

# Applications of Denoising Diffusion Models on Images

Jiaming Song



Chenlin Meng



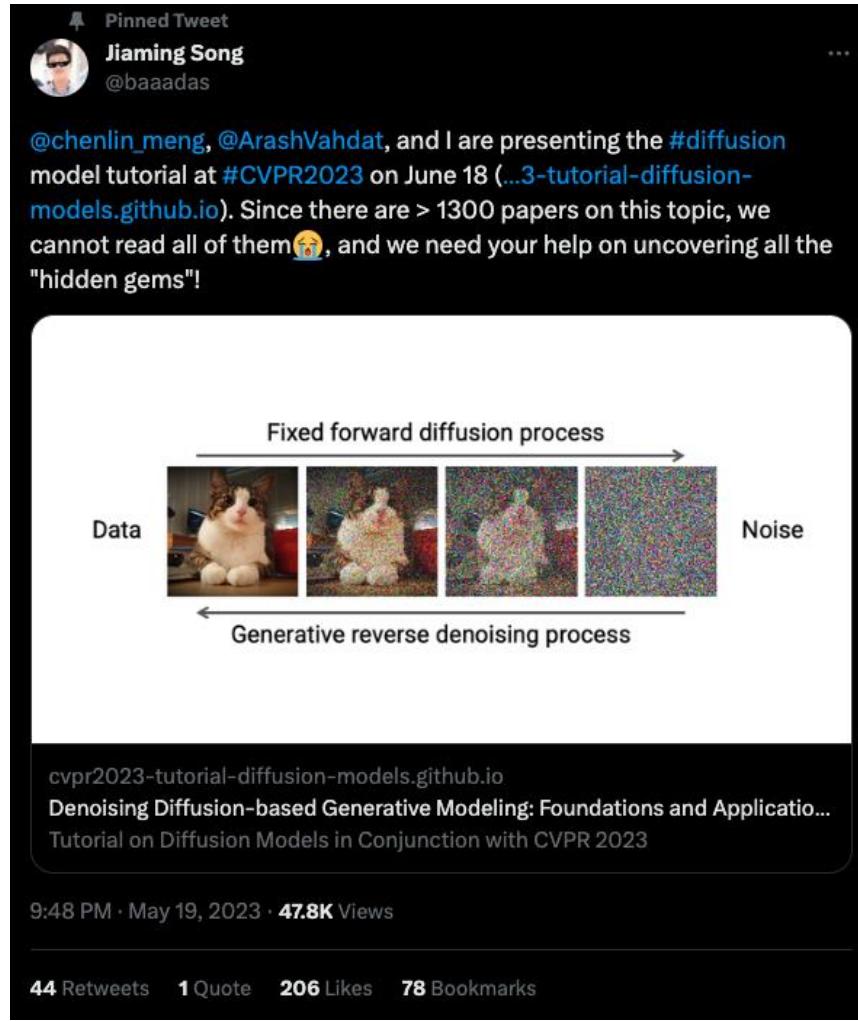
Arash Vahdat



# Outline

- **Diffusion model architectures**
- **Editing and customization with diffusion models**
- **Other applications of diffusion models**

# Thanks everyone for their contribution!



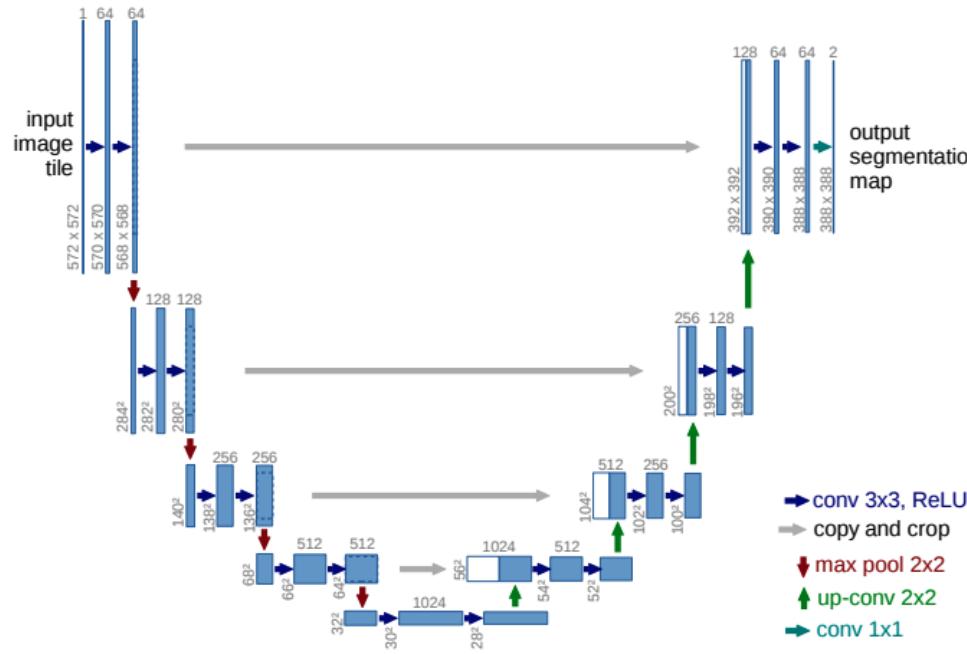
*There are so many amazing works! Due to the time constraint, we are only able to cover part of the papers list on <https://github.com/cvpr2023-tutorial-diffusion-models/papers>*

*Thanks everyone for their contribution! Please check out the entire paper list for more amazing works!*



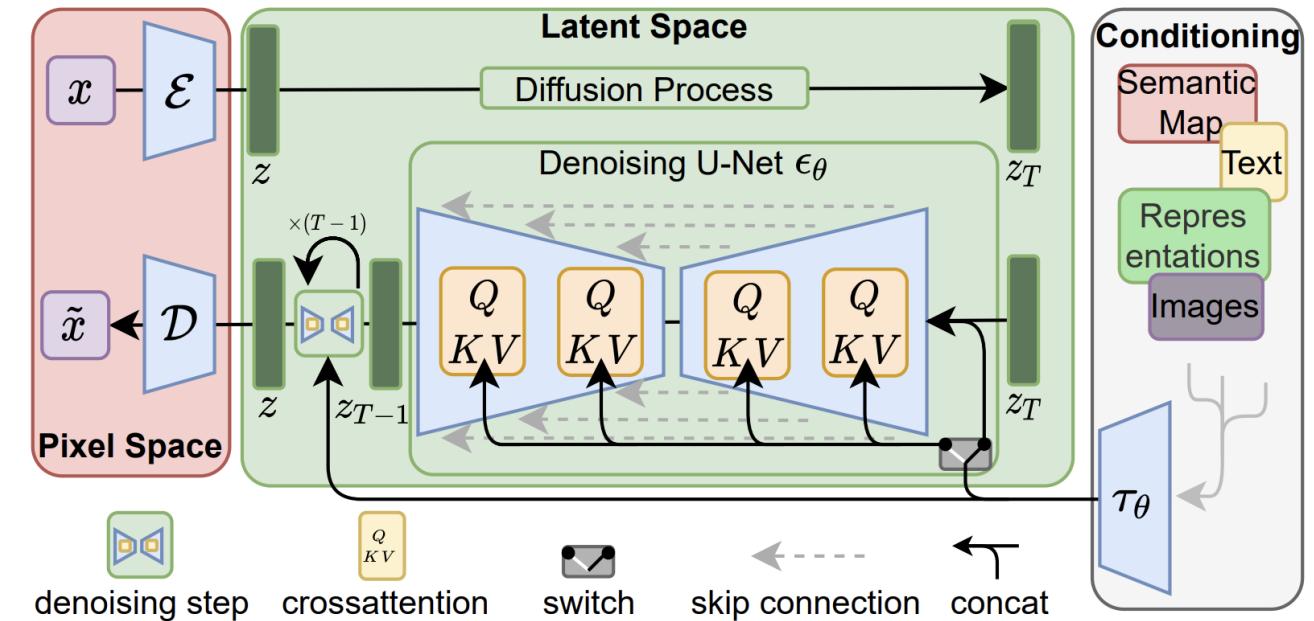
# Architecture

# U-Net Architecture



**U-Net architecture**

Image source: Ronneberger et al.



**U-Net based diffusion architecture**

Image source: Rombach et al.

# U-Net Architecture



**Imagen**  
Saharia et al.



**Stable Diffusion**  
Rombach et al.



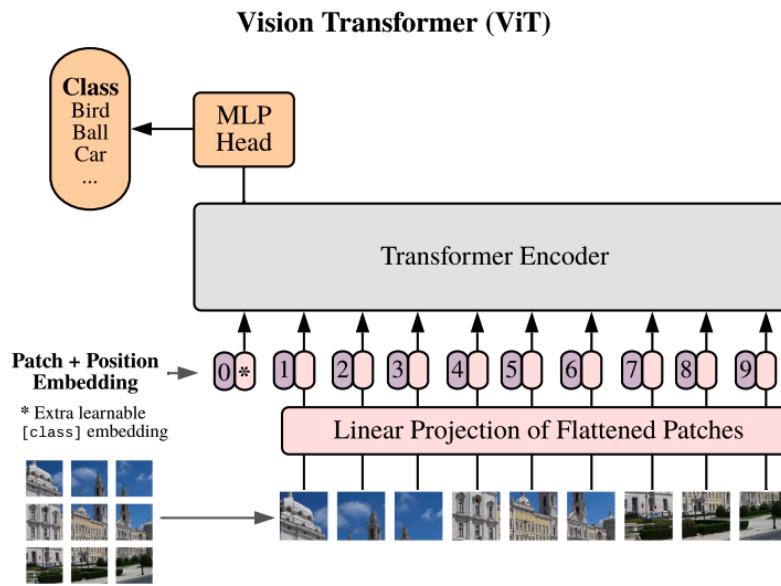
**eDiff-I**  
Balaji et al.

Saharia et al. "[Photorealistic text-to-image diffusion models with deep language understanding](#)", NeurIPS 2022

Rombach et al., "[High-Resolution Image Synthesis with Latent Diffusion Models](#)", CVPR 2022

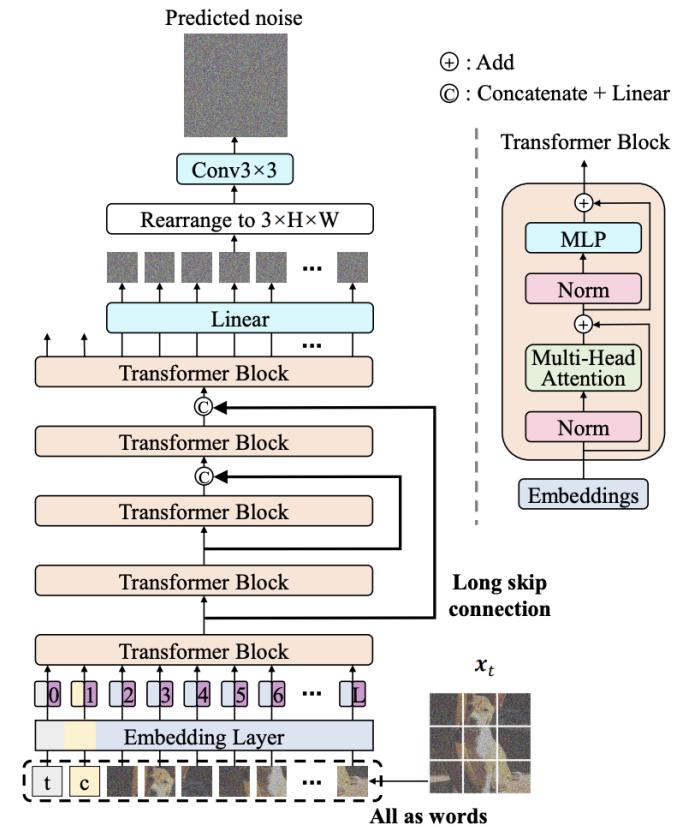
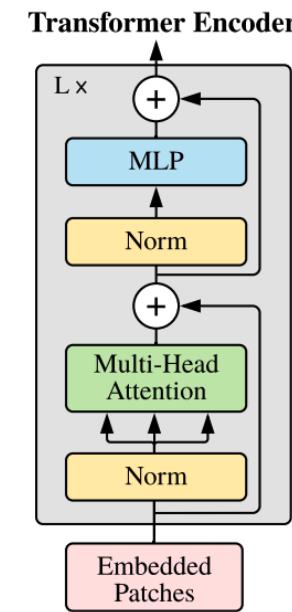
Balaji et al., "[ediffi: Text-to-image diffusion models with an ensemble of expert denoisers](#)", arXiv 2022

# Transformer Architecture

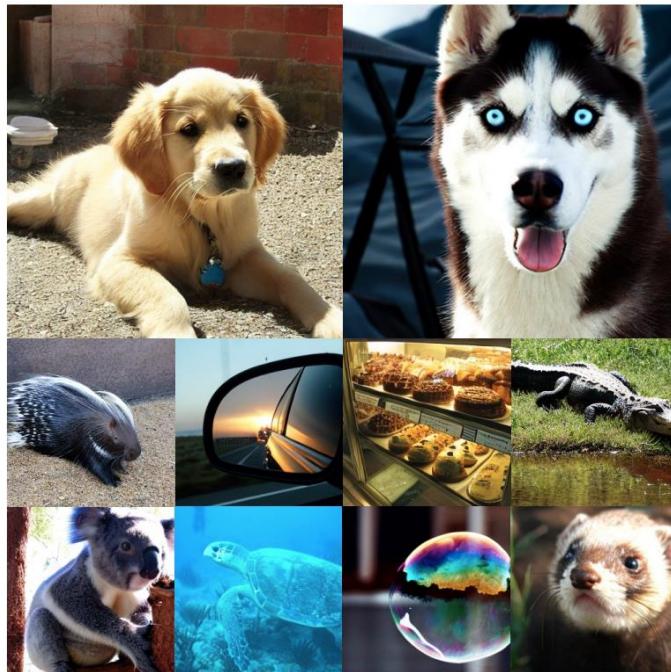


**Vision transformer**

Image source: Dosovitskiy et al.



# Transformer Architecture



**Scalable Diffusion Models  
with Transformers**

**One Transformer Fits All Distributions  
in Multi-Modal Diffusion at Scale**



Beautiful view of the Himalayas

A colorful train passes through the flowers



(a) A render of a bright and colorful city under a dome

(b) A raccoon playing the saxophone

(d) A cartoon of a strawberry drinking a smoothie

(e) A surrealistic painting of a robot riding a skateboard

**simple diffusion: End-to-end diffusion for high resolution images**

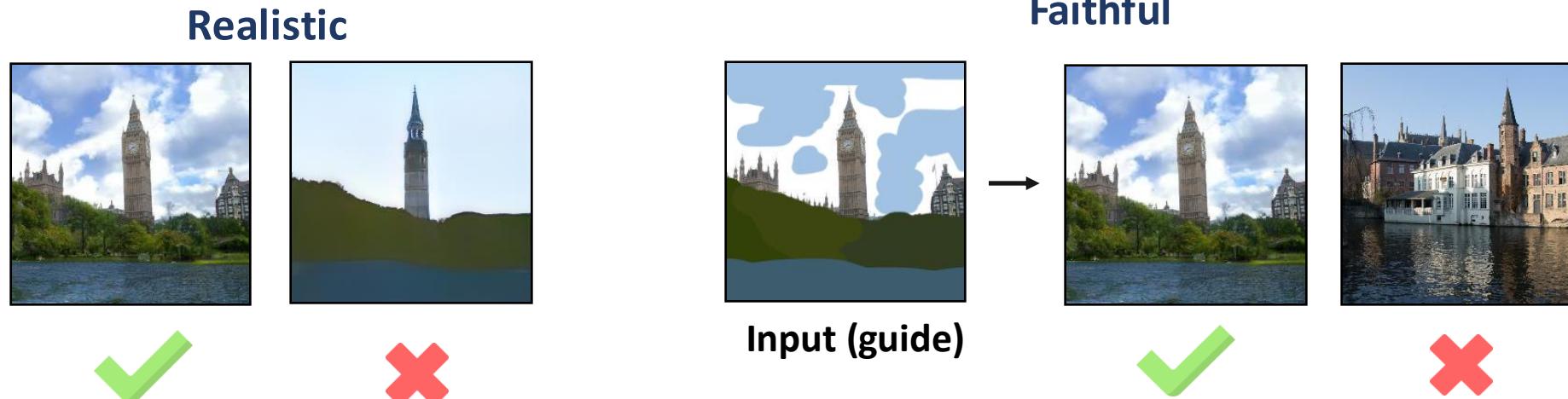
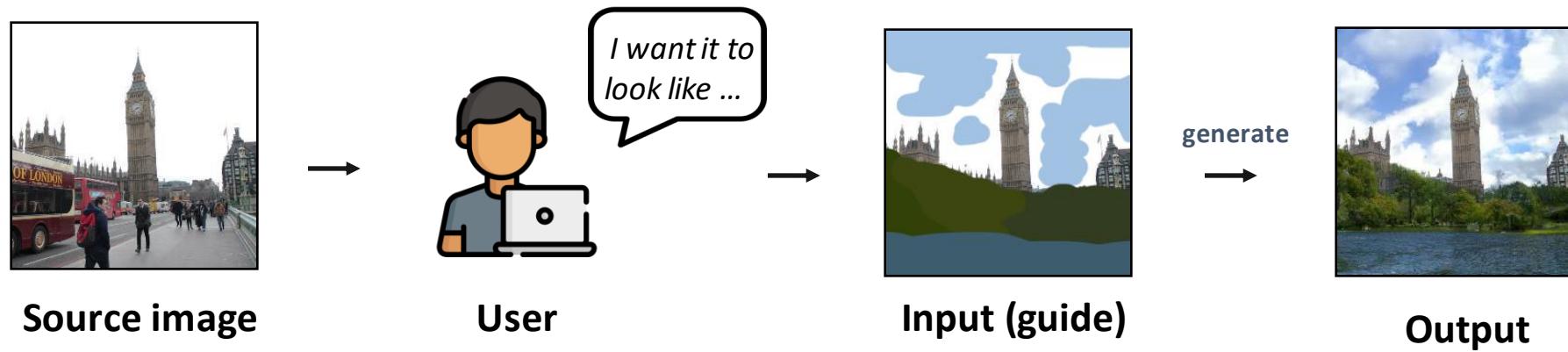


