CENTER FOR RESEARCH IN COMPUTER VISION

Dr. Chen Chen
Associate Professor
Center for Research in Computer Vision
& Department of Computer Science
University of Central Florida
chen.chen@crcv.ucf.edu
https://www.crcv.ucf.edu/chenchen/

Research Overview

A team of 8 Ph.D. students, 1 MS in CV student, and 2 UG students

Computer Vision

- Object detection and tracking
- Action detection and recognition
- Human 2d/3d pose estimation
- Image semantic segmentation
- Image restoration
- 3D Vision

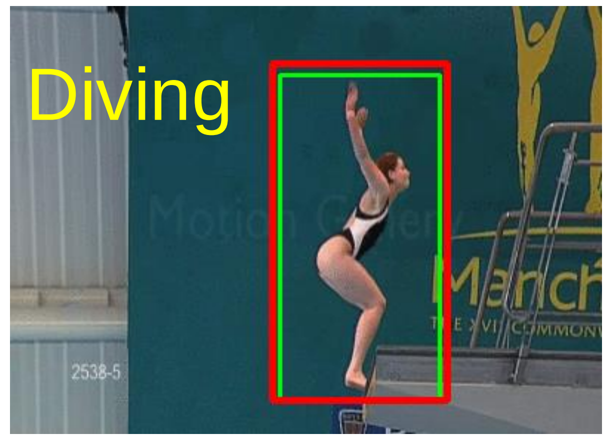
Machine Learning

- Efficient machine learning (computation-, label-, data-efficiency)
- Federated learning
- Multimodal learning
- GenAI

Applications

- Healthcare, medicine
- Remote sensing
- Smart agriculture

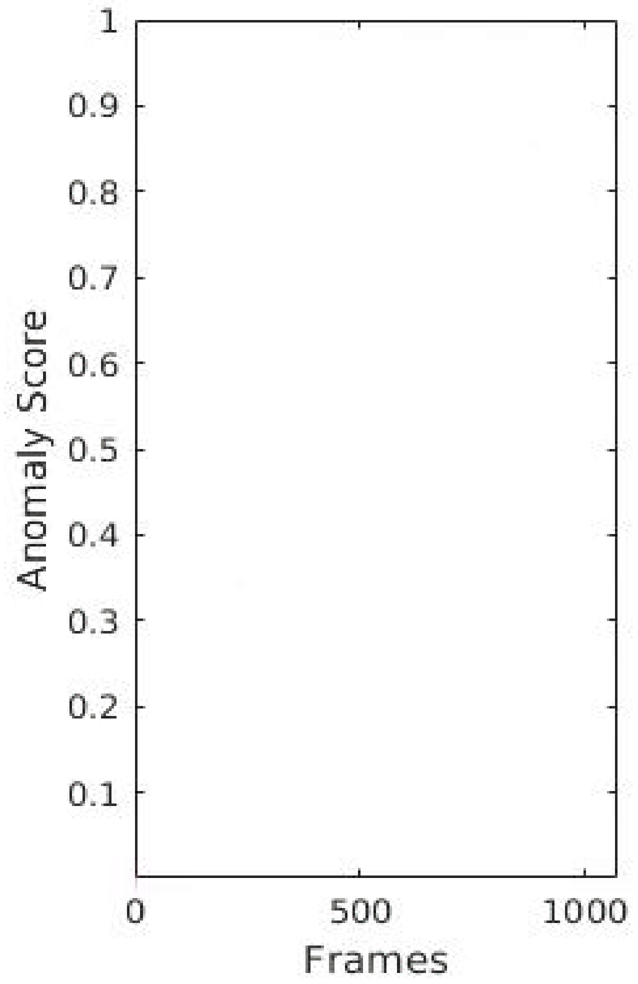
Action detection



Red: Our detection Green: Ground

Video object segmentation

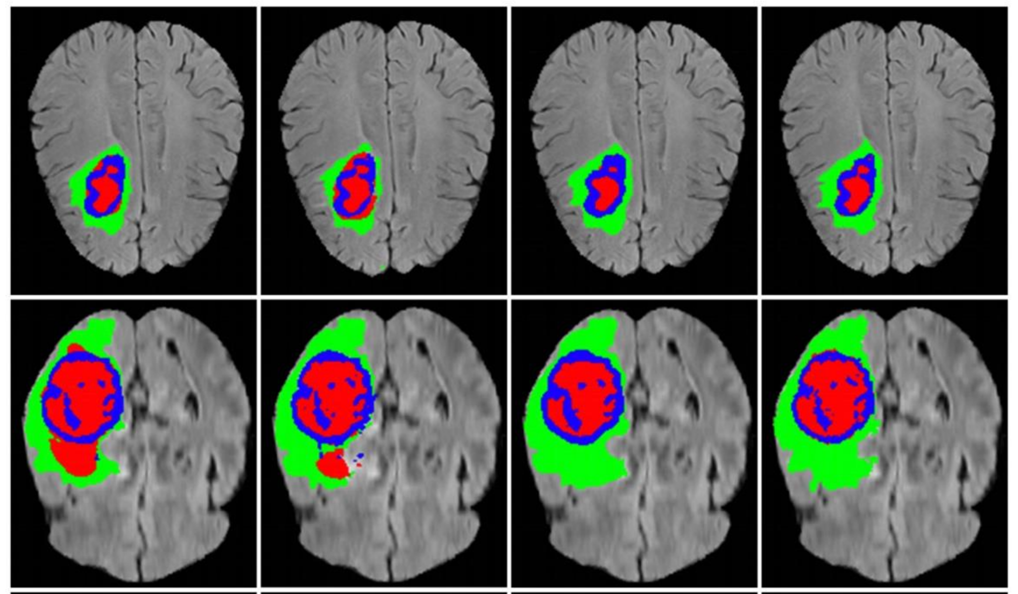
Video anomaly detection



3D human pose/mesh reconstruction

Low-level vision (image enhancement, denoising, super-resolution, etc.)

Medical image computing



3D rendering (NeRF/Gaussian Splatting)

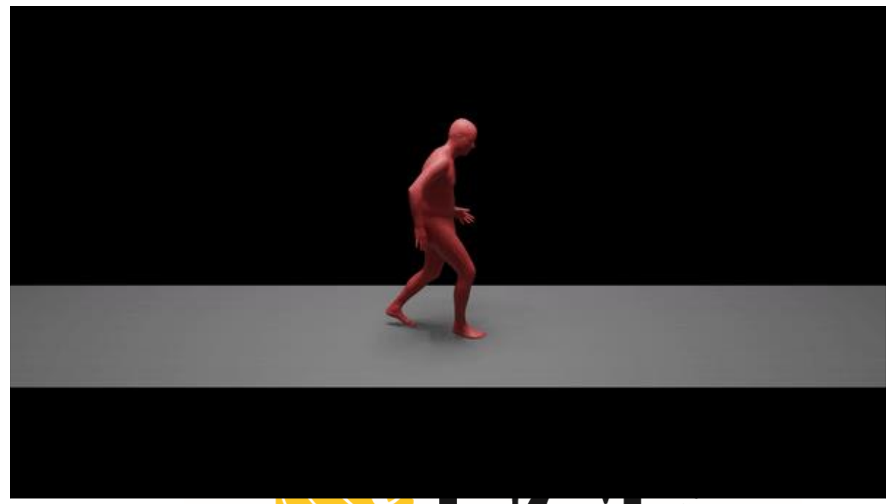
Input: A lion is roaring on the rock

Text driven video editing

Edit: ~~A~~ lion **tiger** is roaring on the rock

Text to Motion

"a person walks forward then turns completely around and does a cartwheel"



ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback

Ming Li[1], Taojiannan Yang[1], Huafeng Kuang[2], Jie Wu[2], Zhaoning Wang[1], Xuefeng Xiao[2], and Chen Chen[1]

[1]University of Central Florida, [2]TikTok, ByteDance Inc

ECCV 2024

https://liming-ai.github.io/ControlNet_Plus_Plus

Outline

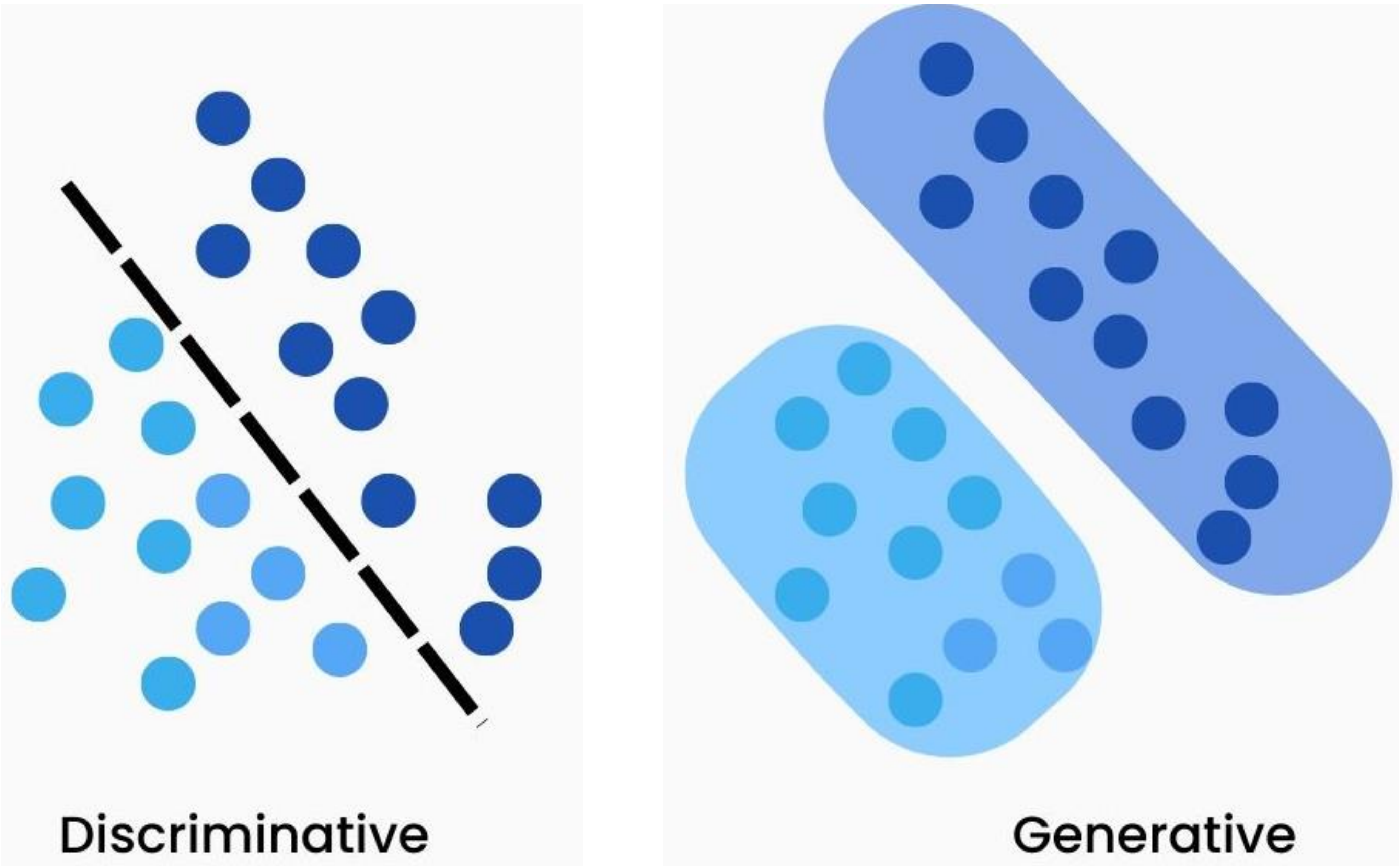
- **Background: Generative Learning for Images**
- Motivation: Do existing methods achieve good controllability?
- Method: Efficient Consistency Feedback
- Experiments: Better Controllability Without Loss of Image Quality and Text Guidance
- Future Plans: More Conditions & Text-to-Image Models; Scaling Up

Discriminative vs. Generative Models

- Generative artificial intelligence (generative AI or GenAI) is artificial intelligence capable of generating text, images, or other media, using generative models.

- The majority of discriminative models aim to separate the data points into different classes and learning the boundaries using probability estimates and maximum likelihood.

- Generative models model the actual data distribution and learn the different data points, rather than model just the decision boundary between classes.

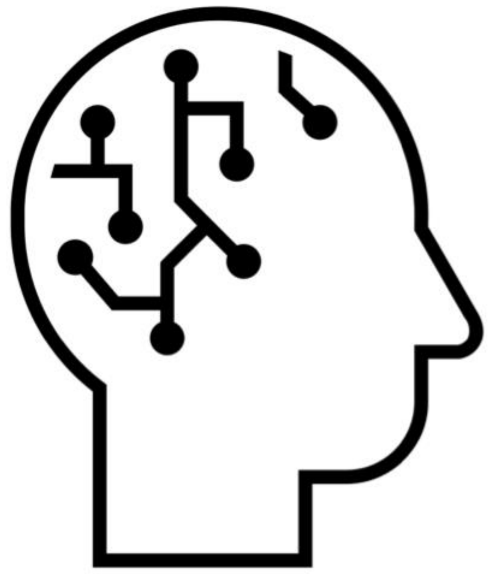


Deep Generative Learning for Image

Learning to generate data

Samples from a Data Distribution

Train



Neural Network

Sample

Application

Art & Design

Content Generation

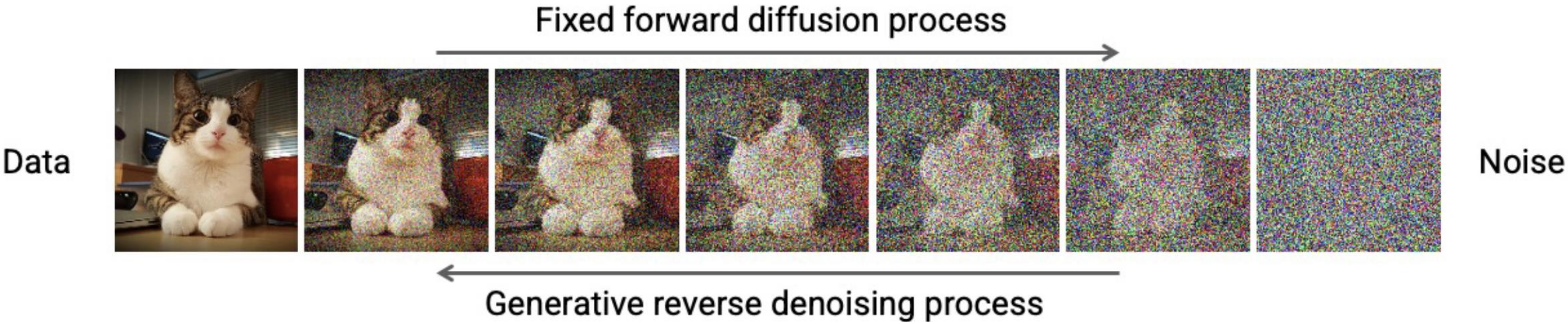
Representation Learning

Entertainment

Diffusion Model

Diffusion models consist of two processes:

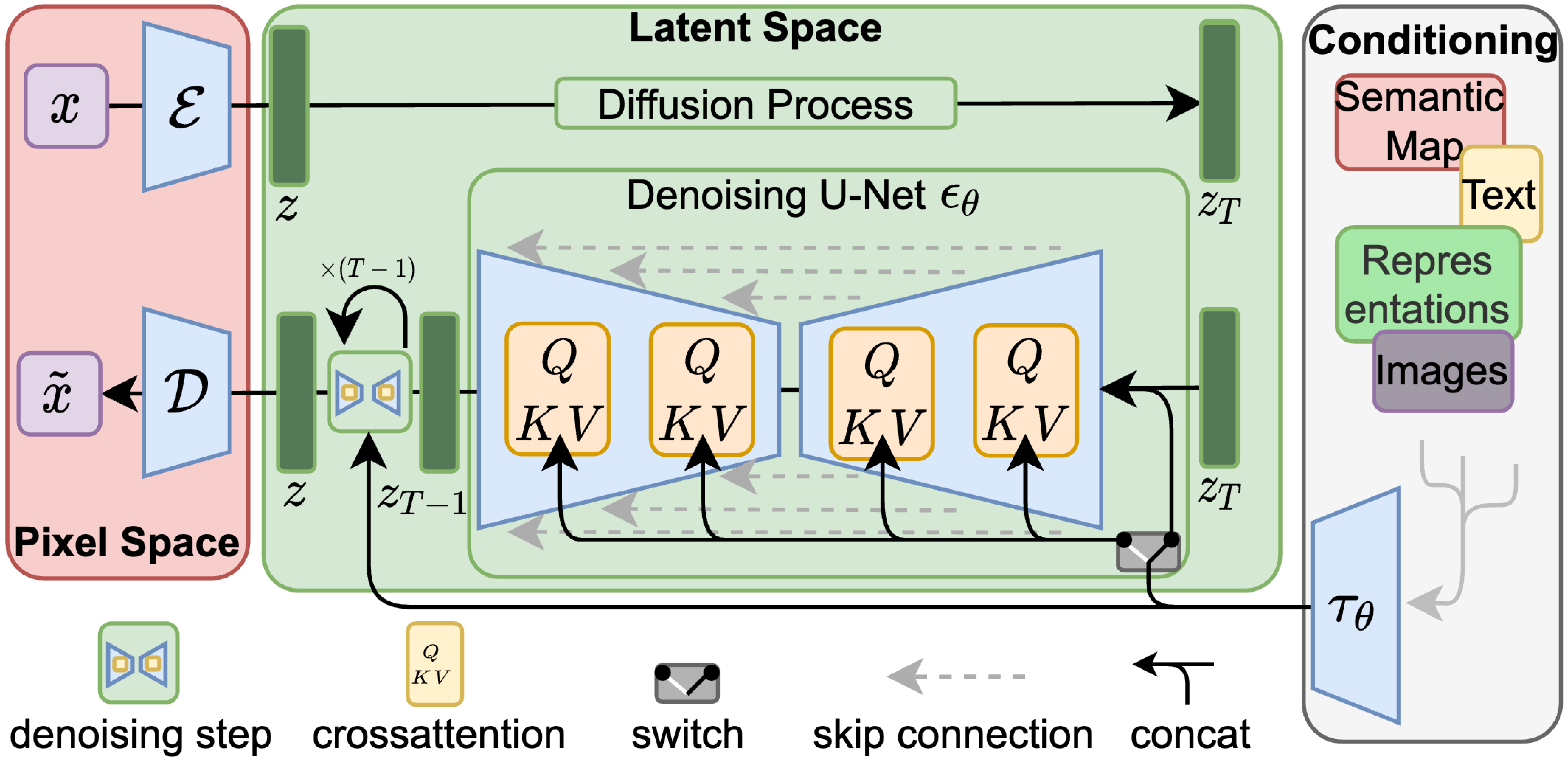
- **Forward diffusion process that gradually adds noise to input**
- **Reverse denoising process that learns to generate data by denoising**



The model learns the reverse of the diffusion process, predicting the distribution of the previous step given the current noisy data.

