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

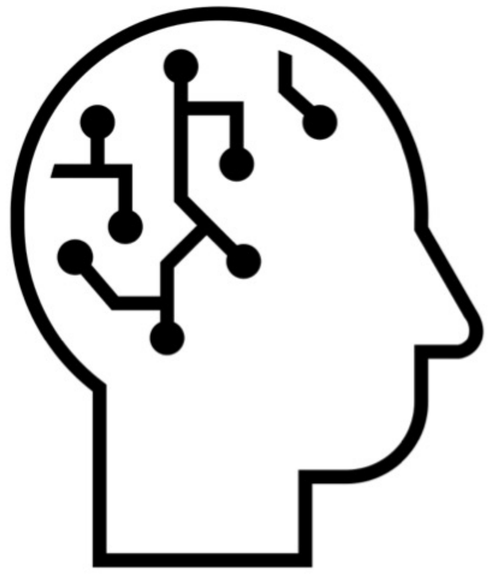
- **Background: Generative Learning for Images**
- Motivation: Do existing methods achieve good controllability?
- Method: Efficient Consistency Feedback
- Experiments: Better Controllability Without Loss of Image Quality and Text Guidance

Deep Generative Learning for Image

Learning to generate data

Samples from a Data Distribution

Train



Neural Network

Sample

Application

Art & Design

Content Generation

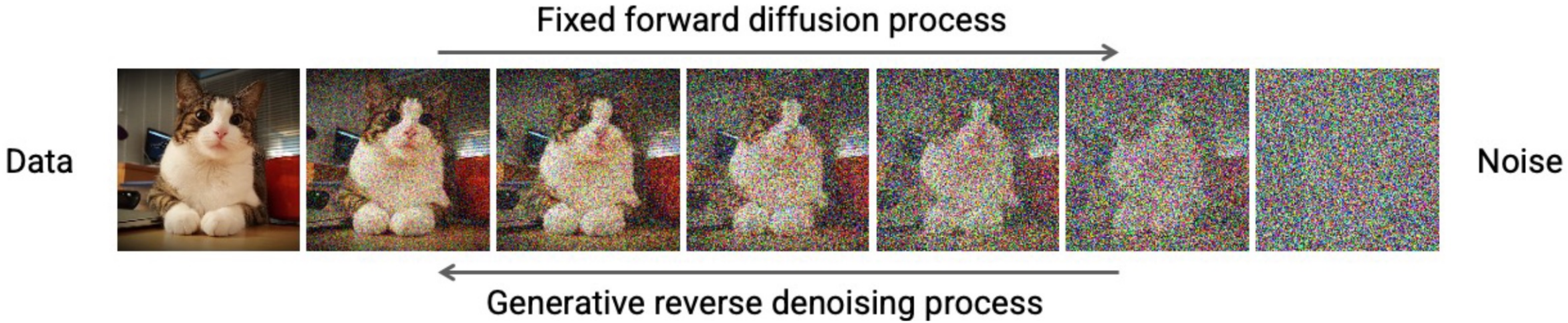
Representation Learning

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**



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.

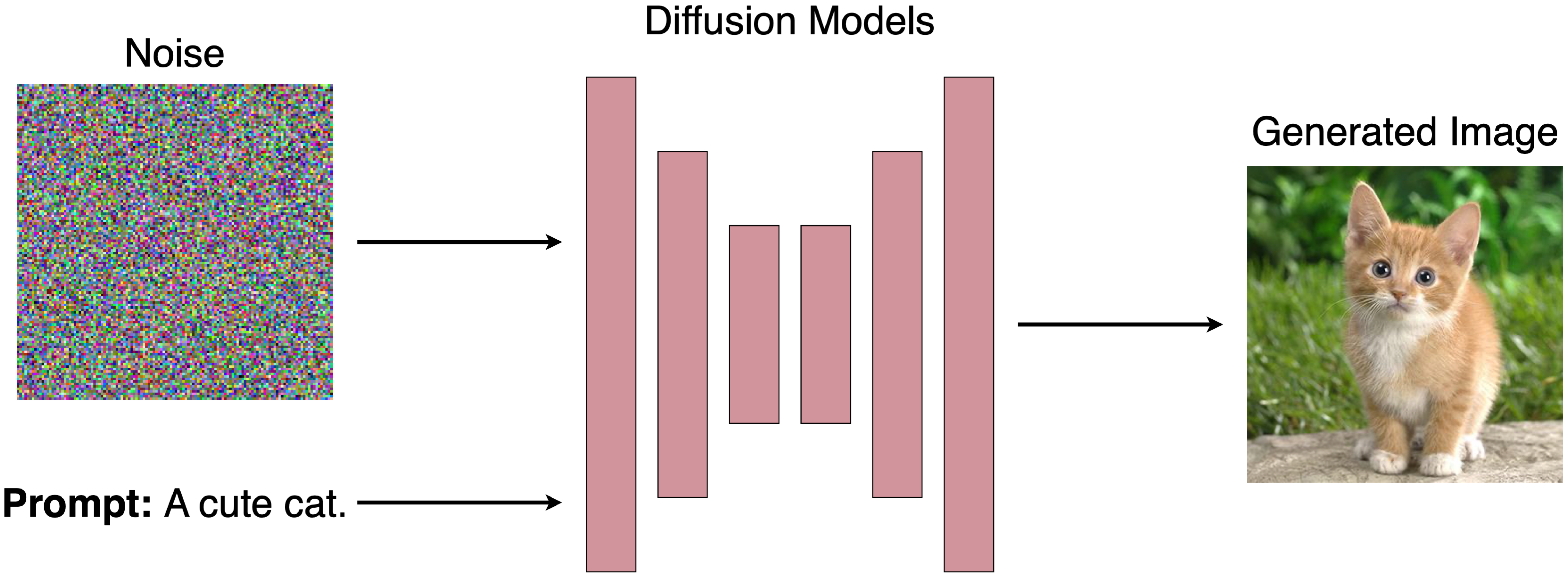


Image Source: https://hanlab.mit.edu/projects/can

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

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 <u>NOT</u> 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

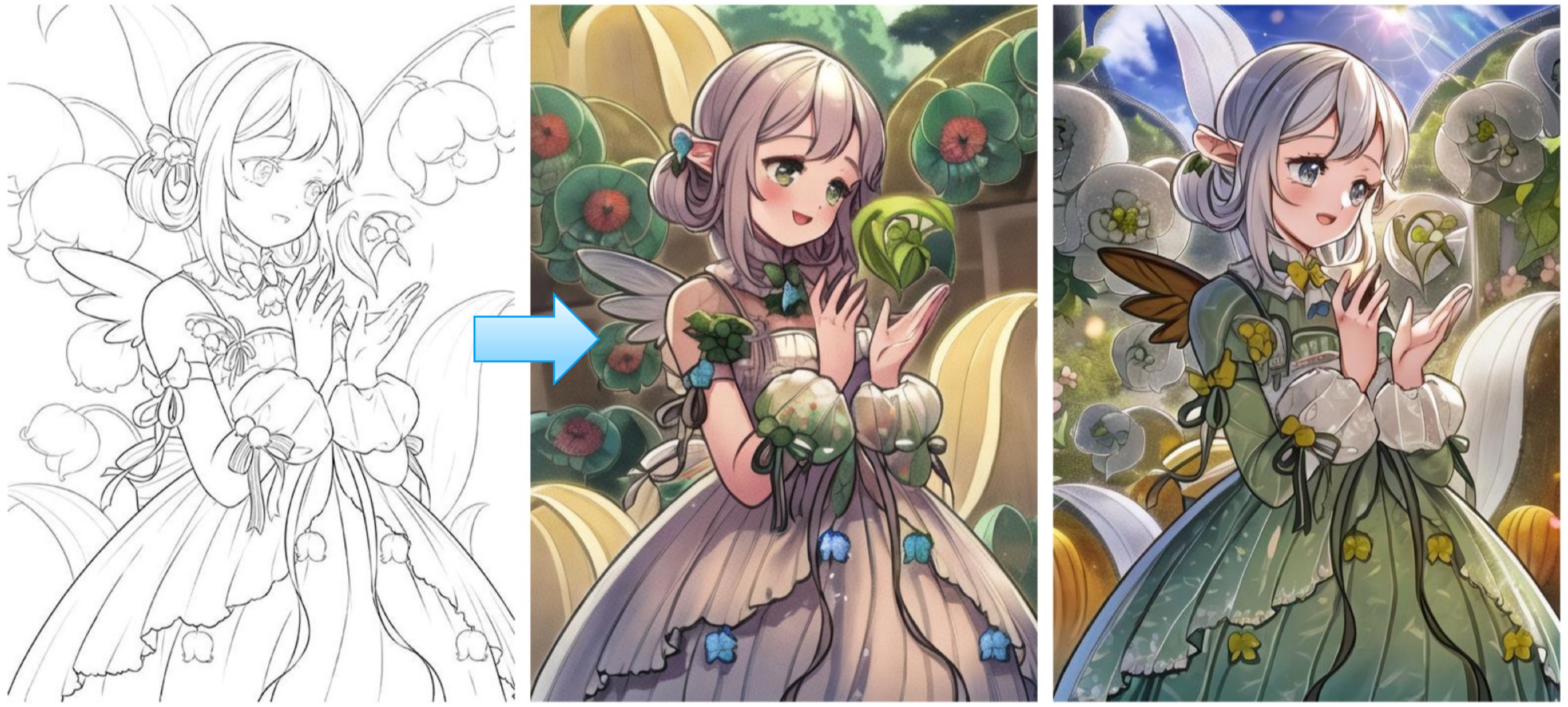
DALL-E 3

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.

