AI6126 Literature Review - Prompt-to-Prompt Image Editing with Cross Attention Control

Group Members:	
Sean Goh Ann Ray G2202190G sean0057@e.ntu.edu.sg	Teng Guang Way G2102434F teng0141@e.ntu.edu.sg

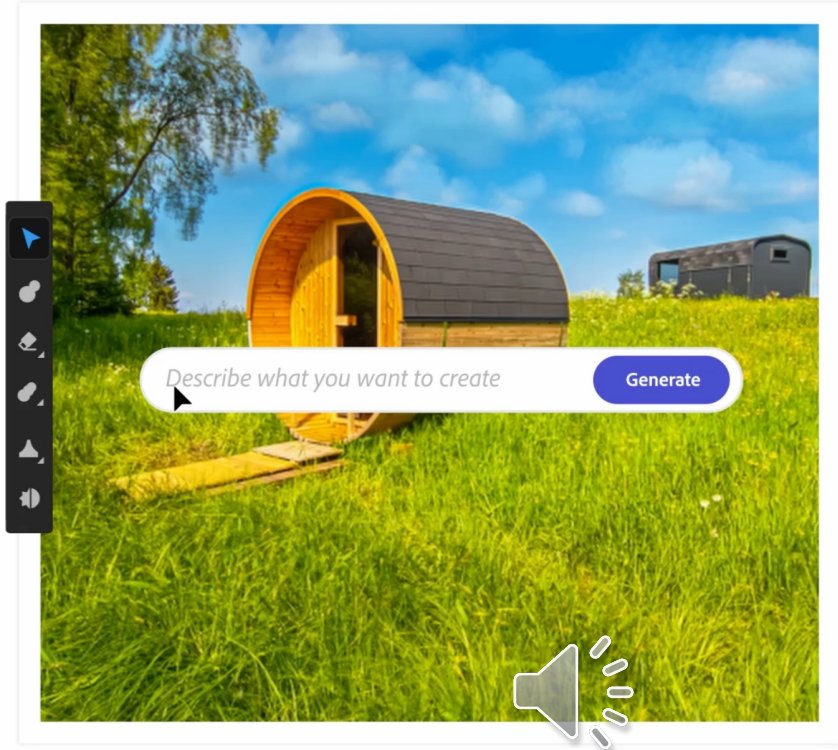
Table of Content

- What is Image Editing
- Paper’s Motivation
- Related Work
- Proposed Method
- Results
- Summary

What is Image Editing in AI

- Conventional Methods (Adobe Photoshop, etc.)
  - Manually adjust pixel values based on observed patterns
  - Macros for repetitive tasks
    - Color balancing, sharpening, etc.
  - Competent image editing skills required
  - (Few could use)

- Shift to AI implementation → Firefly (Adobe)



Paper's Motivation

- Modern Large-scale Language-Image (LLI) models have powerful generative semantic capability, **producing** images from text prompts.
  - Imagen, DALL-E 2, Parti

- However, these LLI models only uses text input as guidance, and provides **little controllability** to the user
  - Adding the adjective 'white' to 'dog' would end up producing an image very different in shape
  - Such problem might be fixed by having the user to provide a mask to control specific spatial region to edit, but hampers quick and intuitive text driven application

- Author's solution: Provide an intuitive and powerful image editing method that supports **localized edit controllability with textual input only**.

Related Work (**GAN** Based)

- Text Driven Image Manipulation using GANs
- Revolutionary works combining GANs and CLIP
- Use Mask to restrict change to specific spatial region
- Development of VQ-GAN trained over diverse data

VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance

Katherine Crowson*[1], Stella Biderman*[1,2], Daniel Kornis[3], Dashiell Stander[1], Eric Hallahan[1], Louis Castricato[1,4], and Edward Raff[2]

[1] EleutherAI
[2] Booz Allen Hamilton
[3] AIDock
[4] Georgia Institute of Technology

Abstract Generating and editing images from open domain text prompts is a challenging task that heretofore has required expensive and specially trained models. We demonstrate a novel methodology for both tasks which is capable of producing images of high visual quality from text prompts of significant semantic complexity without any training by using a multimodal encoder to guide image generations. We demonstrate on a variety of tasks how using CLIP [40] to guide VQGAN [12] produces higher visual quality outputs than prior, less flexible approaches like minDALL-E [20], GLIDE [36] and Open-Edit [26], despite not being trained for the tasks presented. Our code is available in a public repository.

