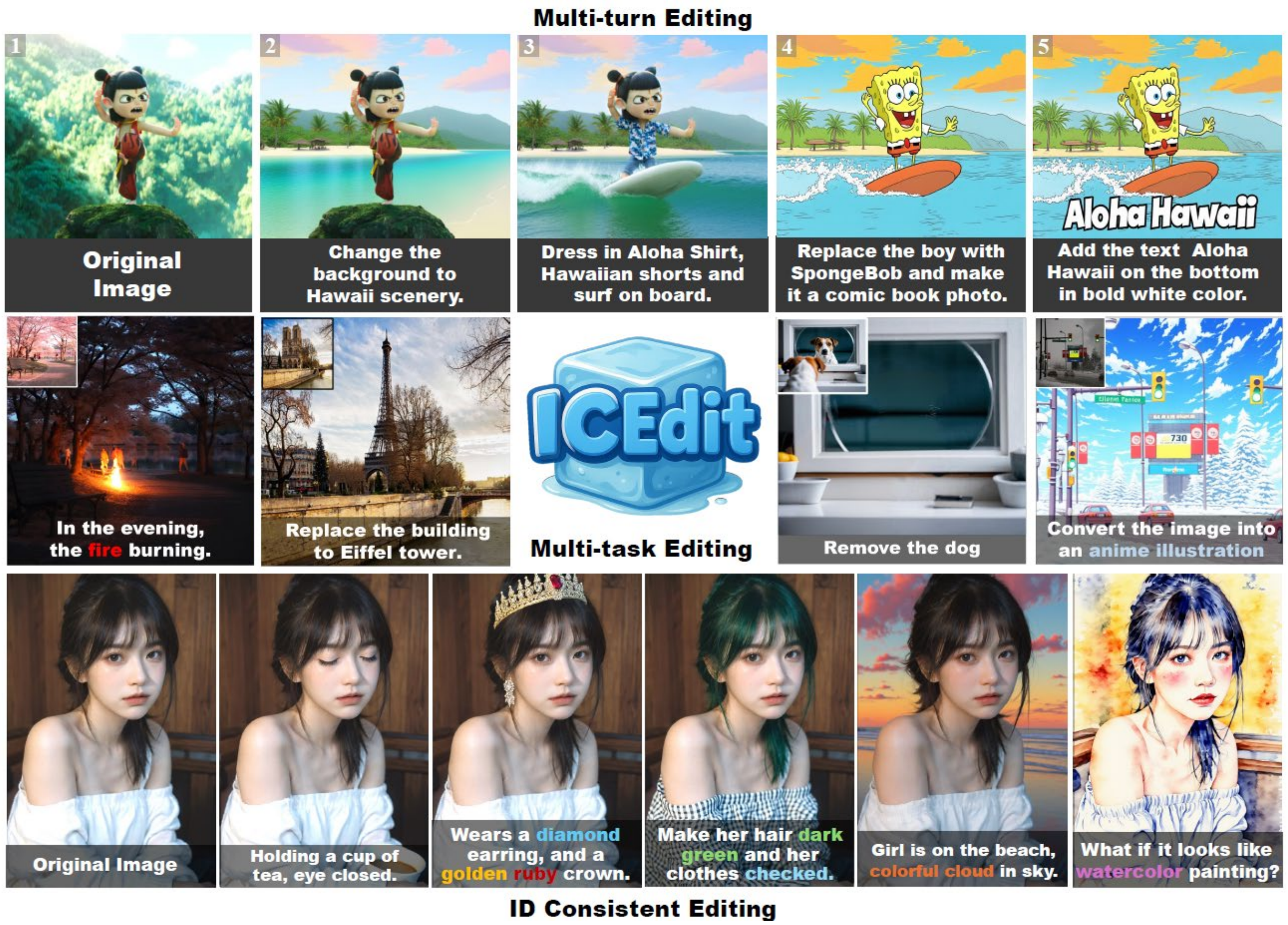
Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer



Zechuan Zhang (张泽川)

I am currently a PhD student at Zhejiang University in Hangzhou, China. My PhD supervisor is Prof. Yi Yang. My research interests lie in the intersection of computer vision, machine learning. I am particularly interested in 3D vision, multi-modal, diffusion models and image generation and editing.

Prior to that, I obtained the B.Sc Degree in Geographical Information Science from Zhejiang University in 2023. I was also a member of Advanced Honor Class of Engineering Education (ACEE) at Chu Kochen Honors College (CKC) of Zhejiang University.