# Image editing and customization with diffusion models

# Image editing and customization with diffusion models

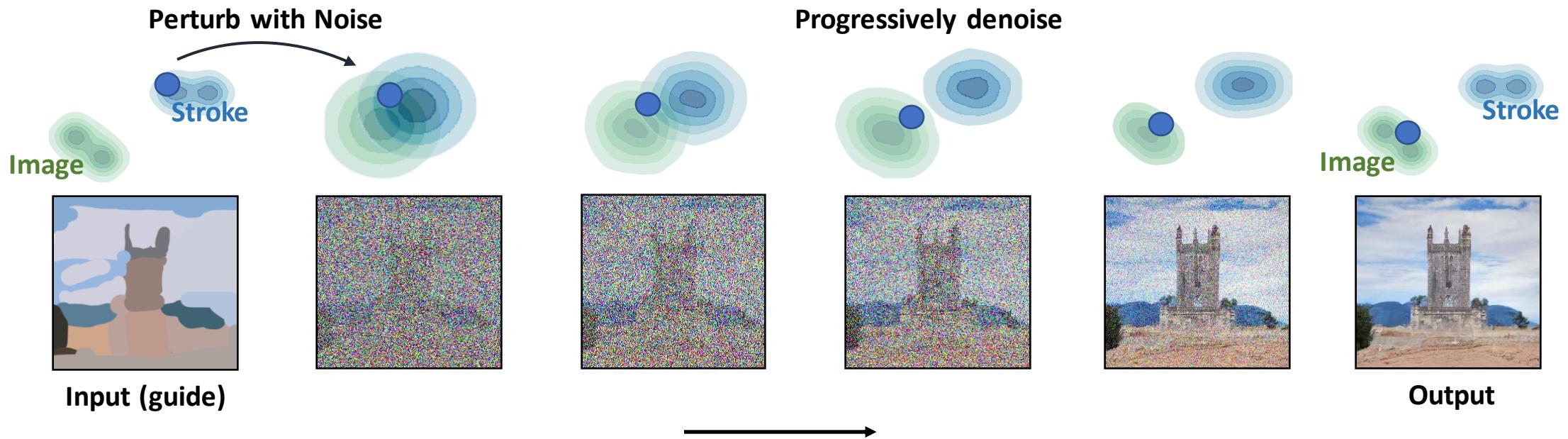
- RGB pixel guidance
- Text guidance
- Reference image guidance

# How to perform guided synthesis/editing?



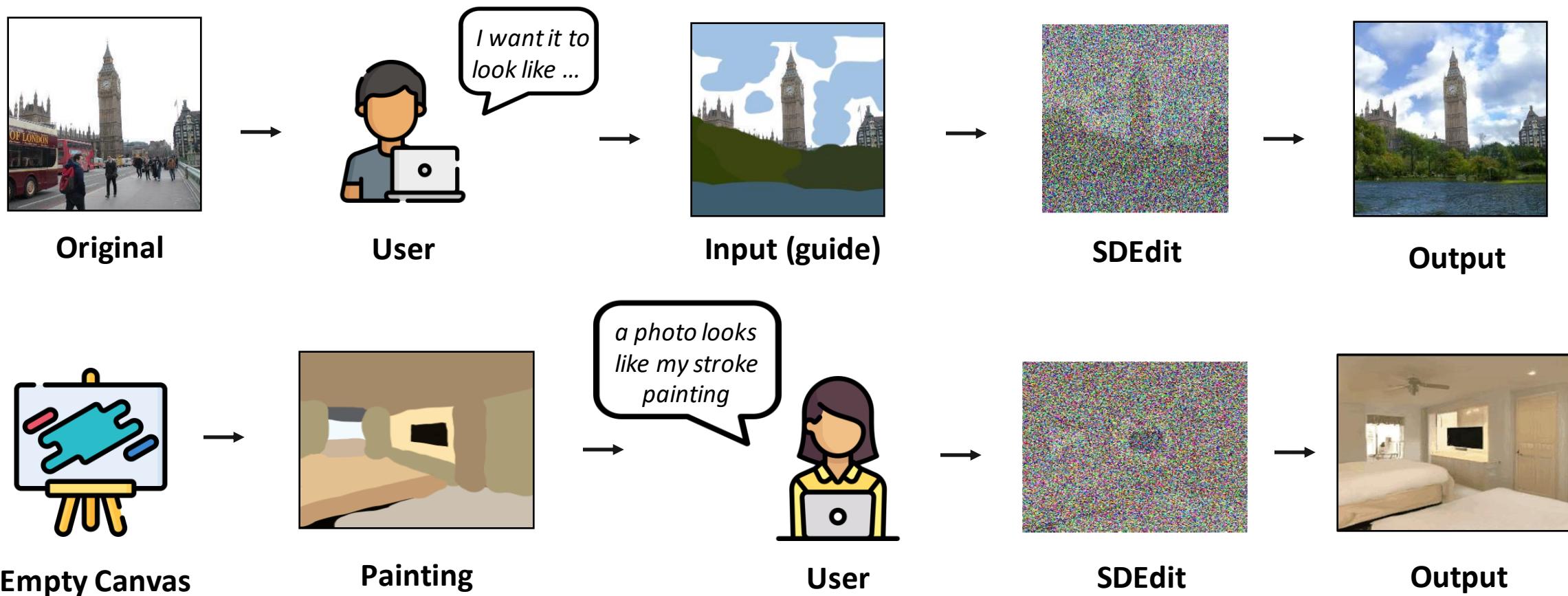
# SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations

First perturb the input with **Gaussian noise** and then progressively remove the noise using a pretrained diffusion model.

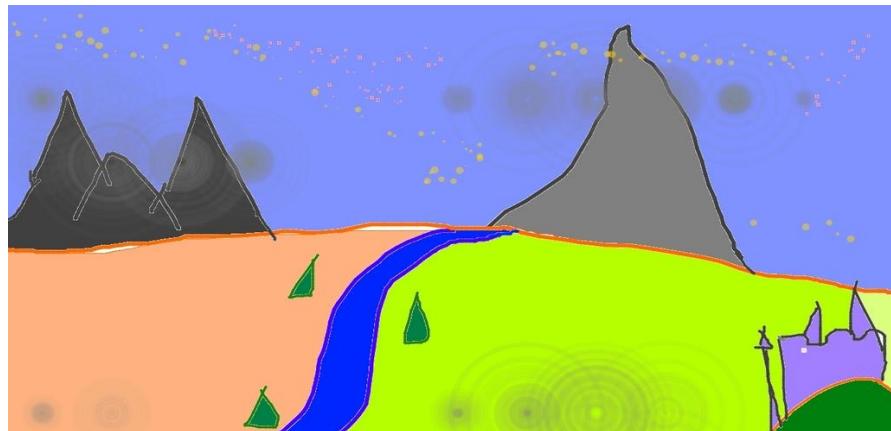


Gradually projects the input to the manifold of natural images.

# Fine-grained control using strokes



# Application to Stable Diffusion (img2img)



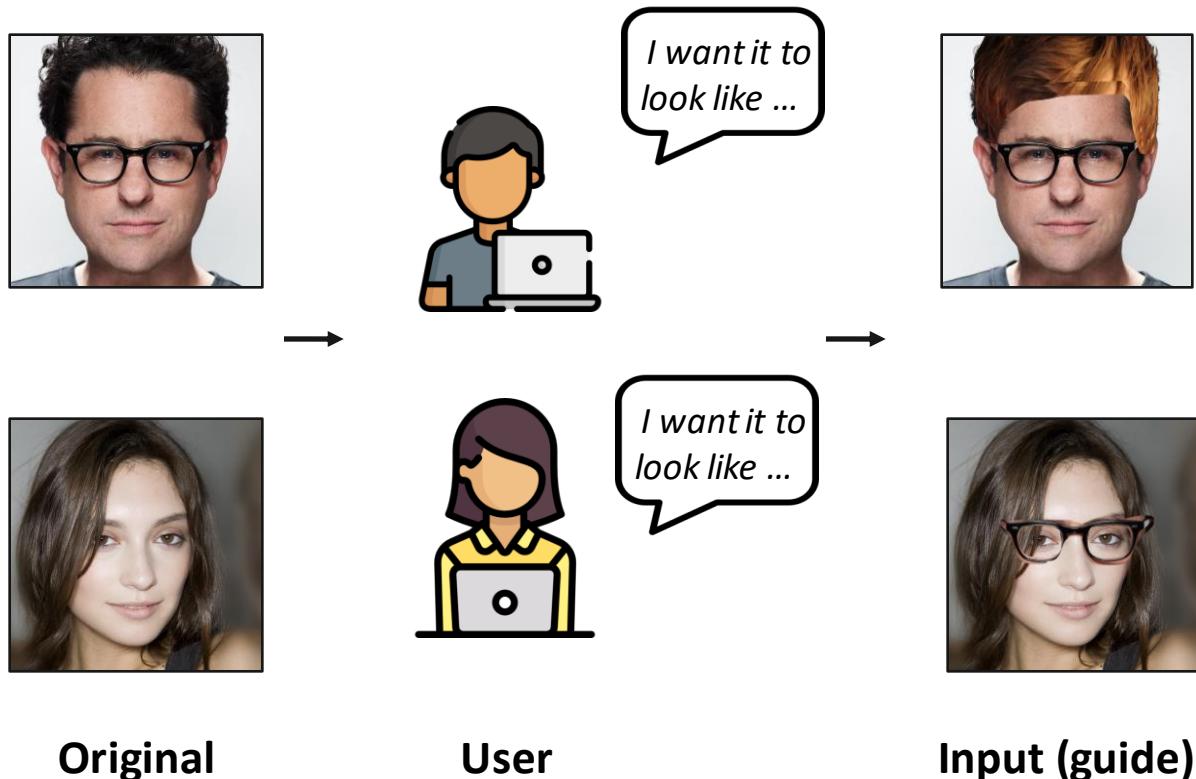
Input

SDEdit/img2img  
→

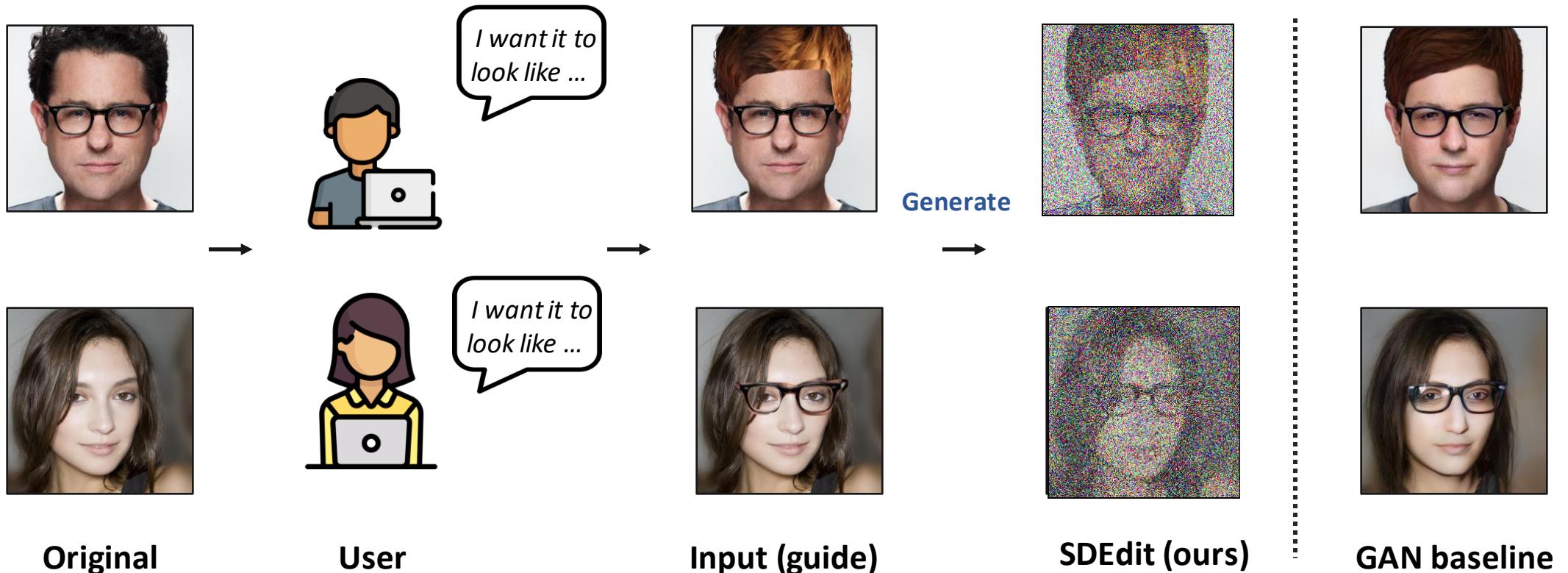


Outputs

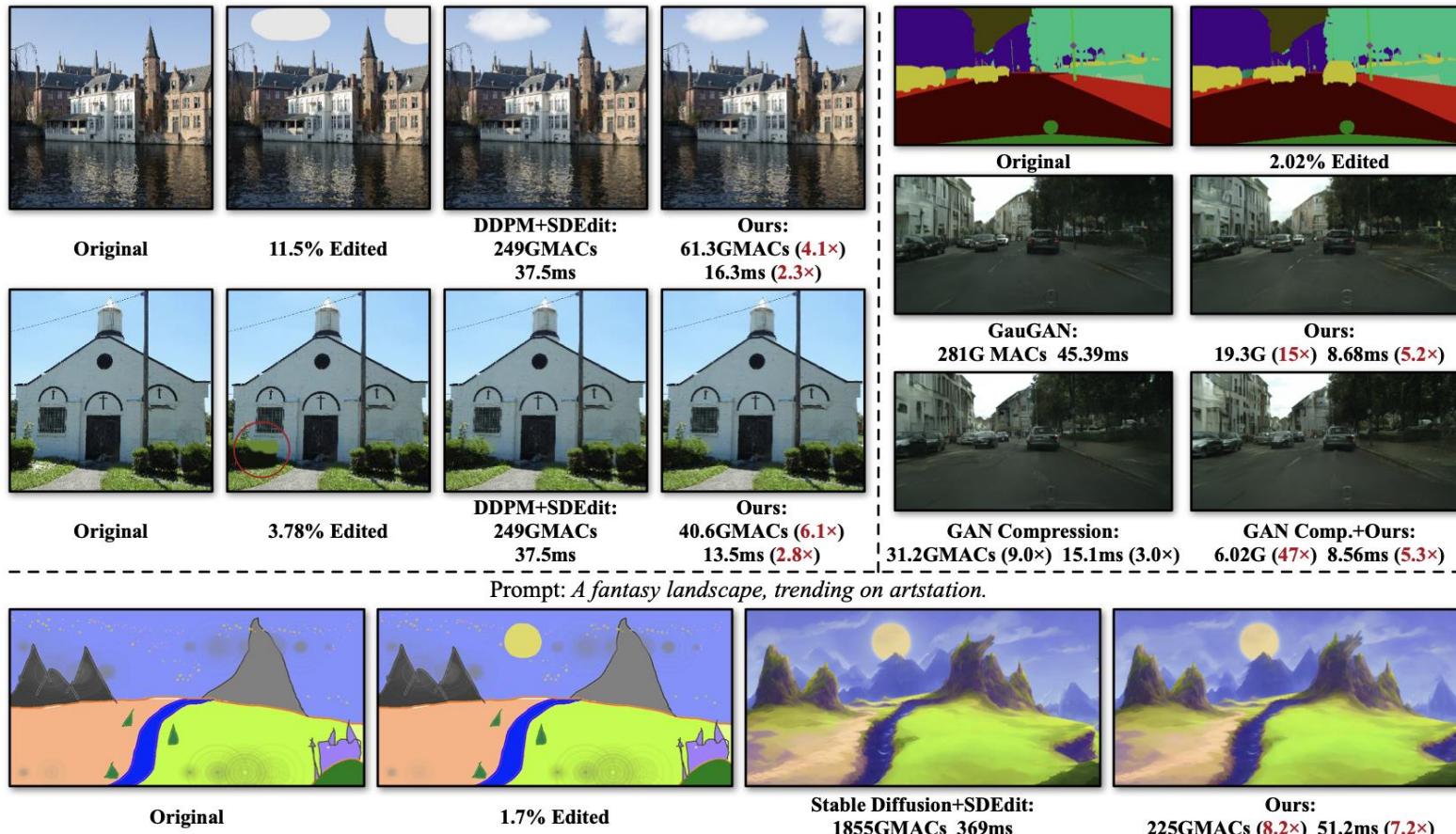
# Image compositing



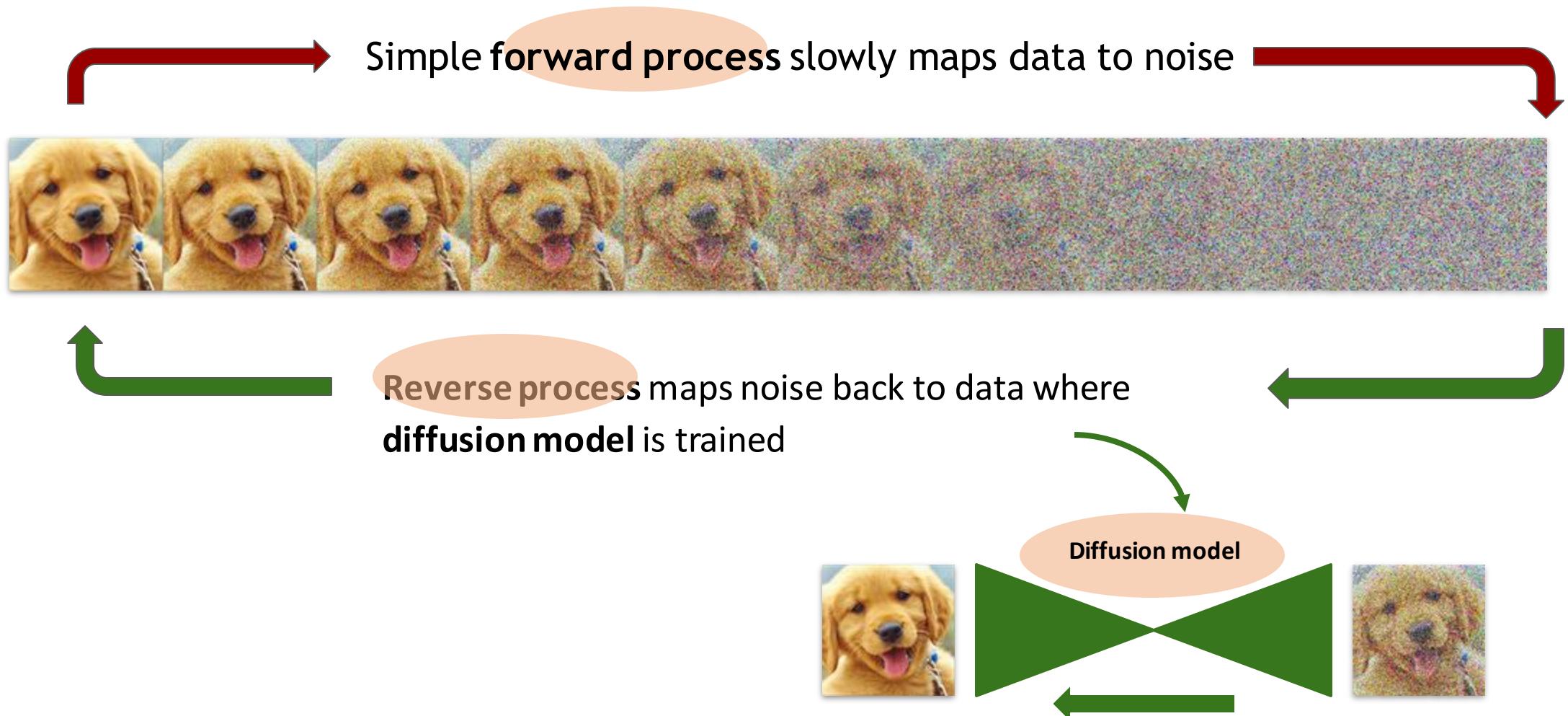
# Image compositing



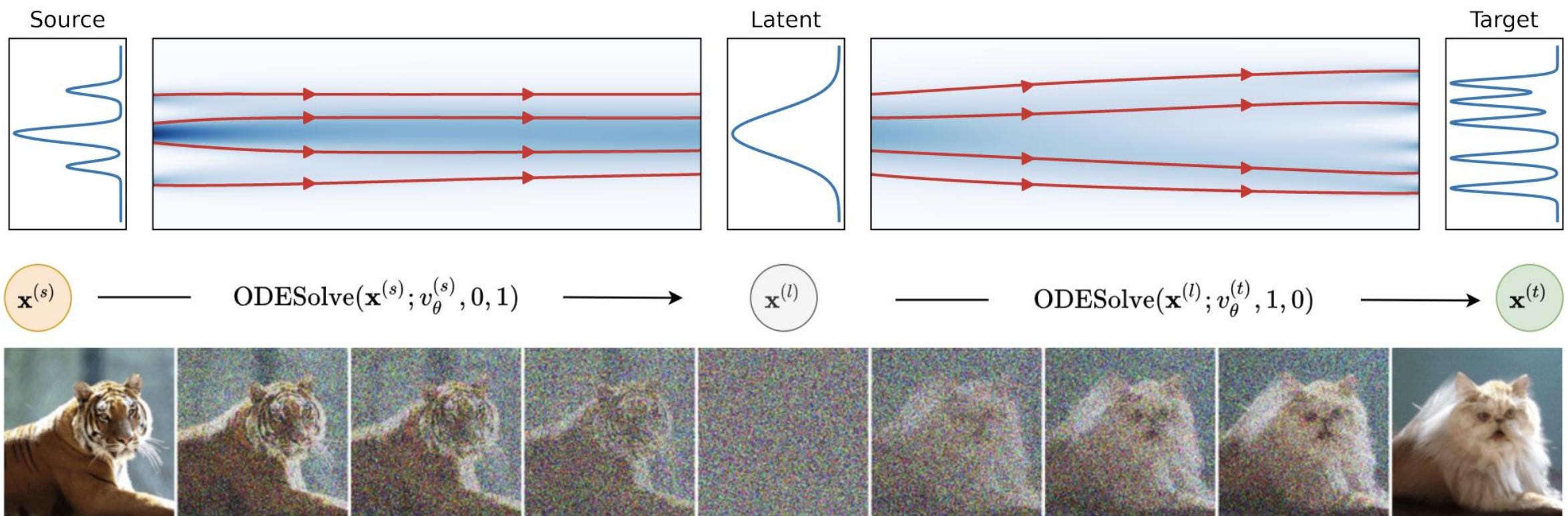
# Efficient Spatially Sparse Inference for Conditional GANs and Diffusion Models