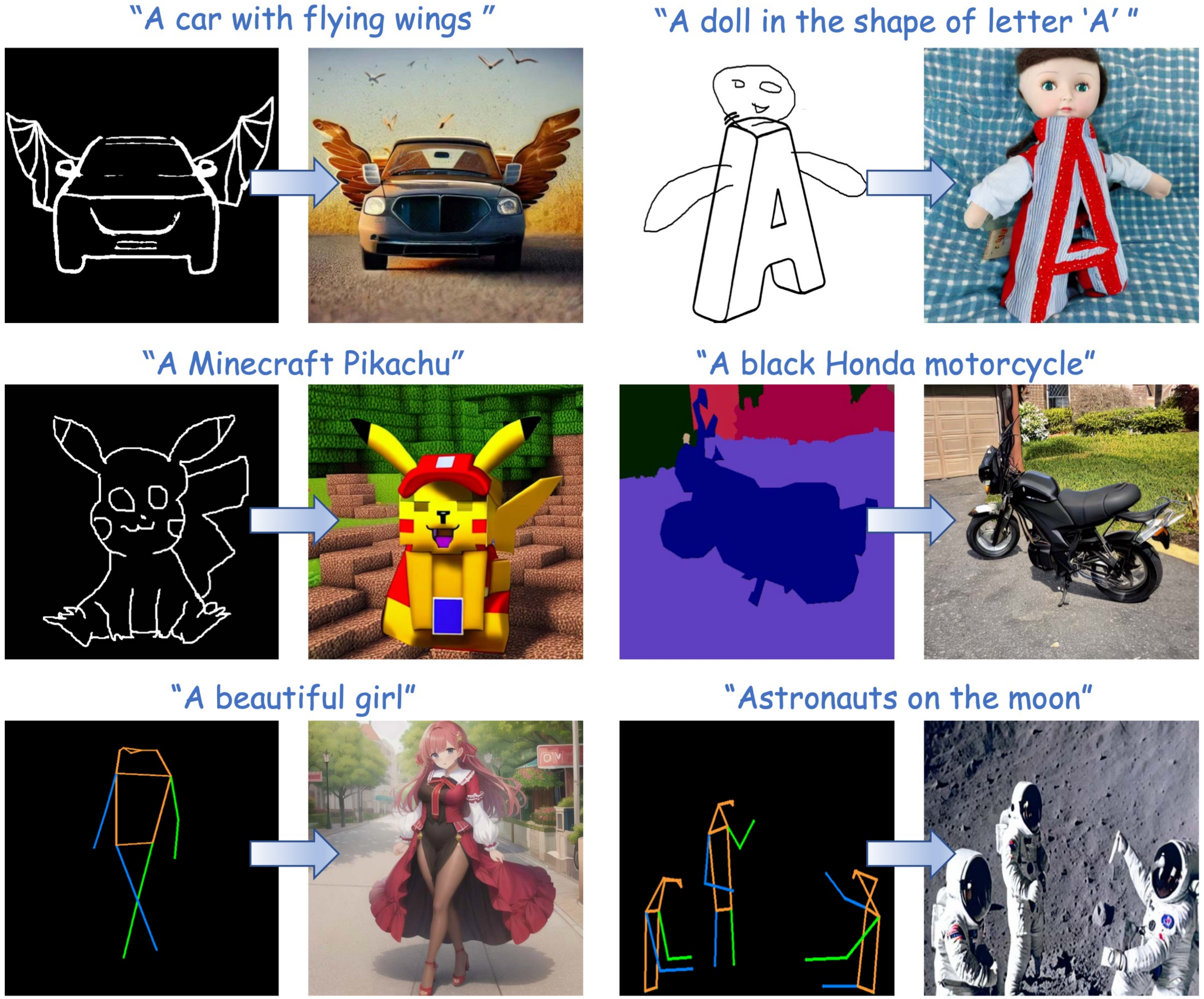
ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment, arXiv 2024

Adding Image Controls Signals for Image Generation

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

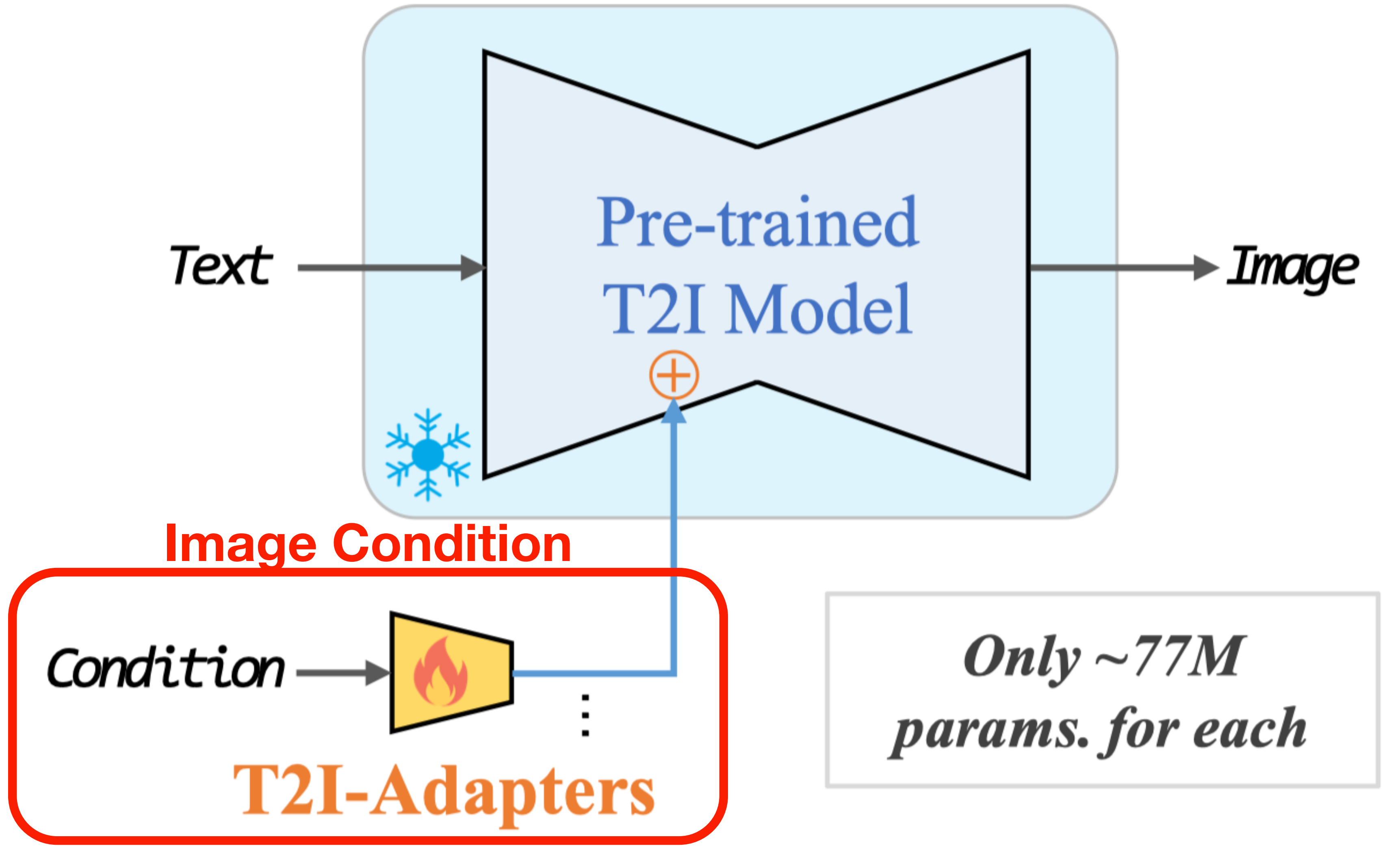


Cartoon line drawing "1girl, masterpiece, best quality, ultra-detailed, illustration"