Keywords: generative adversarial networks; grounded language; image manipulation

1 Introduction

Using free-form text to generate or manipulate high-quality images is a challenging task, requiring a grounded learning between visual and textual representations. Manipulating images in an open domain context was first proposed by the seminal Open-Edit [26], which allowed text prompts to alter an image's content. This was done mostly with semantically simple transformations (e.g., turn a red apple green), and does not allow generation of images. Soon after DALL-E [41] and GLIDE [36] were developed, both of which can perform generation (and inpainting) from arbitrary text prompts, but do not themselves enable image manipulation.

In this work we propose the first a unified approach to semantic image generation and editing, leveraging a pretrained joint image-text encoder [40] to steer an image generative model [12]. Our methodology works by using the multimodal encoder to define a loss function evaluating the similarity of a (text, image) pair and backpropagating to the latent space of the image generator. We iteratively

* Co-first authors

Related Work (**Diffusion** Based)

- State of the art generation quality on highly diverse dataset with Diffusion Models
- Text2Live (Bar-Tal et al.) proposed text based very-basic localized image editing without mask techniques (Such as image texture)
- Numerous works and emergence of Imagen, DALL-E2, however minimal control capability

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh* OpenAI aramesh@openai.com

Prafulla Dhariwal* OpenAI prafulla@openai.com

Alex Nichol* OpenAI alex@openai.com

Casey Chu* OpenAI casey@openai.com

Mark Chen OpenAI mark@openai.com

Abstract

Contrastive models like CLIP have been shown to learn robust representations of images that capture both semantics and style. To leverage these representations for image generation, we propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. We show that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity. Our decoders conditioned on image representations can also produce variations of an image that preserve both its semantics and style, while varying the non-essential details absent from the image representation. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion. We use diffusion models for the decoder and experiment with both autoregressive and diffusion models for the prior, finding that the latter are computationally more efficient and produce higher-quality samples.

Related Work that support Localize Image Editing - Comparison

Localized Image Editing Attributes	GAN Based Image Editing Models	Diffusion Based Image Editing Models
Requires User to provide mask	Paint By Word, VQGAN-CLIP	Blended Diffusion (Avrahami et al.)
Does not requires user to provide mask	CLIP2StyleGAN	DALL-E2, Imagen, Text2Live

Proposed Method – Goal

Proposed Method

Fixed attention maps and random seed

Fixed random seed

Previous Works

Proposed Method – Imagen

- **Imagen** text-guided synthesis model as backbone
- Inject cross-attention maps during diffusion process