# DDIM Inversion



# Style transfer with DDIM inversion



# Style transfer with DDIM inversion



Multi-domain translation



Reference Image



Source Image 1

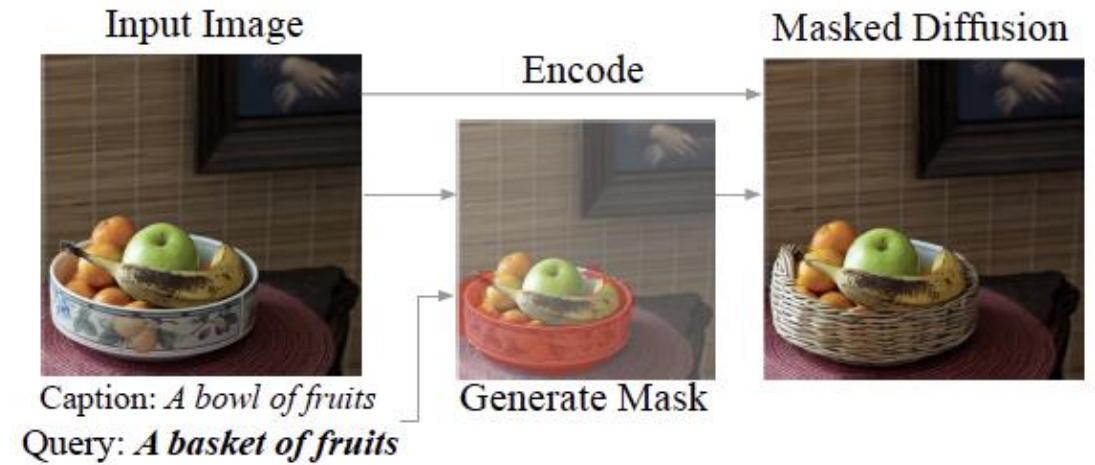
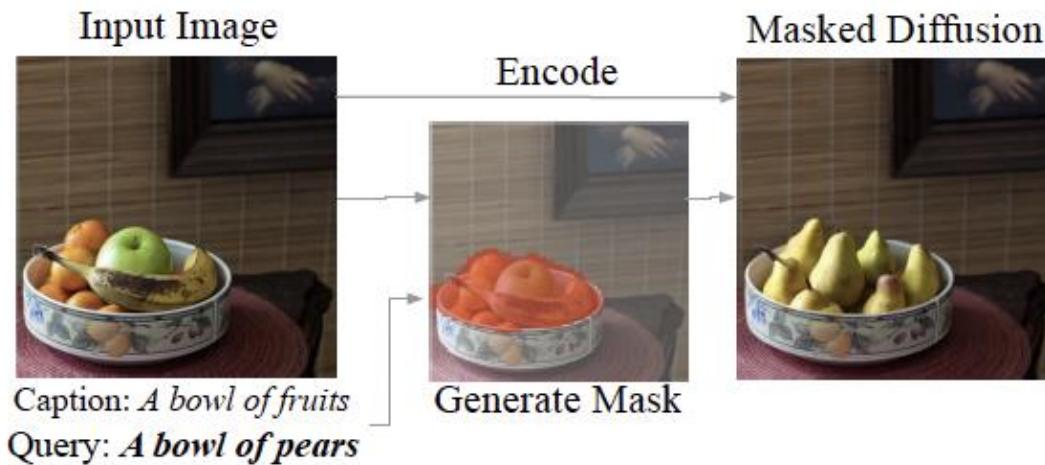


Target Image 1

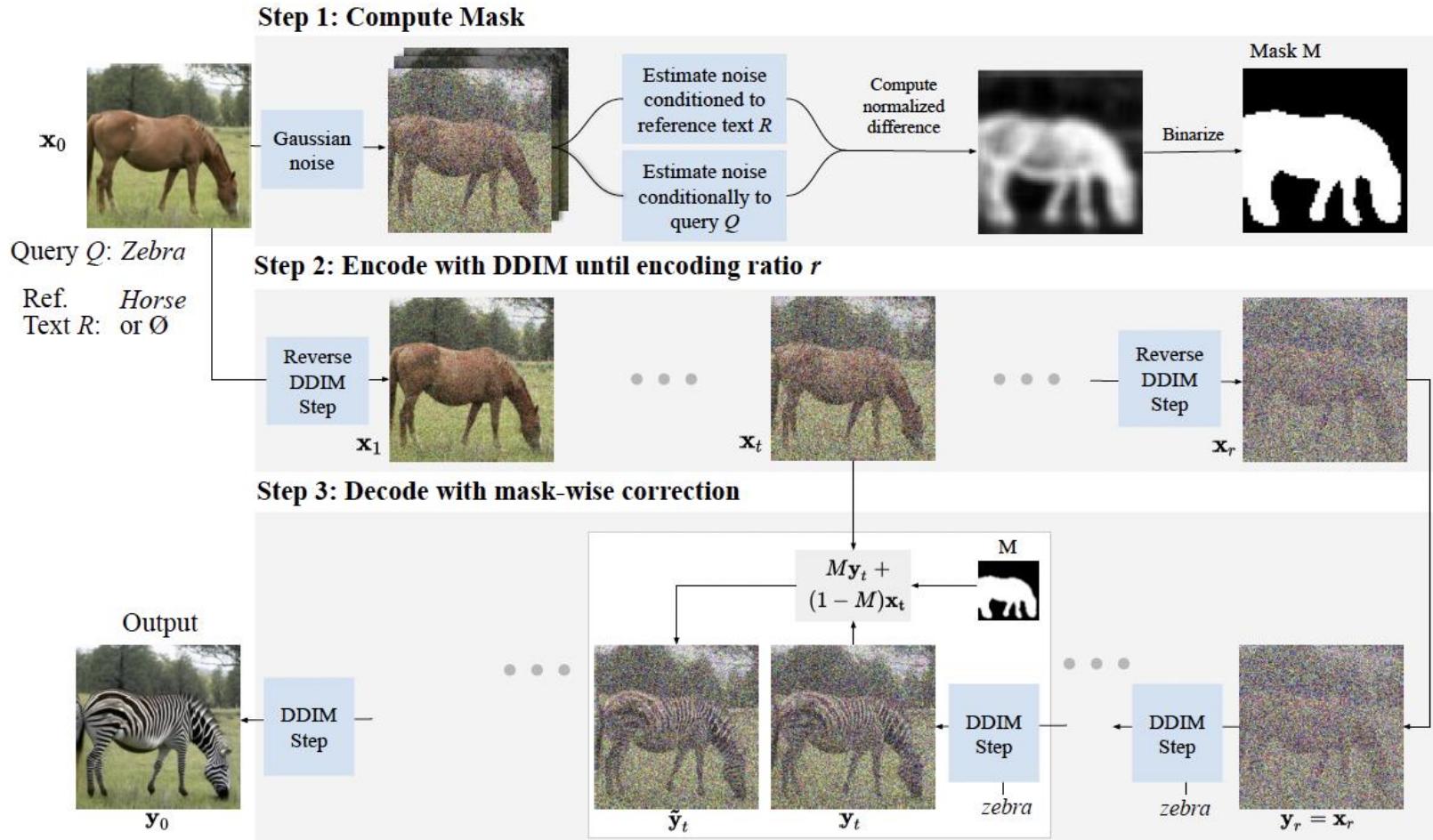
Example-Guided Color Transfer

# DiffEdit: Diffusion-based semantic image editing with mask guidance

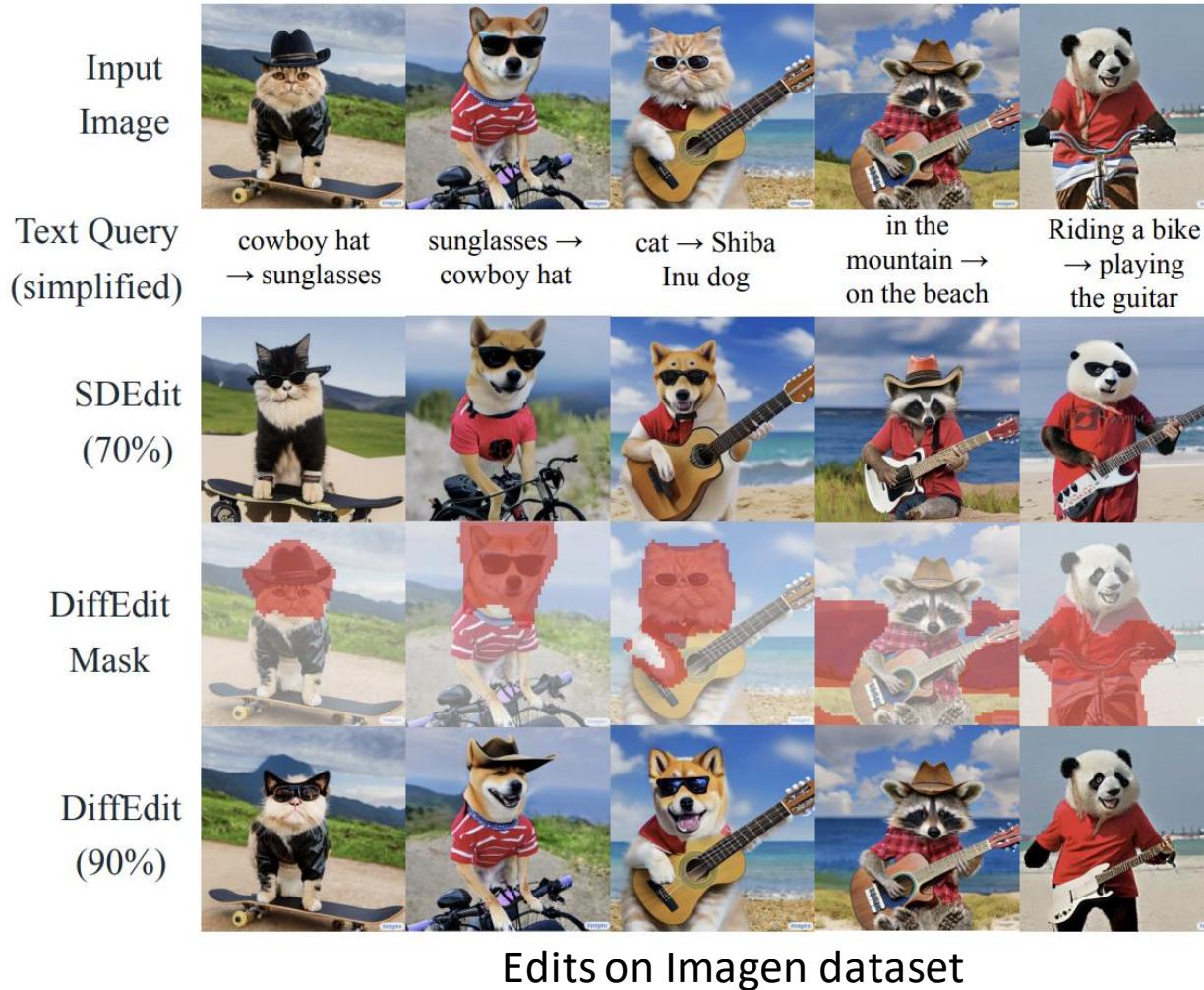
Instead of asking users to provide the mask, the model will generate the mask itself based on the caption and query.



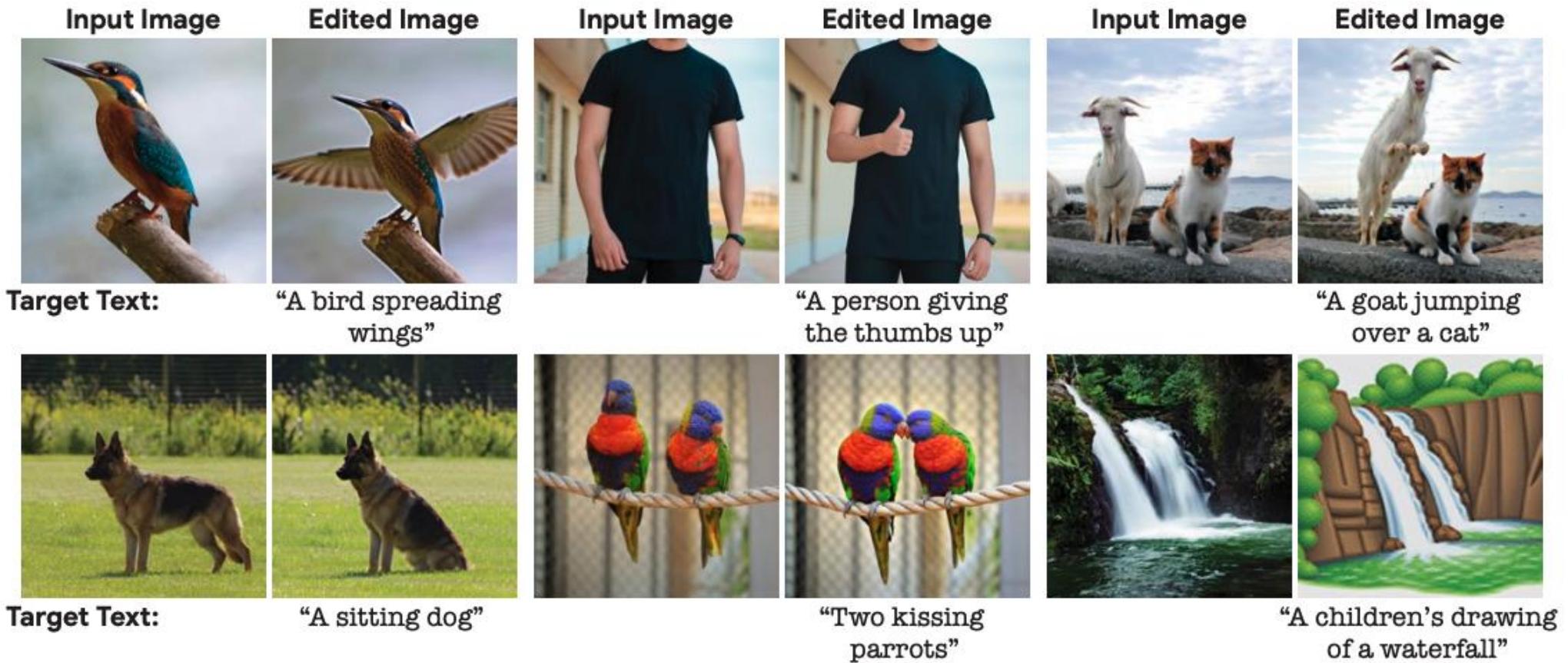
# DiffEdit: Diffusion-based semantic image editing with mask guidance



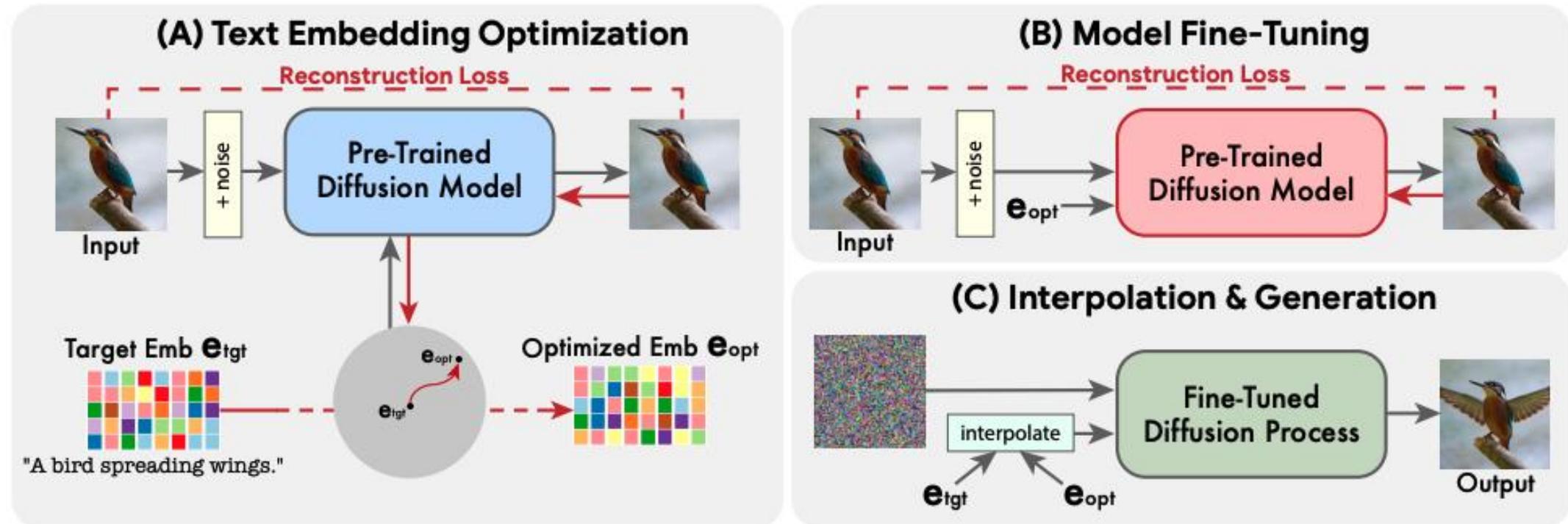
# DiffEdit: Diffusion-based semantic image editing with mask guidance