Latent Diffusion Model (Stable Diffusion)

- Pixel-space Diffusion is too computationally expensive
- We use VAE to map it to latent-space and then perform the Diffusion process



The VAE encoder ϵ map images from pixel-space to latent-space, the denoised image latents will be map back into pixel-space with VAE decoder $\mathcal{D}$

High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

Text-to-Image Diffusion Models

- Adding control over image generation is crucial for the practical application.
- Thanks to large-scale text-image datasets, existing diffusion models are well trained to perform image generation with given text prompt as control signals.

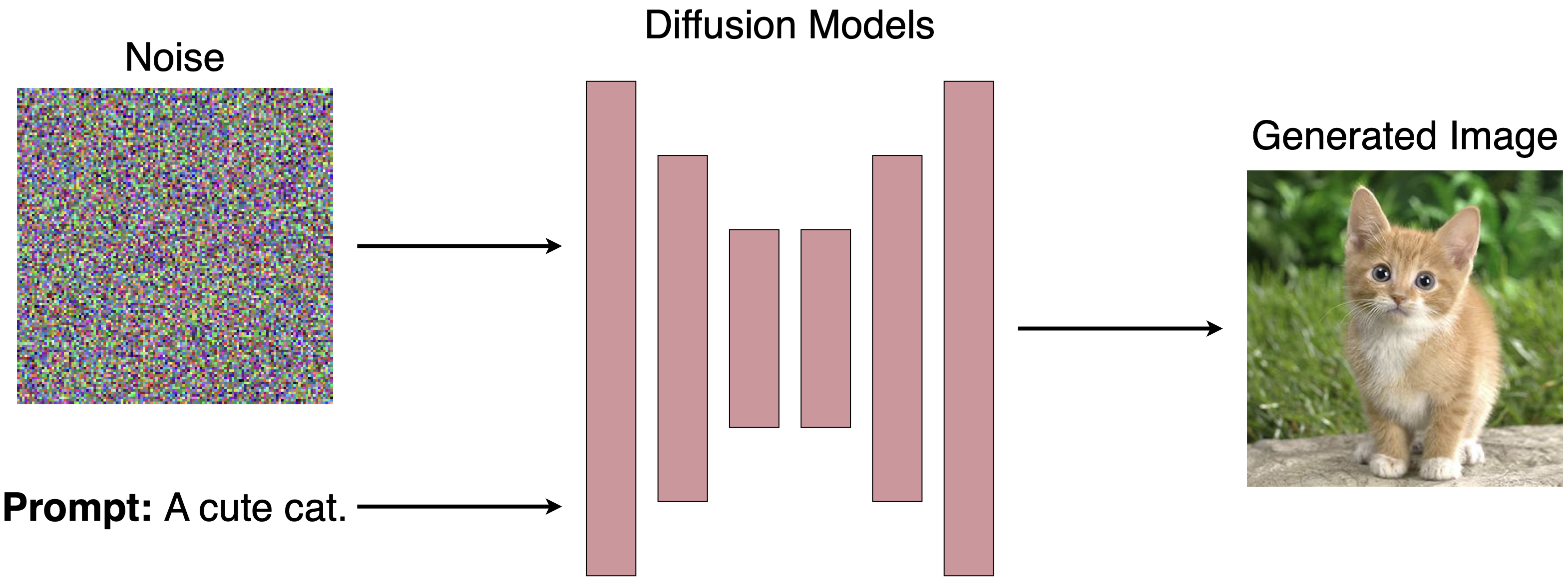


Image Source: https://hanlab.mit.edu/projects/can

Control Image Generation with Text is NOT Enough

- An image is worth a thousand words. It's hard to describe an image with language.

Overall content

The image depicts a majestic deer standing on a grassy and slightly elevated terrain. The deer has a robust body and carries an impressive set of antlers. The background features a misty, mountainous landscape, adding a sense of depth and natural beauty to the scene. The overall ambiance of the image evokes a sense of tranquility and the beauty of wildlife in its natural habitat.

Object properties

1. **Deer**: A large, robust deer with an impressive set of antlers, standing on a grassy and slightly elevated terrain.
2. **Terrain**: The ground is covered with grass and small shrubs, typical of a natural, hilly landscape.
3. **Background**: The background consists of misty mountains, adding depth and a sense of wilderness to the scene.

It's hard to describe:

- **How is the aesthetic of this image?**
- **What the details, textures, and contours of the image look like?**
- **What the location, pose, material, quantity, and size of each object?**

Control Image Generation with Text is NOT Enough

- Even with very detailed text descriptions, existing text-to-image diffusion models still cannot achieve controllable generation based on the given text control signals.

SDXL DALL-E 3

Prompt: a black dog sitting between a bush and a pair of green pants standing up with nobody inside them

SDXL DALL-E 3

Prompt: a spaceship that looks like the Sydney Opera House

SDXL DALL-E 3

Prompt: a panda bear with aviator glasses on its head

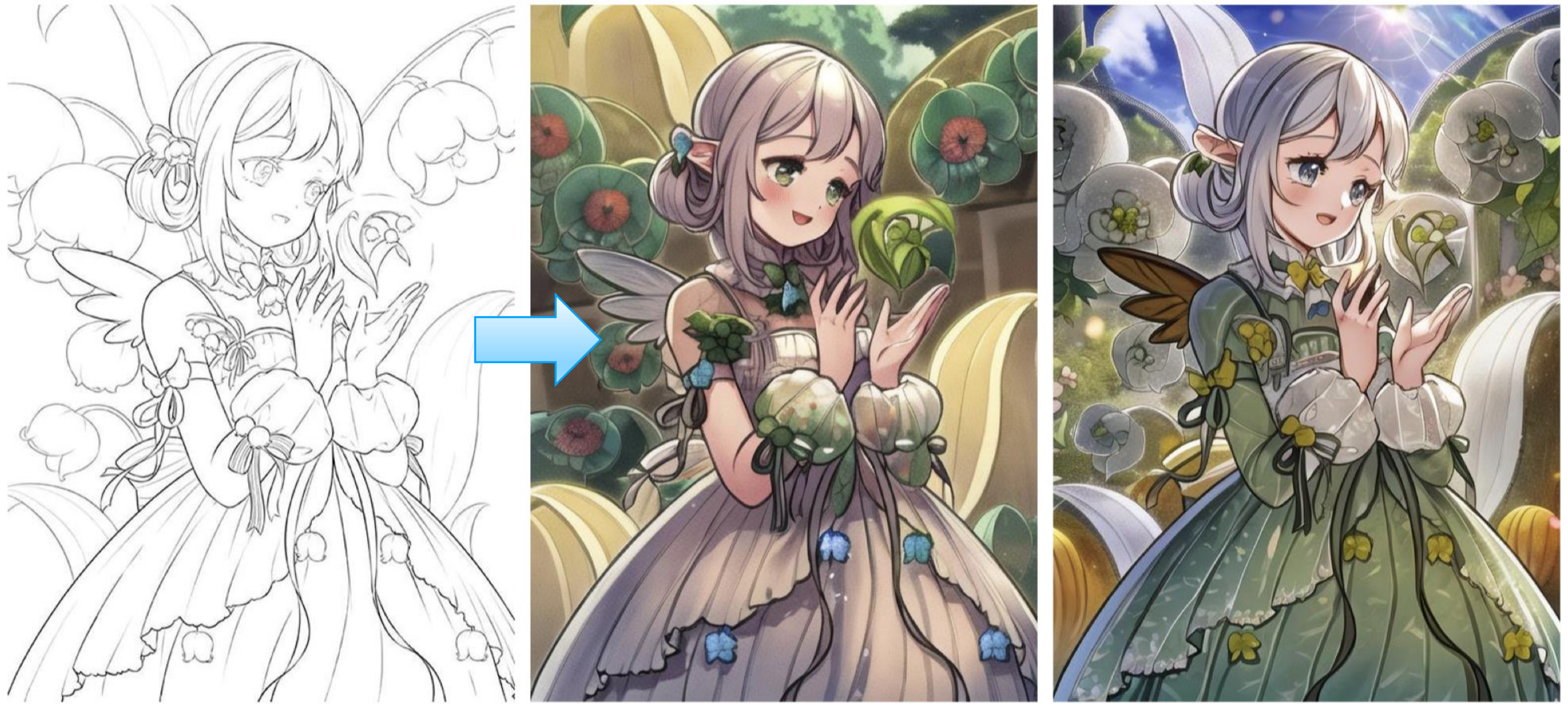
SDXL DALL-E 3

Prompt: An intricately detailed oil painting depicts a raccoon dressed in a black suit with a crisp white shirt and a red bow tie. The raccoon stands upright, donning a black top hat and gripping a wooden cane in one paw, while the other paw clutches a dark garbage bag. The background of the painting features soft, brush-stroked trees and mountains, reminiscent of traditional Chinese landscapes, with a delicate mist enveloping the scene.

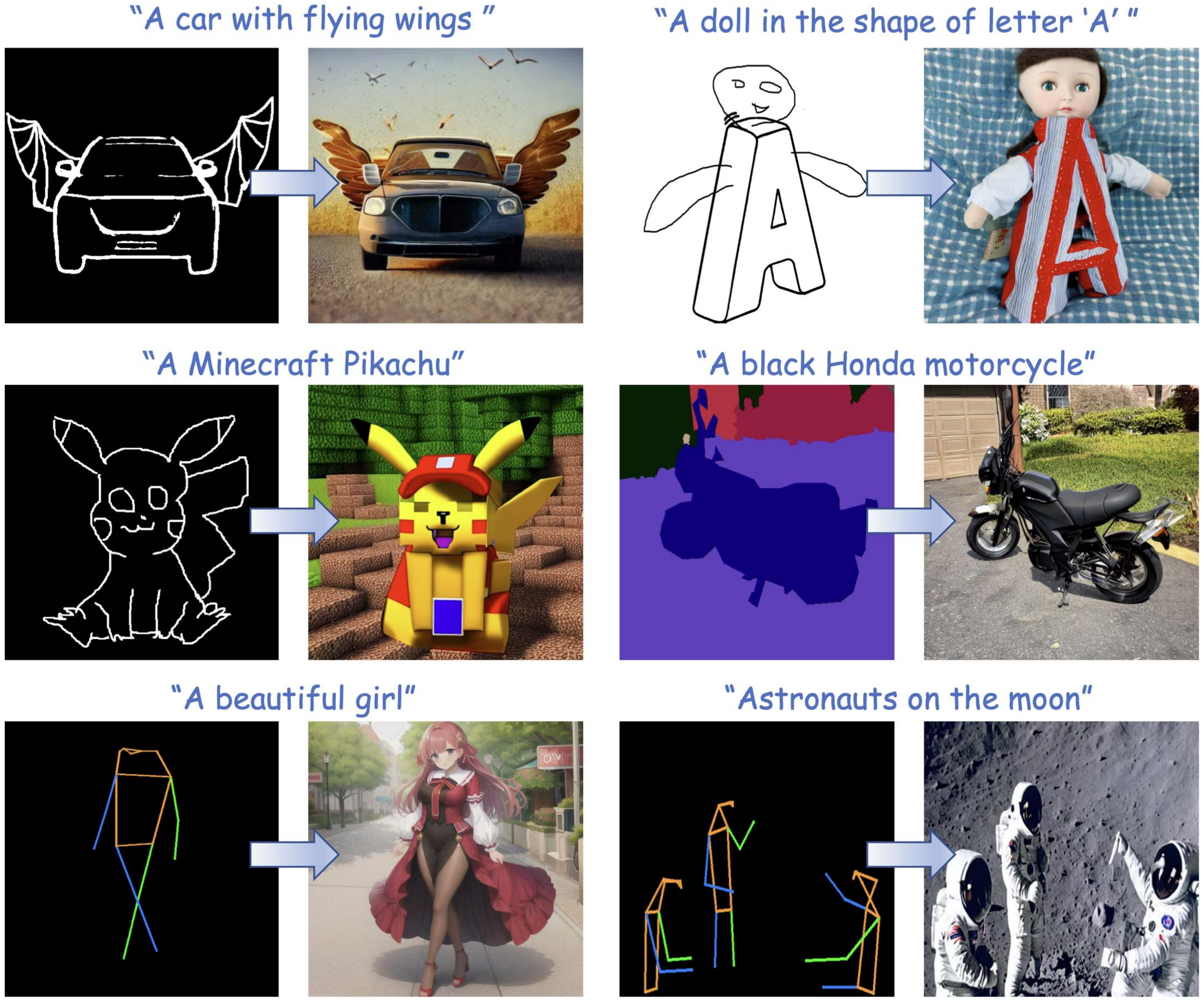
ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment, arXiv 2024

Adding Image Controls Signals for Image Generation

Normal map “Yharnam, the fictional city comes from a 2015 video game”