Email | CV | GitHub | Google Scholar | Twitter

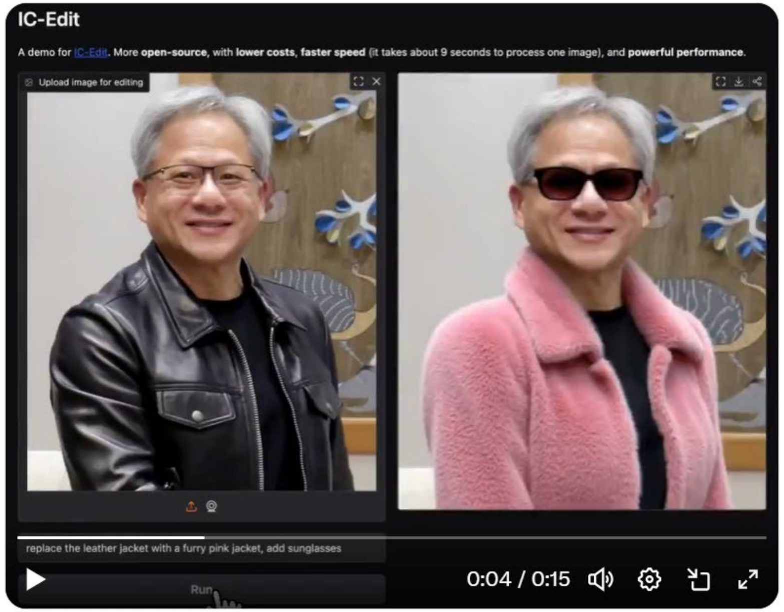
ZHEJIANG UNIVERSITY

Image Editing via In-Context Generation

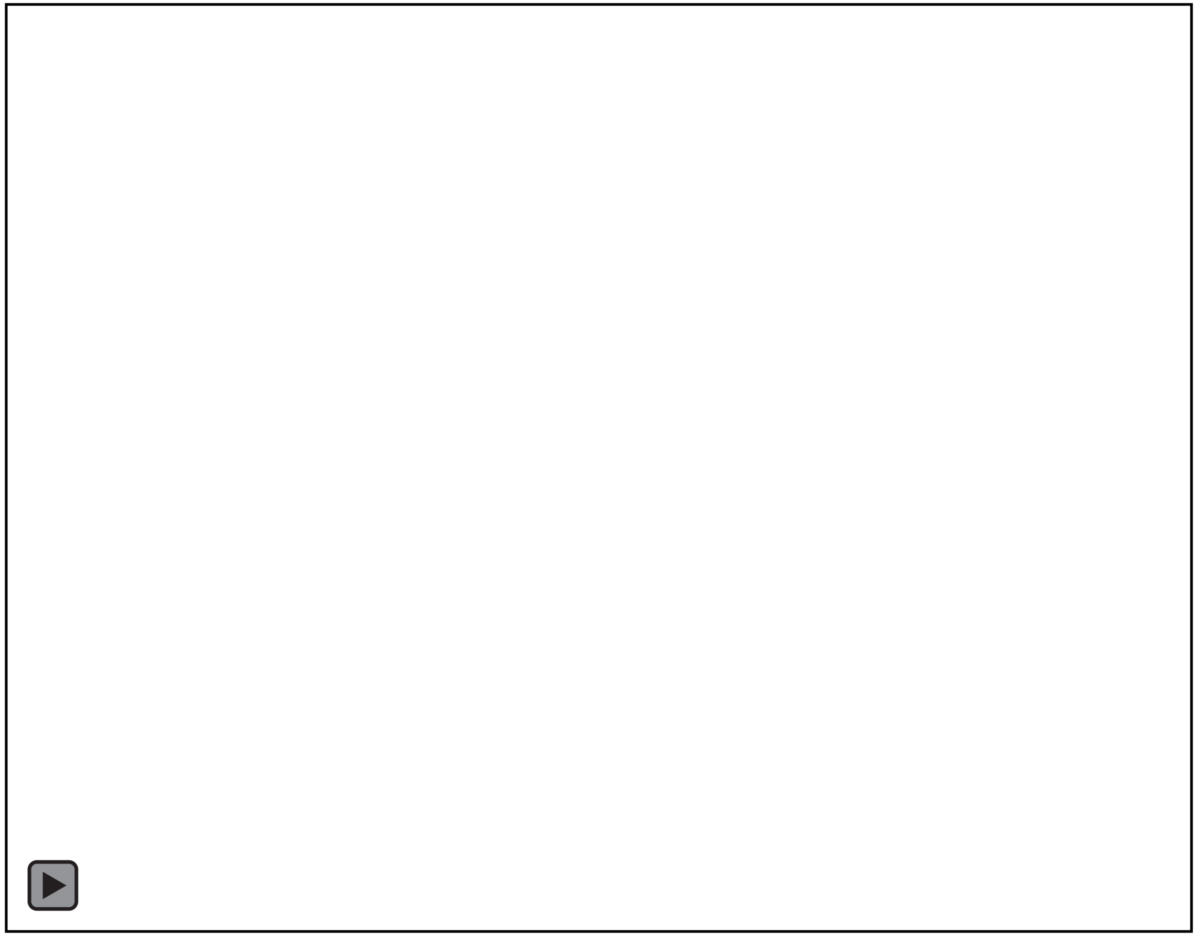
apolinario 🌐
@multimodalart

Whoa, I just tested the IC Edit @huggingface demo and it seems the new 🐐👑 of image editing for