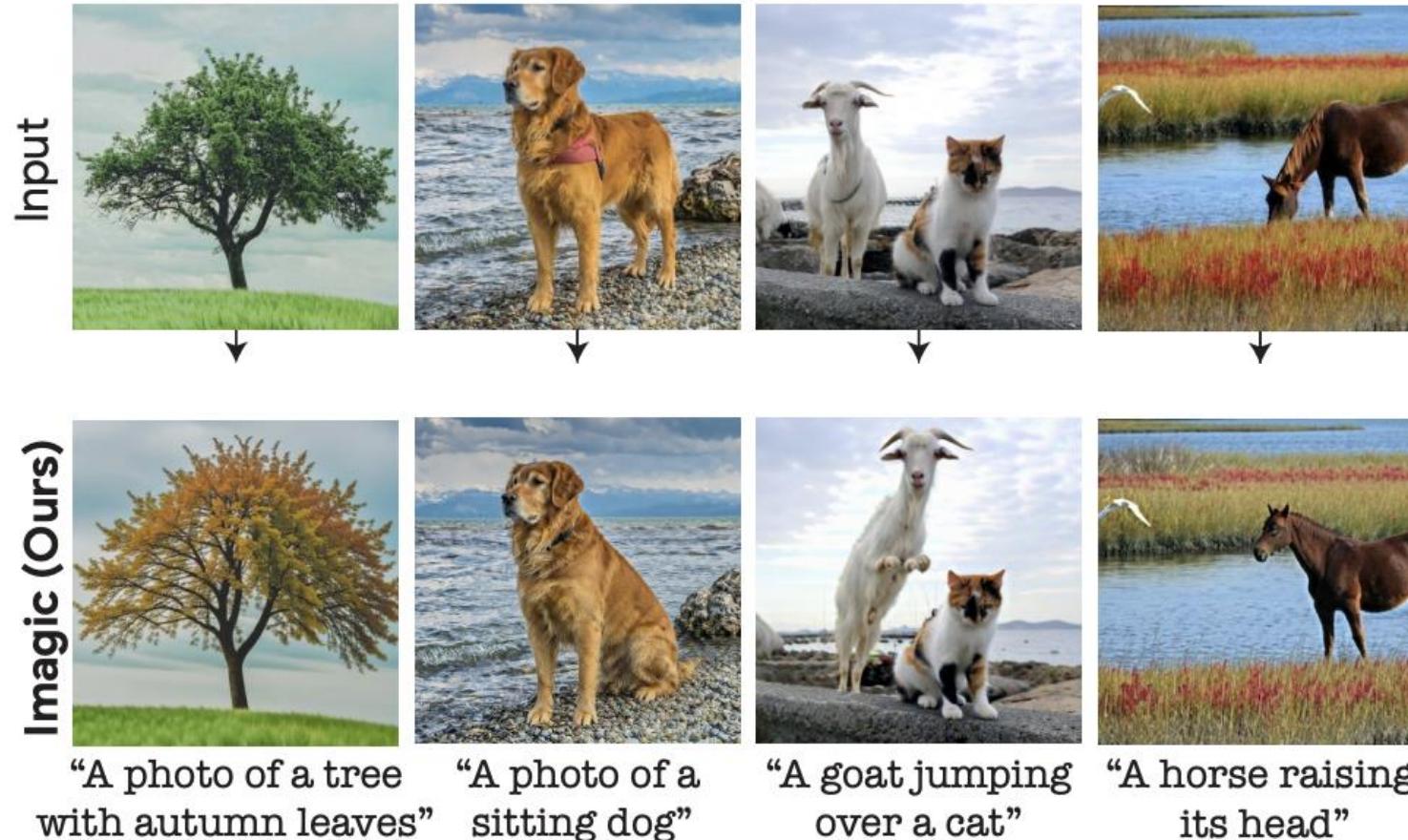
# Imagic: Text-Based Real Image Editing with Diffusion Models



# Imagic: Text-Based Real Image Editing with Diffusion Models



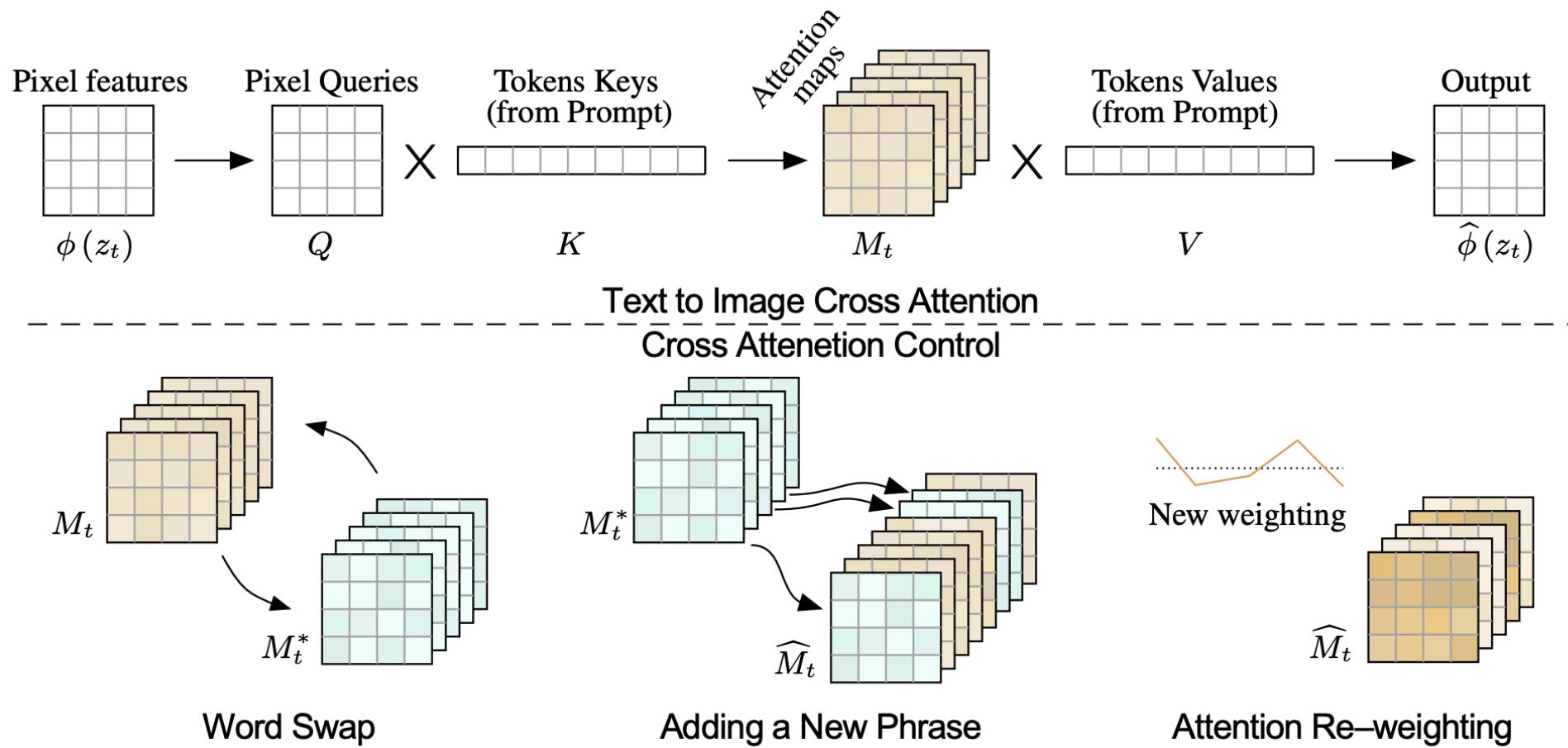
# Imagic: Text-Based Real Image Editing with Diffusion Models



# Prompt-to-Prompt Image Editing with Cross-Attention Control



# Prompt-to-Prompt Image Editing with Cross-Attention Control



# Prompt-to-Prompt Image Editing with Cross-Attention Control



“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.”

# InstructPix2Pix: Learning to Follow Image Editing Instructions

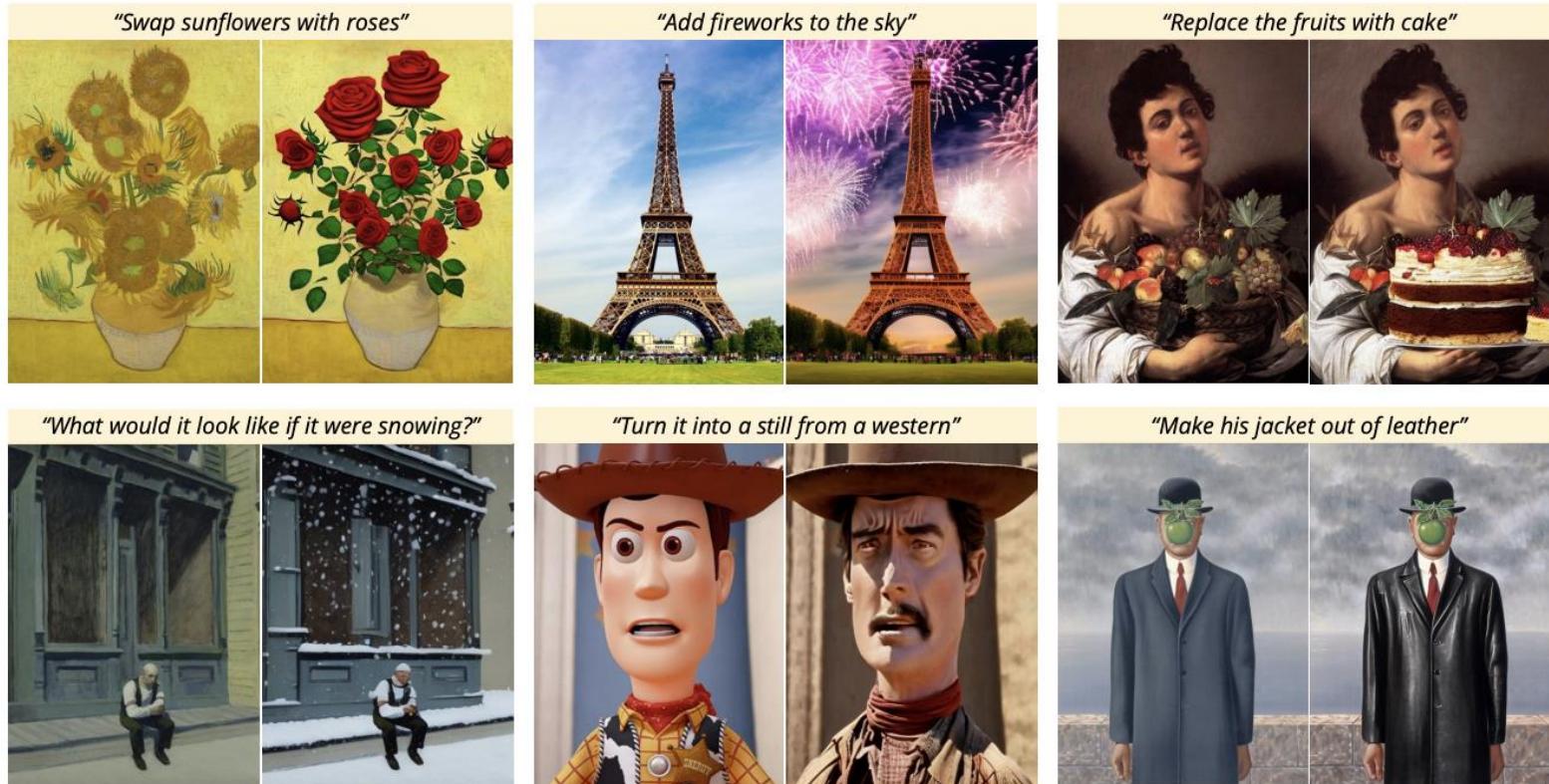
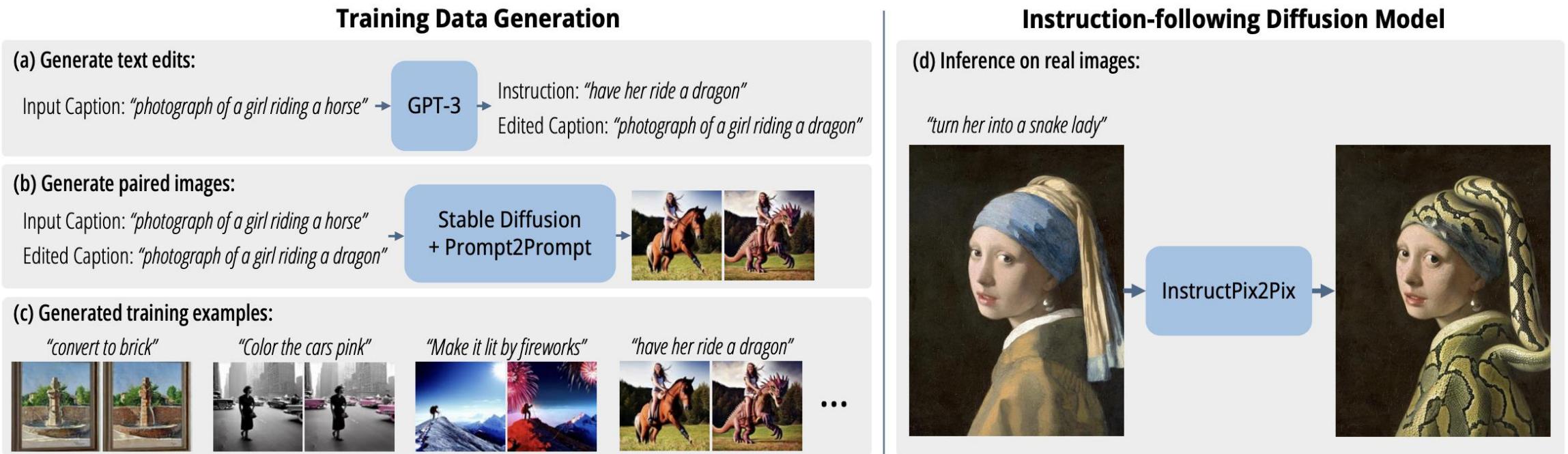


Figure 1. Given an **image** and an **instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

# InstructPix2Pix: Learning to Follow Image Editing Instructions



# Personalization with diffusion models



Input images



in the Acropolis



swimming



sleeping



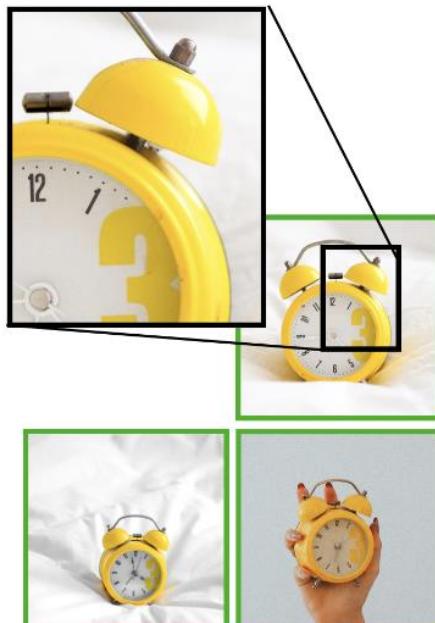
in a bucket



getting a haircut

Generated images

# DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation



Input Images

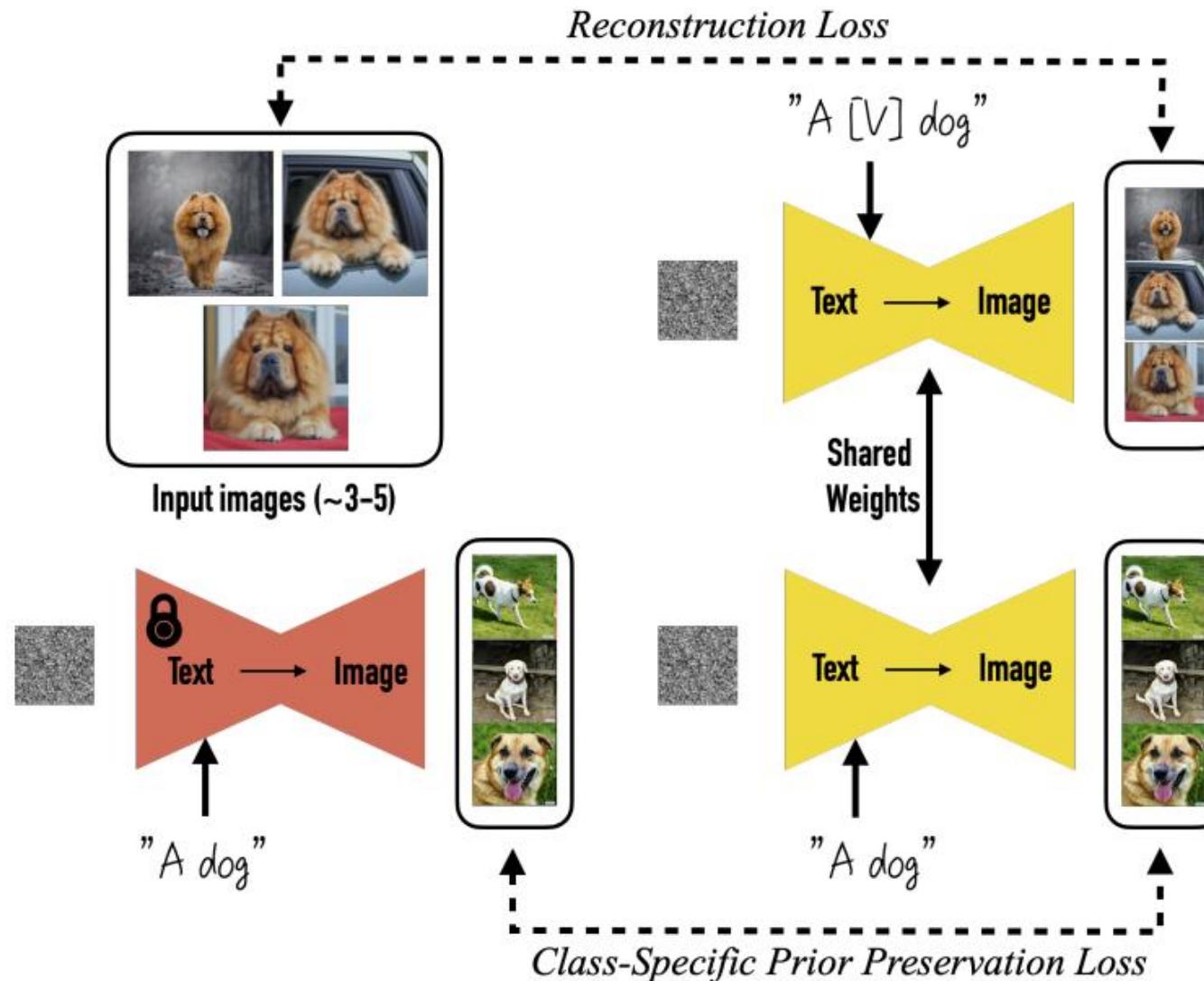


Image-guided, DALL-E2



Text-guided, Imagen

# The DreamBooth Method



# DreamBooth Results



Input images



A [V] backpack in the  
Grand Canyon



A wet [V] backpack  
in water



A [V] backpack in Boston



A [V] backpack with the  
night sky



Input images



A [V] teapot floating  
in milk



A transparent [V] teapot  
with milk inside



A [V] teapot  
pouring tea



A [V] teapot floating  
in the sea

# DreamBooth Applications

## *Text-guided view synthesis*

Input images



Top view ↑



Bottom view ↓

Back view ↘

## *Art Renditions*

Van Gogh



Michelangelo



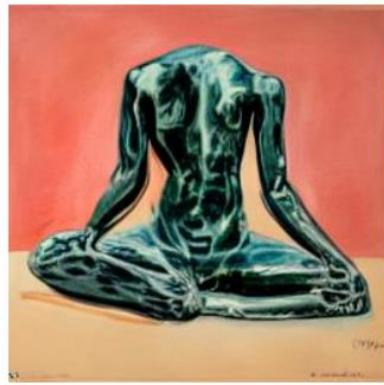
Vermeer



# Textual Inversion: Optimizing Text Embedding



→



Input samples  $\xrightarrow{\text{invert}}$  “ $S_*$ ”