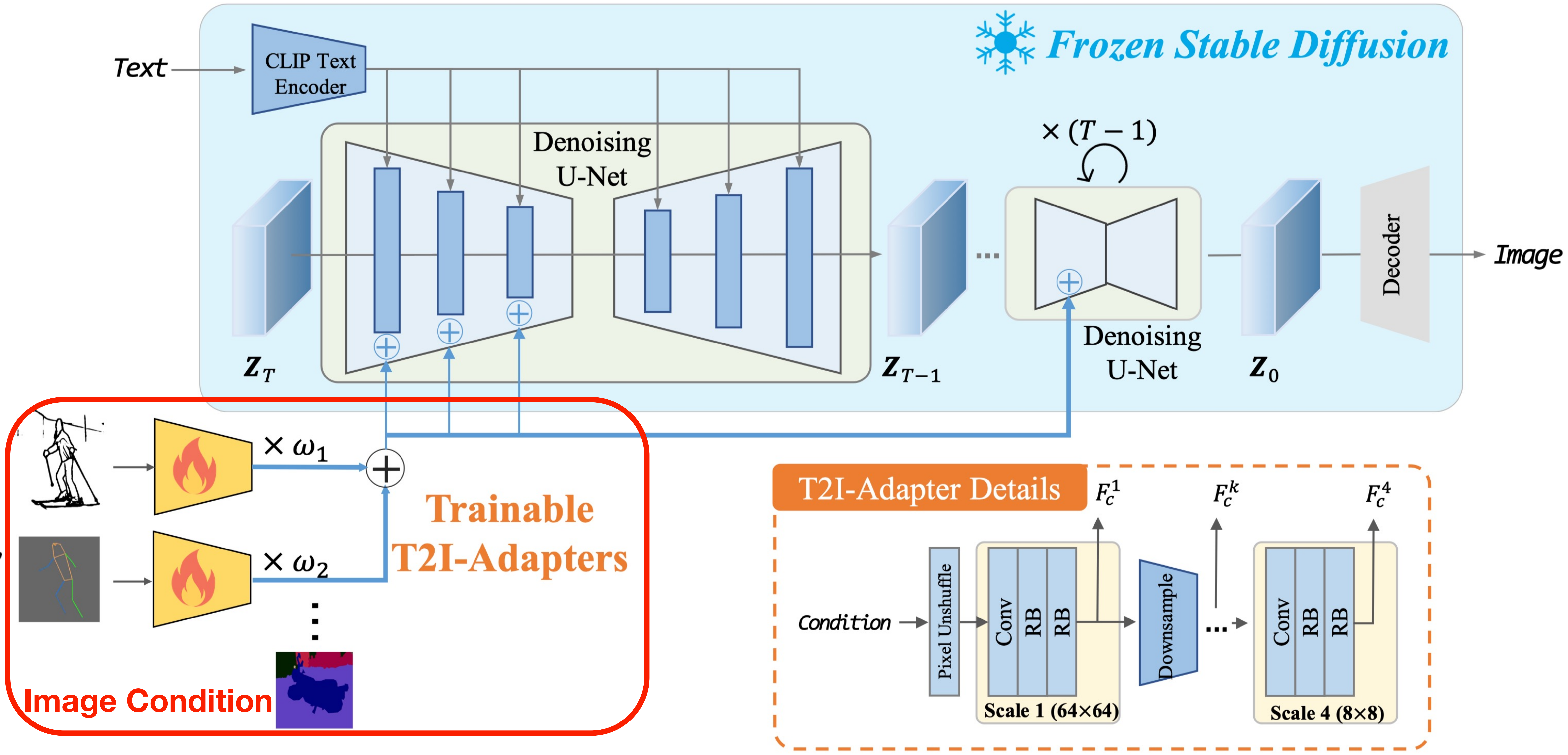


Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2023 Best Paper
T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, AAAI 2024

Encode the Image Features as the Condition for Denoising Training



Only ~77M
params. for each



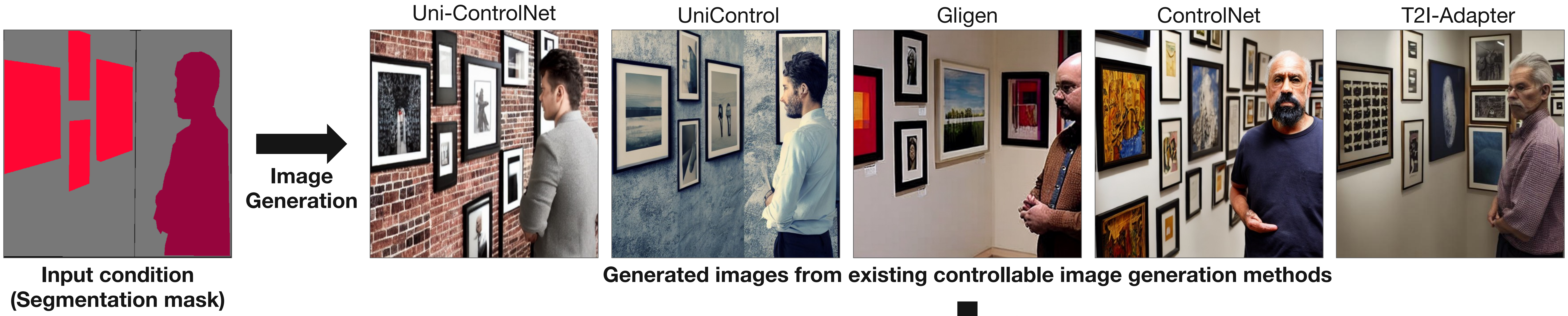
- ✓ ***Plug-and-play.*** *Not affect original network topology and generation ability*
- ✓ ***Simple and small.*** *~77M parameters and ~300M storage*
- ✓ ***Flexible.*** *Various adapters for different control conditions*
- ✓ ***Composable.*** *Several adapters to achieve multi-condition control*
-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

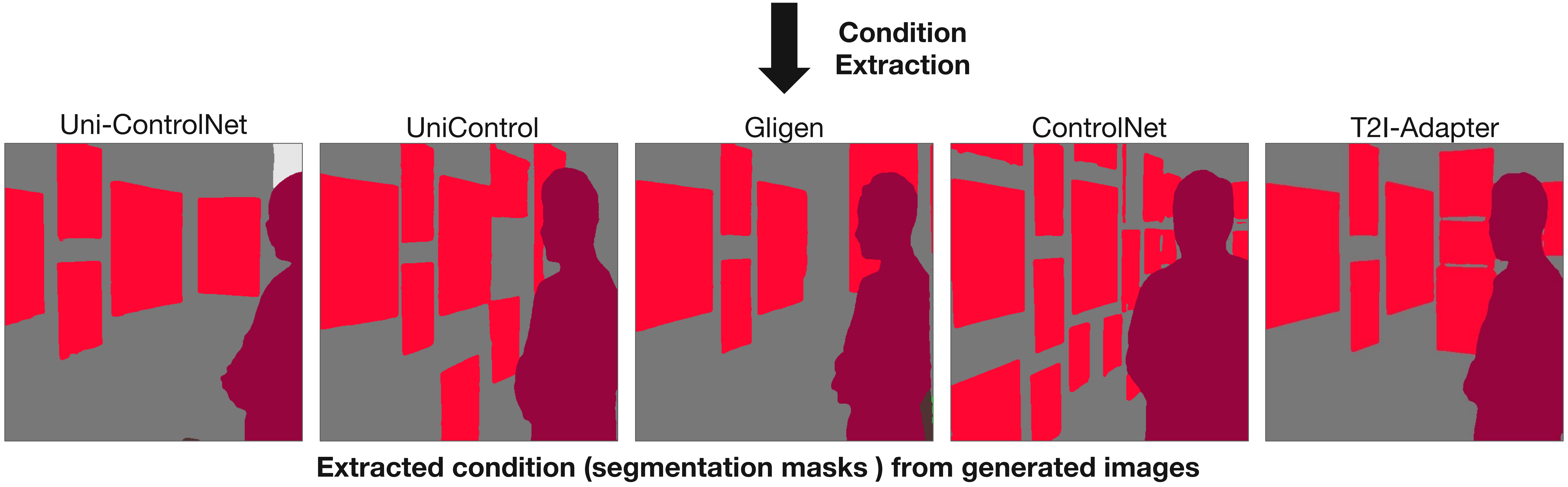
T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, AAAI 2024