It's an image editing LoRA for FLUX featuring:
🧑‍🎨 Identity preservation (beating GPT-4o)
✏️✏️ Does multiple edits
🐎 10s image editing
🌌 style support



5:05 PM · May 2, 2025 · **41.7K** Views



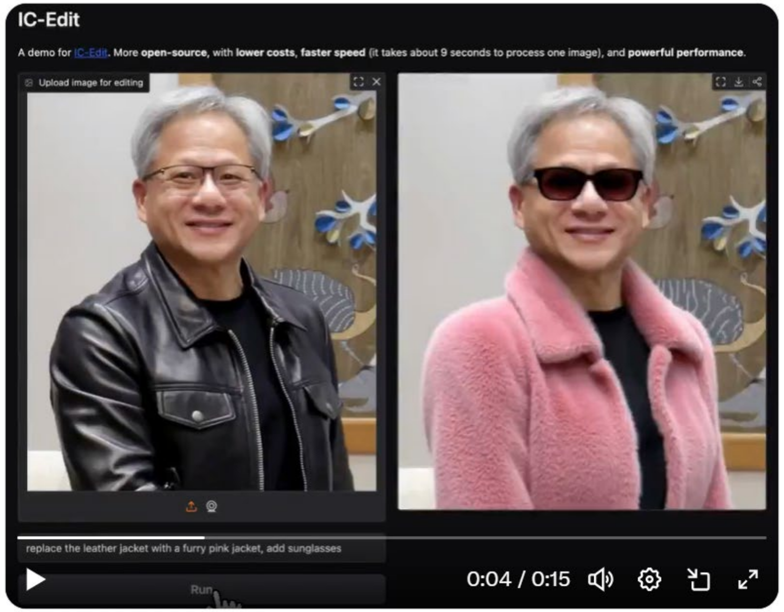
Impact on community

Gaining massive traction on Twitter, Reddit, and the ComfyUI community!

apolinario @multimodalart

Whoa, I just tested the IC Edit @huggingface demo and it seems the new 🐐👑 of image editing for

It's an image editing LoRA for FLUX featuring:
🧑‍🎨 Identity preservation (beating GPT-4o)
✏️✏️ Does multiple edits
🐎 10s image editing
🌌 style support



5:05 PM · May 2, 2025 · **41.7K** Views

r/StableDiffusion • 2 days ago
Some_Smile5927 Top 1% Poster

ICEdit, I think it is more consistent than GPT4-o.

Workflow Included

In-Context Edit, a novel approach that achieves state-of-the-art instruction-based editing **using just 0.5% of the training data and 1% of the parameters required by prior SOTA methods**.
https://river-zhang.github.io/ICEdit-gh-pages/

I tested the three functions of image deletion, addition, and attribute modification, and the results were all good.

317 · 77 · Share

Impact on community

- Attracted over **500,000 followers** across all platforms (first month)
- Gained more than **1,500 stars** on GitHub
- Remained ranked second on the Hugging Face trending list for a week.

Analytics Hugging face

Last update: 3 hours ago · Next update: in 20 hours

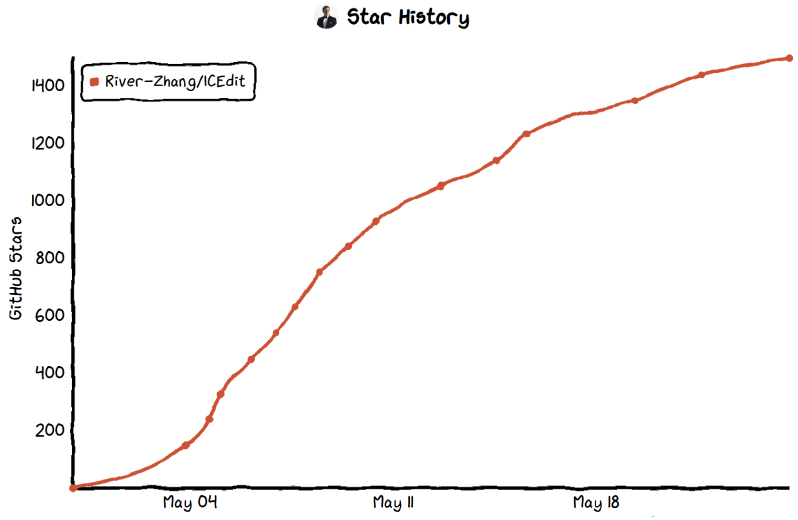
All time visits	Last month visits
352,268	352,268