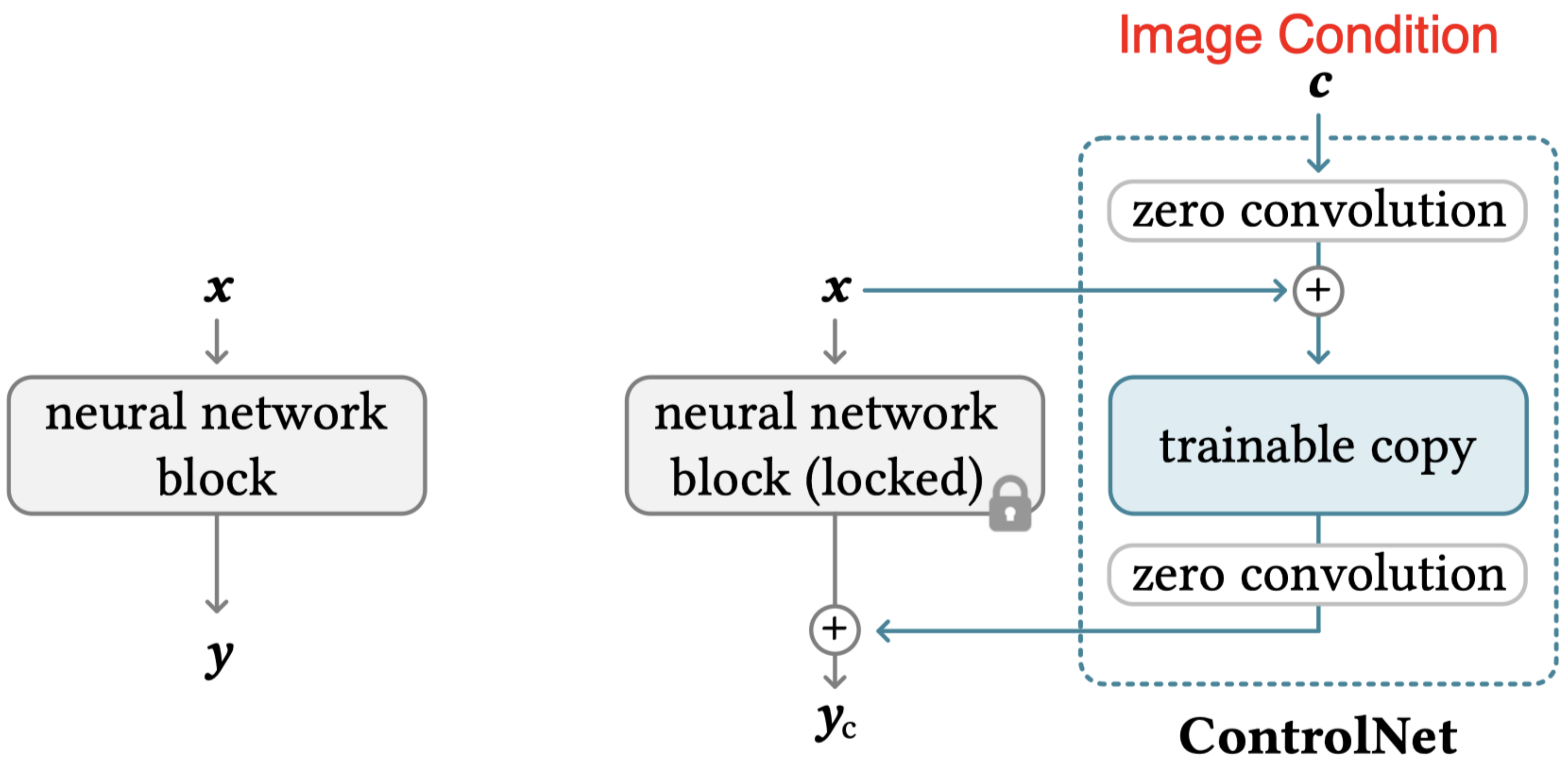


Cartoon line drawing “1girl, masterpiece, best quality, ultra-detailed, illustration”

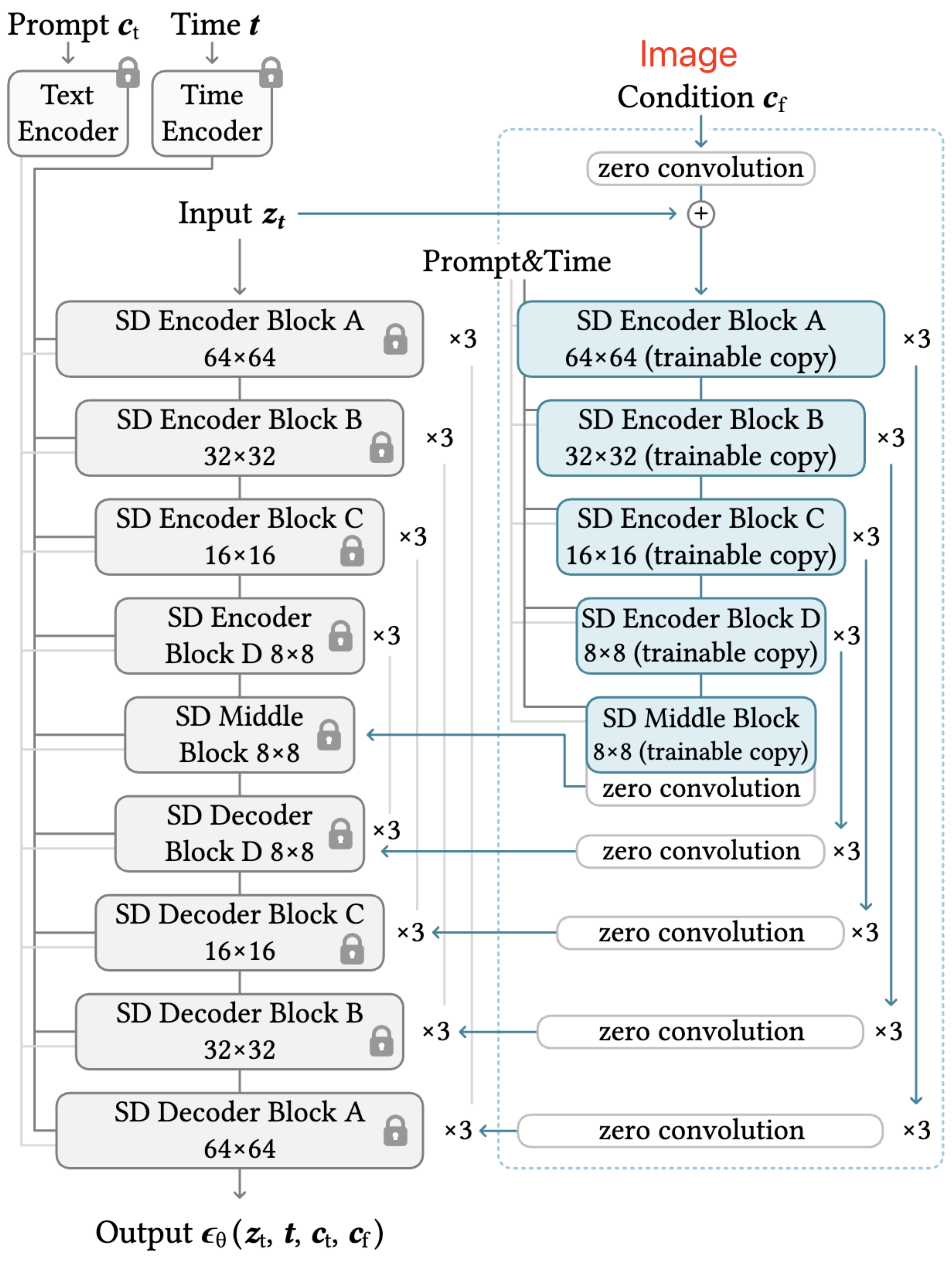


Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2023 Best Paper
T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, AAAI 2024

ControlNet



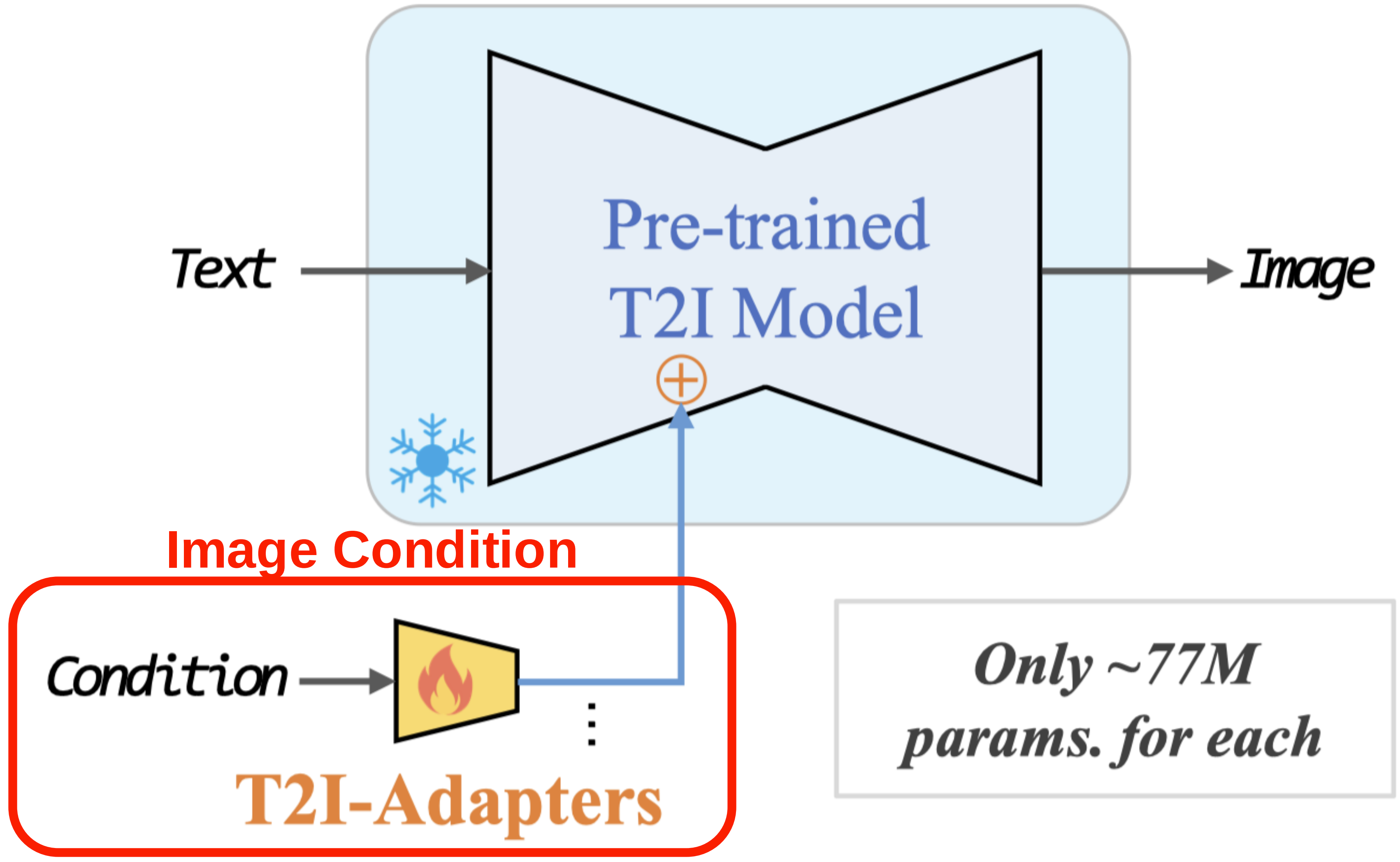
(a) Before (b) After

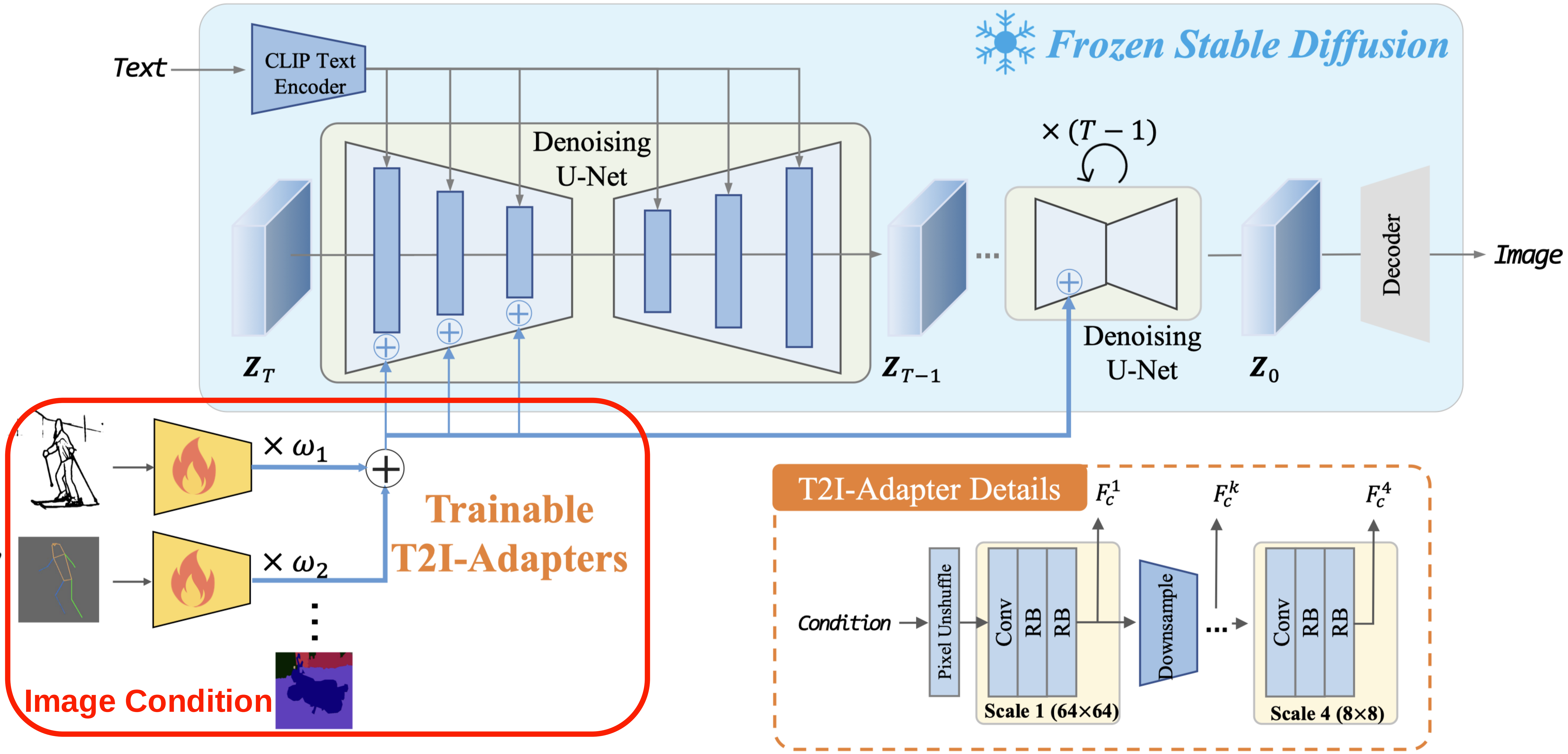


(a) Stable Diffusion (b) ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2023

Encode the Image Features as the Condition for Denoising Training





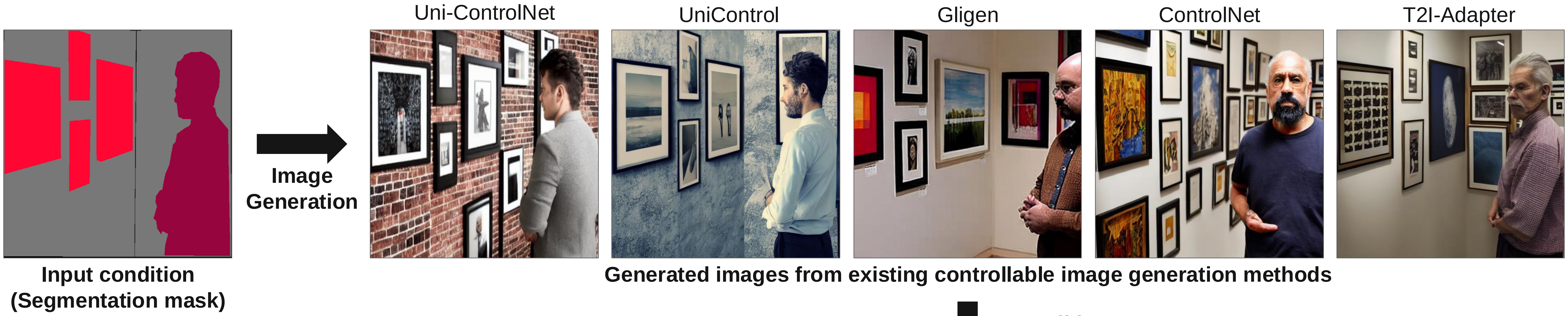
- ✓ *Plug-and-play.* ***Not affect original network topology and generation ability***
- ✓ *Simple and small.* ***~77M parameters and ~300M storage***
- ✓ *Flexible.* ***Various adapters for different control conditions***
- ✓ *Composable.* ***Several adapters to achieve multi-condition control***
-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