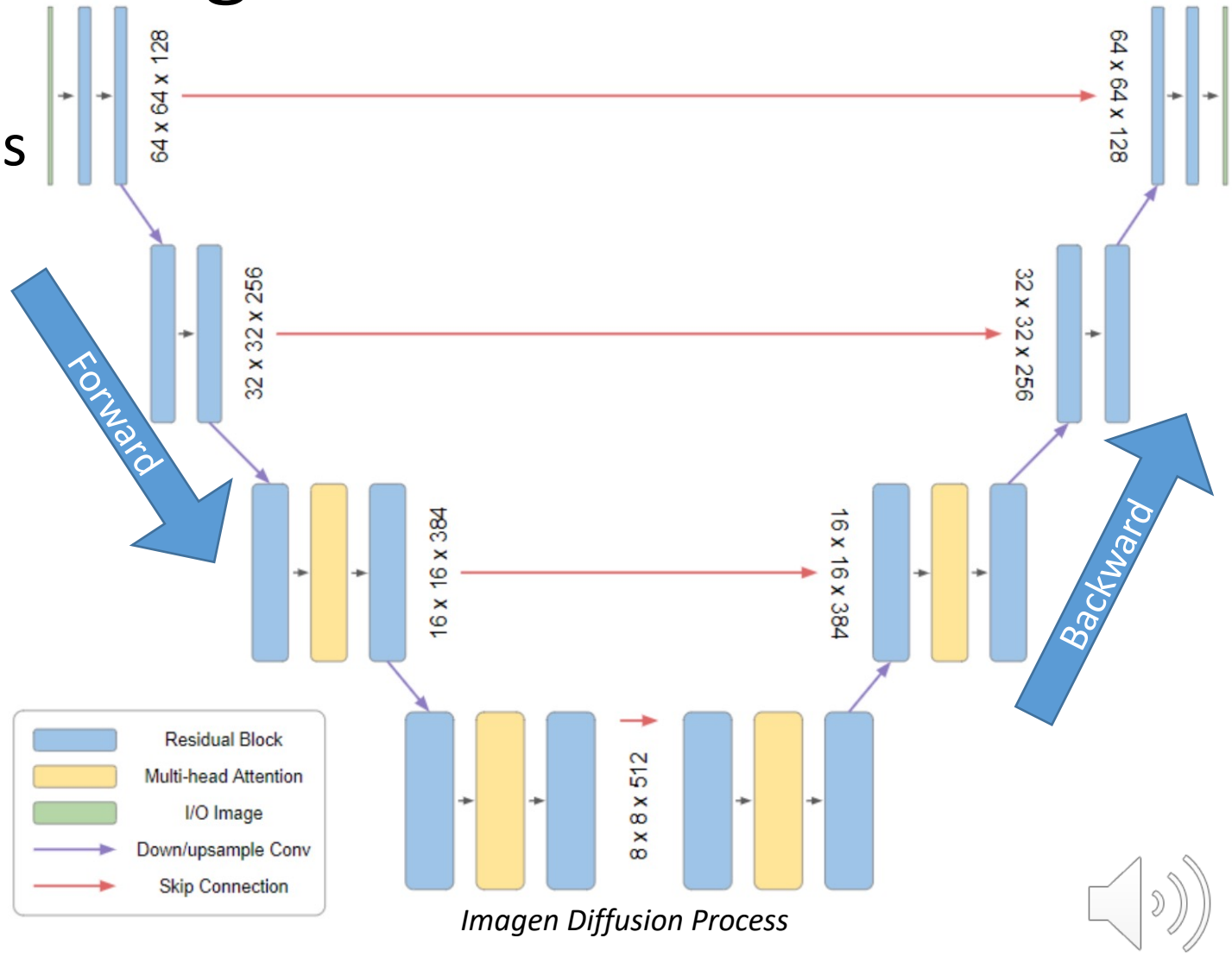
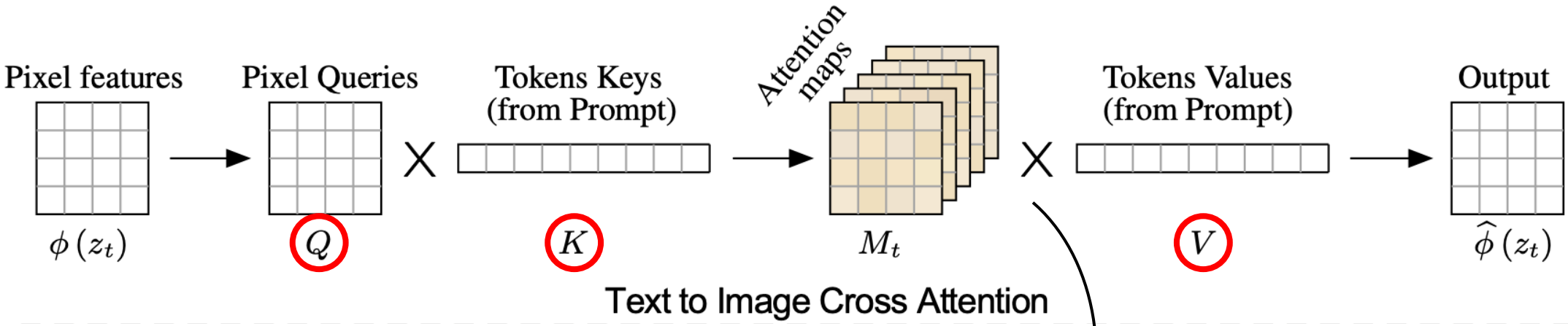