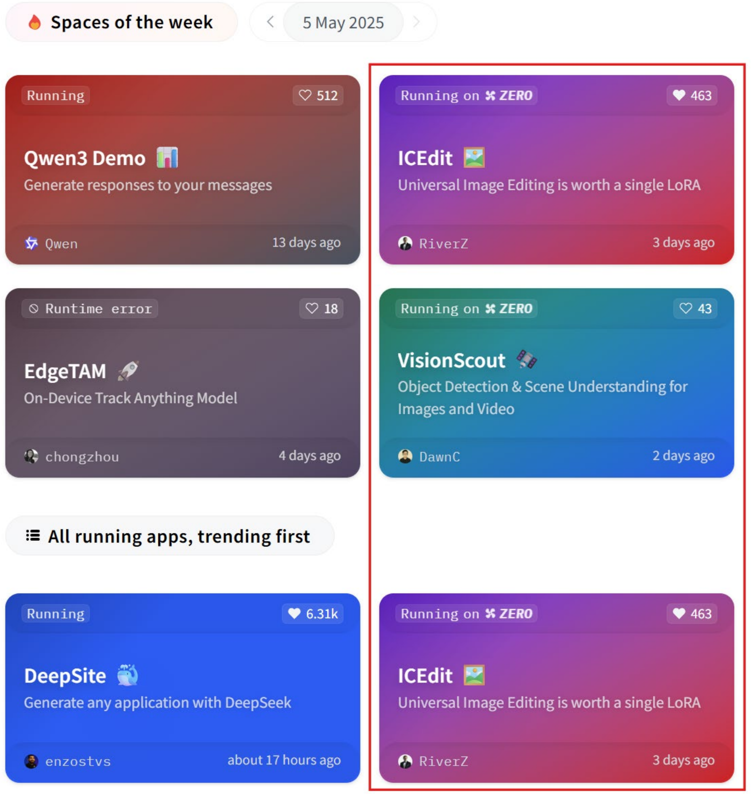
Project Page

River-Zhang/ICEdit

Image editing is worth a single LoRA! 0.1% training data for fantastic image editing! Training released! Surpasses GPT-4o in ID persistence! Official ComfyUI workflow release! Only 4GB VRAM is enou...

Python ★ 1.5k





https://github.com/River-Zhang/ICEdit

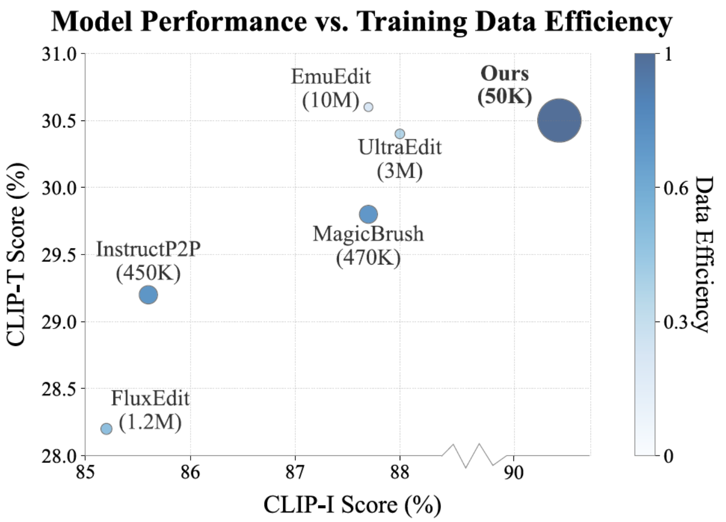
Introduction

Background:

- **Instruction-based image editing has garnered significant attention for its ability to transform and process images using natural language prompts。**

Challenge:

- Existing methods struggle to **balance precision and efficiency**.
- **Training-intensive approaches** achieve high-fidelity instruction execution but sacrifice efficiency, relying on: 10M+ training examples for instruction understanding; Large-scale model parameters (e.g., billions)
- **Training-free methods** save costs but suffer from poor editing precision.



- ➢ Our approach uses only **1% training parameters and 0.1% training data** for better performance.

- ➢ Training-free methods cannot directly understand instructional prompts and often fail to handle complex tasks