“An oil painting of  $S_*$ ”

“App icon of  $S_*$ ”

“Elmo sitting in  
the same pose as  $S_*$ ”

“Crochet  $S_*$ ”



→



Input samples  $\xrightarrow{\text{invert}}$  “ $S_*$ ”

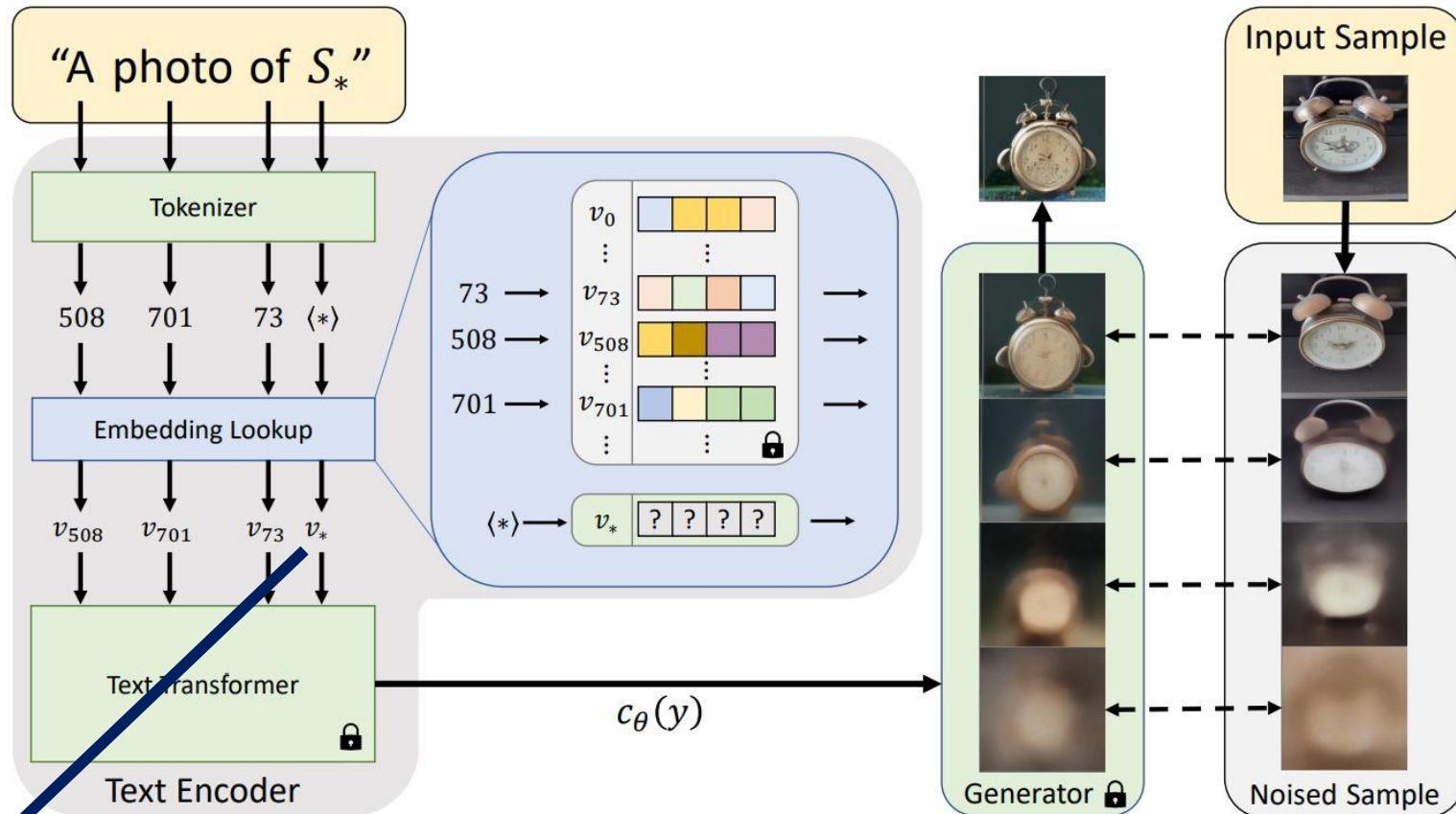
“Painting of two  $S_*$   
fishing on a boat”

“A  $S_*$  backpack”

“Banksy art of  $S_*$ ”

“A  $S_*$  themed lunchbox”

# Textual Inversion: Optimizing Text Embedding



$$v_* = \arg \min_v \mathbb{E}_{z \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0, 1), t} \left[ \|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2 \right]$$

# Textual Inversion Results

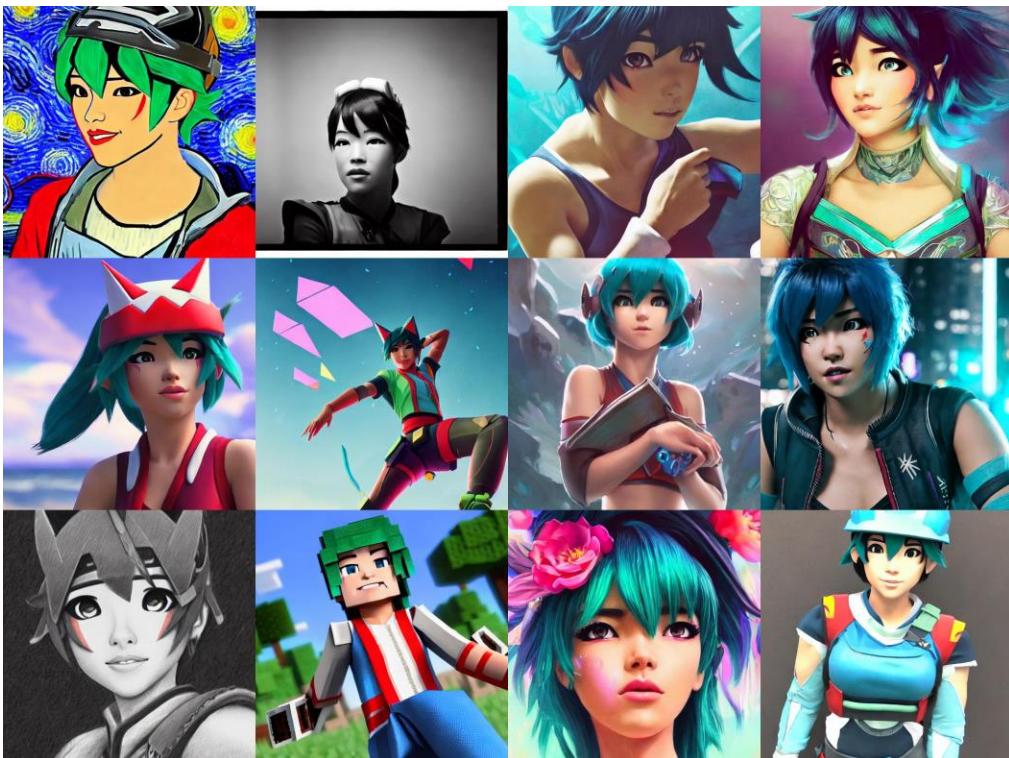


# Works well for artistic styles



# Low-rank Adaptation (LoRA)

- Lora: Low-rank adaptation of large language models



Original weights

$$W = W_0 + BA$$

Low-rank difference

Lora [Edward J. Hu\*, Yelong Shen\*, et al., ICLR 2022]

Lora + Dreambooth (by Simo Ryu): <https://github.com/cloneofsimo/lora>

# Low-rank Adaptation (LoRA)



Finetuned with only 9 images  
Visualized every 500 steps

Lora [Edward J. Hu\*, Yelong Shen\*, et al., ICLR 2022]

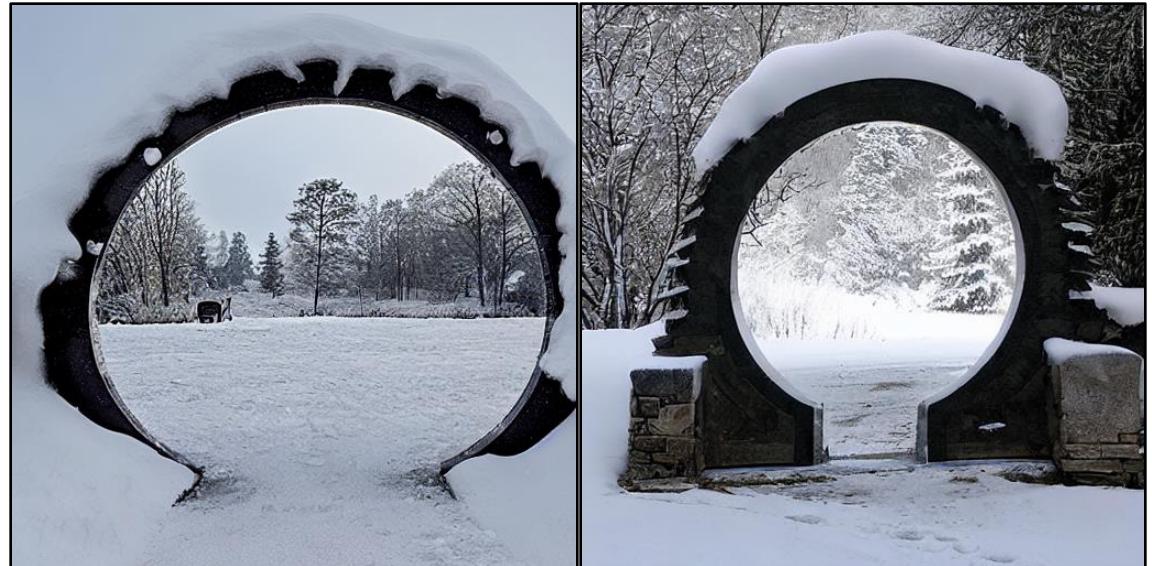
Lora + Dreambooth (by Simo Ryu): <https://github.com/cloneofsimo/lora>

# Fine-tuning all model weights

Photo of a [moongate](#)



[Moongate](#) in snowy ice



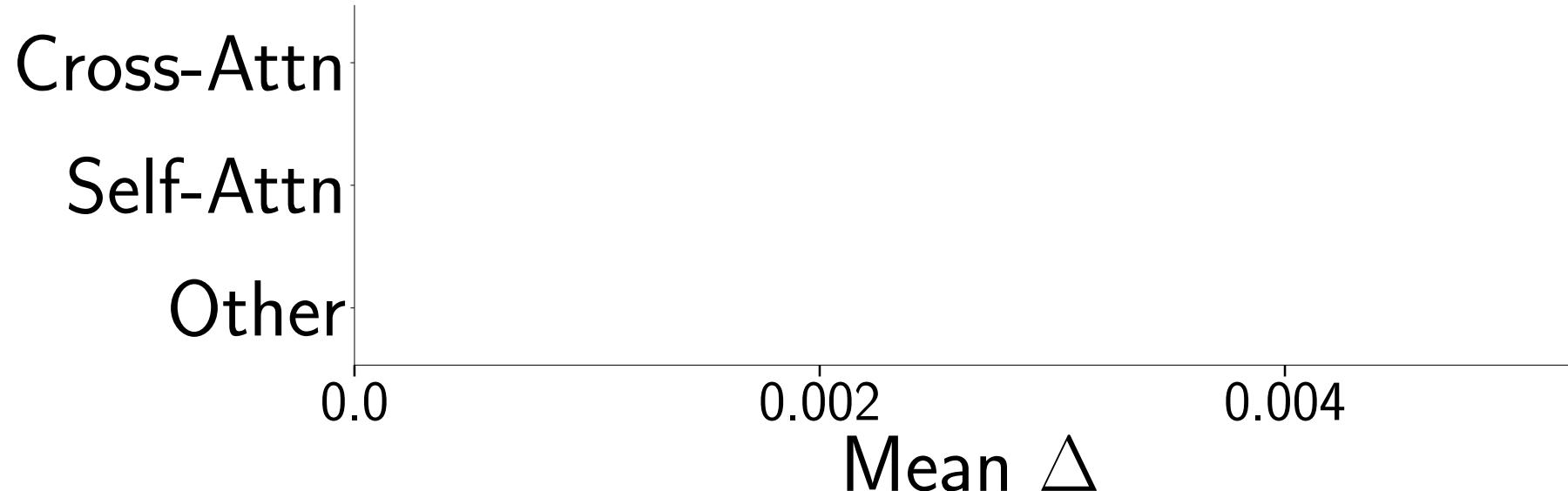
**Storage requirement.** 4GB storage for each fine-tuned model.

**Compute requirement.** It requires more VRAM/training time.

**Compositionality.** Hard to combine multiple models.

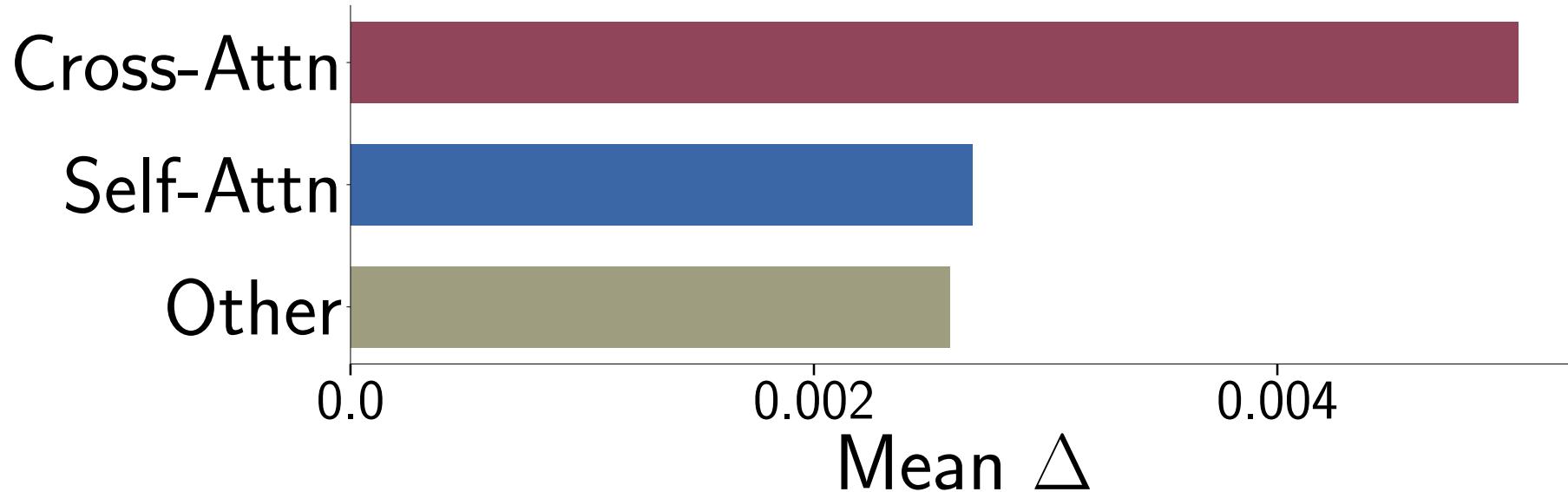
# Analyze change in weights

$$\Delta_l = \frac{||\theta'_l - \theta_l||}{||\theta_l||} \quad \text{where} \quad \begin{aligned} \theta'_l &: \text{updated weights} \\ \theta_l &: \text{pretrained weights} \end{aligned}$$

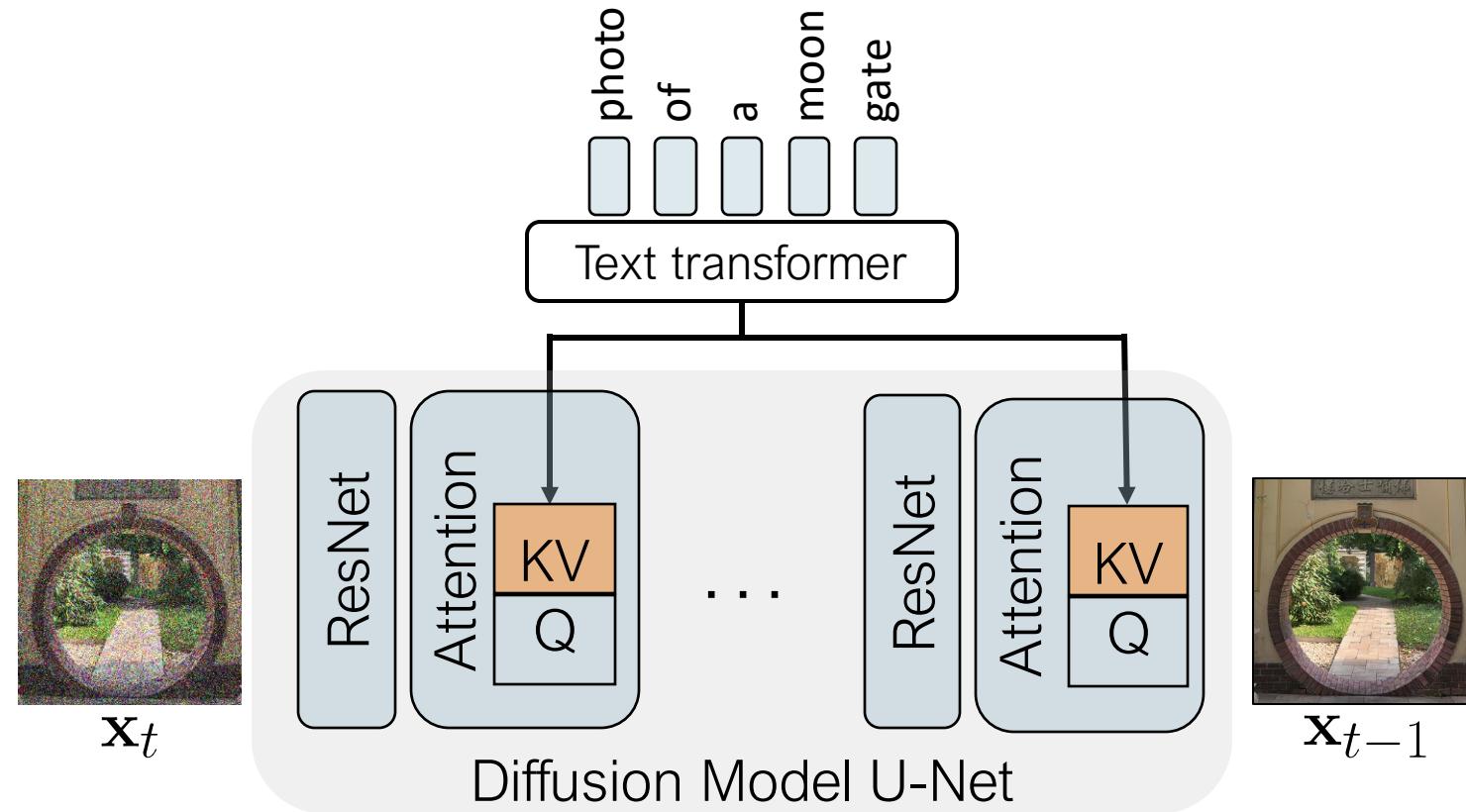


# Analyze change in weights

$$\Delta_l = \frac{||\theta'_l - \theta_l||}{||\theta_l||} \quad \text{where} \quad \begin{aligned} \theta'_l &: \text{updated weights} \\ \theta_l &: \text{pretrained weights} \end{aligned}$$