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

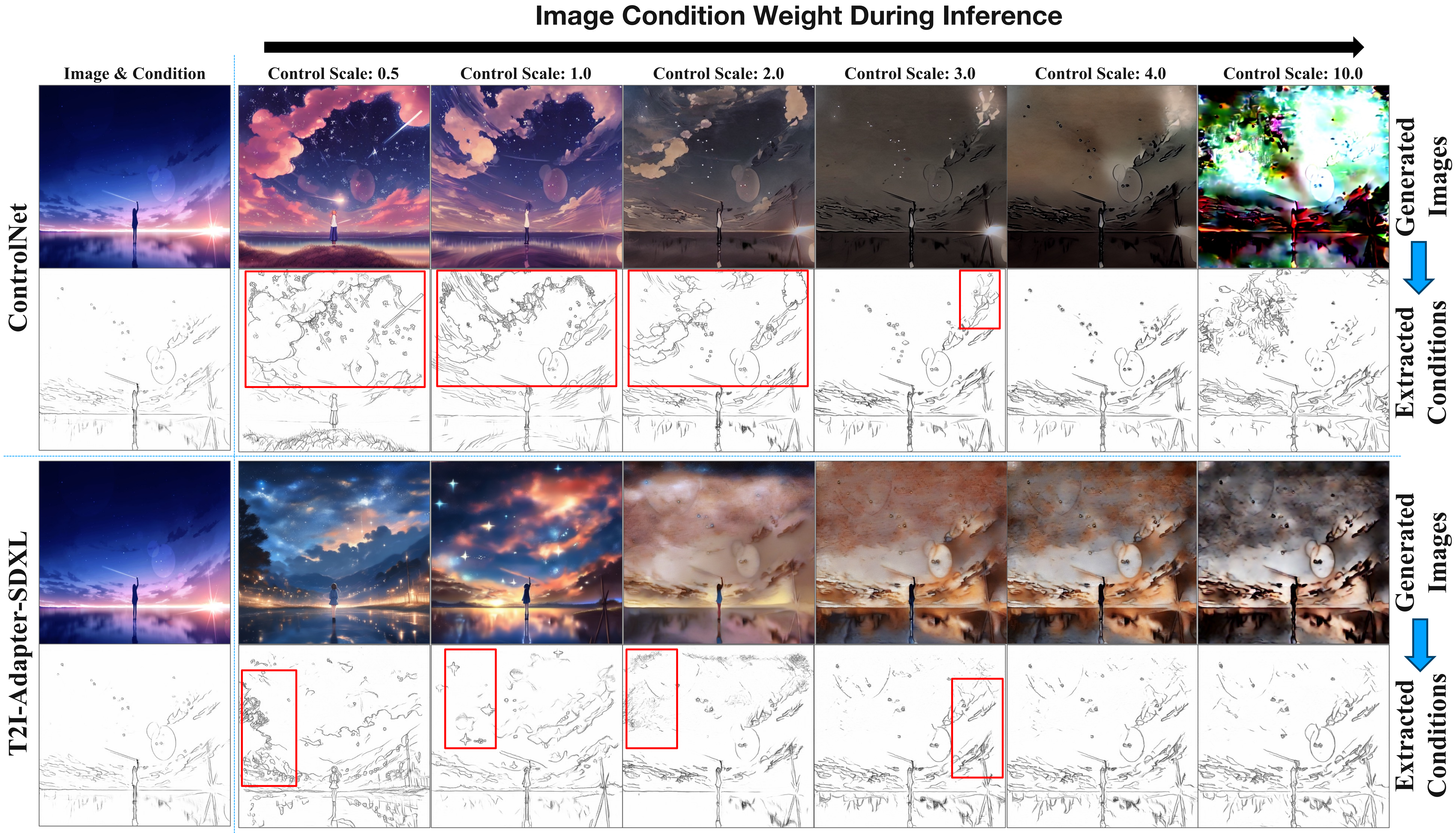
Existing Methods Still Cannot Accurately Control Image Generation





Inconsistencies between input and extracted condition

Controllability Cannot Be Improved by Emphasizing Image Condition



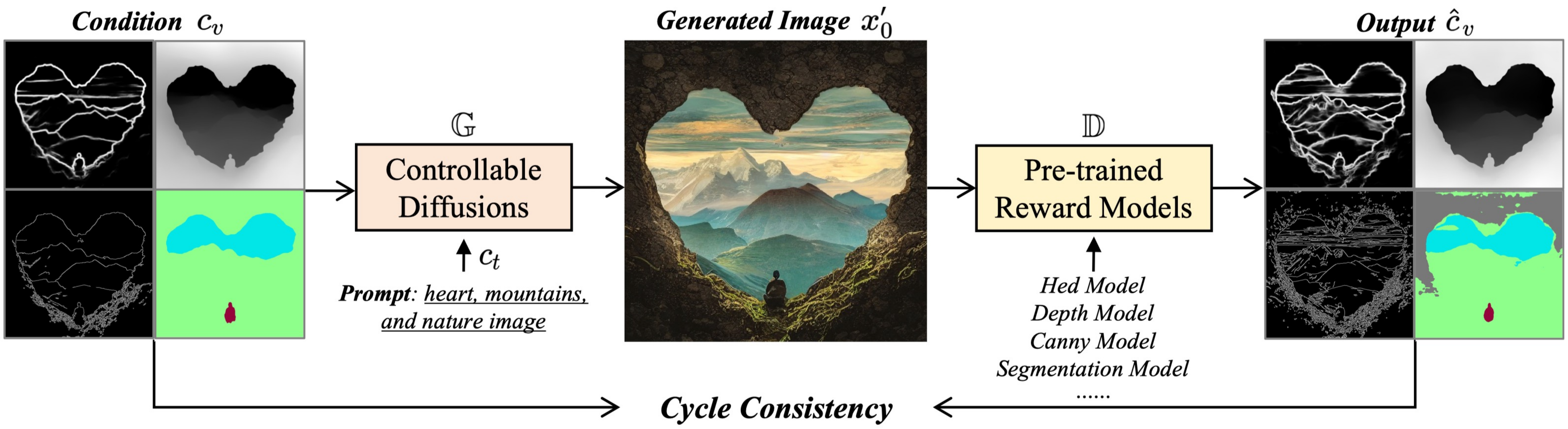
Outline

- 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

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 ($\hat{c}_v = c_v$).

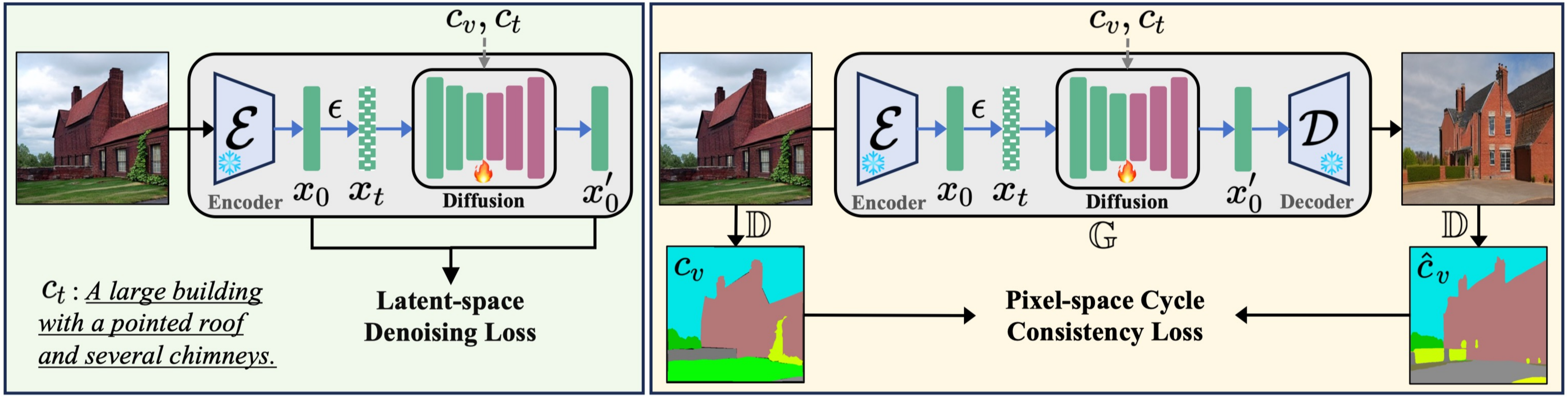


ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback, ECCV 2024
Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks, ICCV 2017

What Makes Our ControlNet++ More Controllable?