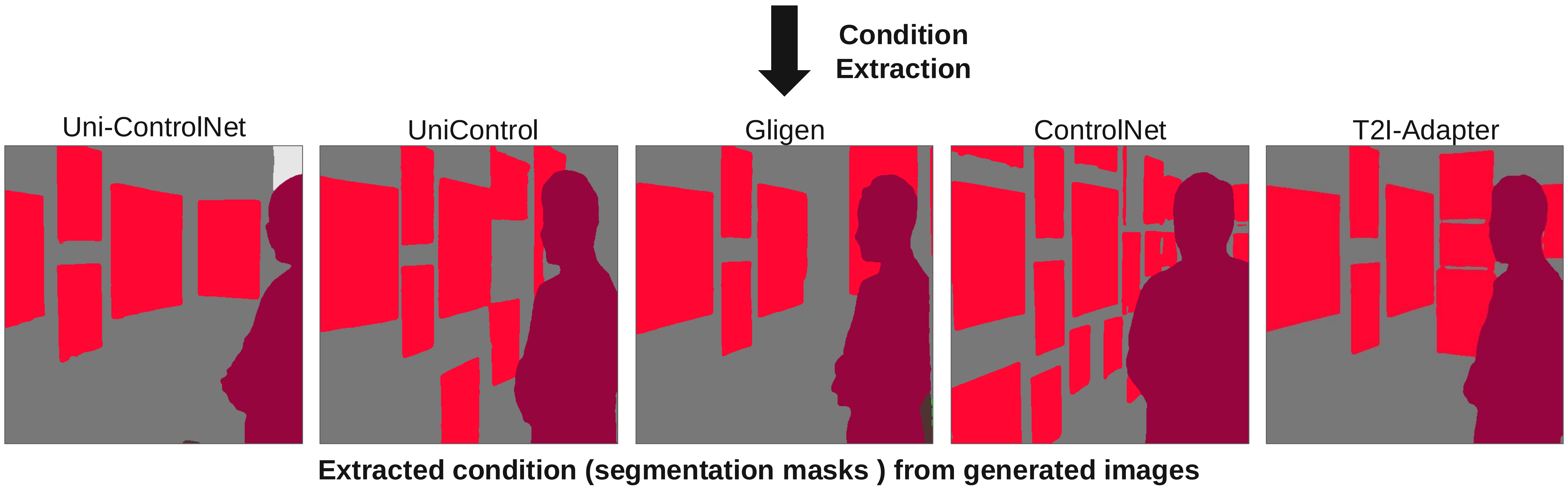
T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, AAAI 2024

Outline

- Background: Generative Learning for Images
- **<u>Motivation: Do existing methods achieve good controllability?</u>**
- Method: Efficient Consistency Feedback
- Experiments: Better Controllability Without Loss of Image Quality and Text Guidance
- Future Plans: Future Plans: More Conditions & Text-to-Image Models; Scaling Up

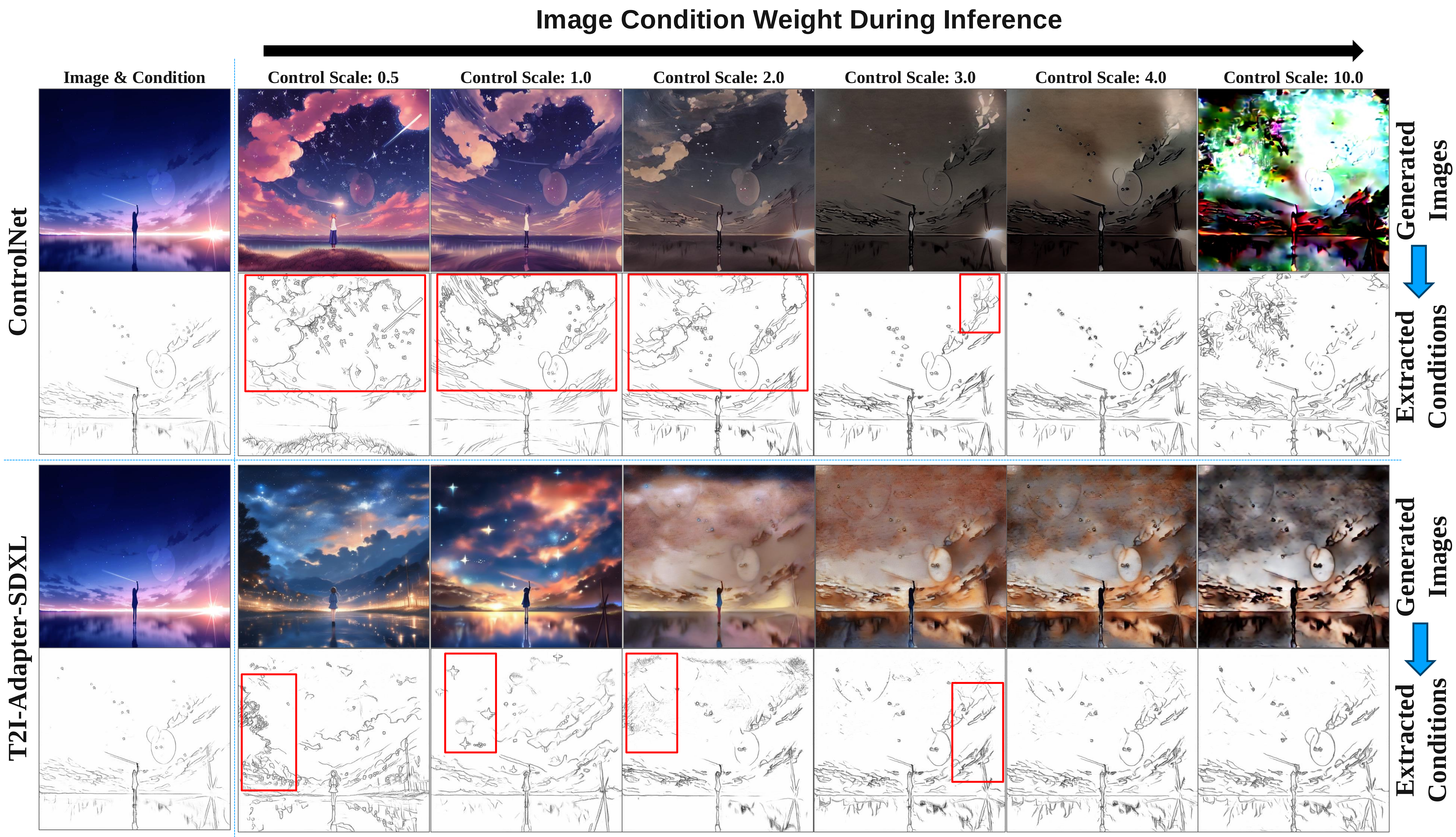
Existing Methods Still Cannot Accurately Control Image Generation





Inconsistencies between input and extracted condition

Controllability Cannot Be Improved by Emphasizing Image Condition



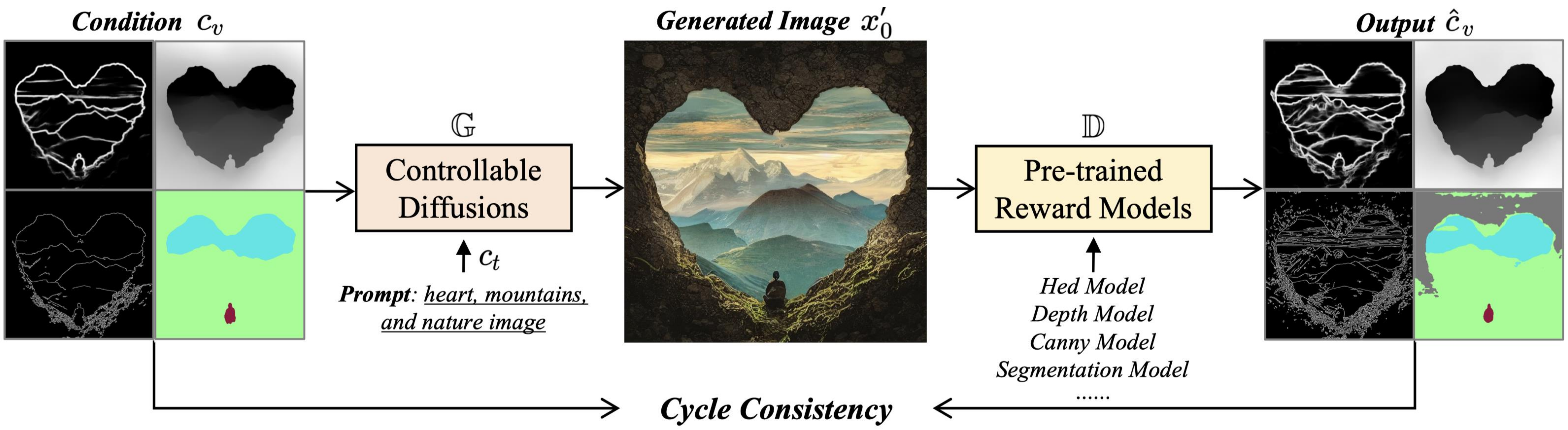
Outline

- Background: Generative Learning for Images
- Motivation: Do existing methods achieve good controllability?
- **Method: Efficient Consistency Feedback**
- Experiments: Better Controllability Without Loss of Image Quality and Text Guidance
- Future Plans: Future Plans: More Conditions & Text-to-Image Models; Scaling Up

Improving Controllability by Cycle Consistency

- **Definition**: We model controllable generation as an image translation task from input conditions to output generated images, the controllability can be defined as the consistency between them.

- **Optimization**: If we translate images from one domain to the other (condition $c_v \rightarrow$ generated image x_0'), and back again (generated image $x_0' \rightarrow$ condition $\hat{c}_v$) we should arrive where we started ($\hat{c}_v$ = c_v).

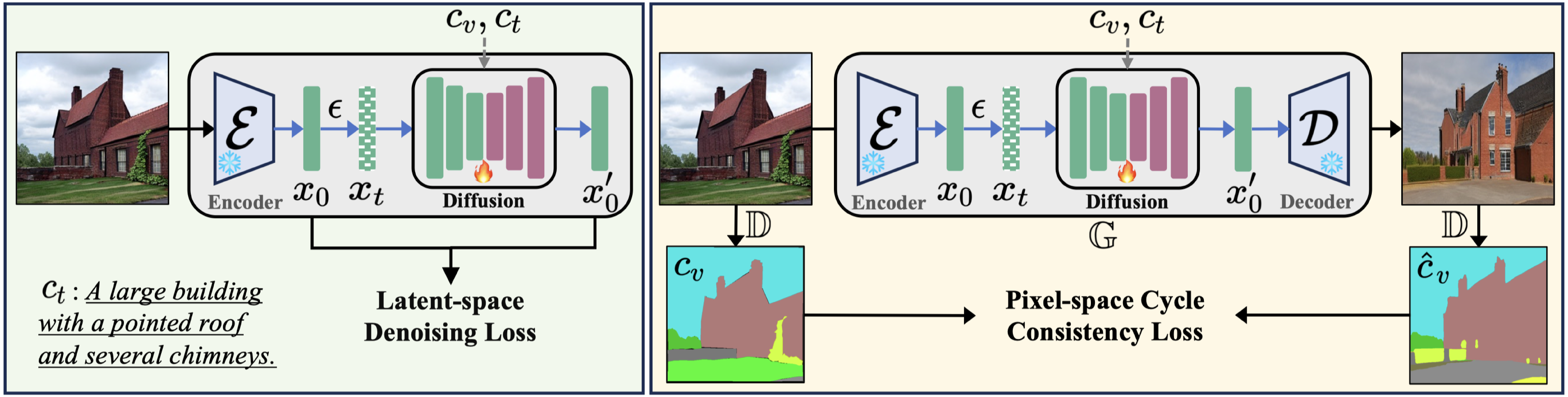


ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback, ECCV 2024
Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

What Makes Our ControlNet++ More Controllable?

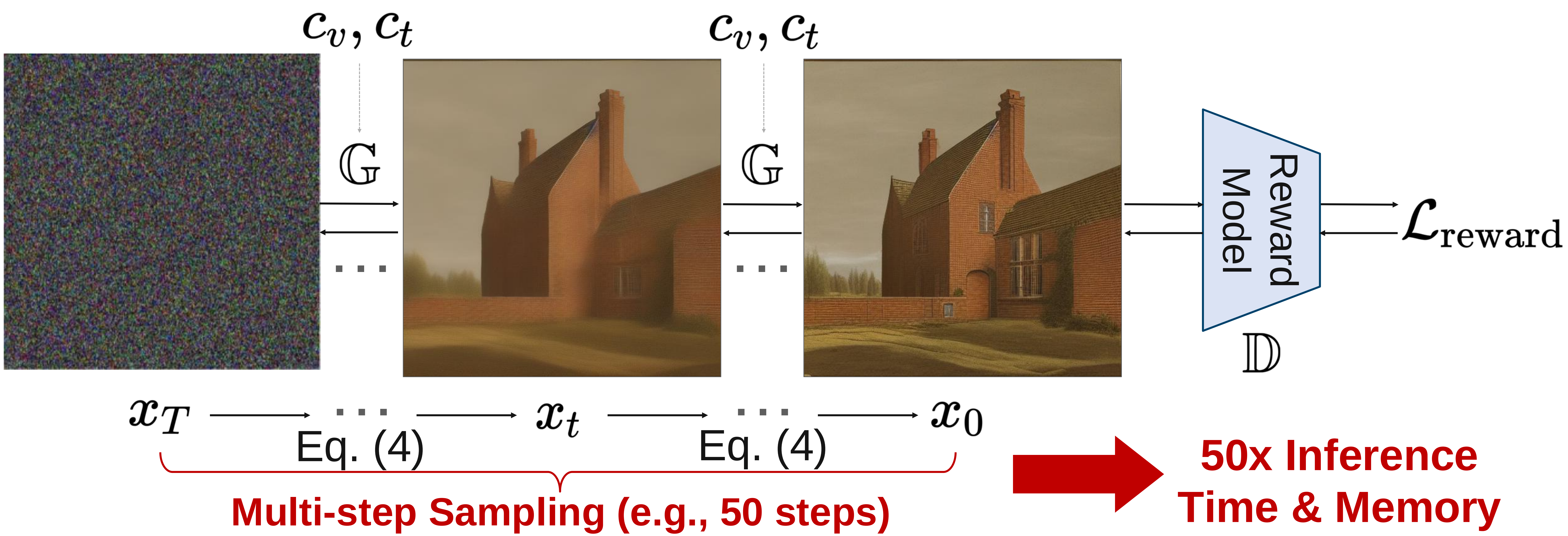
a. Existing methods achieve **implicit** controllability by introducing image-based conditional control c_v into the denoising process of diffusion models, with the guidance of latent-space denoising loss.

b. We utilize discriminative reward models $\boldsymbol{D}$ to **explicitly** optimize the controllability of the diffusion model $\boldsymbol{G}$ via pixel-level cycle consistency loss.



(a) Existing Methods **(b) Our Solution**

Default Step-by-Step Reward Strategy



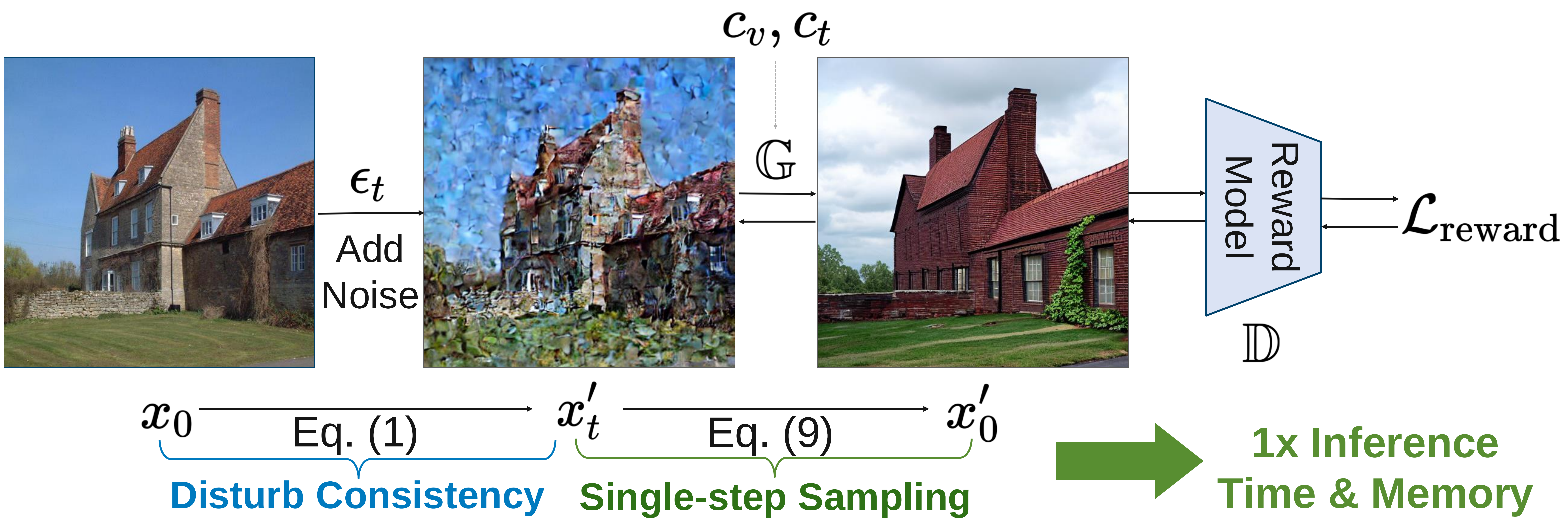
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \epsilon$$

Eq. (4)

step-by-step denoising process

$$\begin{aligned}
\mathcal{L}_{\text{reward}} &= \mathcal{L}(c_v, \hat{c}_v) \\
&= \mathcal{L}\big(c_v, \mathbb{D}(x_0')\big) \\
&= \mathcal{L}\big(c_v, \mathbb{D}[\mathbb{G}^T(c_t, c_v, x_T, t)]\big),
\end{aligned}$$

Our Efficient Reward Strategy

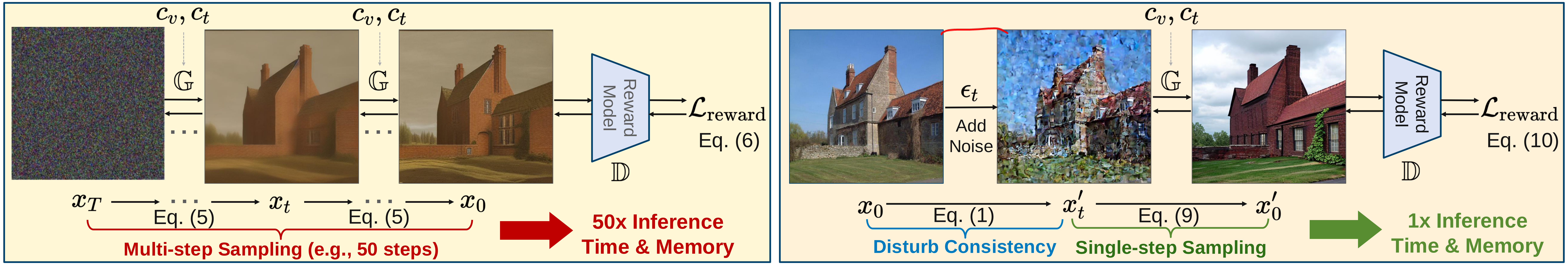


$$x_0 \approx x'_0 = \frac{x'_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x'_t, c_v, c_t, t-1)}{\sqrt{\alpha_t}}$$

$$\begin{aligned}\mathcal{L}_{\text{reward}} &= \mathcal{L}(c_v, \hat{c}_v) \\ &= \mathcal{L}\big(c_v, \mathbb{D}(x'_0)\big) \\ &= \mathcal{L}\Big(c_v, \mathbb{D}\big[\mathbb{G}\big(c_t, c_v, x'_t, t\big)\big]\Big),\end{aligned}$$

Directly Optimizing All Timesteps is Computationally Infeasible

The core idea of **(b)** is to use the single-step denoised image to estimate the step-by-step sampled image for reward loss, thus avoiding the sampling progress and gradient storage.