Imagen Diffusion Process

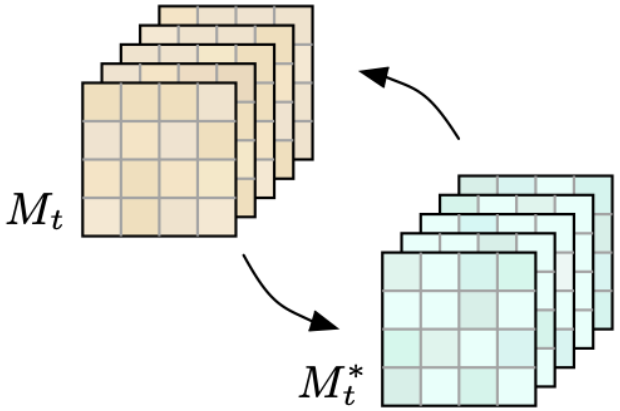
Proposed Method – Theory



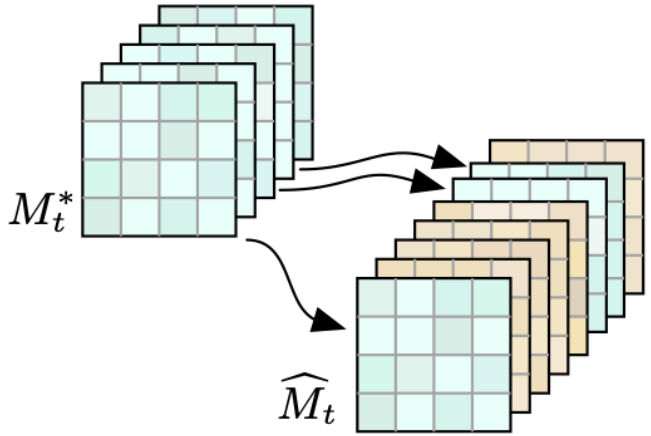
$$M = Softmax(\frac{QK^T}{\sqrt{d}})$$

Proposed Method – Theory

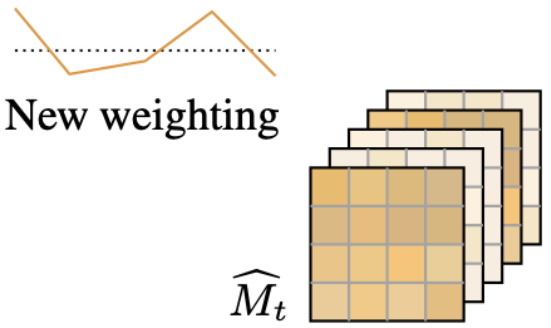
- By controlling these cross attention maps, different image editing tasks can be made by modifying the text prompt
  - Change local/global style of image (via word/phrase swap/addition/removal)
  - Amplify/attenuate semantic (via attention re-weighting)