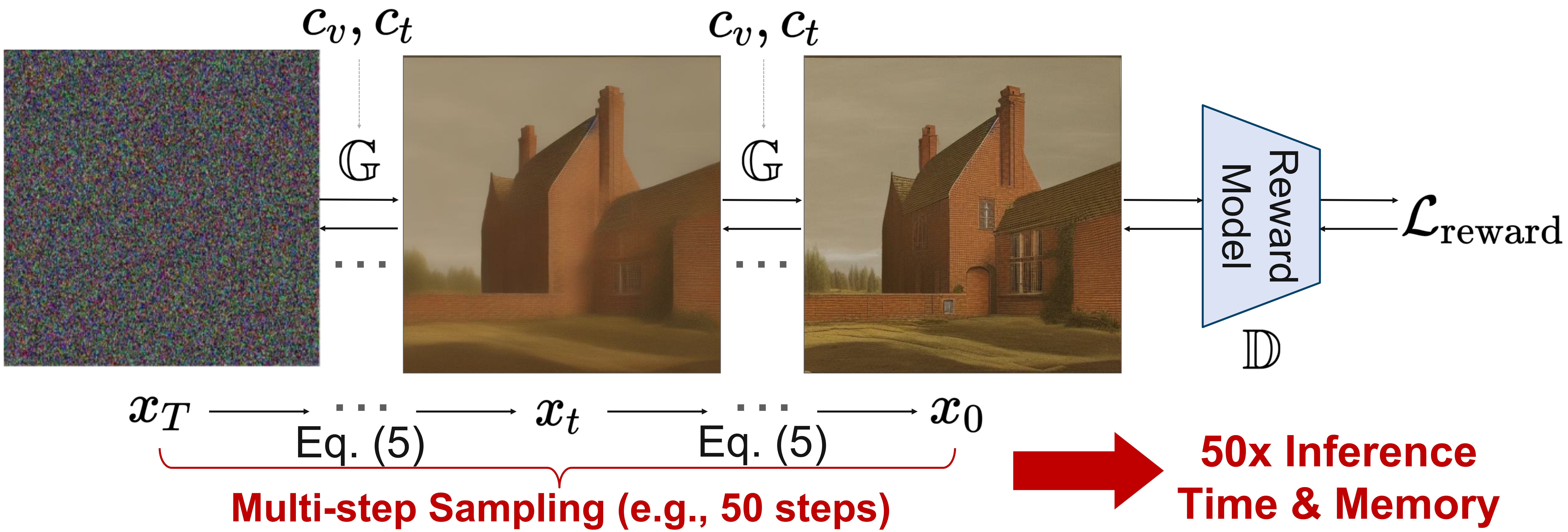
a. Existing methods achieve **implicit** controllability by introducing image-based conditional control c_v into the denoising process of diffusion models, with the guidance of latent-space denoising loss.

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

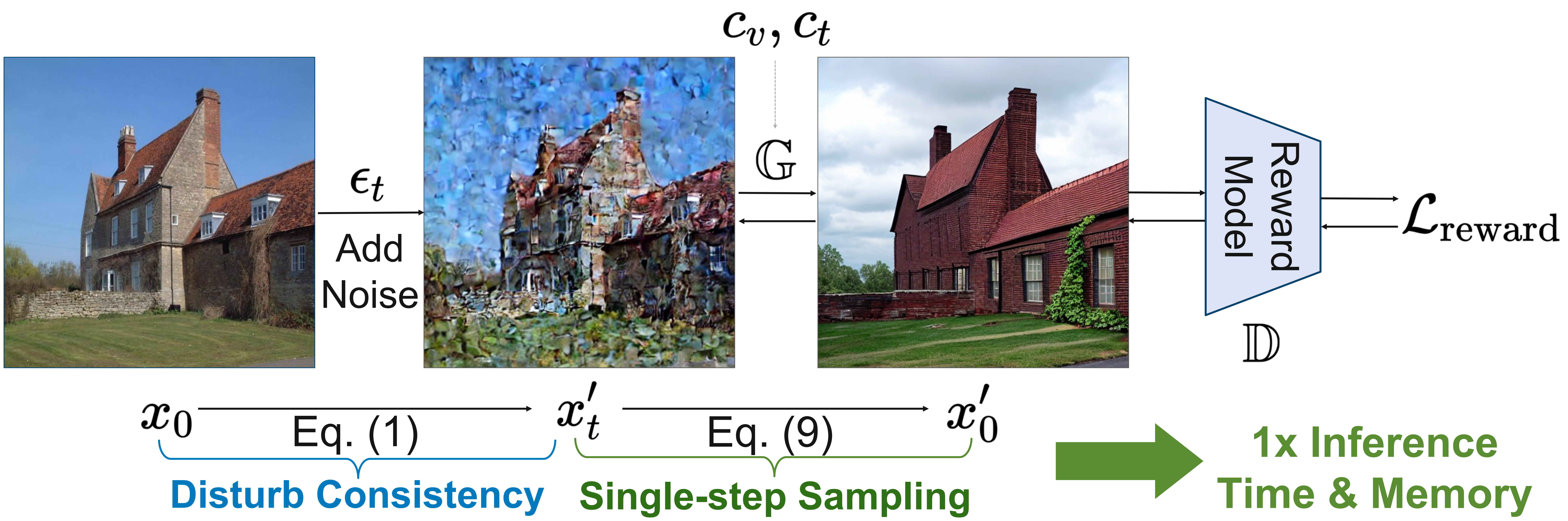


$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \epsilon$$

Eq. (5)

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

Our Efficient Reward Strategy

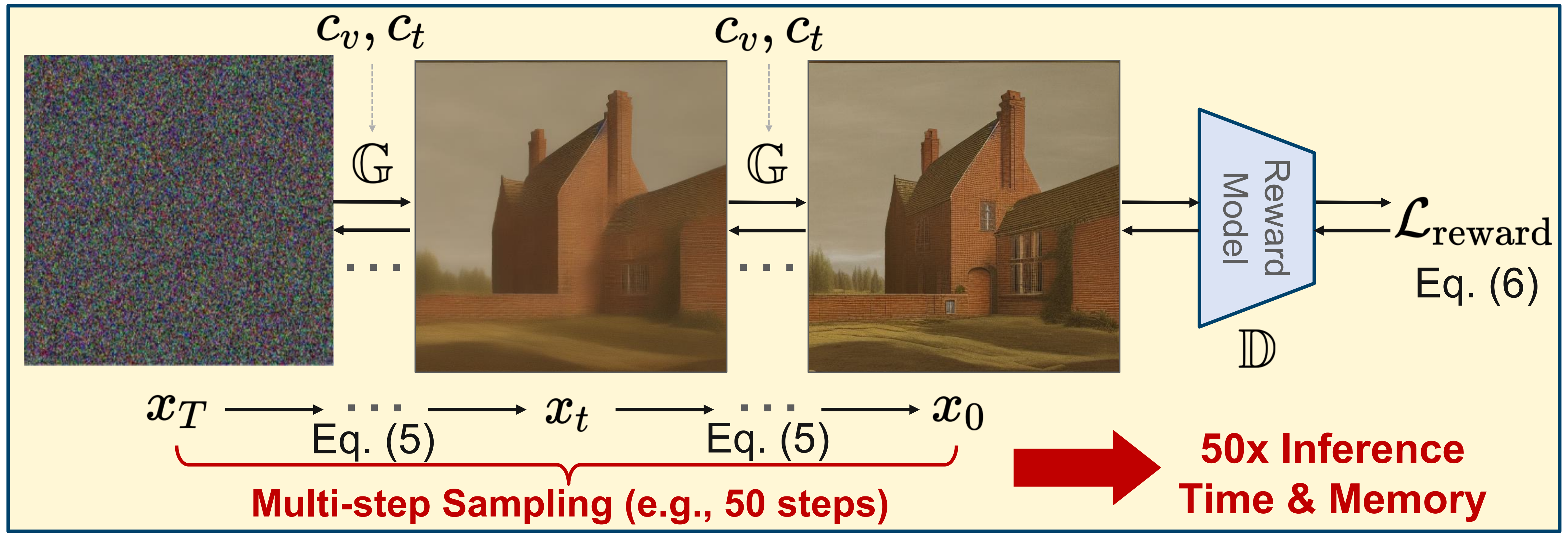


$$x_0 \approx x'_0 = \frac{x'_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x'_t, c_v, c_t, t-1)}{\sqrt{\alpha_t}}$$