(a) Default Reward Strategy

(b) Efficient Reward Strategy (Ours)

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \epsilon$$

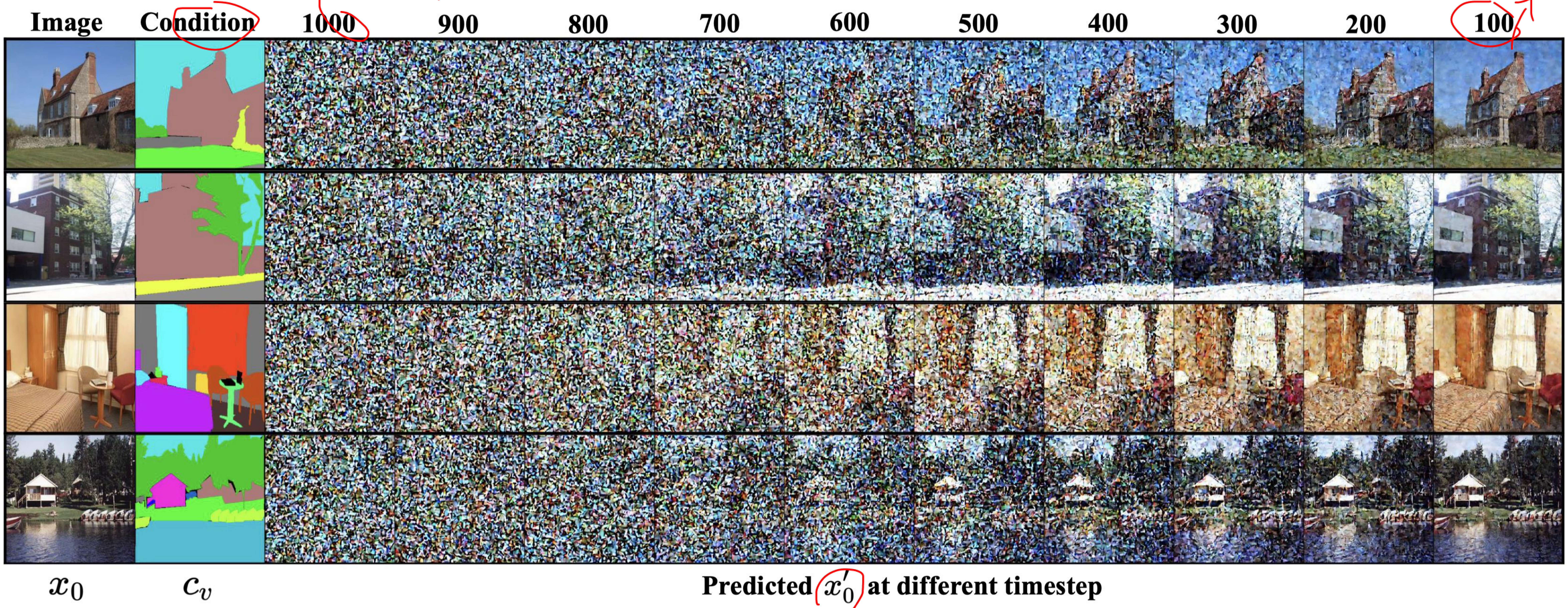
$$x_0 \approx x'_0 = \frac{x'_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x'_t, c_v, c_t, t-1)}{\sqrt{\alpha_t}}$$

$$\begin{aligned}\mathcal{L}_{\text{reward}} &= \mathcal{L}(c_v, \hat{c}_v) \\ &= \mathcal{L}\big(c_v, \mathbb{D}(x'_0)\big) \\ &= \mathcal{L}\big(c_v, \mathbb{D}[\underbrace{\mathbb{G}^T(c_t, c_v, x_T, t)}_{\text{step-by-step sampled image}}]\big),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\text{reward}} &= \mathcal{L}(c_v, \hat{c}_v) \\ &= \mathcal{L}\big(c_v, \mathbb{D}(x'_0)\big) \\ &= \mathcal{L}\big(c_v, \mathbb{D}[\underbrace{\mathbb{G}(c_t, c_v, x'_t, t)}_{\text{single-step denoised image}}]\big),\end{aligned}$$

Such Estimation is Reasonable When Timestep is Small Enough

Single-step Denoising Visualizations



ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback, ECCV 2024

Outline

- Background: Generative Learning for Images
- Motivation: Do existing methods achieve good controllability?
- Method: Efficient Consistency Feedback
- **Experiments: Better Controllability Without Loss of Image Quality and Text Guidance**
- Future Plans: More Conditions & Text-to-Image Models; Scaling Up

Evaluation Metrics

- **Controllability**
  - The consistency between the input condition and the condition extracted from the generated image.
  - The specific metric depends on each image condition
- **Image Quality**
  - FID, a metric used to evaluate the feature distance between generated images and real images. A lower FID score indicates that the generated images are more similar to the real images in terms of their visual features.
- **Text-Image Alignment**
  - CLIP-Score, measuring the image-text alignment between the input text and the generated image.

Better Controllability Than Other Methods

Table 1: Controllability comparison with state-of-the-art methods under different conditional controls and datasets. ↑ denotes higher result is better, while ↓ means lower is better. ControlNet++ achieves significant controllability improvements. '-' indicates that the method does not provide a public model for testing. We generate four groups of images in png format and report the average result to reduce random errors.

Condition (Metric)	**T2I Model**	**Seg. Mask (mIoU ↑)**		**Canny Edge (F1 Score ↑)**	**Hed Edge (SSIM ↑)**	**LineArt Edge (SSIM ↑)**	**Depth Map (RMSE ↓)**
Dataset		**ADE20K**	**COCO-Stuff**	**MultiGen-20M**	**MultiGen-20M**	**MultiGen-20M**	**MultiGen-20M**
ControlNet	SDXL	-	-	-	-	-	40.00
T2I-Adapter	SDXL	-	-	28.01	-	0.6394	39.75
T2I-Adapter	SD1.5	12.61	-	23.65	-	-	48.40
Gligen	SD1.4	23.78	-	26.94	0.5634	-	38.83
Uni-ControlNet	SD1.5	19.39	-	27.32	0.6910	-	40.65
UniControl	SD1.5	25.44	-	30.82	0.7969	-	39.18
ControlNet	SD1.5	32.55	27.46	34.65	0.7621	0.7054	35.90
Ours	SD1.5	**43.64**	**34.56**	**37.04**	**0.8097**	**0.8399**	**28.32**

No Loss of Image Quality (FID) and Text-Image Alignment (CLIP Score)

Table 2: FID (↓) comparison with state-of-the-art methods under different conditional controls and datasets. All the results are conducted on 512×512 image resolution with Clean-FID implementation [33] for fair comparisons. '-' indicates that the method does not provide a public model for testing. We generate four groups of images in png format and report the average result to reduce random errors.

Method	T2I Model	Seg. Mask		Canny Edge	Hed Edge	LineArt Edge	Depth Map
		ADE20K	COCO	MultiGen-20M	MultiGen-20M	MultiGen-20M	MultiGen-20M
Gligen	SD1.4	33.02	-	18.89	-	-	18.36
T2I-Adapter	SD1.5	39.15	-	15.96	-	-	22.52
UniControlNet	SD1.5	39.70	-	17.14	17.08	-	20.27
UniControl	SD1.5	46.34	-	19.94	15.99	-	18.66
ControlNet	SD1.5	33.28	21.33	**14.73**	15.41	17.44	17.76
Ours	SD1.5	**29.49**	**19.29**	18.23	**15.01**	**13.88**	**16.66**

No Loss of Image Quality (FID) and Text-Image Alignment (CLIP Score)

Table 2: FID (↓) comparison with state-of-the-art methods under different conditional controls and datasets. All the results are conducted on 512×512 image resolution with Clean-FID implementation [33] for fair comparisons. '-' indicates that the method does not provide a public model for testing. We generate four groups of images in png format and report the average result to reduce random errors.

Method	T2I Model	Seg. Mask ADE20K	Seg. Mask COCO	Canny Edge MultiGen-20M	Hed Edge MultiGen-20M	LineArt Edge MultiGen-20M	Depth Map MultiGen-20M
Gligen	SD1.4	33.02	-	18.89	-	-	18.36
T2I-Adapter	SD1.5	39.15	-	15.96	-	-	22.52
UniControlNet	SD1.5	39.70	-	17.14	17.08	-	20.27
UniControl	SD1.5	46.34	-	19.94	15.99	-	18.66
ControlNet	SD1.5	33.28	21.33	**14.73**	15.41	17.44	17.76
Ours	SD1.5	**29.49**	**19.29**	18.23	**15.01**	**13.88**	**16.66**

Table 3: CLIP-score (↑) comparison with state-of-the-art methods under different conditional controls and datasets. '-' indicates that the method does not provide a public model for testing. We generate four groups of images in png format and report the average result to reduce random errors.

Method	T2I Model	Seg. Mask ADE20K	Seg. Mask COCO	Canny Edge MultiGen-20M	Hed Edge MultiGen-20M	LineArt Edge MultiGen-20M	Depth Map MultiGen-20M
Gligen	SD1.4	31.12	-	31.77	-	-	31.75
T2I-Adapter	SD1.5	30.65	-	31.71	-	-	31.46
UniControlNet	SD1.5	30.59	-	31.84	31.94	-	31.66
UniControl	SD1.5	30.92	-	31.97	32.02	-	**32.45**
ControlNet	SD1.5	31.53	**13.31**	**32.15**	**32.33**	**32.46**	**32.45**
Ours	SD1.5	**31.96**	13.13	31.87	32.05	31.95	32.09

ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback, ECCV 2024

Controllable Generative Models in Return Help Discriminative Models!

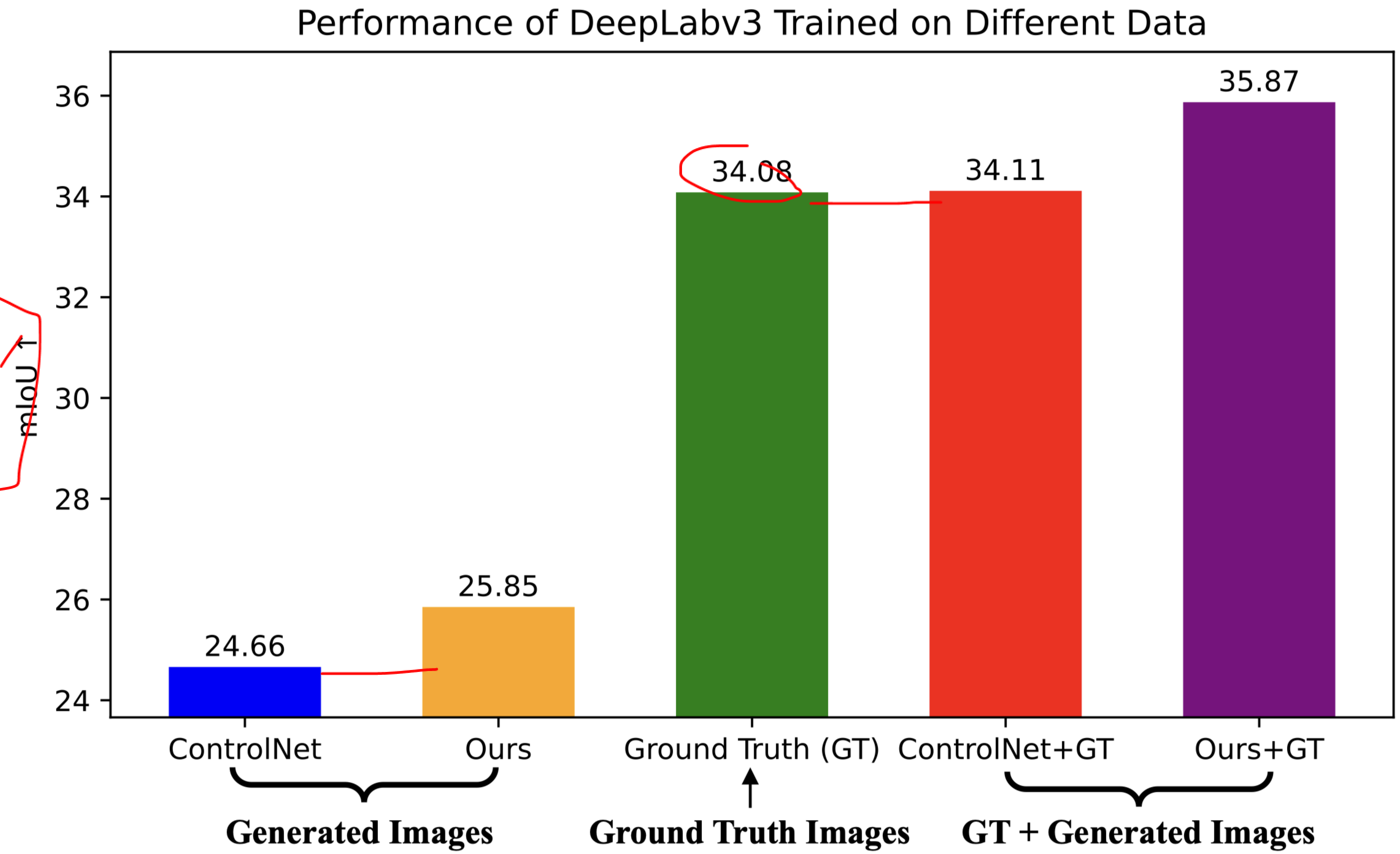


Fig. 5: Training DeepLabv3 (MobileNetv2) from scratch with different images, including ground truth images from ADE20K, and the generated images from ControlNet and ours. All the labels (i.e., segmentation masks) are ground truth labels in ADE20K. **Please note improvements here are non-trivial for semantic segmentation.**