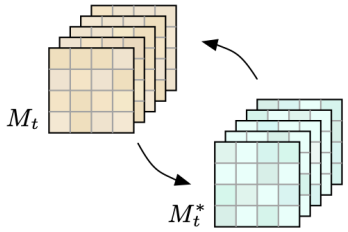
Word Swap



Adding a New Phrase



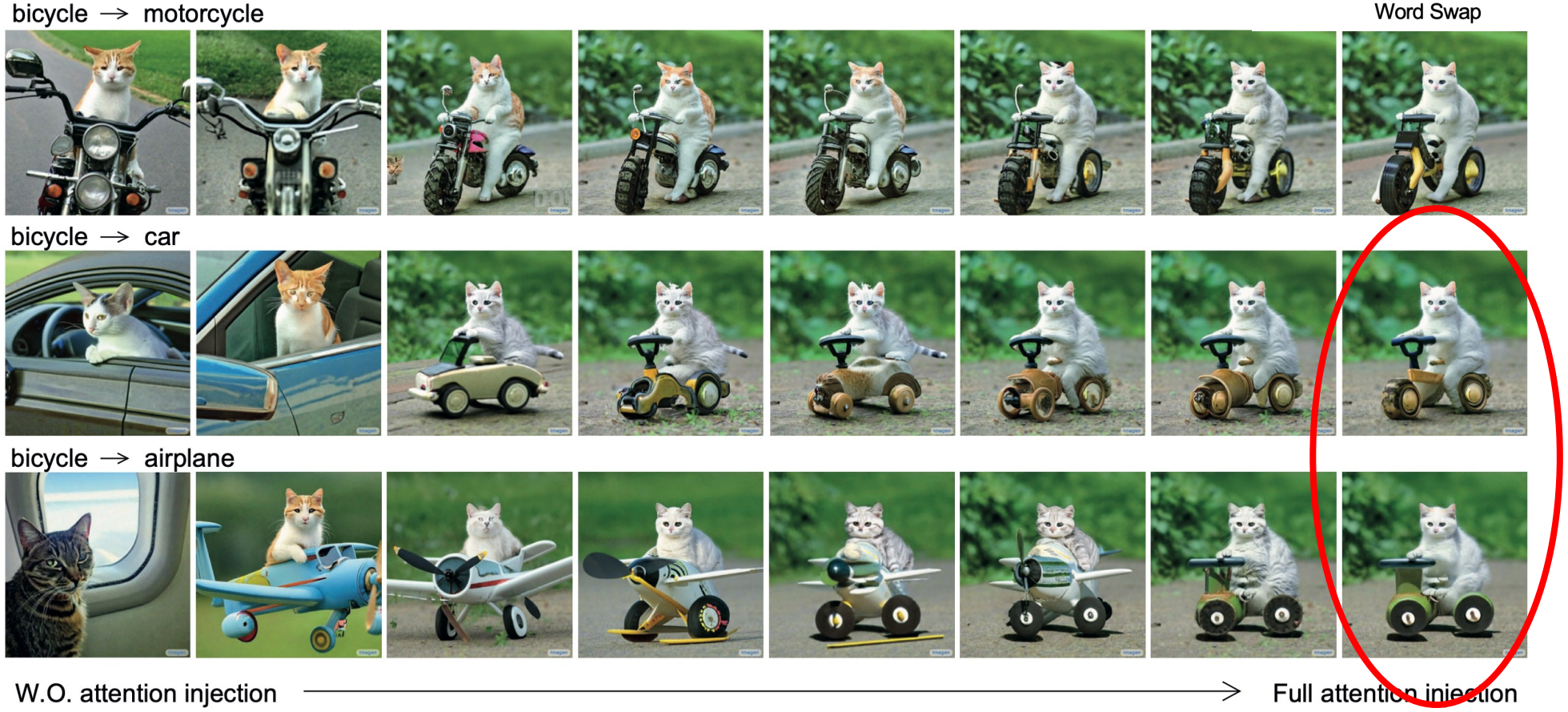
Attention Re–weighting



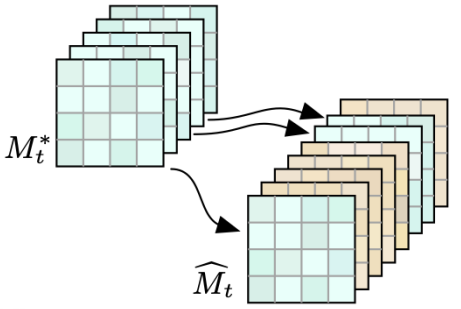
Results – Word Swap

Source Image:

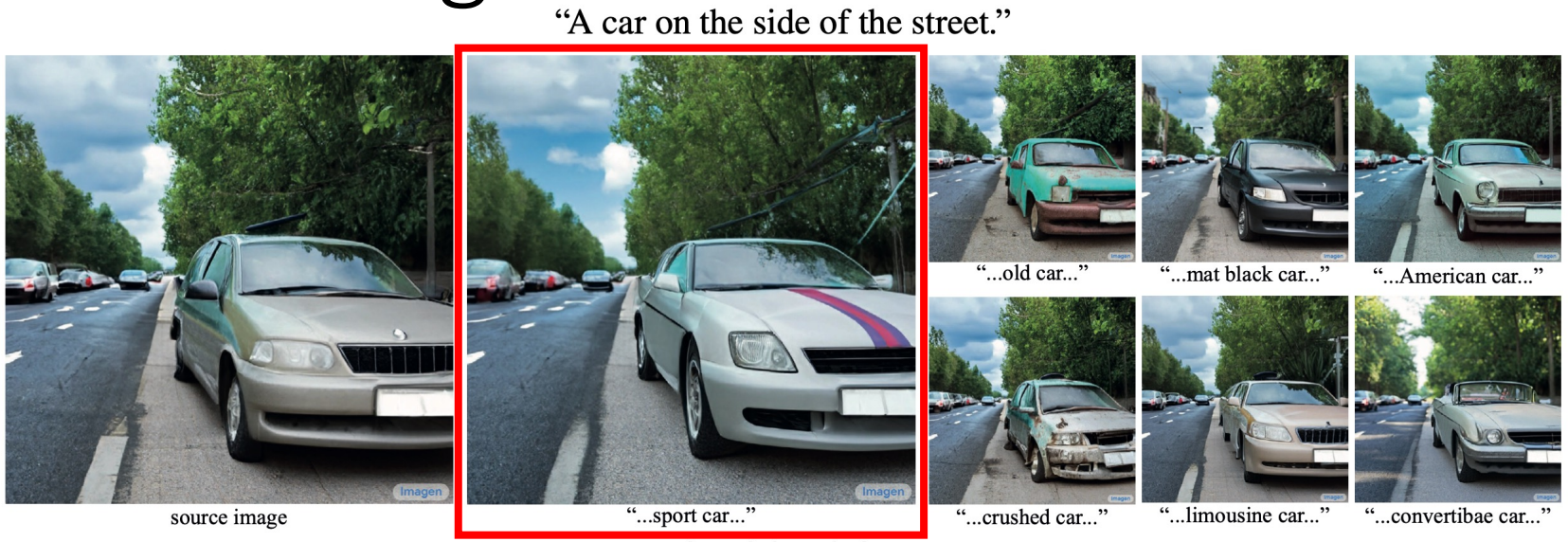
Source Prompt:
Photo of a cat riding on a *bicycle*



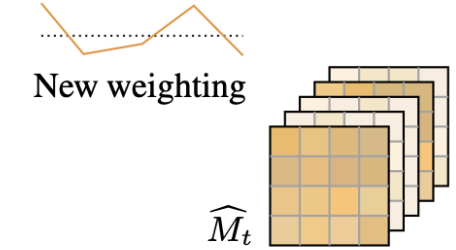
Results – Adding New Phrase



Adding a New Phrase



Results – Attention Re-weighting



"A photo of a birthday(↓) cake next to an apple."

"The picnic is ready under a blossom(↓) tree."

"A photo of a house on a snowy(↑) mountain."

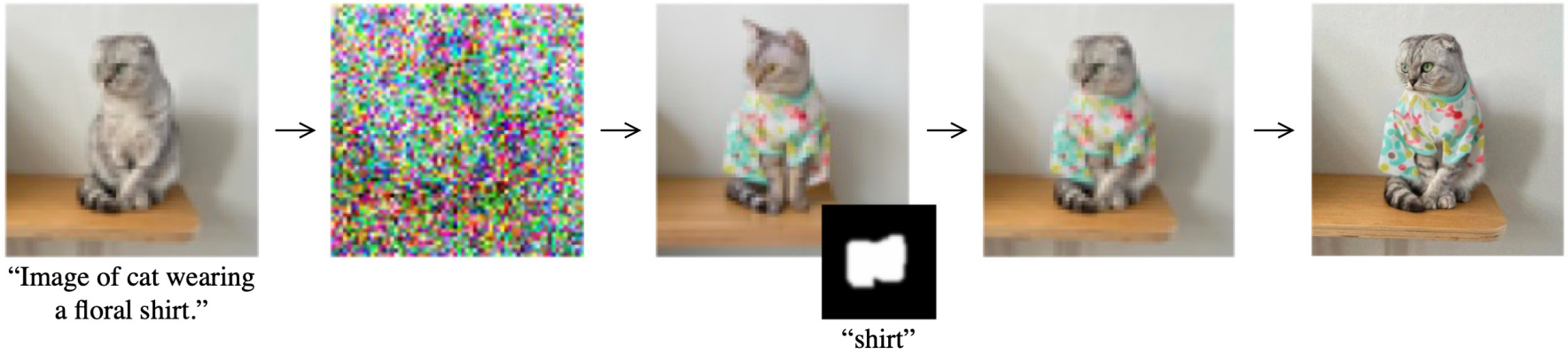
"My fluffy(↑) bunny doll.