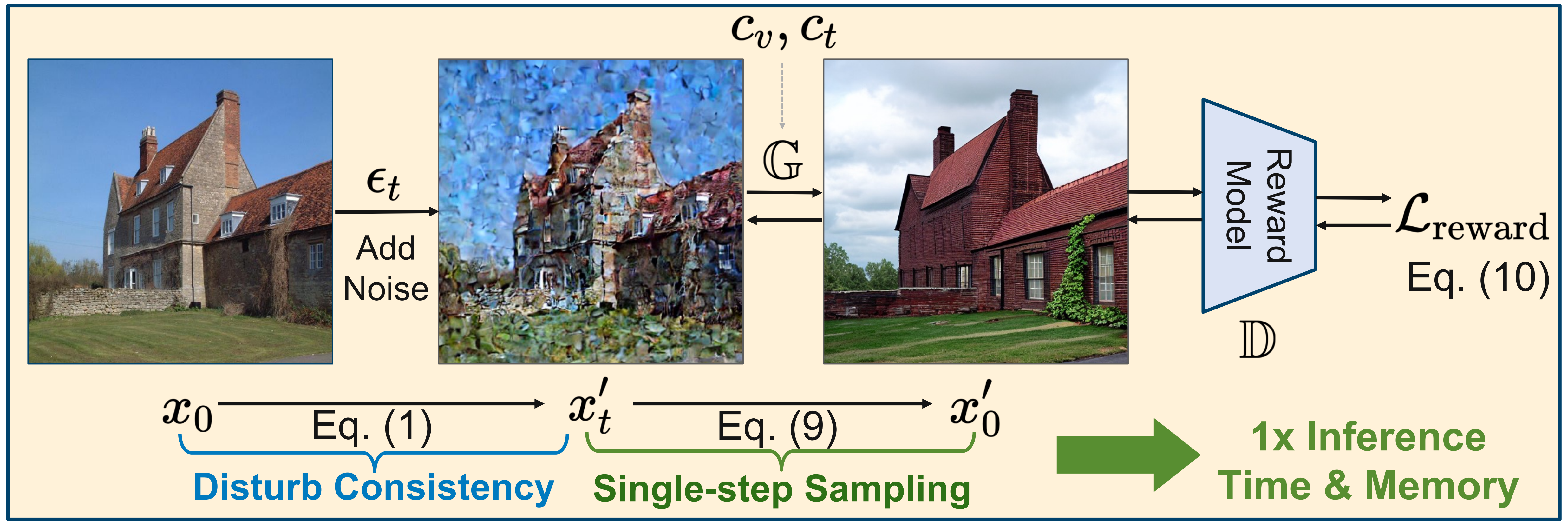
$$\begin{aligned}
\mathcal{L}_{\text{reward}} &= \mathcal{L}(c_v, \hat{c}_v) \\
&= \mathcal{L}\big(c_v, \mathbb{D}(x'_0)\big) \\
&= \mathcal{L}\big(c_v, \mathbb{D}[\mathbb{G}(c_t, c_v, x'_t, t)]\big),
\end{aligned}$$

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)

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

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

step-by-step sampled image

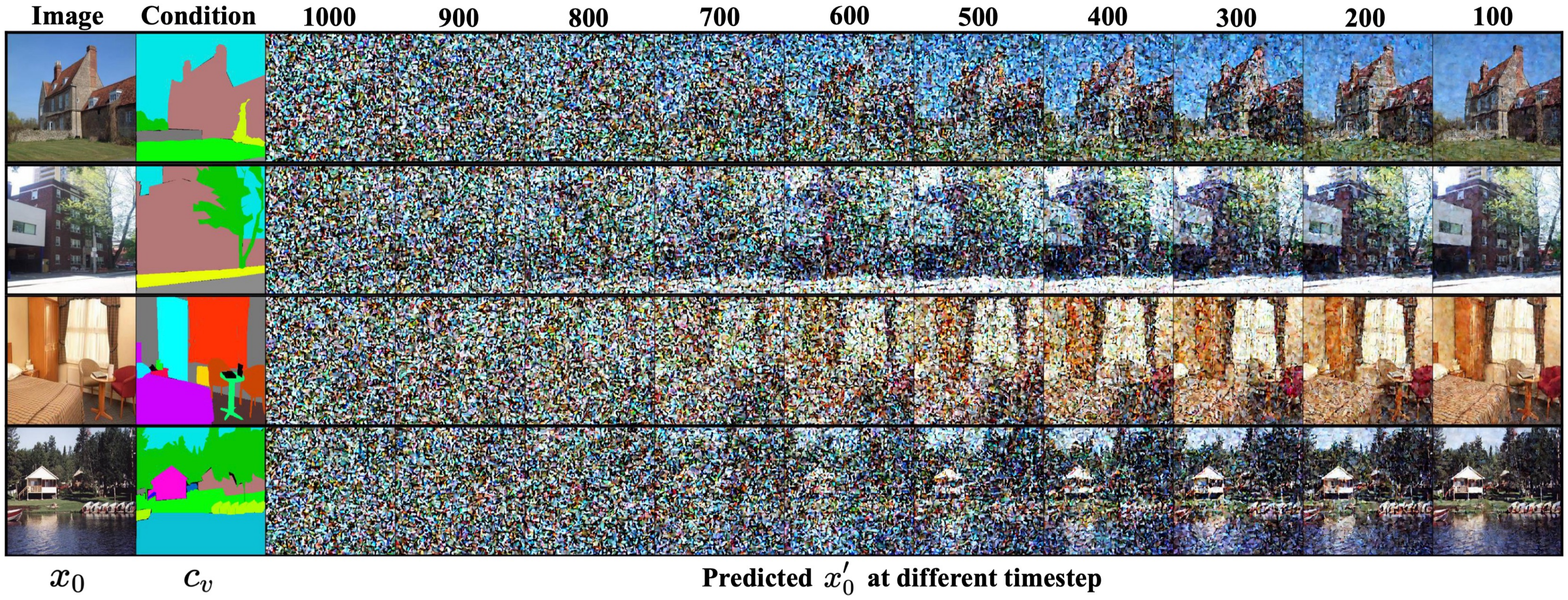
$$x_0 \approx x'_0 = \frac{x'_t - \sqrt{1-\alpha_t}\epsilon_\theta(x'_t, c_v, c_t, t-1)}{\sqrt{\alpha_t}}$$