# Only fine-tune cross-attention layers



# How to prevent overfitting?



Photo of a {moongate}



Photo of a {moongate}

+



sky full of stars and the  
moon



Blood moon

...  
Target images

...  
Add regularization images

# Personalized concepts



Jun-Yan's **dog**, Stark

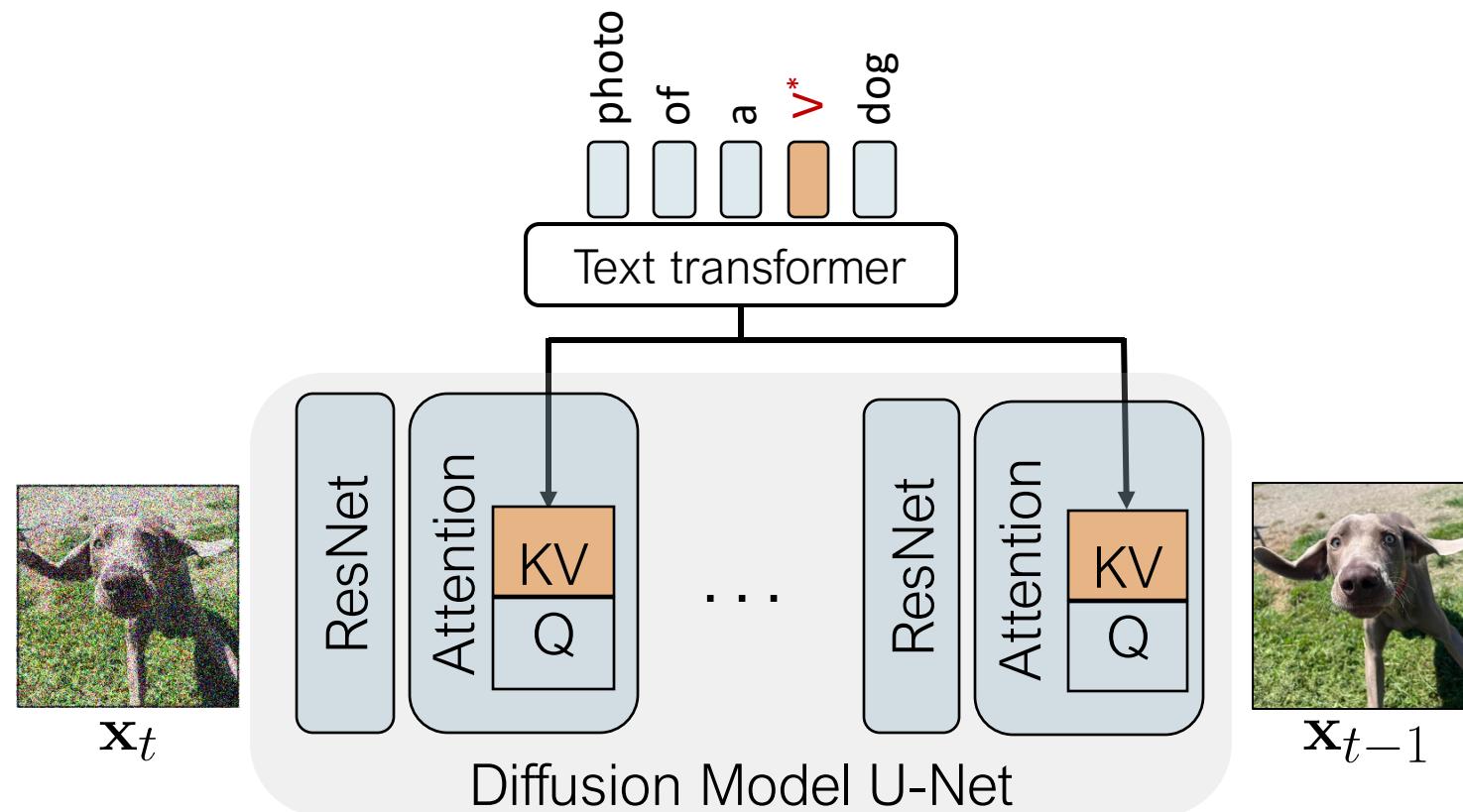
How to describe personalized concepts?

**V\*** dog

Where **V\*** is a modifier token in the text embedding space

# Personalized concepts

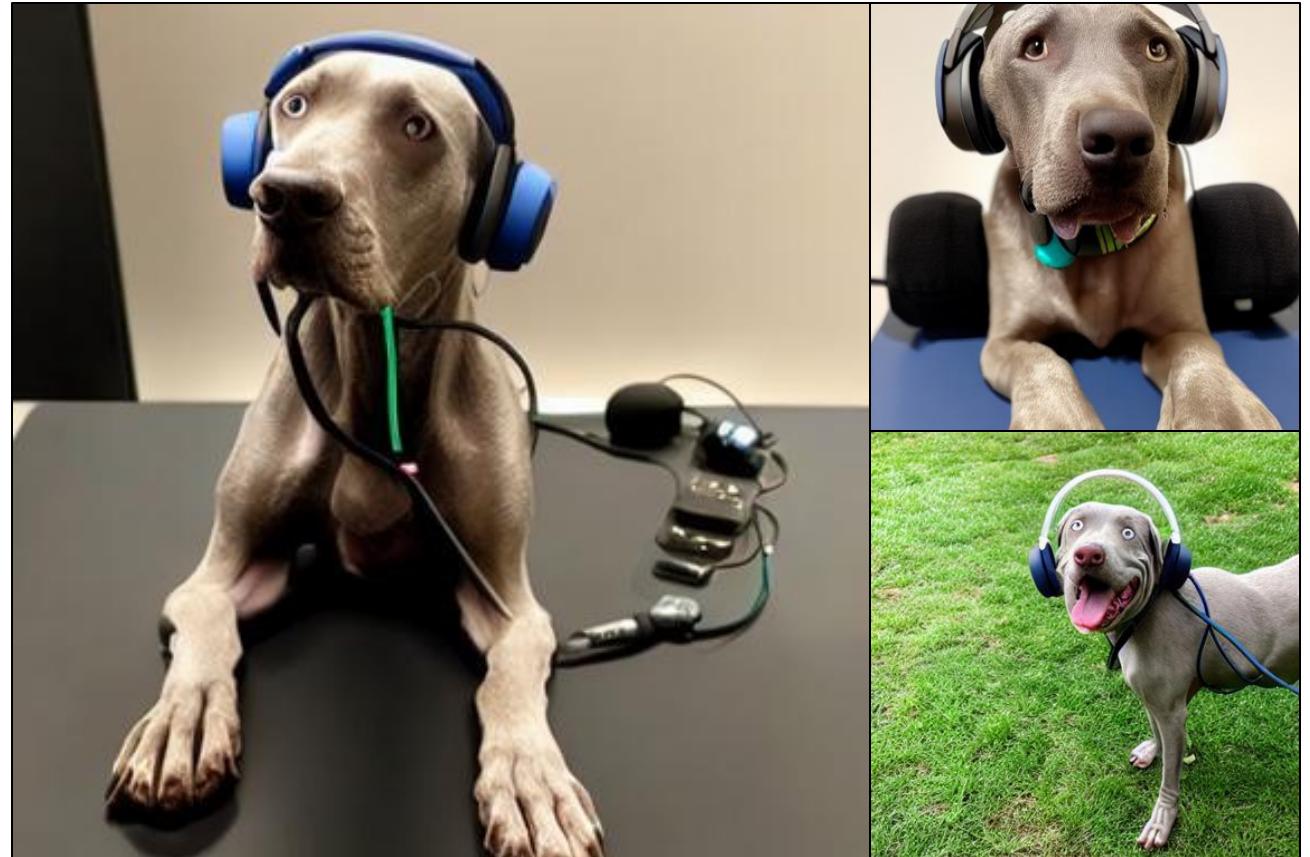
Also fine-tune the modifier token  $V^*$  that describes the personalized concept



Trainable

Frozen

# Single concept results

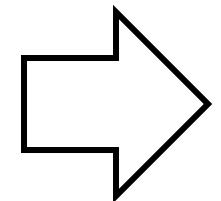


V\* dog wearing headphones

# Multiple new concepts?



+



?

# Joint training

1. Combine the training dataset of multiple concepts

Target images



$V^*$  dog

Regularization images



Dog

Cute dog



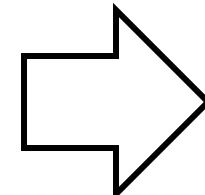
Moongate



Wisdom moon

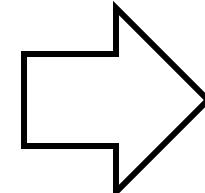
Gated entry

# Two concept results



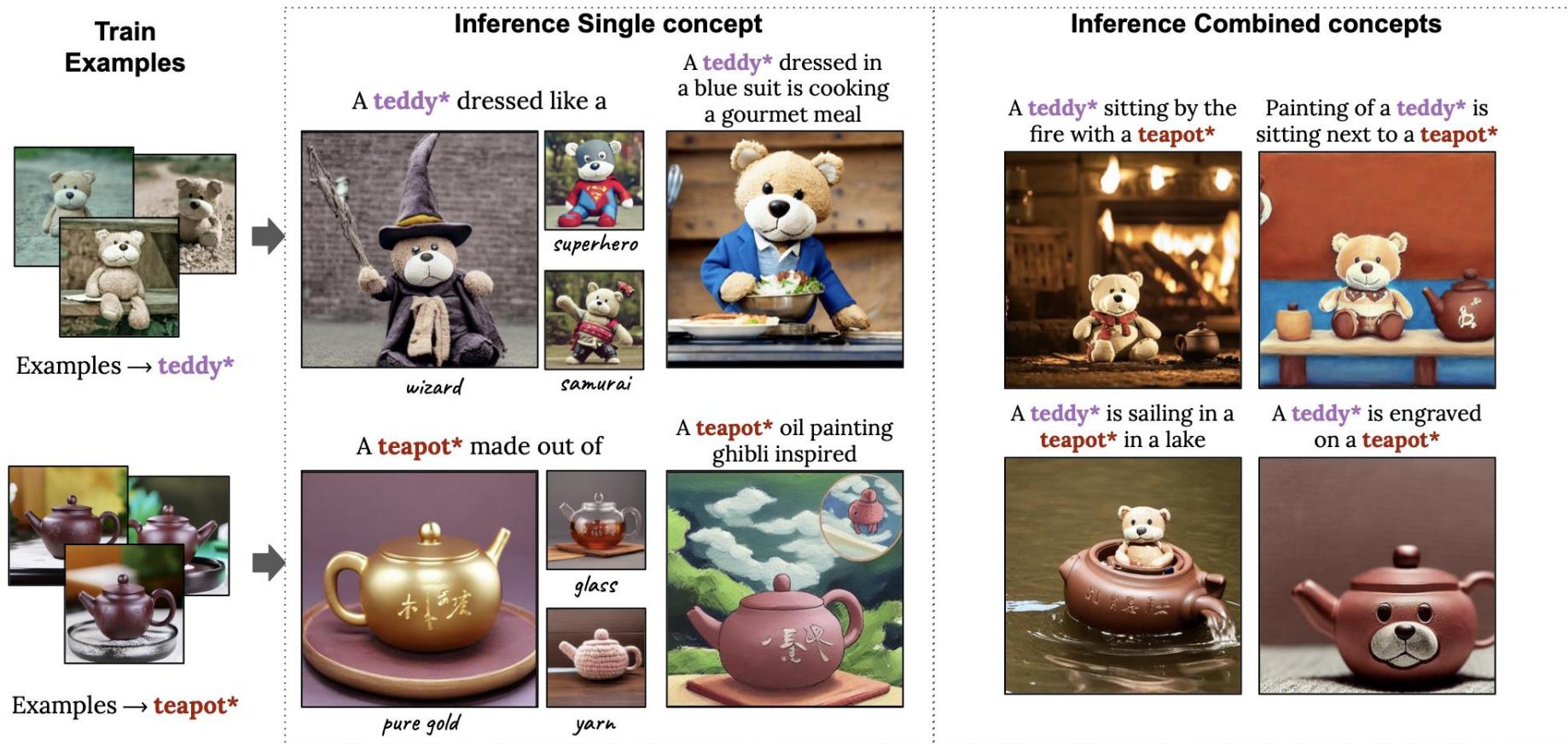
The  $V_1^*$  cat is sitting inside a  $V_2^*$  wooden pot and looking up

# Two concept results

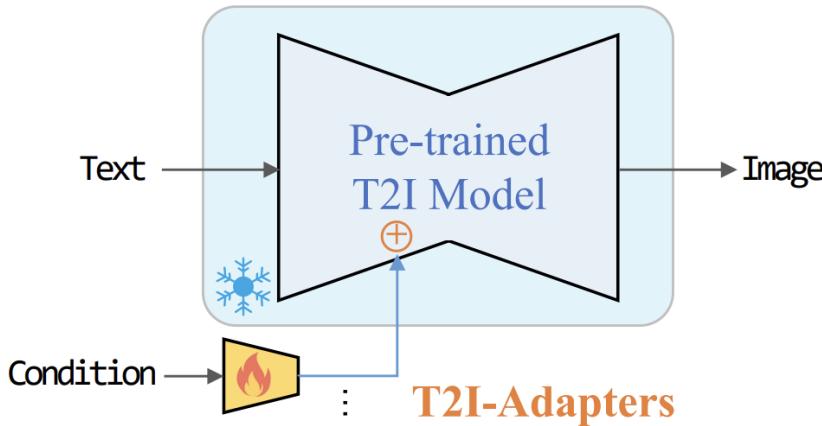


$V_1^*$  chair with the  $V_2^*$  cat sitting on it  
near a beach

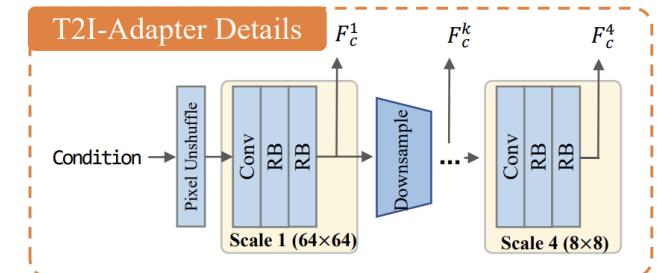
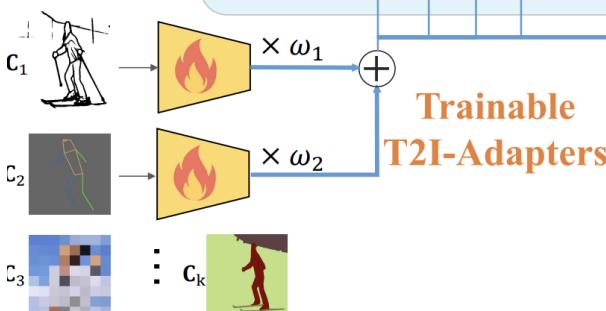
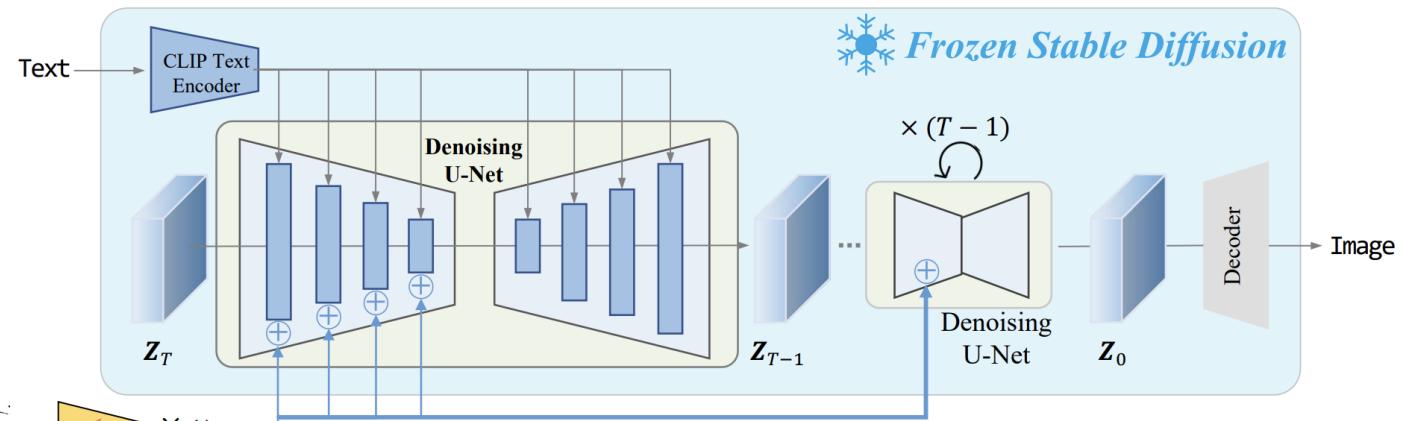
# Key-Locked Rank One Editing for Text-to-Image Personalization