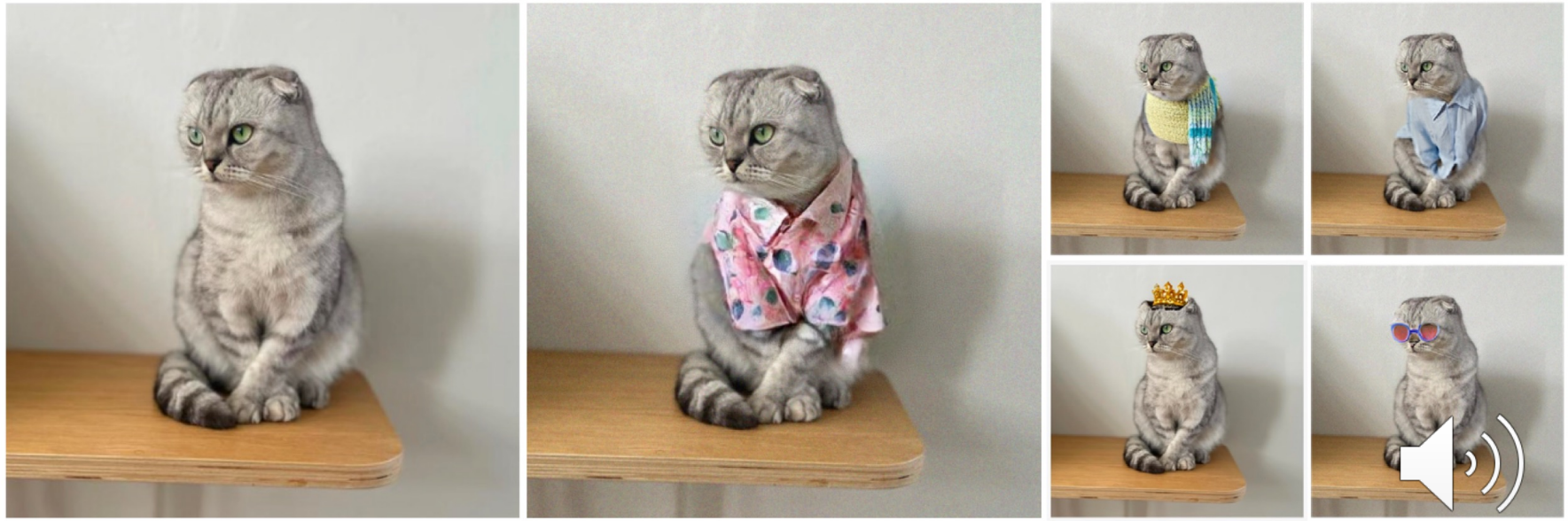
Results – Mask Based Editing

Real image Reconstructed

Real image Reconstructed

Real image Reconstructed





real image "...wearing a colorful shirt..." More dresses

Summary

- Pros:
  - Semantically accurate
  - Intuitive and easy for users

- Cons:
  - Some images not reconstructed properly during diffusion process
  - Attention maps are of low resolution
  - Cannot move existing objects within images

Thank You

References

- Tero Karras, Miika Aittala, Samuli Laine, Erik Ha¨rko¨nen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.
- Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip- guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.
- Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castri- cato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kam- yar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems,* 27, 2014.
- Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv preprint arXiv:2112.05219,* 2021.
- David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word, 2021.
- 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.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical textconditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.