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

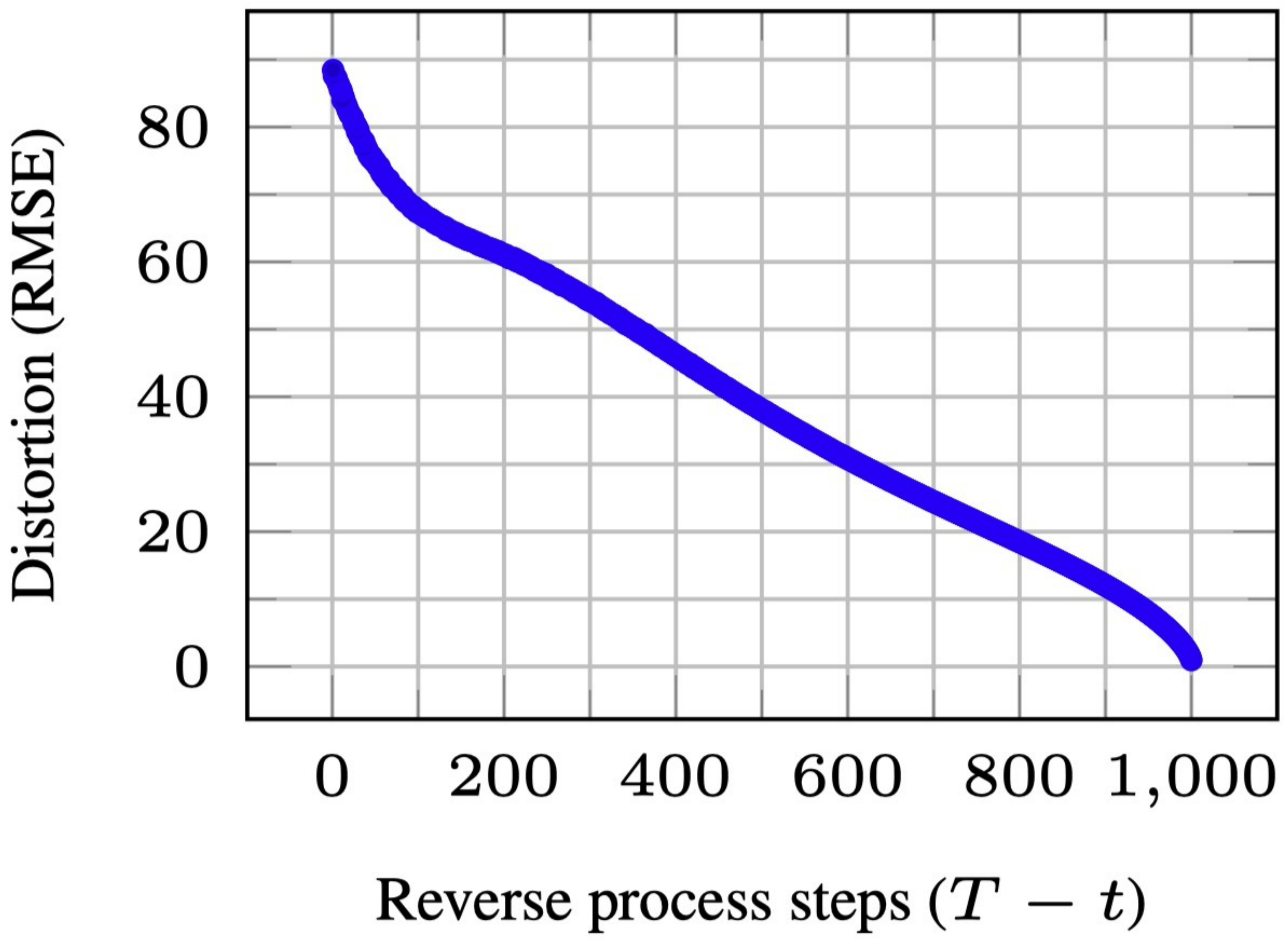
single-step denoised image

Such Estimation is Reasonable When Timestep is Small Enough

Single-step Denoising Visualizations



Estimation Errors



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**

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.
- **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 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!

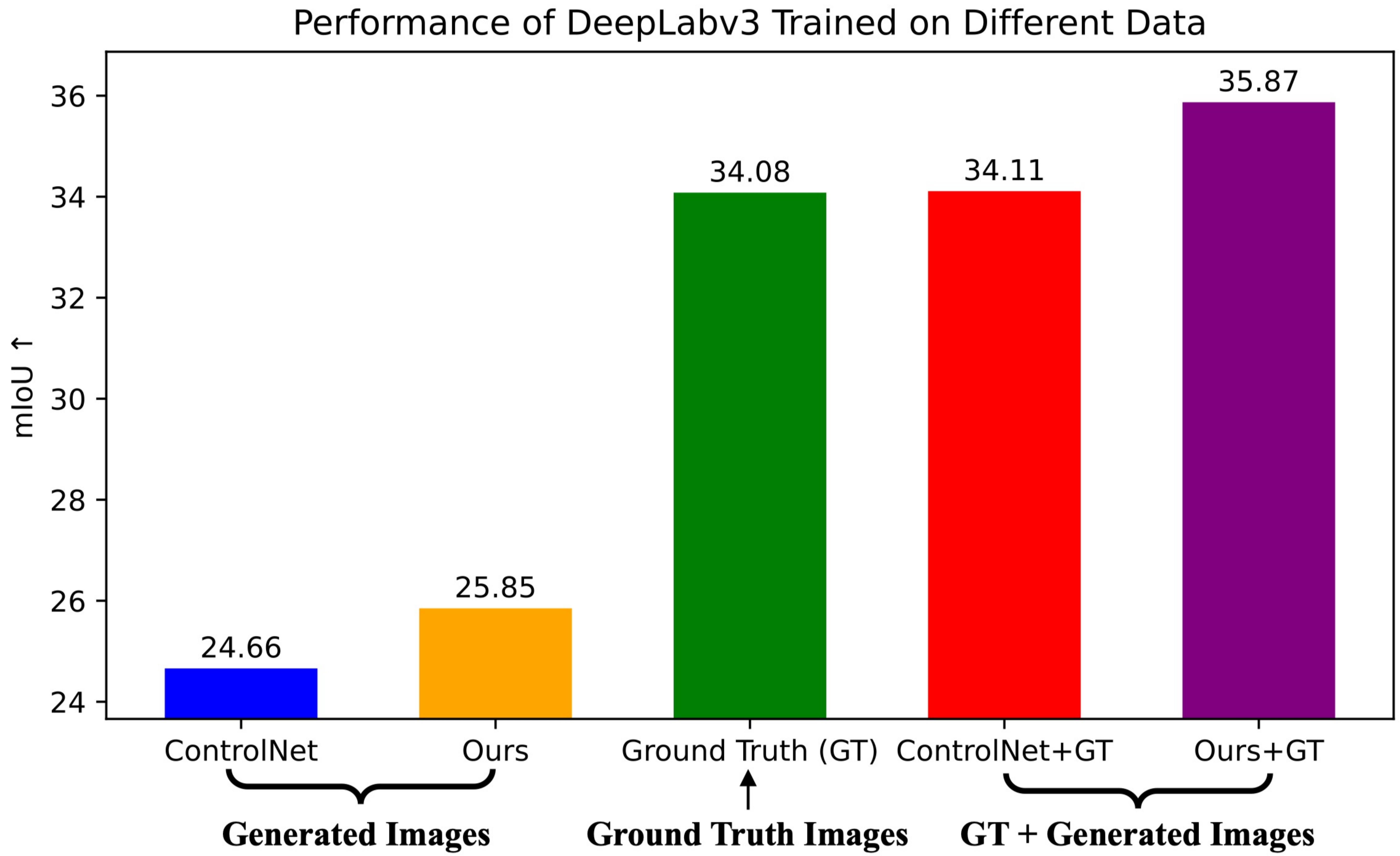


Fig. 5: Training DeepLabv3 (MobileNetv2) from scratch with different images, including ground truth images from ADE20K, and the generated images from ControlNet and ours. All the labels (i.e., segmentation masks) are ground truth labels in ADE20K. **Please note improvements here are non-trivial for semantic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

Ablation Studies

Reward Loss can be applied to time steps that are not explicitly optimized

Table 4: The impact of efficient reward fine-tuning on different timesteps.

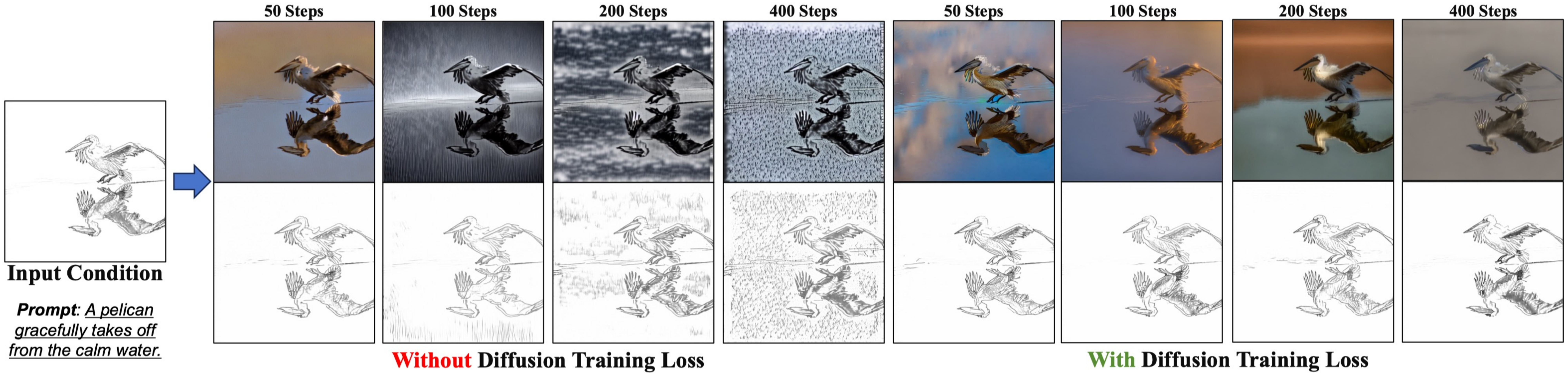
Unoptimized $[T, t_{thre}]$	**Optimized** $[t_{thre}, 1]$	**ADE20K mIoU (↑)**
ControlNet	ControlNet	32.55
ControlNet	Ours	38.03
Ours	ControlNet	41.46
Ours	**Ours**	**43.64**