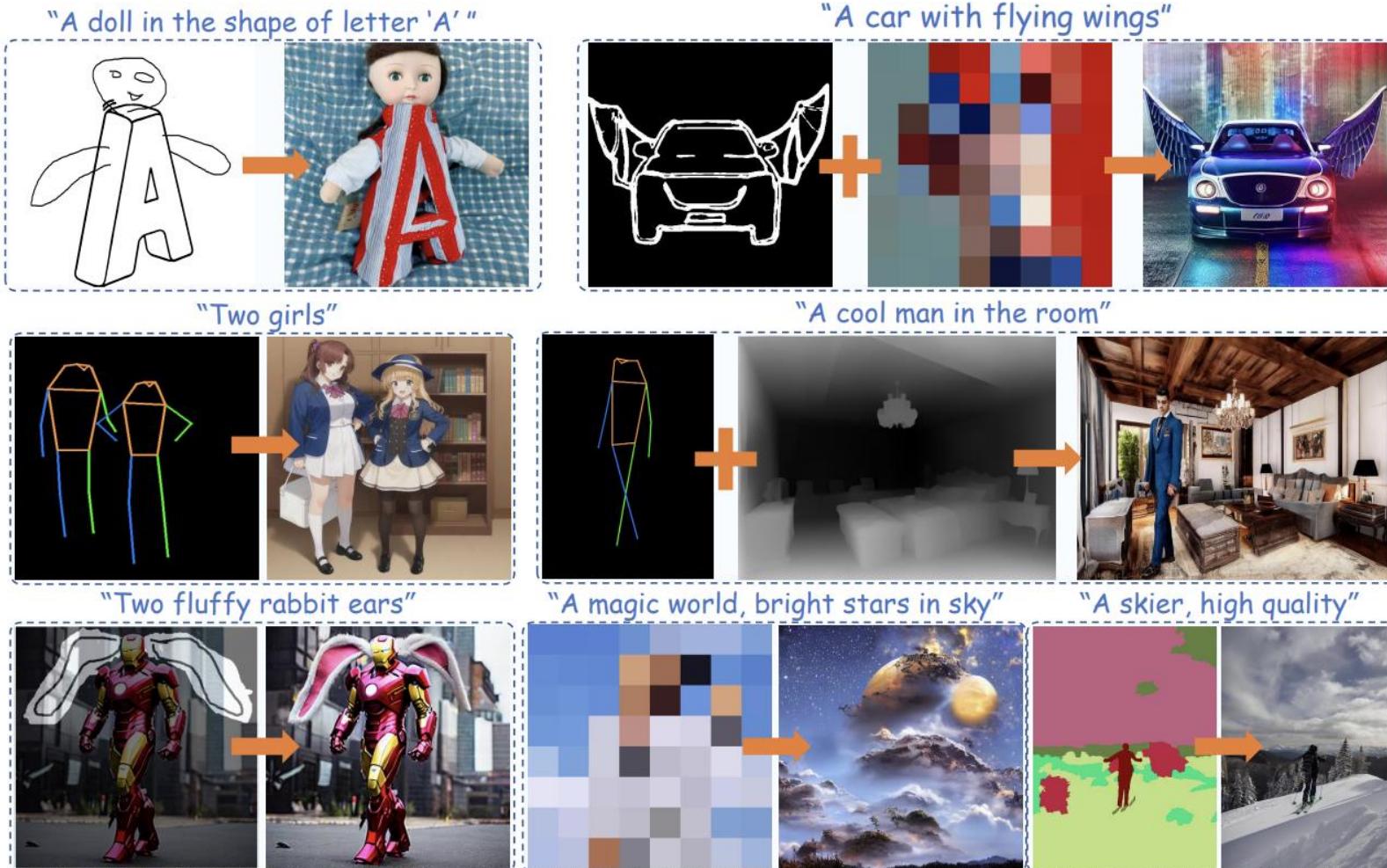
# T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models



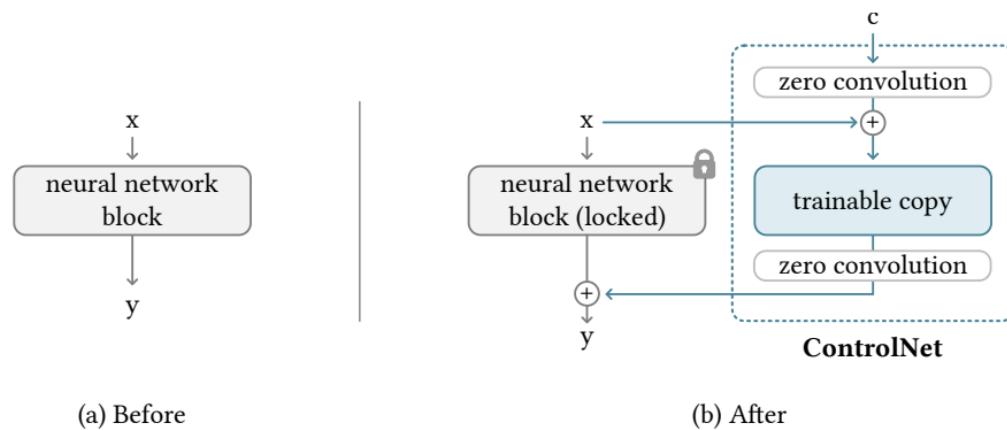
- ✓ **Plug-and-play.** Not affect original network topology and generation ability
- ✓ **Simple and small.** ~77M parameters and ~300M storage
- ✓ **Flexible.** Various adapters for different control conditions
- ✓ **Composable.** Several adapters to achieve multi-condition control
- ✓ **Generalizable.** Can be directly used on customized models



# T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models

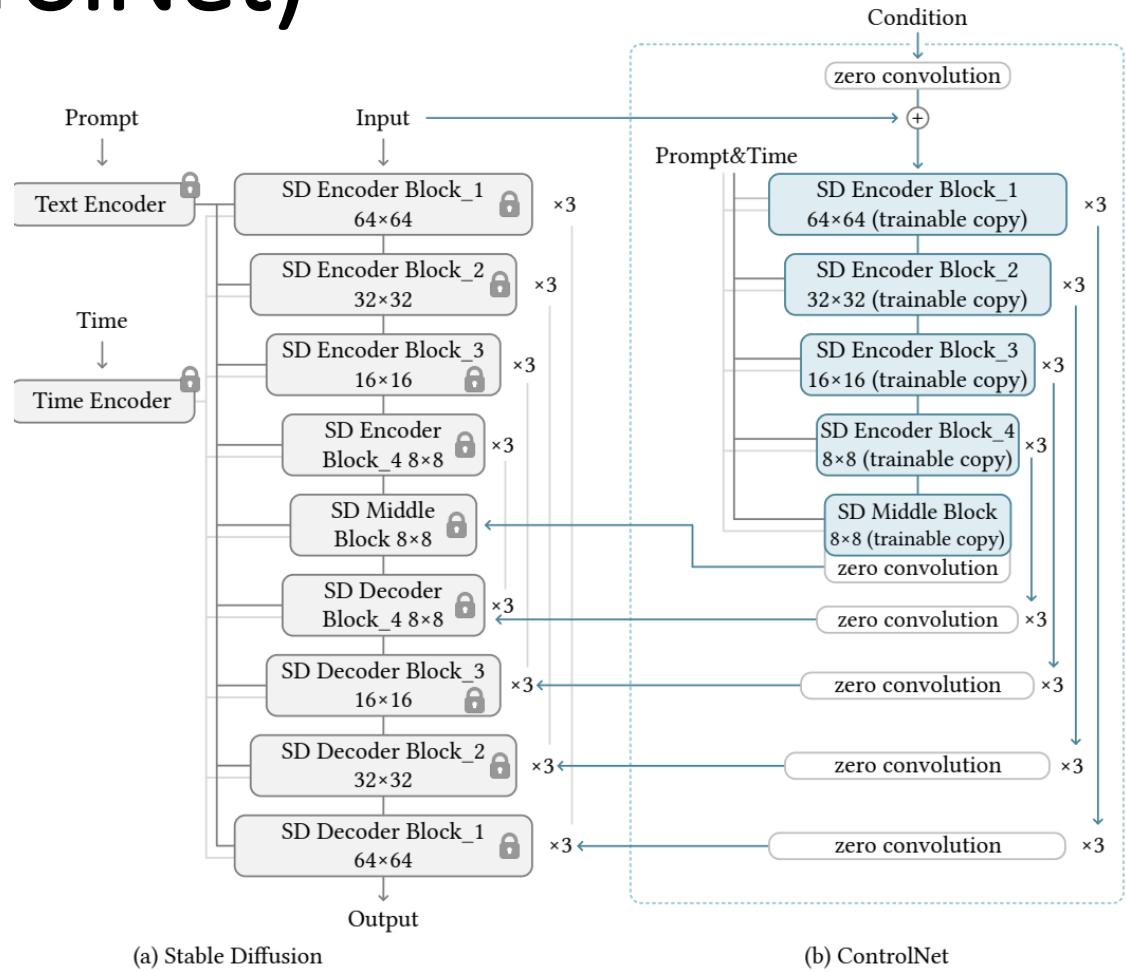
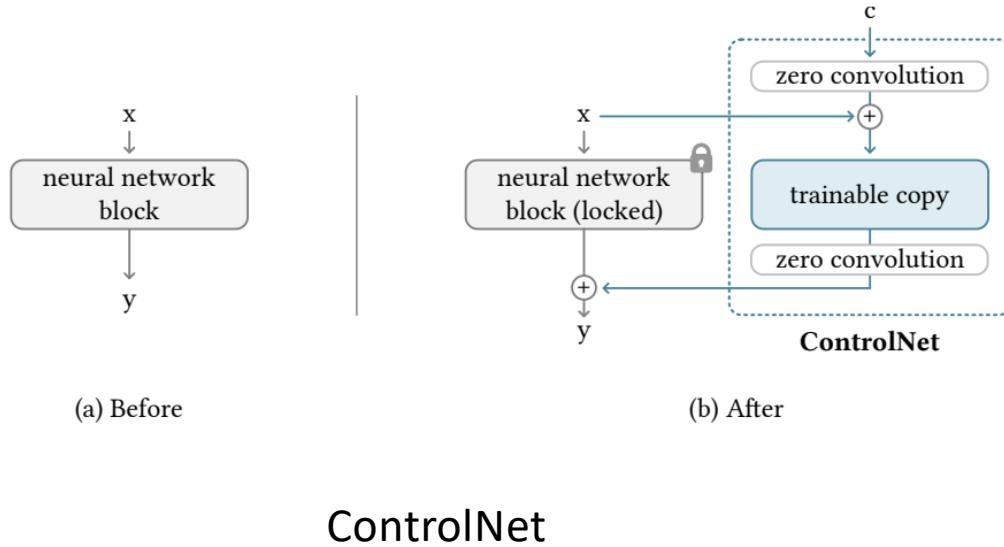


# Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)



ControlNet

# Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)



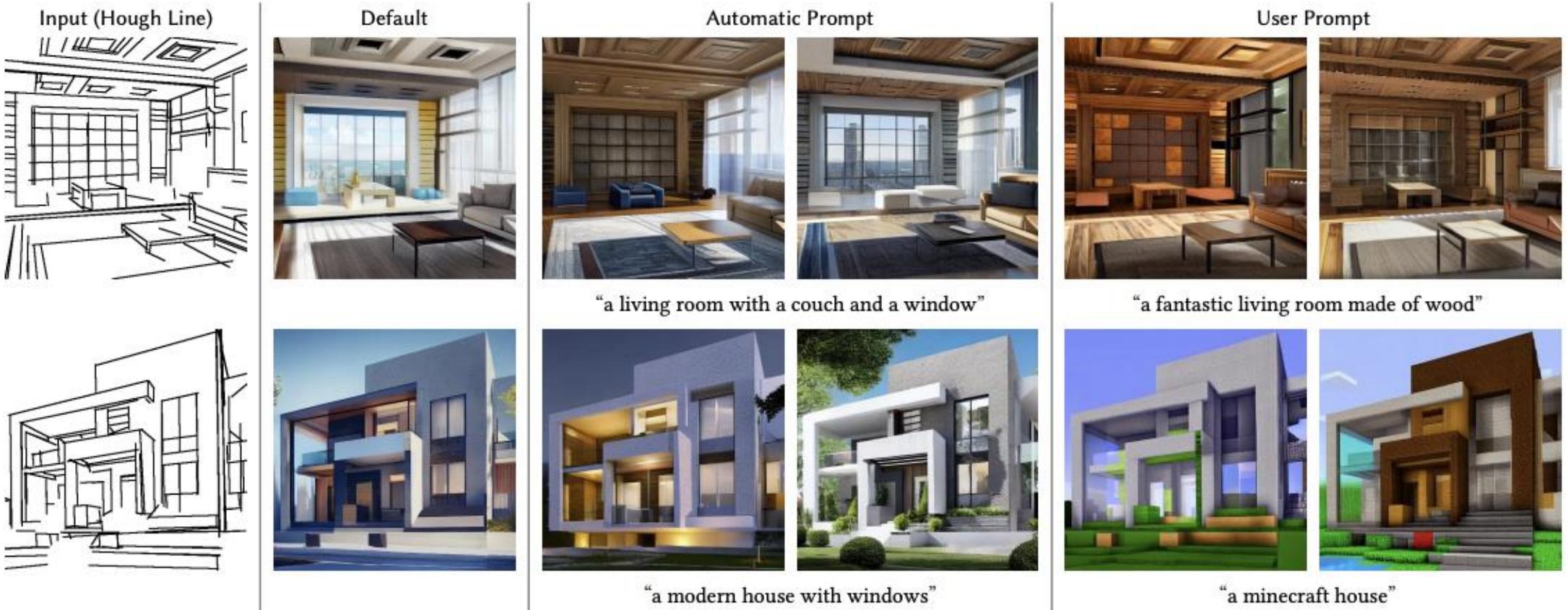
# Adding Conditional Control to Text-to-Image Diffusion Models

Train objective

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0, 1)} \left[ \|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2 \right]$$

where  $t$  is the time step,  $c_t$  is the text prompts,  $c_f$  is the task-specific conditions

# ControlNet

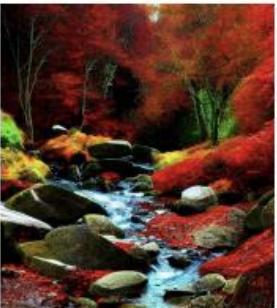
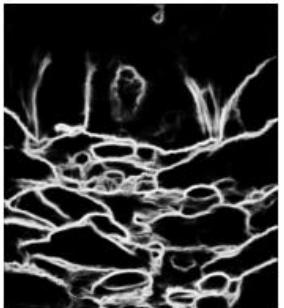


# ControlNet



"a bird on a branch of a tree"

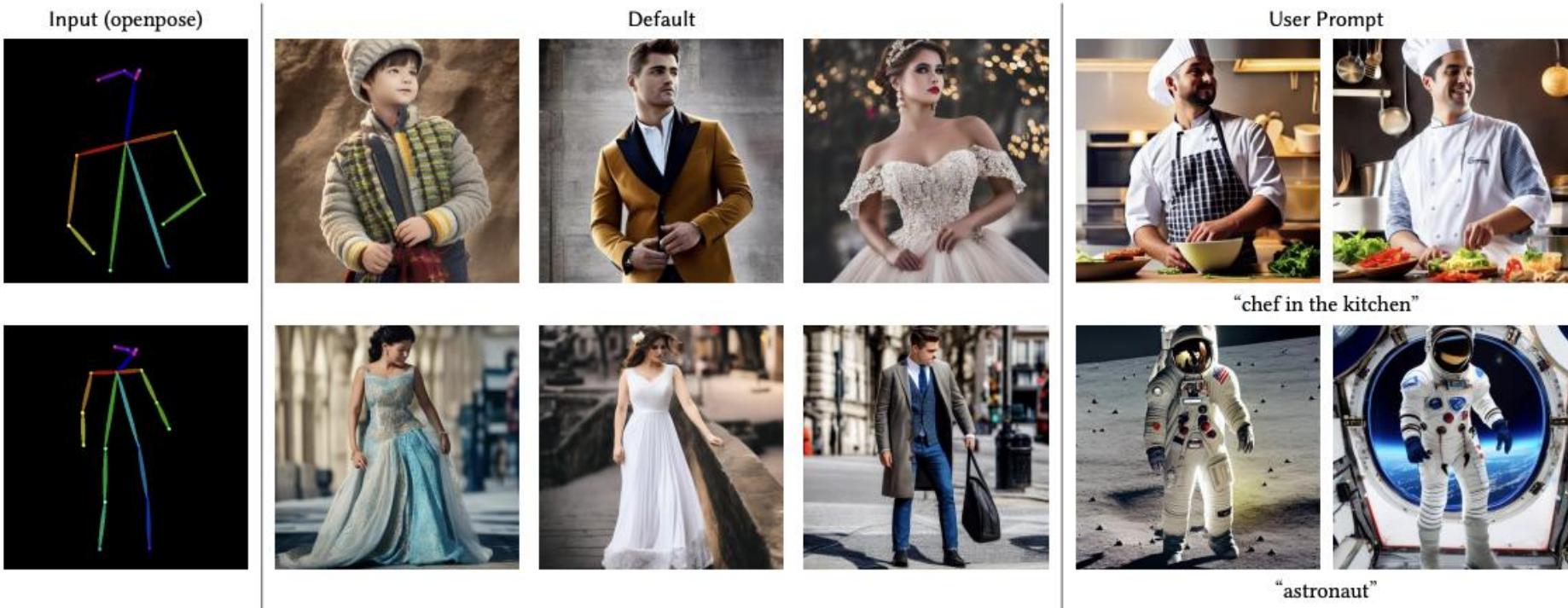
"white sparrow"



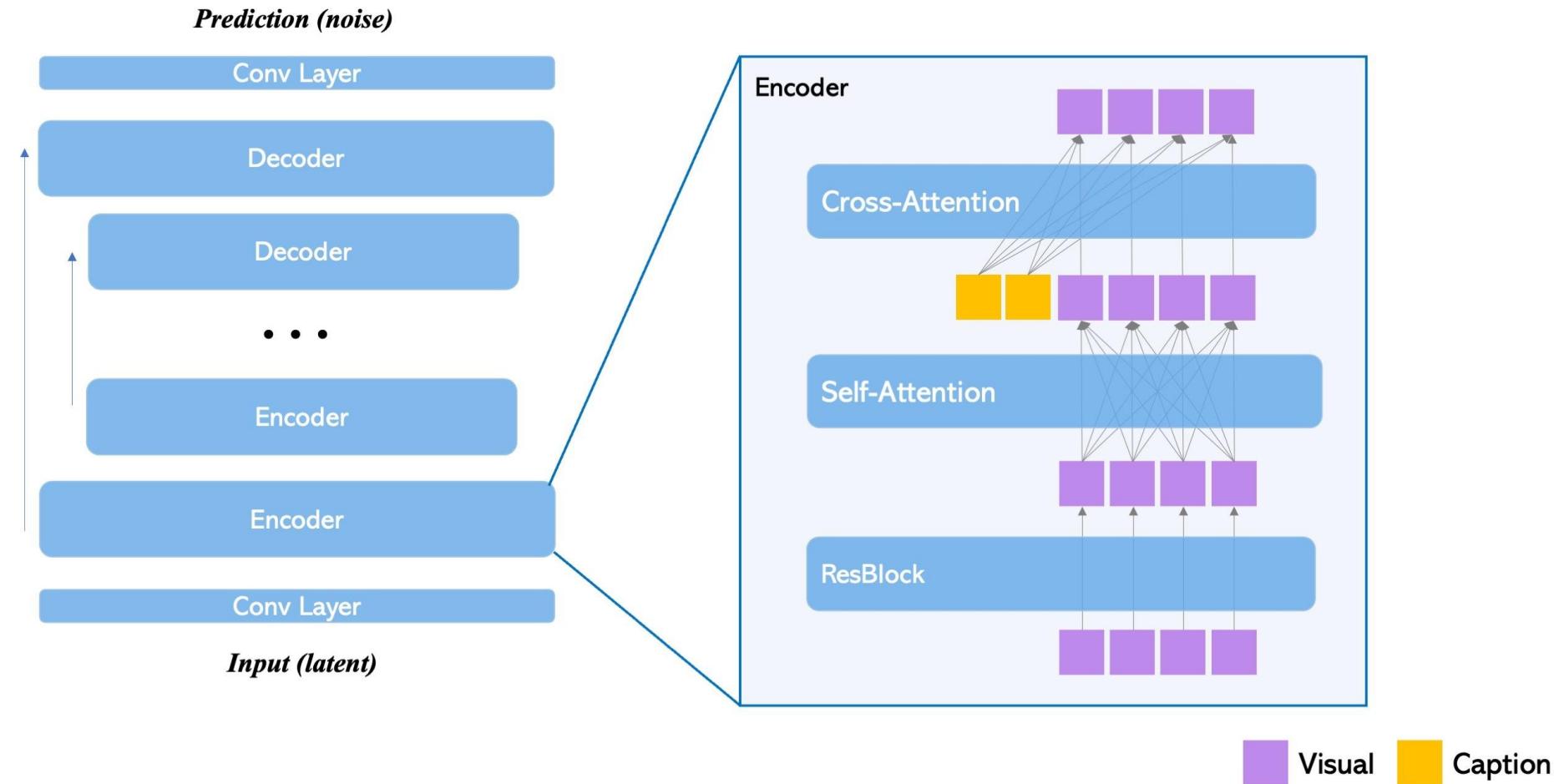
"a stream running through a forest"