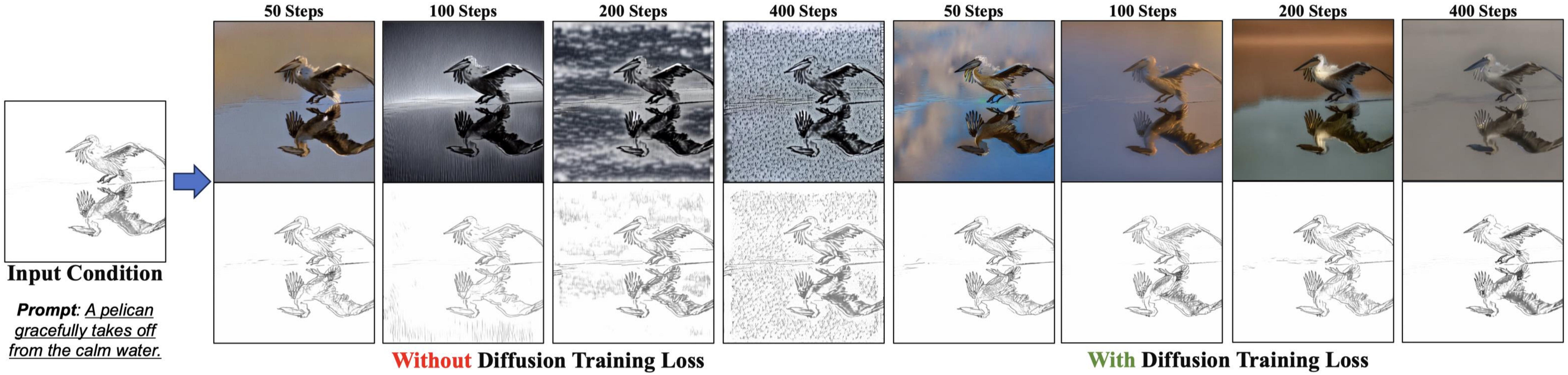
Ablation Studies

More powerful reward model leads to better controllable diffusion models

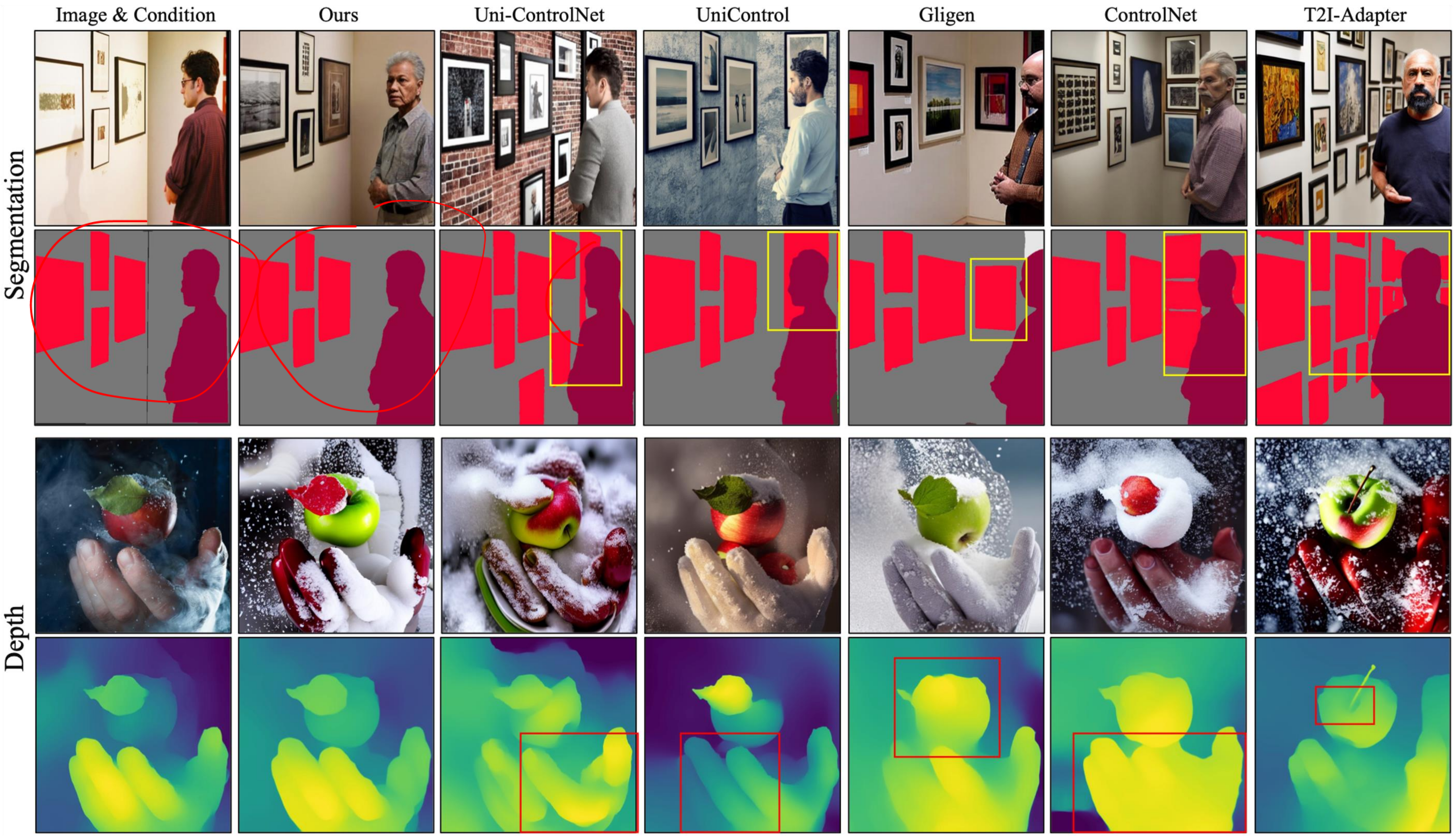
Table 5: Stronger reward model (UperNet-R50) leads to better controllability than the weaker reward model (DeepLabv3-MBv2).

Reward Model (RM)	RM mIoU↑	Eval mIoU↑
-	-	32.55
DeepLabv3-MBv2	34.02	31.96
FCN-R101	39.91	40.44
UperNet-R50	**42.05**	**43.64**

Reward Loss should be used together with Diffusion Training Loss

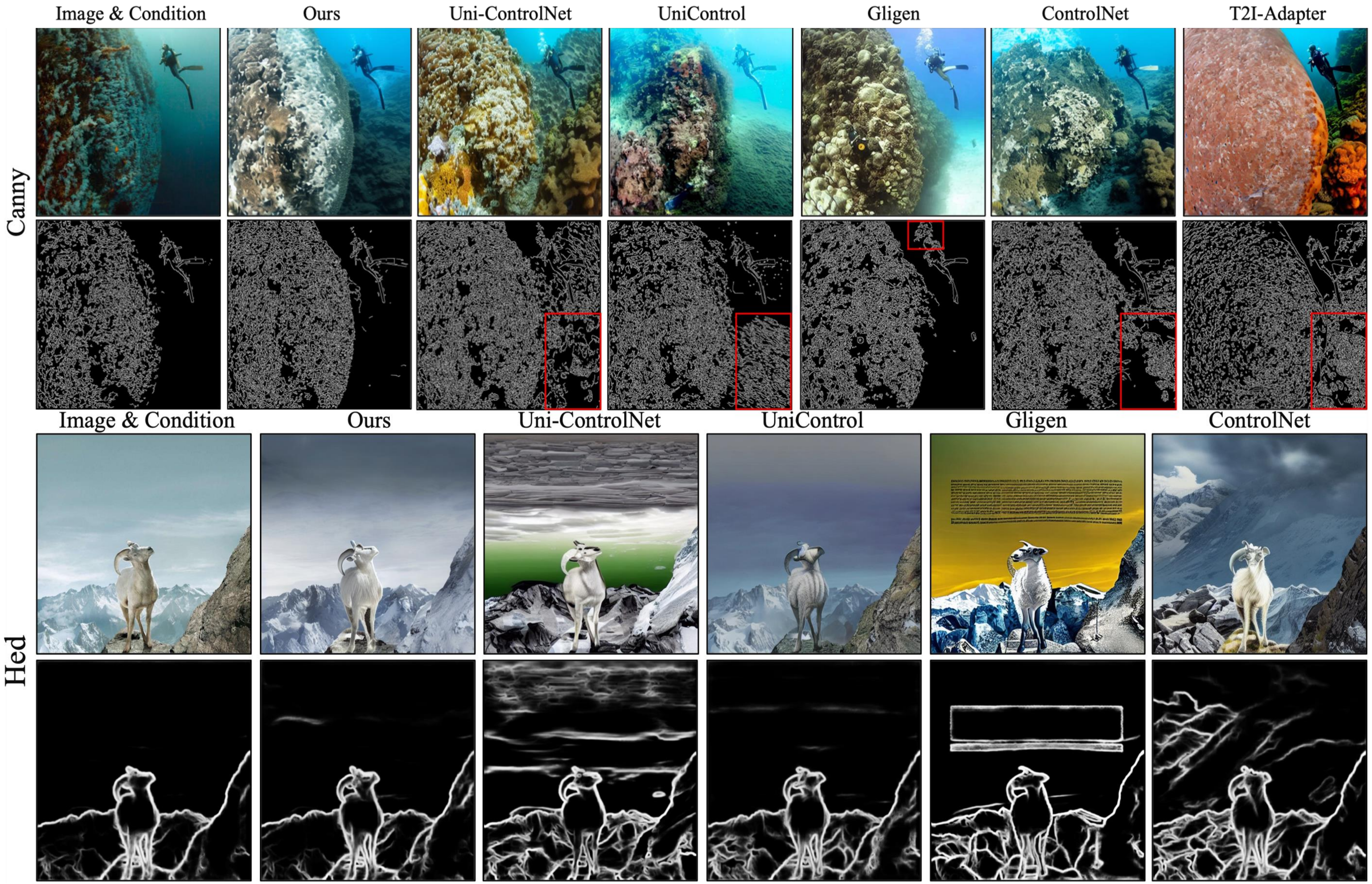


Visualization Comparison



ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback, ECCV 2024

Visualization Comparison



Visualization Results of Our ControlNet++ (Line Drawing)

Visualization Results of Our ControlNet++ (Depth Map)

Visualization Results of Our ControlNet++ (Hed Edge)

Visualization Results of Our ControlNet++ (Canny Edge)

Visualization Results of Our ControlNet++ (Segmentation Mask)

Code and Online Demo

Code: https://github.com/liming-ai/ControlNet_Plus_Plus

Online Demo: https://huggingface.co/spaces/limingcv/ControlNet-Plus-Plus

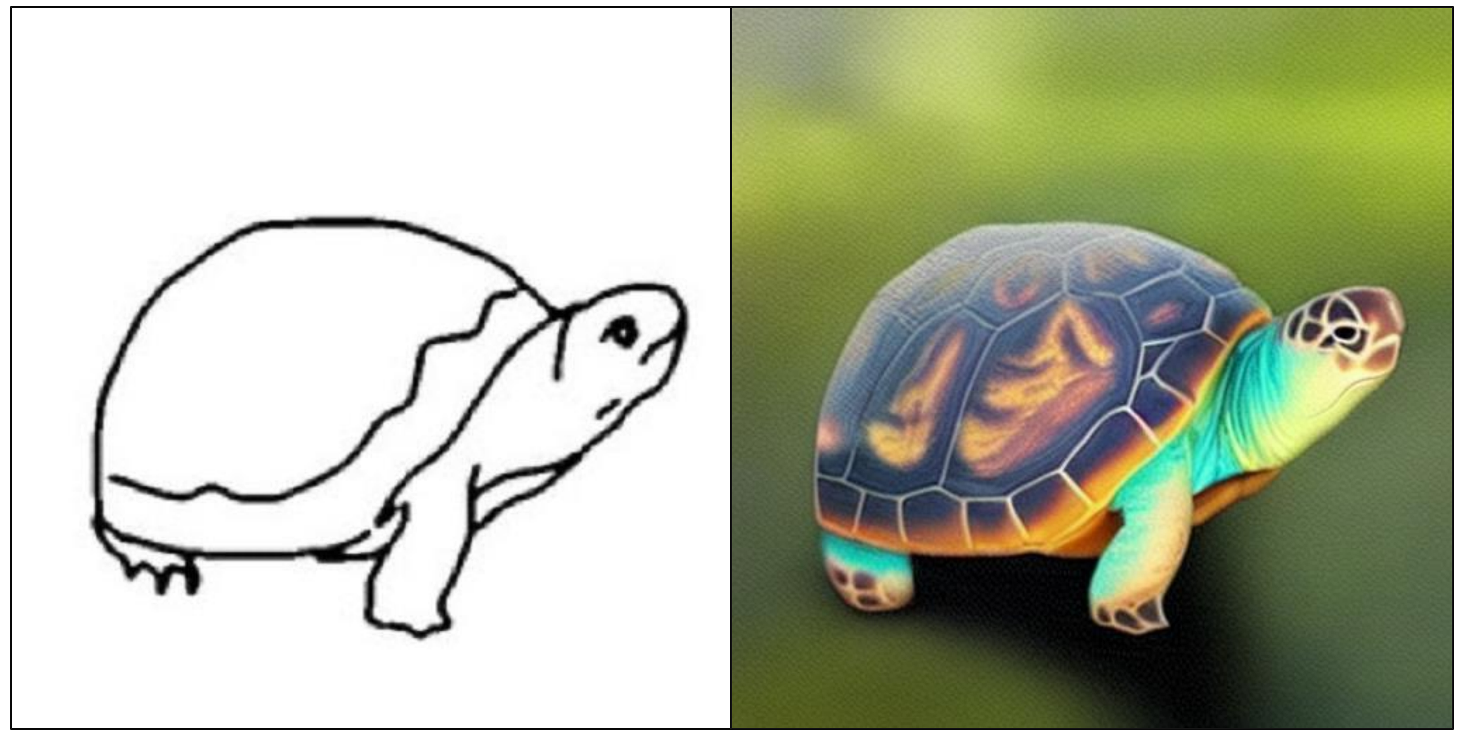
Outline

- Background: Generative Learning for Images
- Motivation: Do existing methods achieve good controllability?
- Method: Efficient Consistency Feedback
- Experiments: Better Controllability Without Loss of Image Quality and Text Guidance
- **Future Plans: More Conditions & Text-to-Image Models; Scaling Up**

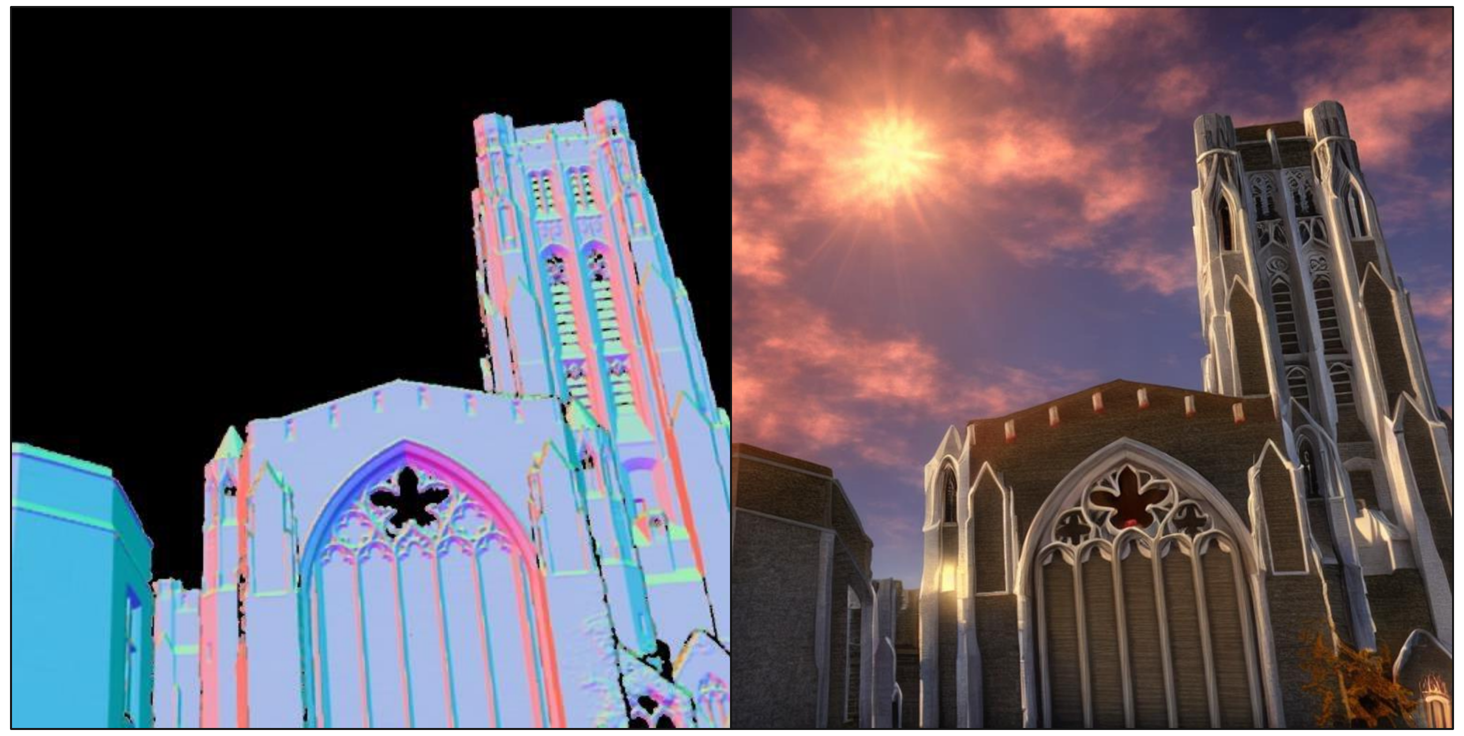
Future Plans: Support More Condition & More Text-to-Image Models