Motivation

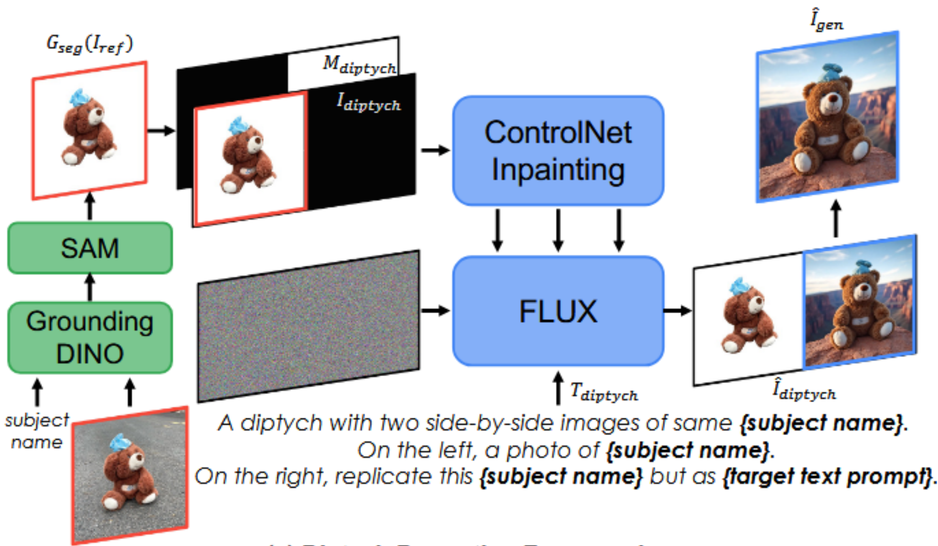
浙江大學 ZHEJIANG UNIVERSITY

How to **balance precision and efficiency:** motivation from previous work based on Large Scale DIT

- ➢ No additional structure modification
- ➢ No fully tuning (e.g. LoRA)

Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator

Chaehun Shin[1] Jooyoung Choi[1] Heeseung Kim[1] Sungroh Yoon[1,2,*]
[1]Data Science and AI Laboratory, ECE, Seoul National University
[2]AIIS, ASRI, INMC, ISRC, and Interdisciplinary Program in AI, Seoul National University



(a) Diptych Prompting Framework

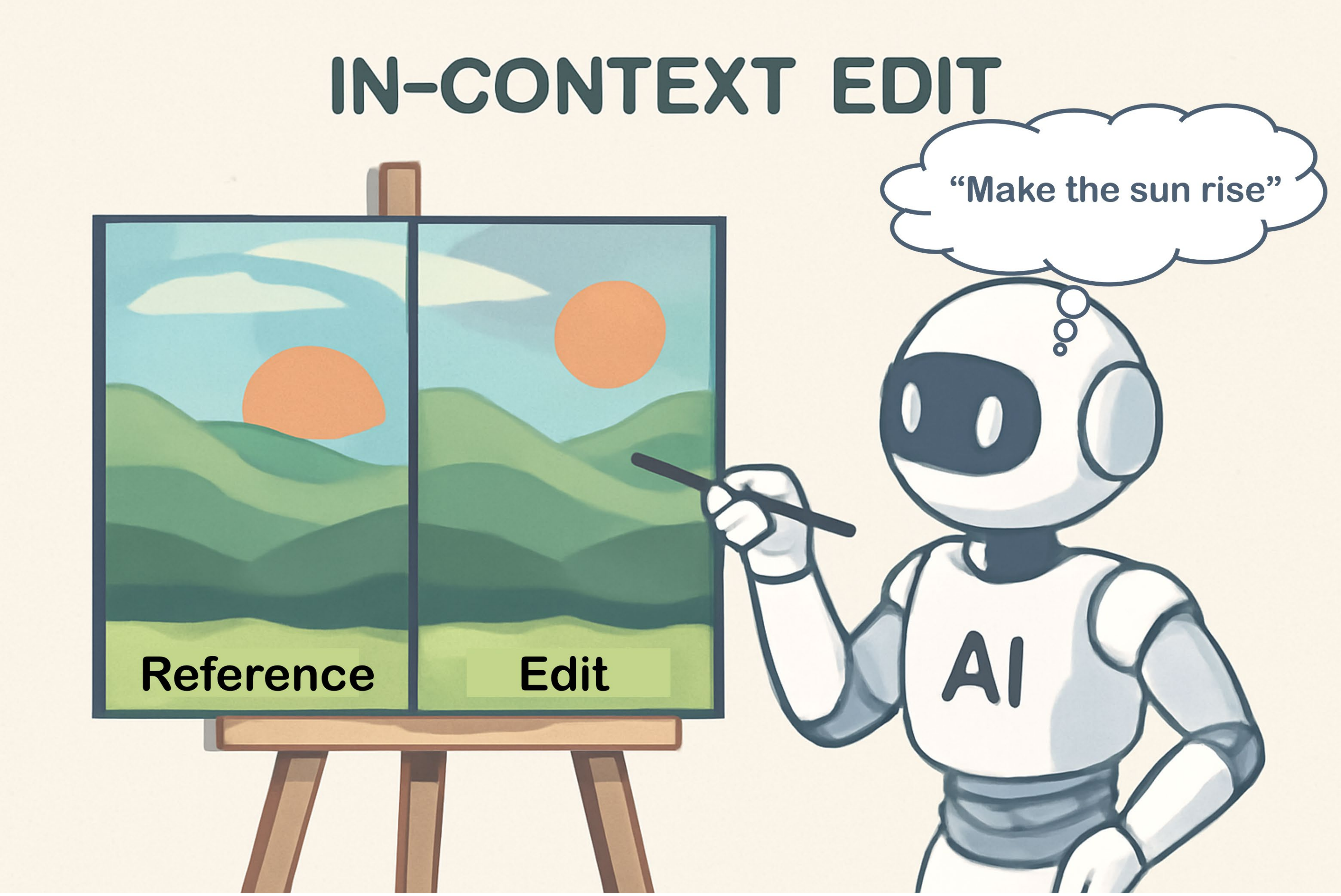
OminiControl: Minimal and Universal Control for Diffusion Transformer