"river in forest, winter, snow"

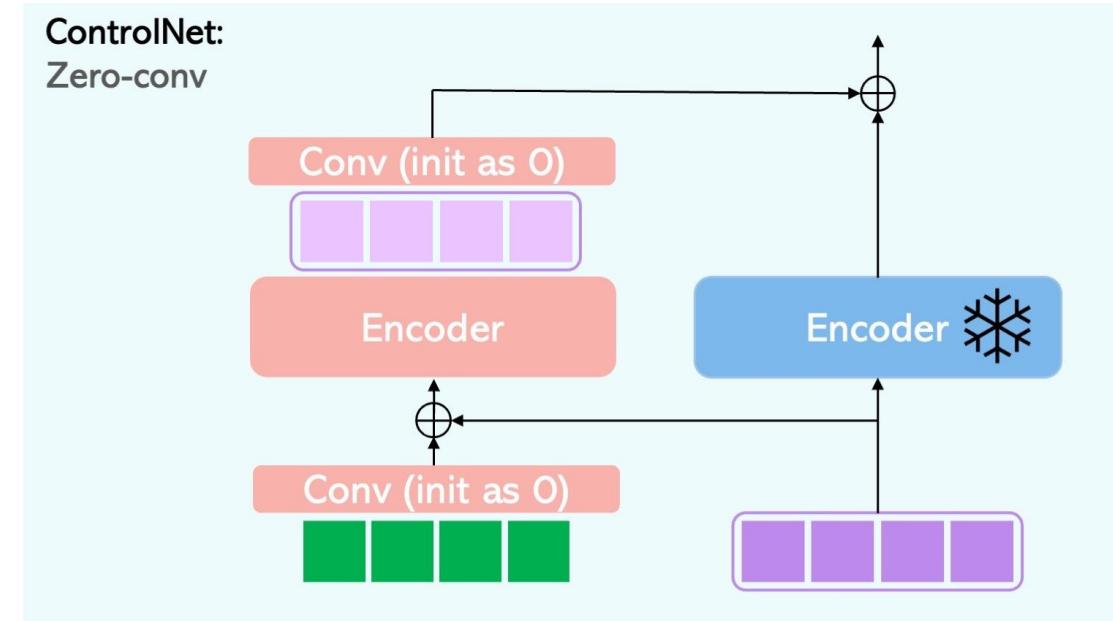
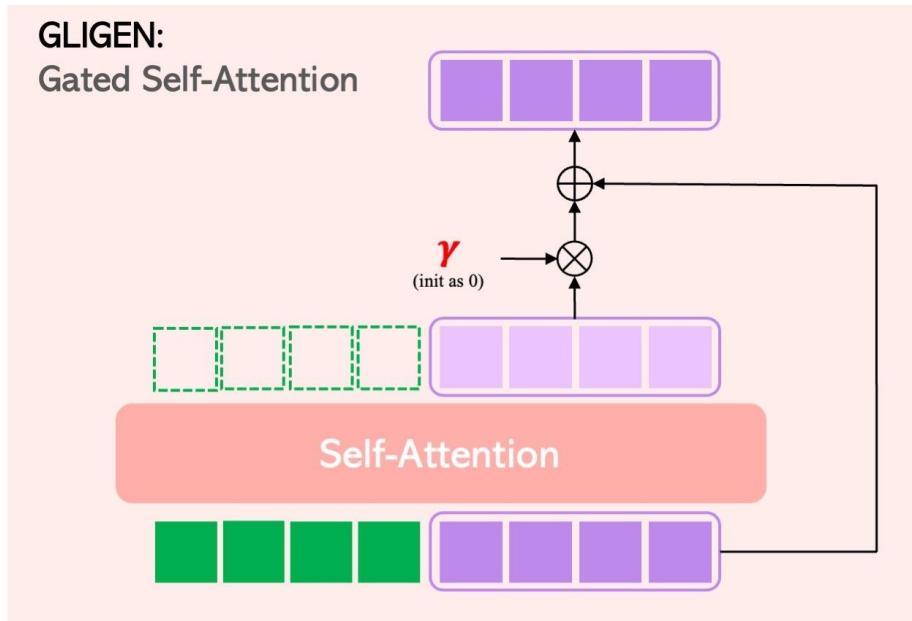
# ControlNet



# GLIGEN: Open-Set Grounded Text-to-Image Generation

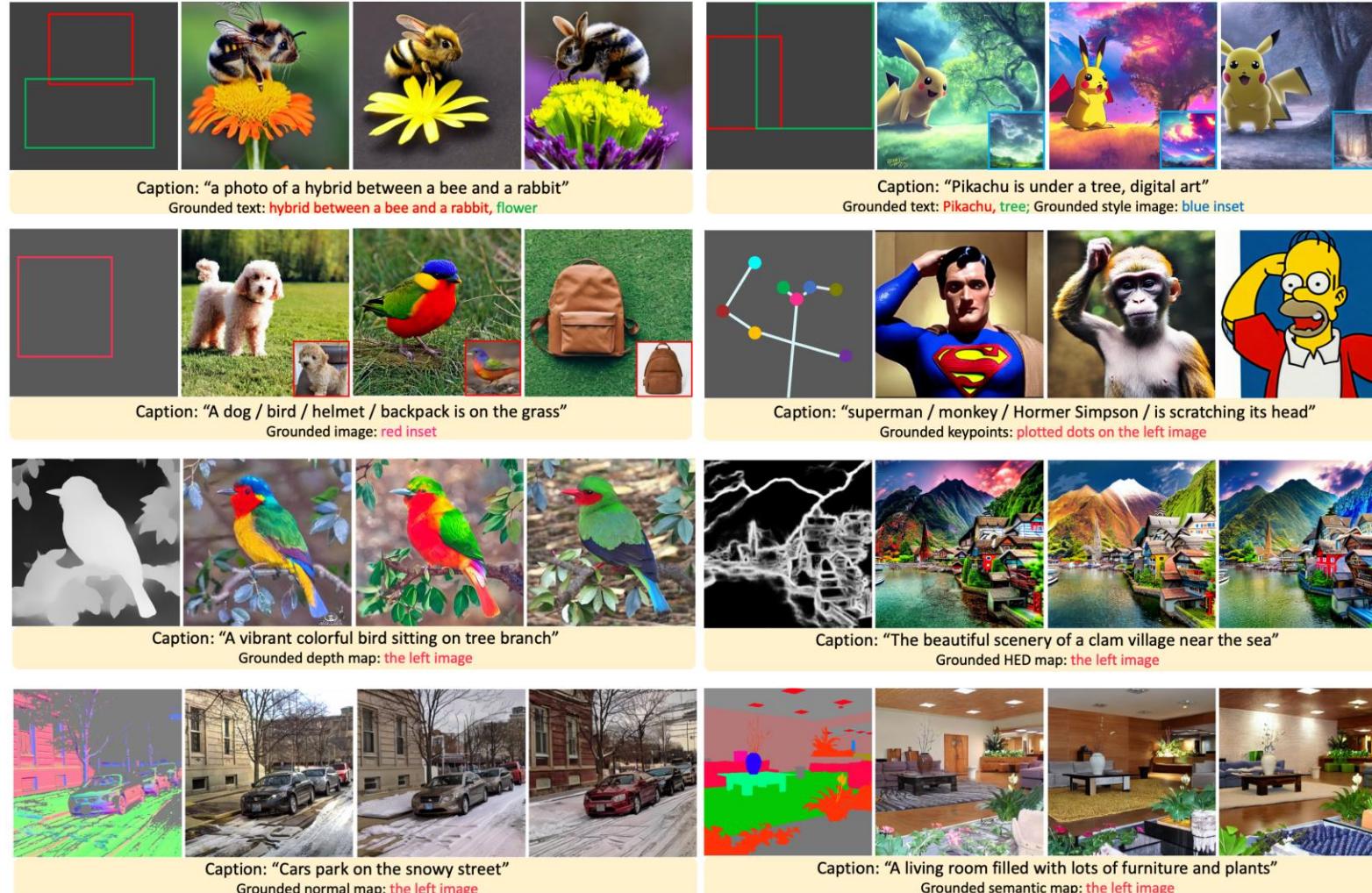


# GLIGEN: Open-Set Grounded Text-to-Image Generation



Trainable params      Grounding      Visual

# GLIGEN: Open-Set Grounded Text-to-Image Generation

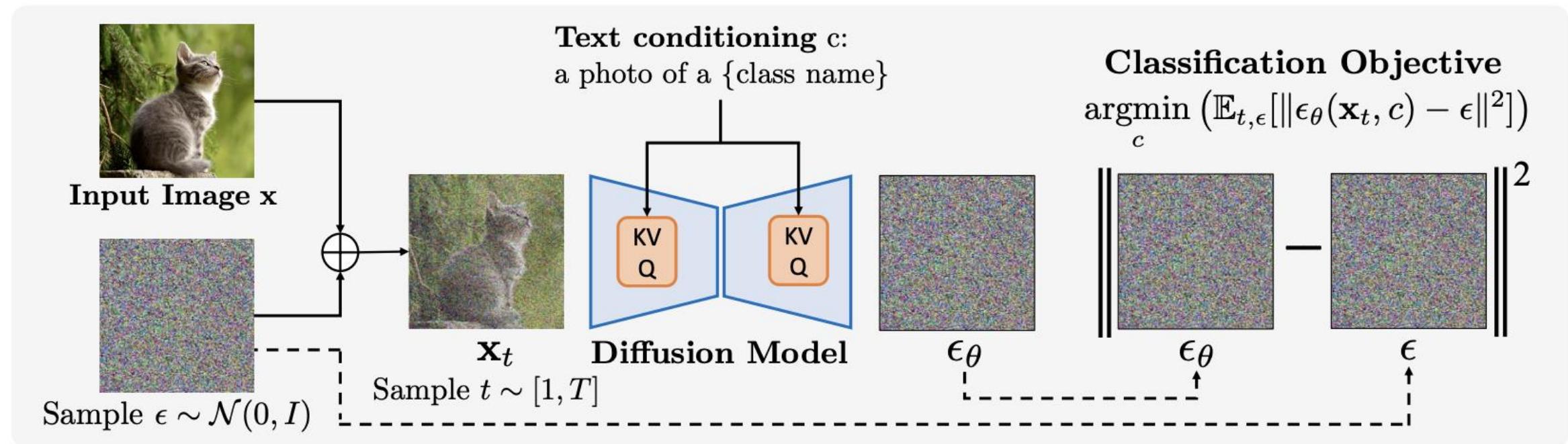


## Diverse form of input

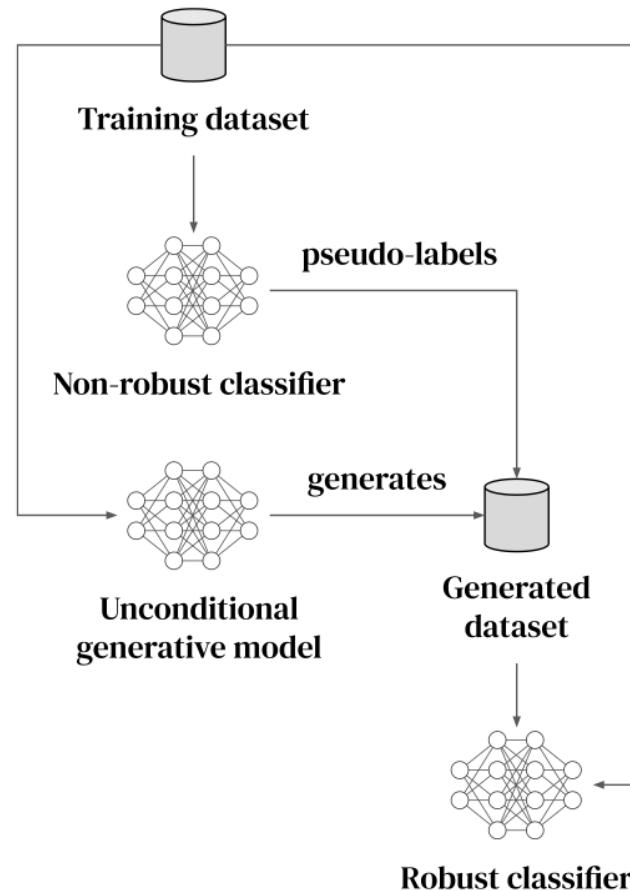
- Box+Text
- Box+Text+Image
- Keypoint
- Box+Text
- Box+Text+Image
- Hed map
- Canny map
- Depth map
- Semantic map
- Normal map

# Other applications

# Your Diffusion Model is Secretly a Zero-Shot Classifier



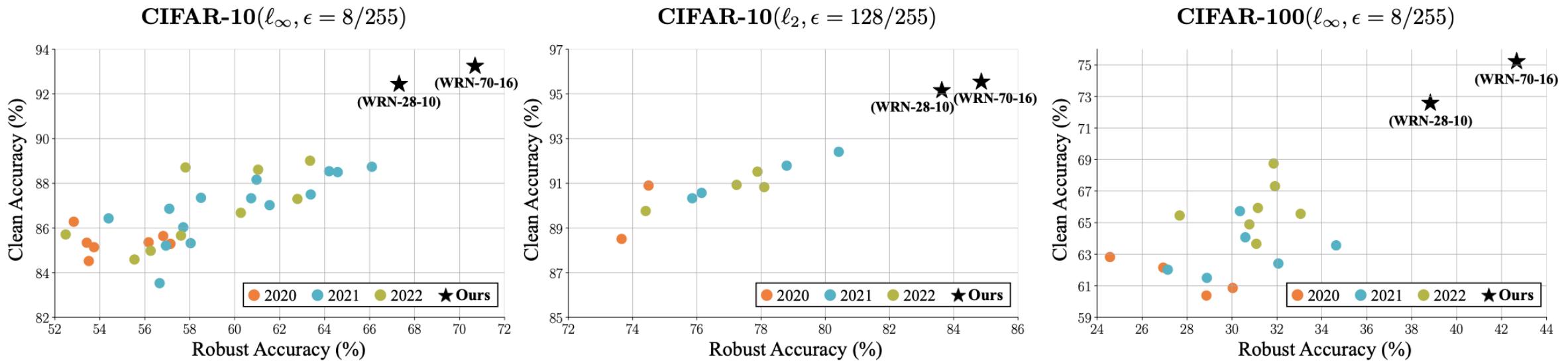
# Improving Robustness using Generated Data



## Overview of the approach:

1. train a generative model and a non-robust classifier, which are used to provide pseudo-labels to the generated data.
2. The generated and original training data are combined to train a robust classifier.

# Better Diffusion Models Further Improve Adversarial Training



# Acknowledgement

- Thanks everyone for suggesting the papers. Topics of the tutorial are based on <https://github.com/cvpr2023-tutorial-diffusion-models/papers>
- Part of the slides are adapted from Jun-Yan Zhu's talk on image customization.

# Reference

- Bao et al., "[All are Worth Words: a ViT Backbone for Score-based Diffusion Models](#)", arXiv 2022
- Peebles and Xie, "[Scalable Diffusion Models with Transformers](#)", arXiv 2022
- Bao et al., "[One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale](#)", arXiv 2023
- Jabri et al., "[Scalable Adaptive Computation for Iterative Generation](#)", arXiv 2022
- Hoogeboom et al., "[simple diffusion: End-to-end diffusion for high resolution images](#)", arXiv 2023
- Meng et al., "[SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations](#)", ICLR 2022
- Li et al., "[Efficient Spatially Sparse Inference for Conditional GANs and Diffusion Models](#)", NeurIPS 2022
- Avrahami et al., "[Blended Diffusion for Text-driven Editing of Natural Images](#)", CVPR 2022
- Hertz et al., "[Prompt-to-Prompt Image Editing with Cross-Attention Control](#)", ICLR 2023
- Kawar et al., "[Imagic: Text-Based Real Image Editing with Diffusion Models](#)", CVPR 2023
- Couairon et al., "[DiffEdit: Diffusion-based semantic image editing with mask guidance](#)", ICLR 2023

- Sarukkai et al., "[Collage Diffusion](#)", arXiv 2023
- Bar-Tal et al., "[MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation](#)", ICML 2023
- Gal et al., "[An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion](#)", ICLR 2023
- Ruiz et al., "[DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation](#)", CVPR 2023
- Kumari et al., "[Multi-Concept Customization of Text-to-Image Diffusion](#)", CVPR 2023
- Tewel et al., "[Key-Locked Rank One Editing for Text-to-Image Personalization](#)", SIGGRAPH 2023
- Zhao et al., "[A Recipe for Watermarking Diffusion Models](#)", arXiv 2023
- Hu et al., "[LoRA: Low-Rank Adaptation of Large Language Models](#)", ICLR 2022
- Li et al., "[GLIGEN: Open-Set Grounded Text-to-Image Generation](#)", CVPR 2023
- Avrahami et al., "[SpaText: Spatio-Textual Representation for Controllable Image Generation](#)", CVPR 2023
- Zhang and Agrawala, "[Adding Conditional Control to Text-to-Image Diffusion Models](#)", arXiv 2023

- Mou et al., "[T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models](#)", arXiv 2023
- Orgad et al., "[Editing Implicit Assumptions in Text-to-Image Diffusion Models](#)", arXiv 2023
- Han et al., "[SVDiff: Compact Parameter Space for Diffusion Fine-Tuning](#)", arXiv 2023
- Xie et al., "[DiffFit: Unlocking Transferability of Large Diffusion Models via Simple Parameter-Efficient Fine-Tuning](#)", arXiv 2023
- Saharia et al., "[Palette: Image-to-Image Diffusion Models](#)", SIGGRAPH 2022
- Whang et al., "[Deblurring via Stochastic Refinement](#)", CVPR 2022
- Xu et al., "[Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models](#)", arXiv 2023
- Saxena et al., "[Monocular Depth Estimation using Diffusion Models](#)", arXiv 2023
- Li et al., "[Your Diffusion Model is Secretly a Zero-Shot Classifier](#)", arXiv 2023
- Gowal et al., "[Improving Robustness using Generated Data](#)", NeurIPS 2021
- Wang et al., "[Better Diffusion Models Further Improve Adversarial Training](#)", ICML 2023