- More Conditions (such as Pose, Sketch, Normal, etc)

Pose-to-Image Generation

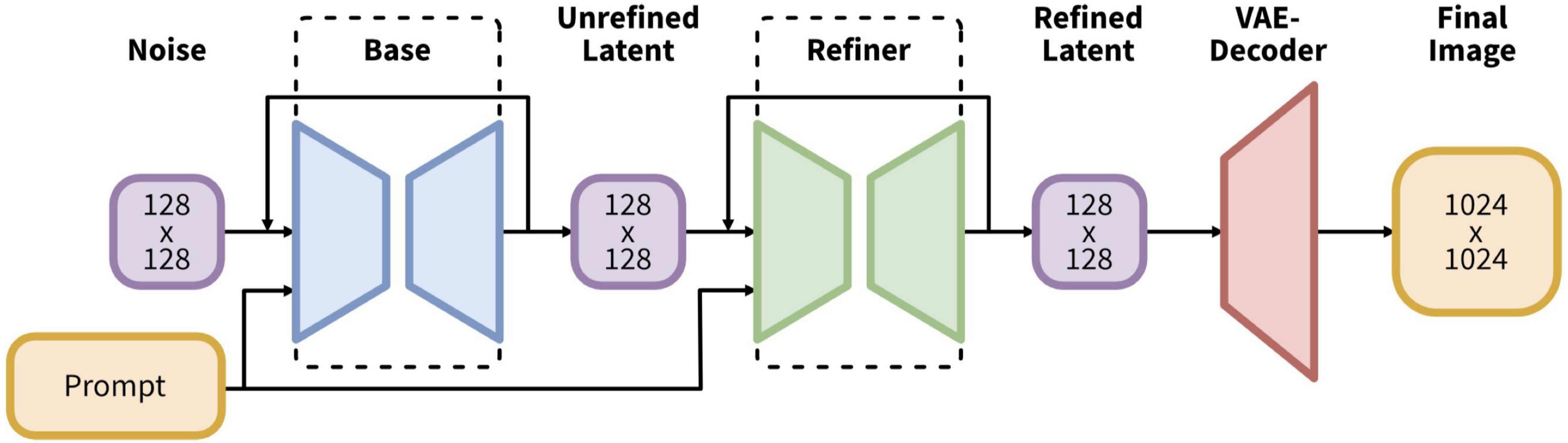


Sketch-to-Image Generation



Normal-to-Image Generation

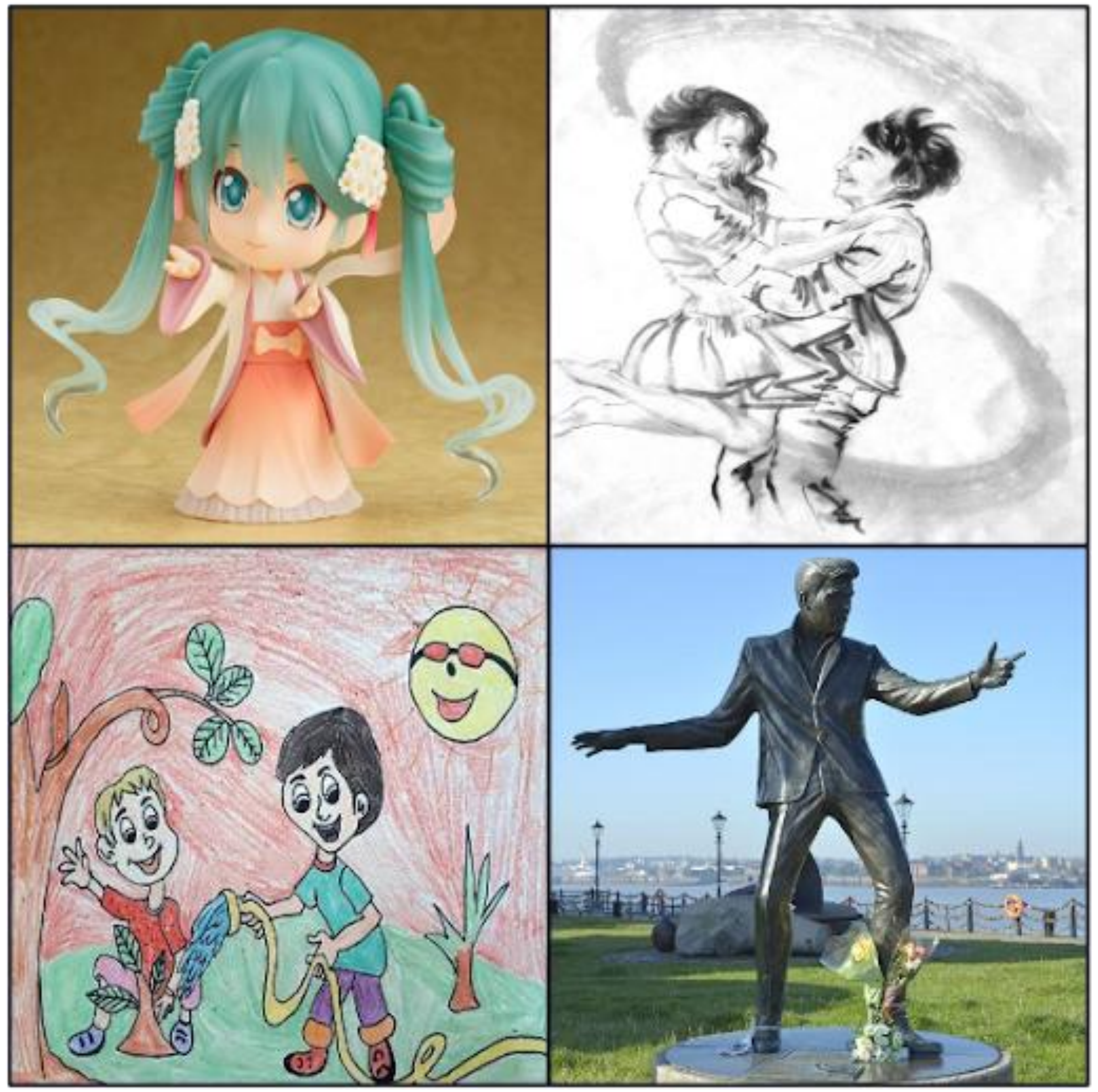
- More Models (such as SDXL, SD3, FLUX, etc)



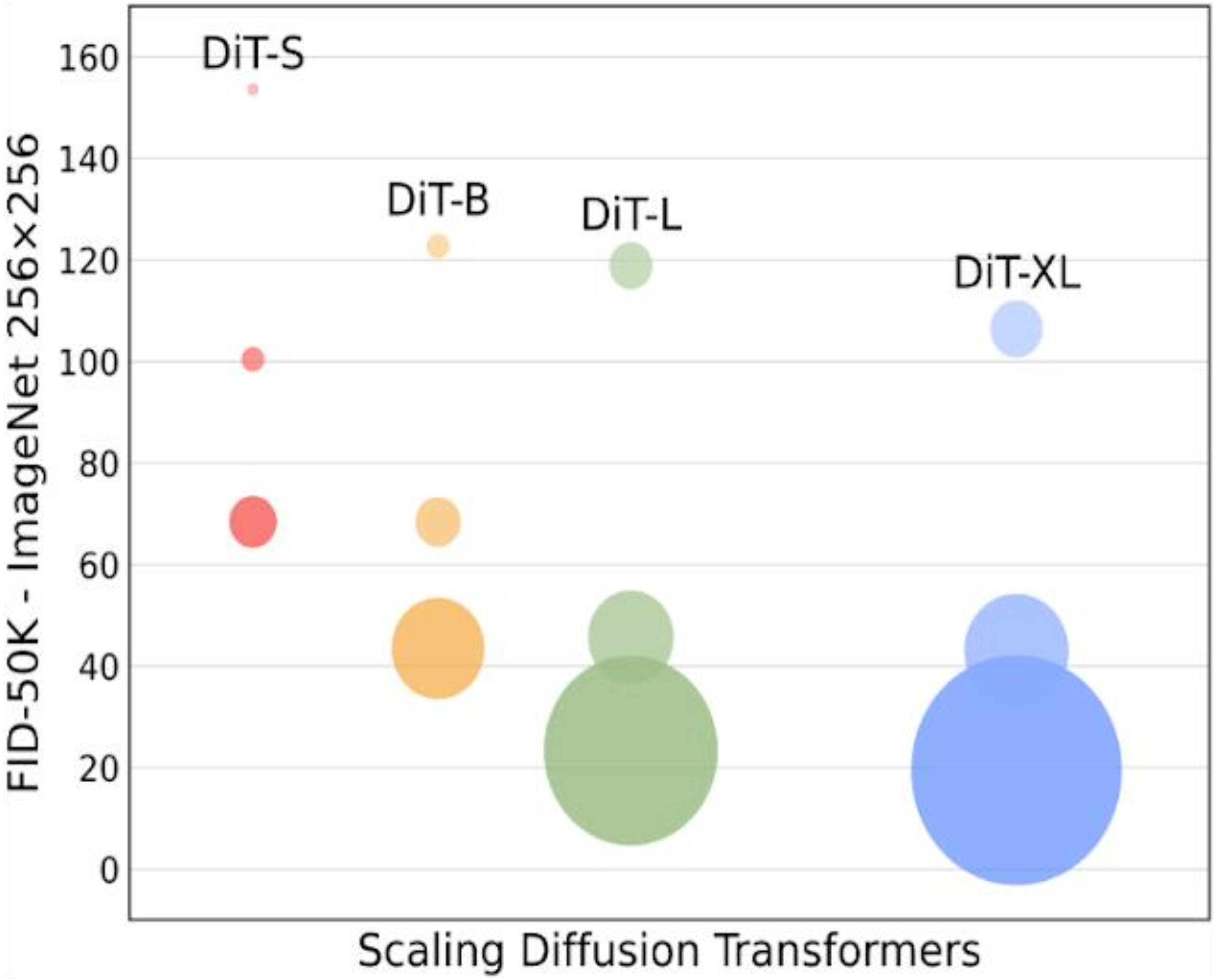
SDXL Pipeline

FLUX Images

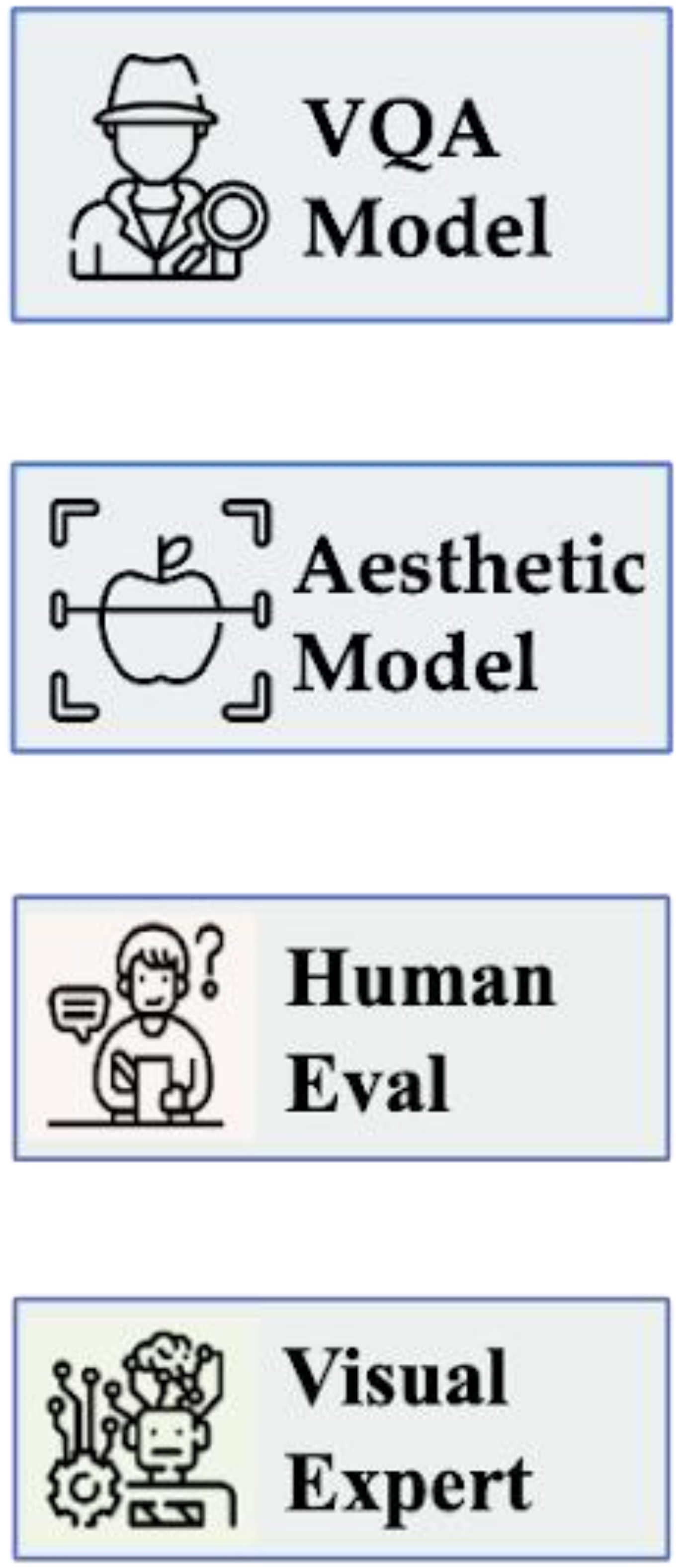
Future Plans: Scaling Data, Model and Rewards



