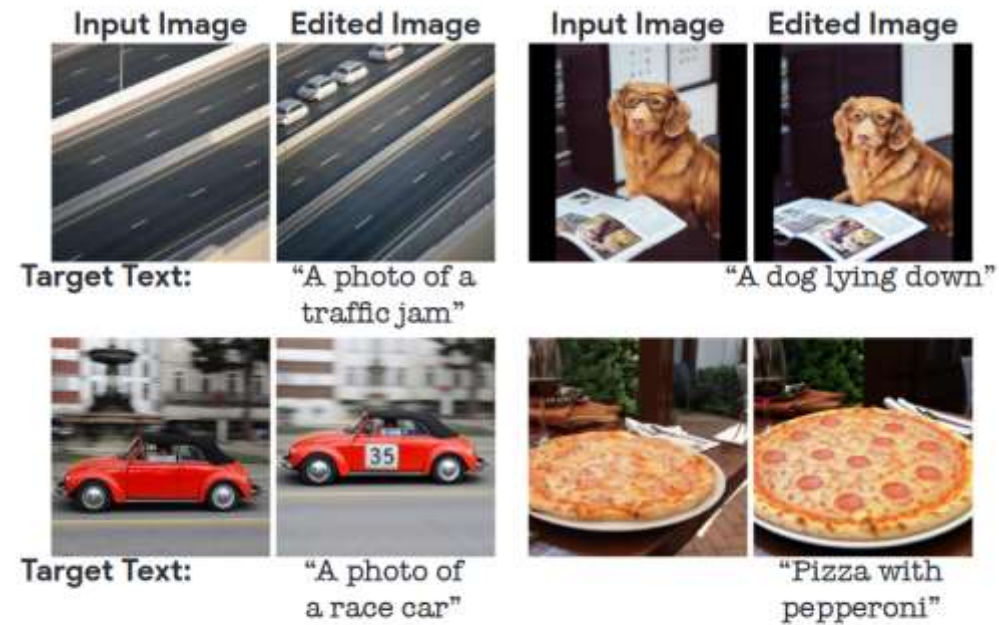


Figure 10. **Failure cases.** *Insufficient consistency with the target text (top), or changes in camera viewing angle (bottom).*