Zhenxiong Tan Songhua Liu Xingyi Yang Qiaochu Xue Xinchao Wang
National University of Singapore
{zhenxiong, songhua.liu, xyang, e1352520}@u.nus.edu xinchao@nus.edu.sg

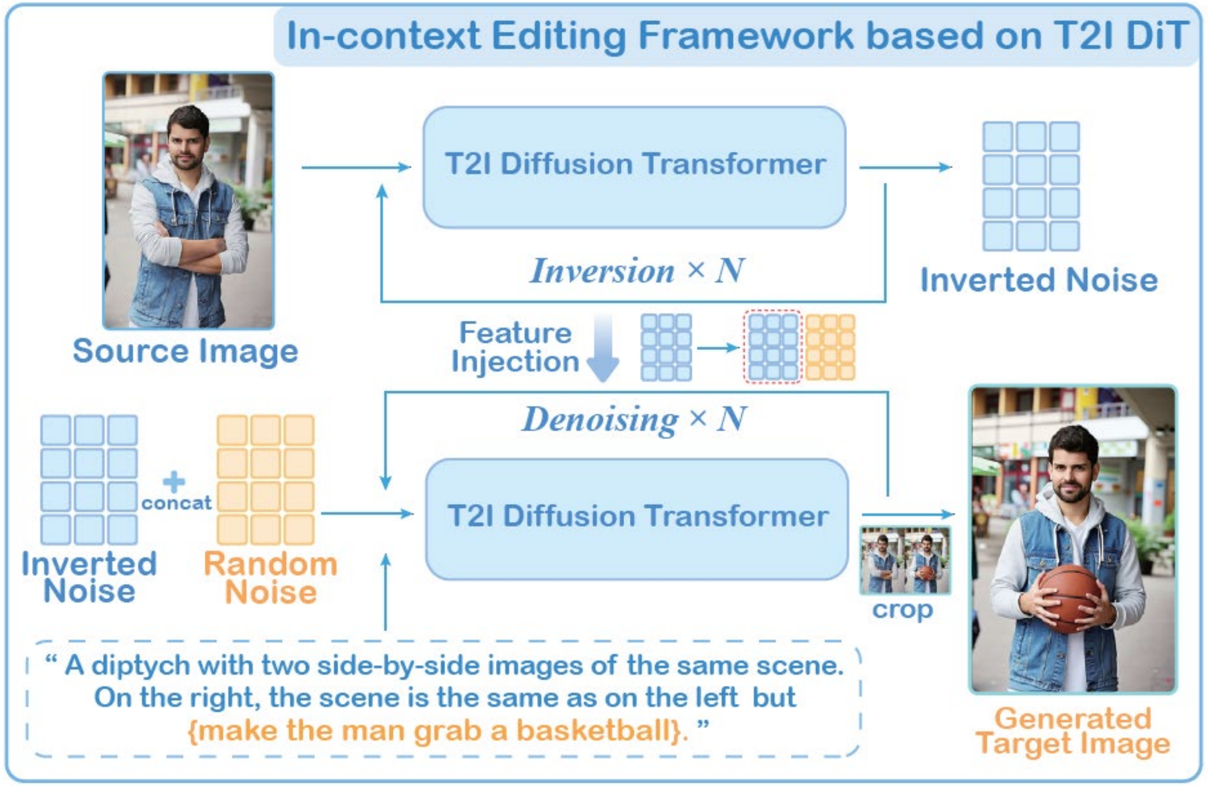
求是創新 Can large scale DiT's generation capacity and contextual awareness directly address instruction-based image editing?