More Data



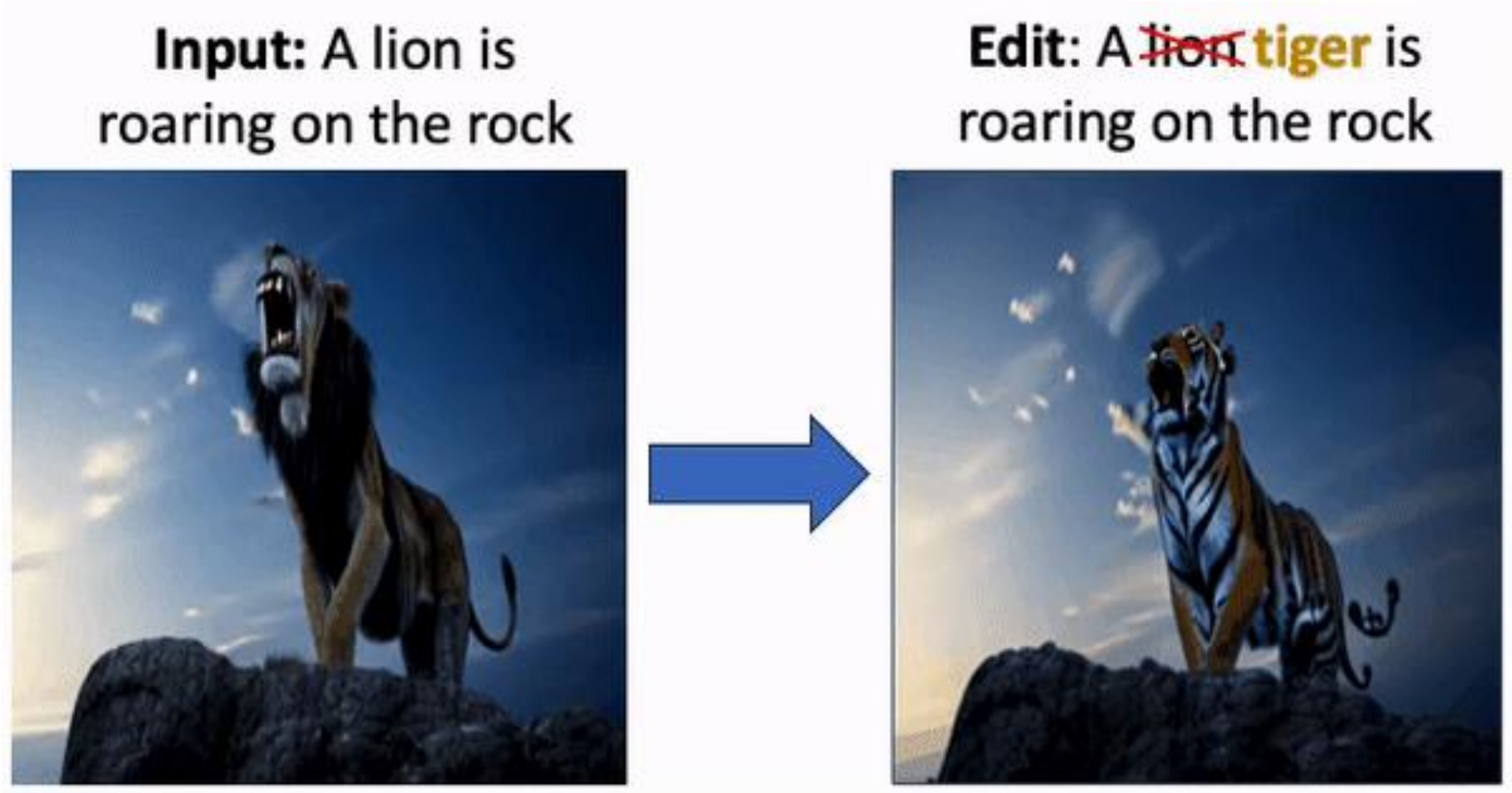
Stronger Model



Diverse Rewards

Other GenAI Related Research

Text guided video editing

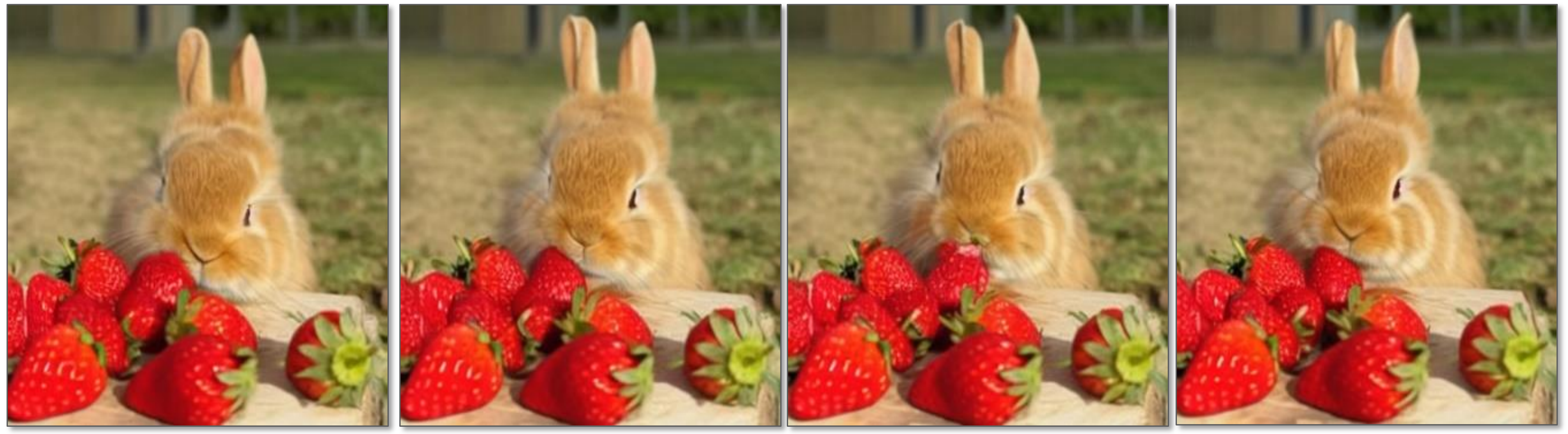


SAVE: Spectral-Shift-Aware Adaptation of Image Diffusion Models for Text-guided Video Editing Nazmul Karim, Umar Khalid, Mohsen Joneidi, Chen Chen, Nazanin Rahnavard arXiv:2305.18670

Text guided video editing

Multi-Object Editing

Original: A rabbit is eating strawberries.



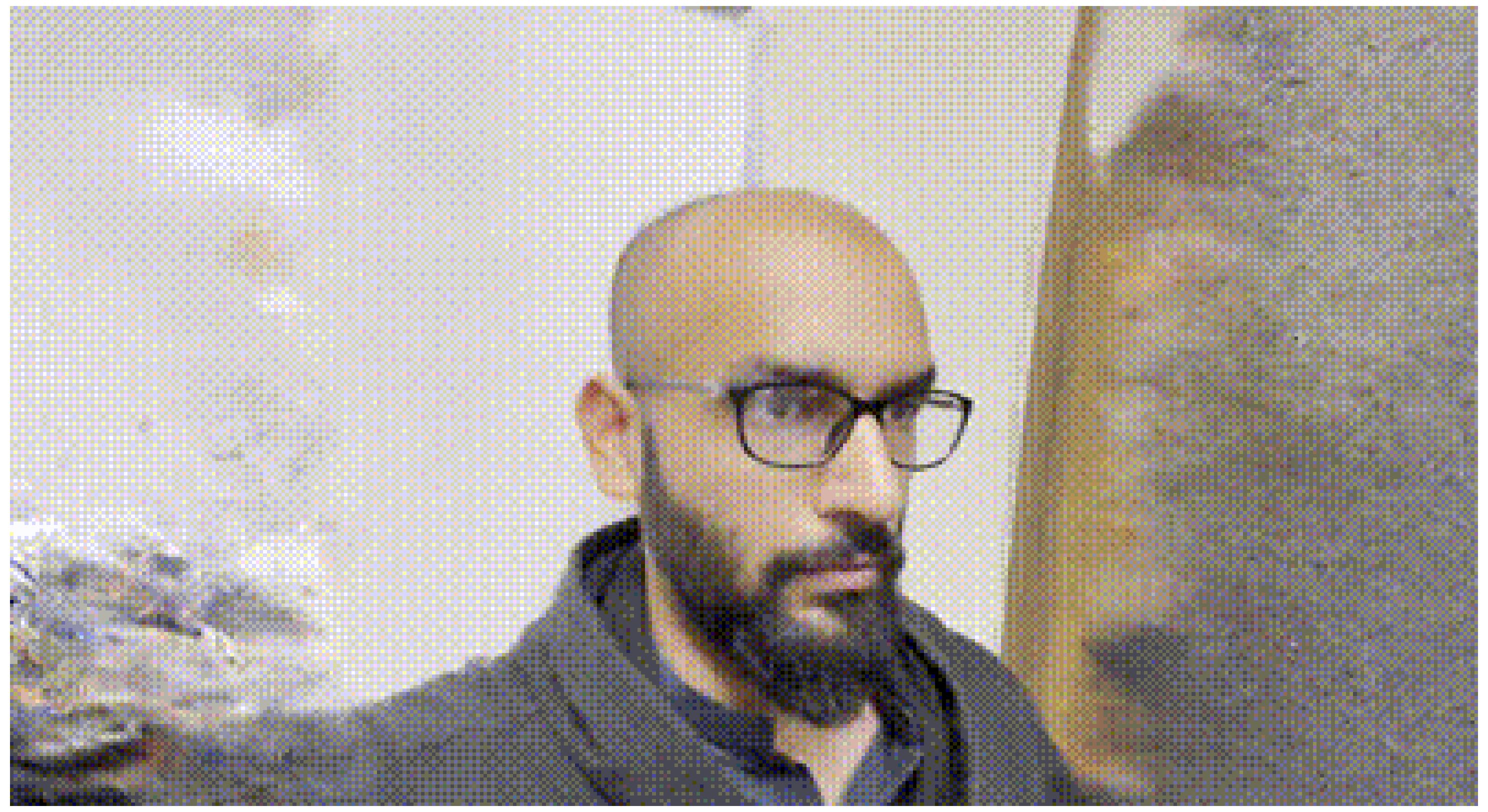
Edited (Ours): A dog is eating leaves.

Original: A squirrel is eating a carrot.

Edited (Ours): A cat is eating an eggplant.

SAVE: Spectral-Shift-Aware Adaptation of Image Diffusion Models for Text-guided Video Editing Nazmul Karim, Umar Khalid, Mohsen Joneidi, Chen Chen, Nazanin Rahnavard arXiv:2305.18670

Text-driven 3D (NeRF/Gaussian Splatting) editing



Original Scene

Turn him into cartoon

Turn him into Joker

Turn into Modigliani

1. 3DEgo: 3D Editing on the Go! Umar Khalid, Hasan Iqbal, Azib Farooq, Jing Hua, Chen Chen, European Conference on Computer Vision **(ECCV), 2024**
2. LatentEditor: Text Driven Local Editing of 3D Scenes Umar Khalid, Hasan Iqbal, Muhammad Tayyab, Md Nazmul Karim, Jing Hua, Chen Chen, European Conference on Computer Vision **(ECCV), 2024**
3. Free-Editor: Zero-shot Text-driven 3D Scene Editing Md Nazmul Karim, Hasan Iqbal, Umar Khalid, Chen Chen, Jing Hua European Conference on Computer Vision **(ECCV), 2024**

Controllable Object-Centric 3D Generation

Framework

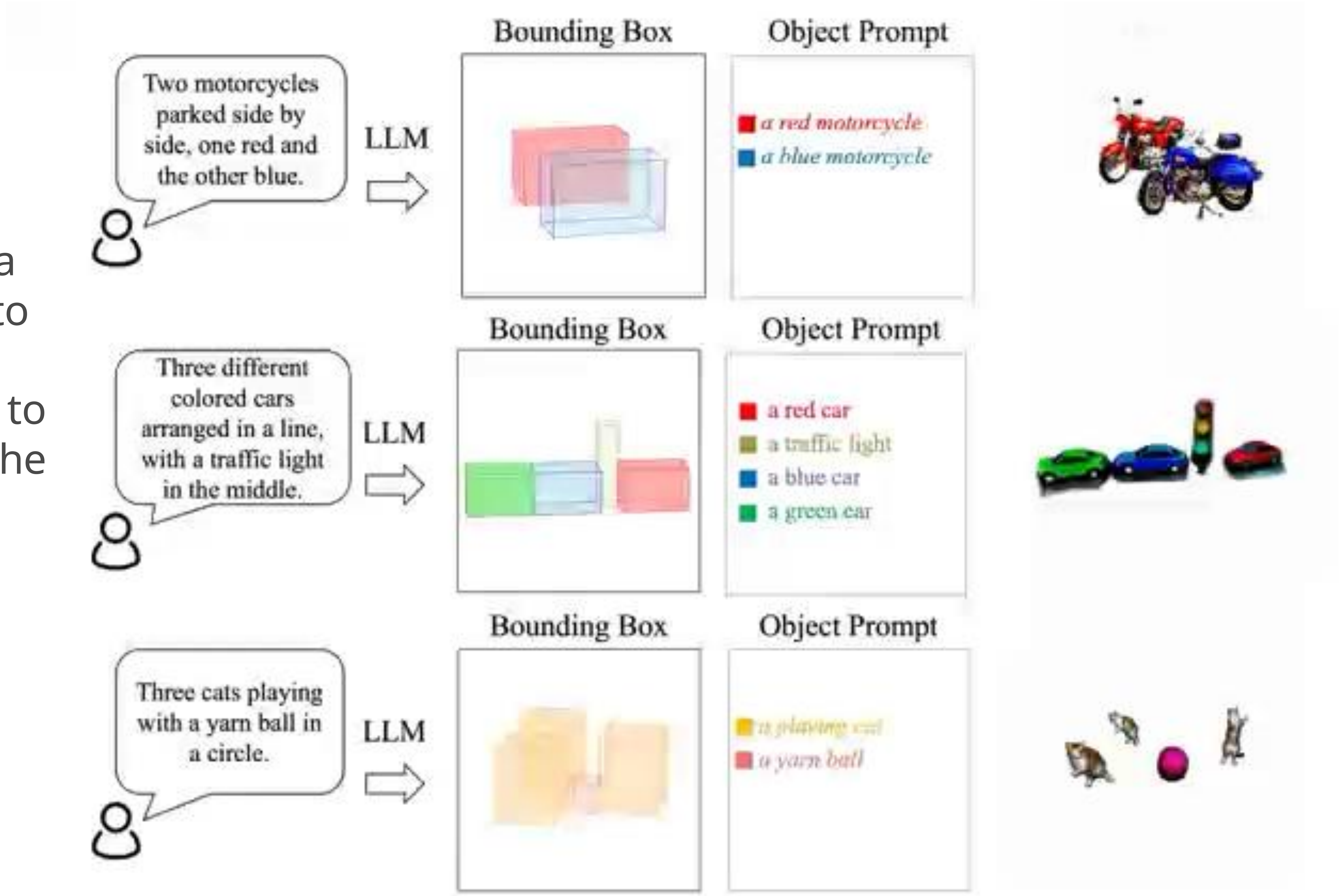
A high-level overview of **LucidDreaming** pipeline, controlling prompts are decomposed into 3D bounding boxes with LLMs, such as GPT4. Then in LucidDreaming, object-centric density bias and clipped ray sampling are used with Score Distillation Sampling (SDS) loss to align the generation with the user's control.

LucidDreaming: Controllable Object-Centric 3D Generation Zhaoning Wang, Ming Li, Chen Chen, ECCV Workshop 2024.

Controllable Object-Centric 3D Generation

Text-to-3D

Given a text prompts, we utilize a Language Model to convert it into bounding boxes and individual prompts. Then we can use them to generate 3D content align with the user's specifications.



LucidDreaming: Controllable Object-Centric 3D Generation Zhaoning Wang, Ming Li, Chen Chen, ECCV Workshop 2024.

Image-to-3D demos

Our framework can also adapt to Image-to-3D generation, given bounding boxes and image conditioning.

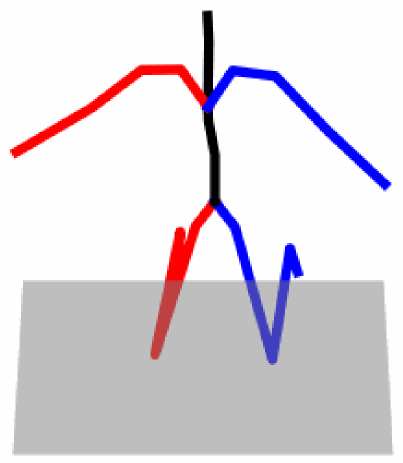
Text to Human Motion Generation

Text prompt : "the person crouches and walks forward."

1. BAMM: Bidirectional Autoregressive Motion Model Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, Chen Chen European Conference on Computer Vision **(ECCV), 2024**
2. MMM: Generative Masked Motion Model Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, Chen Chen IEEE Conference on Computer Vision and Pattern Recognition **(CVPR), 2024**

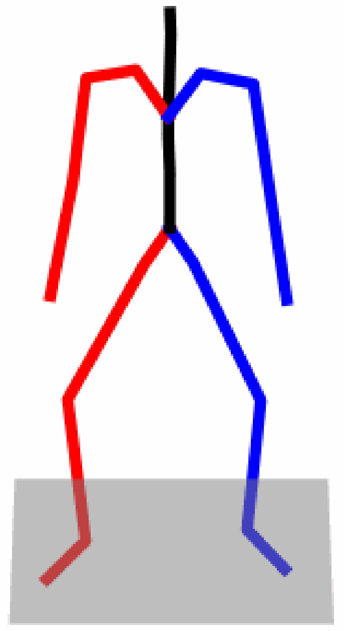
Text to Human Motion Generation

Original text:
A person uses his right arm to help himself to stand up.



T2M-GPT

Perturbed text:
A human utilizes his right arm to help himself to stand up.



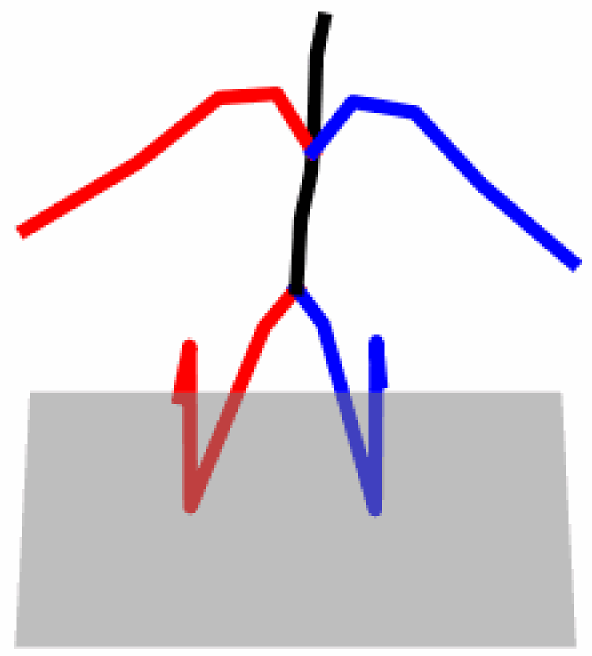
T2M-GPT

1. SATO: Stable Text-to-Motion Framework Wenshuo Chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, Chen Chen ACM Multimedia **(ACM MM), 2024**

Text to Human Motion Generation

Perturbed text:

A human utilizes his right arm to help himself to stand up.



Ours – SATO (T2M-GPT)

1. SATO: Stable Text-to-Motion Framework Wenshuo Chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, Chen Chen ACM Multimedia **(ACM MM), 2024**

Thank you!