More powerful reward model leads to better controllable diffusion models

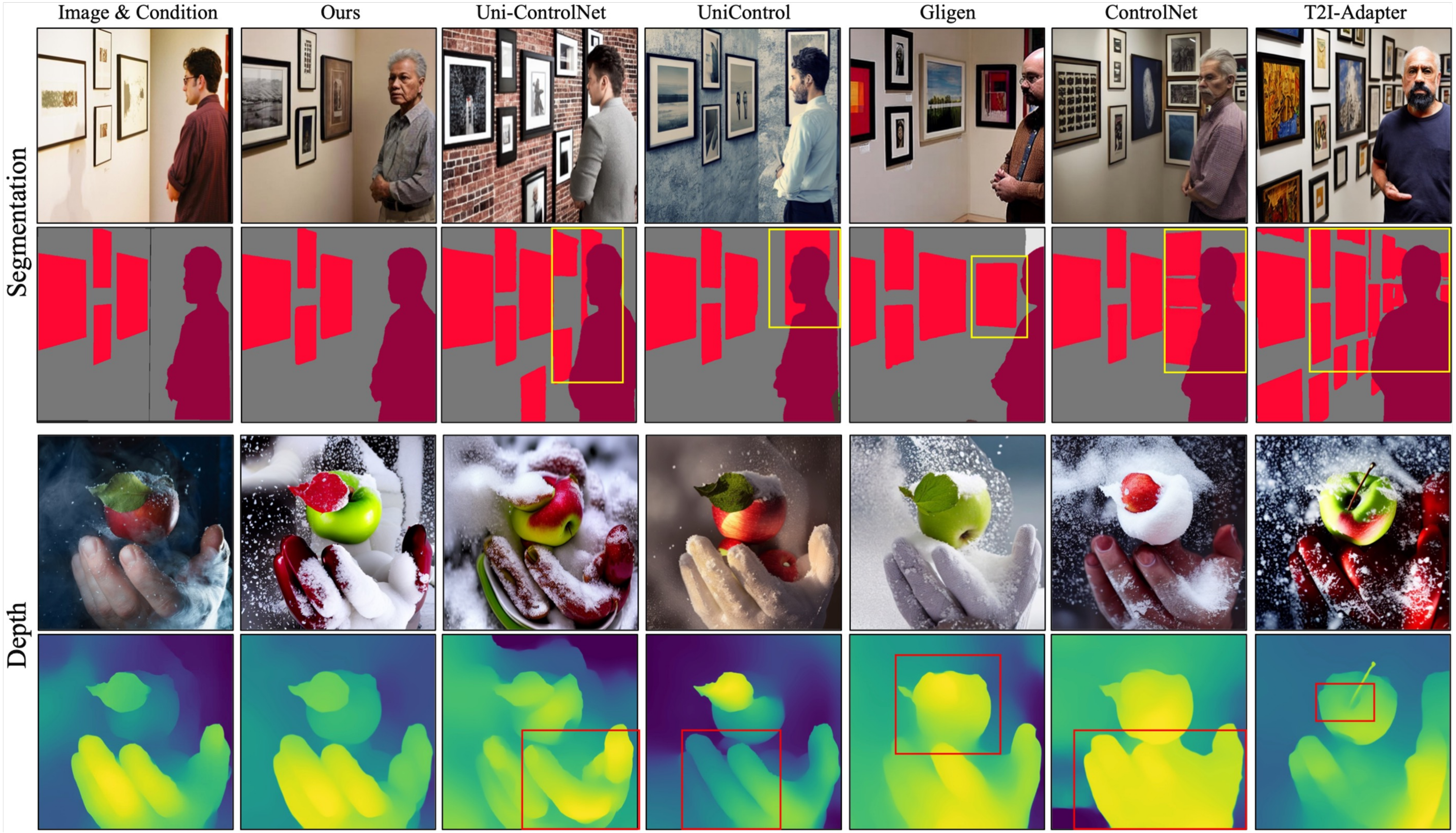
Table 5: Stronger reward model (UperNet-R50) leads to better controllability than the weaker reward model (DeepLabv3-MBv2).

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

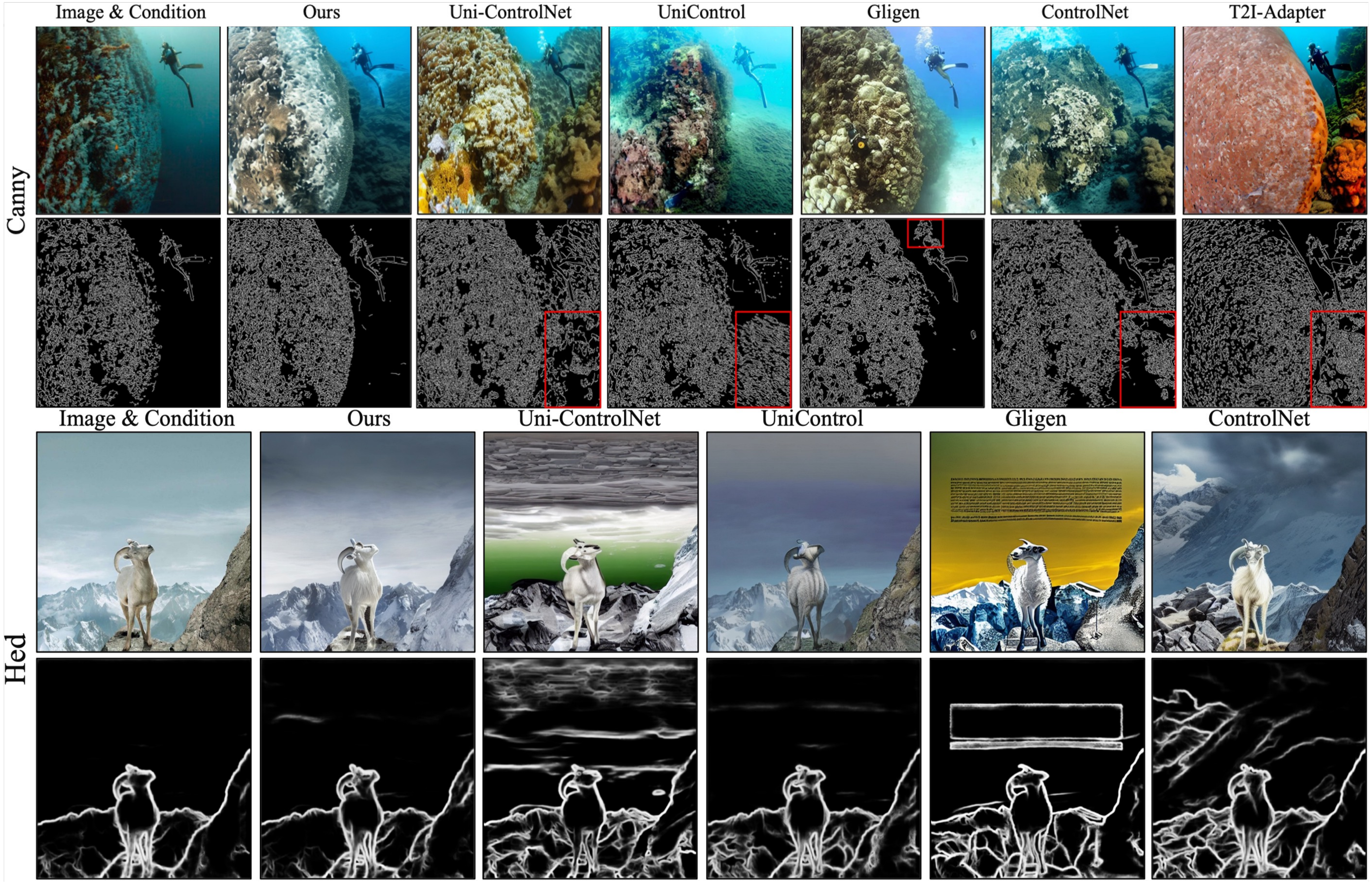


Visualization Comparison



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

Visualization Comparison



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

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)

Thanks