Motivation

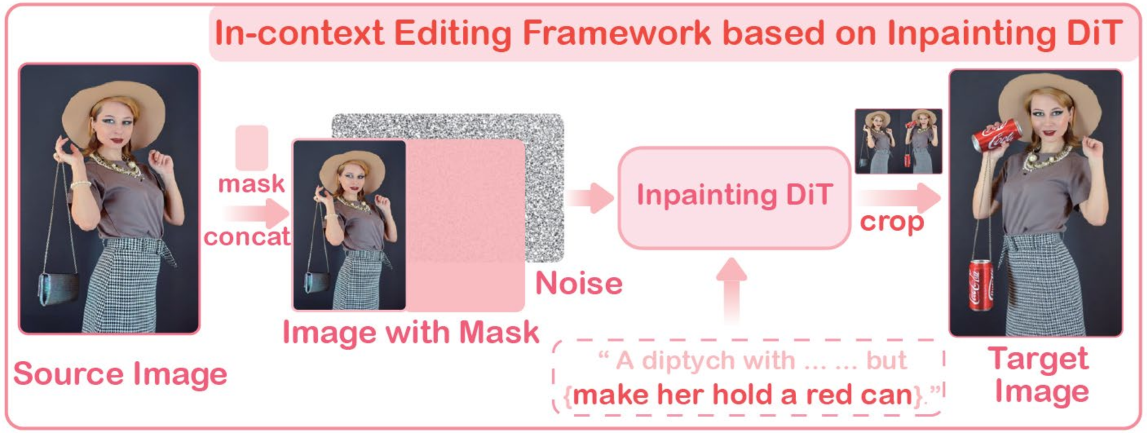
浙江大学 ZHEJIANG UNIVERSITY



Exploration



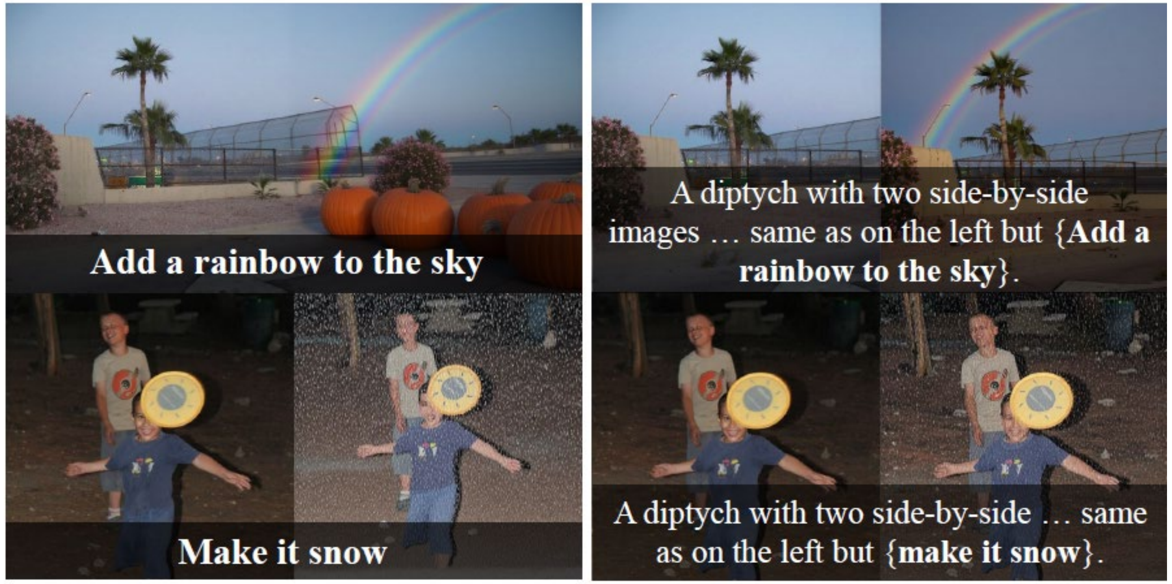
T2I model based framework



Inpainting model based framework

Exploration-Input Prompt Variants

- In-context Prompt – instructions embedded in structure "A diptych with... On the right, the same scene but {instruction}";
- Global Descriptive Prompt - uses full input/output captions ("On the left {input} On the right {output}").



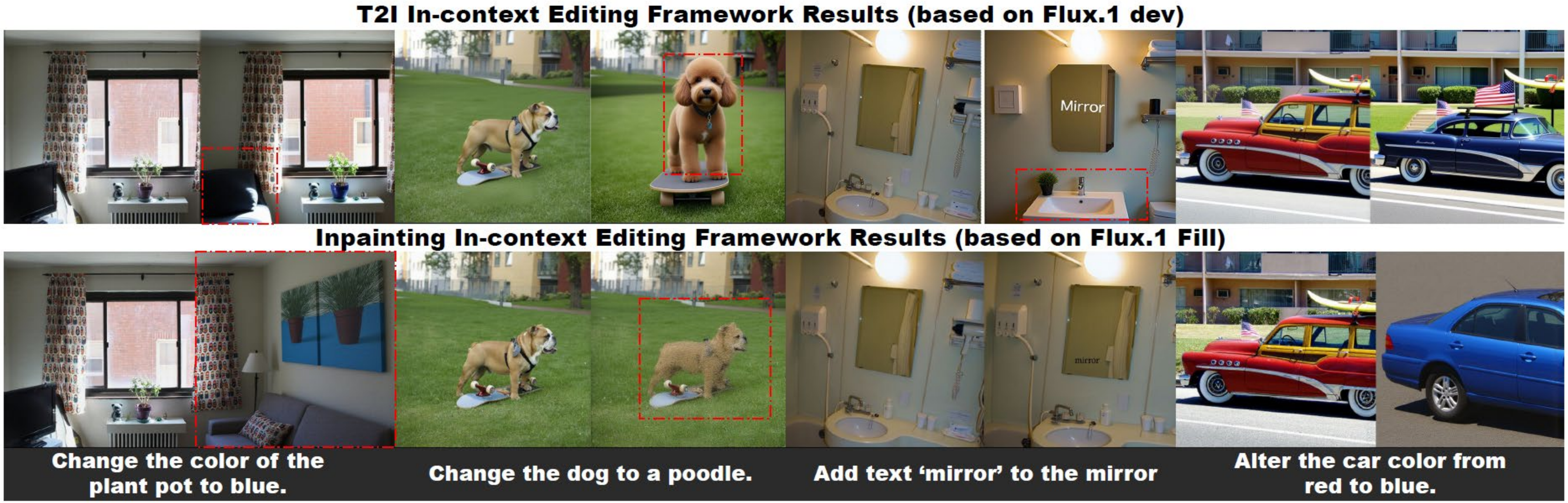
Direct Edit Instruction   In-Context Edit Prompt

Global Descriptive Prompt

Exploration

Training-Free Methods Show Limited Performance.

- Both T2I and inpainting DiT frameworks (based on Flux) yield suboptimal results.
- Despite these shortcomings, both demonstrate potential in following instructions and modifying edited regions



Exploration

浙江大學 ZHEJIANG UNIVERSITY

Discussion of the two training-free framework

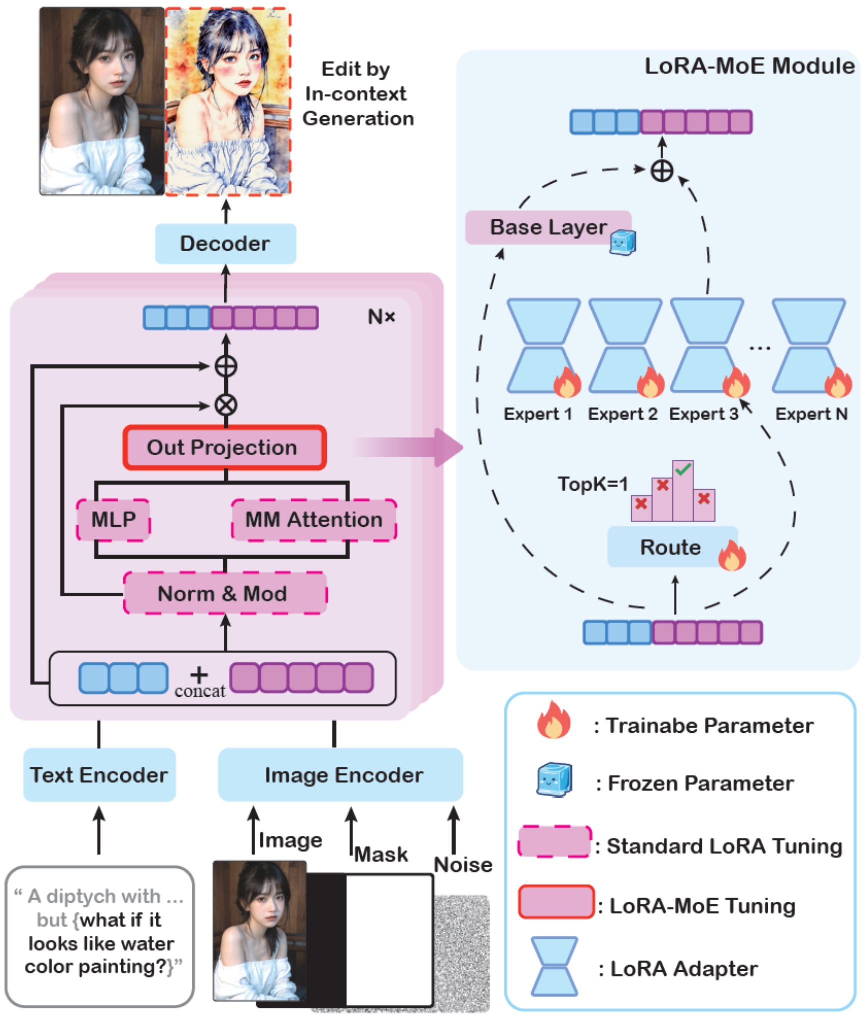
- While both frameworks demonstrate some editing capability, their zero-shot performance is unsatisfactory
- We attribute this to the lack of learned image-to-image editing priors. This limitation could be mitigated through lightweight adjustments, such as finetuning or test-time scaling.
- Given that the T2I DiT framework requires time-consuming image inversion, we favor the inpainting-based framework for its straightforward operation, which facilitates further finetuning.

(a) **Ablation study on model structure (§4.2).**

Settings	Params	CLIP-I ↑	CLIP-T ↑	GPT ↑
Training-free w/o IC prompt	-	0.681	0.258	0.14
Training-free w/ IC prompt	-	0.794	0.273	0.24
Only MoE module	130M	**0.929**	0.300	0.51
LoRA (r=64) w/ IC prompt	240M	<u>0.911</u>	<u>0.301</u>	0.60
Ours w/o IC prompt	214M	0.896	0.300	<u>0.62</u>
Ours	214M	0.907	**0.305**	**0.68**

求是創新

Finetuning



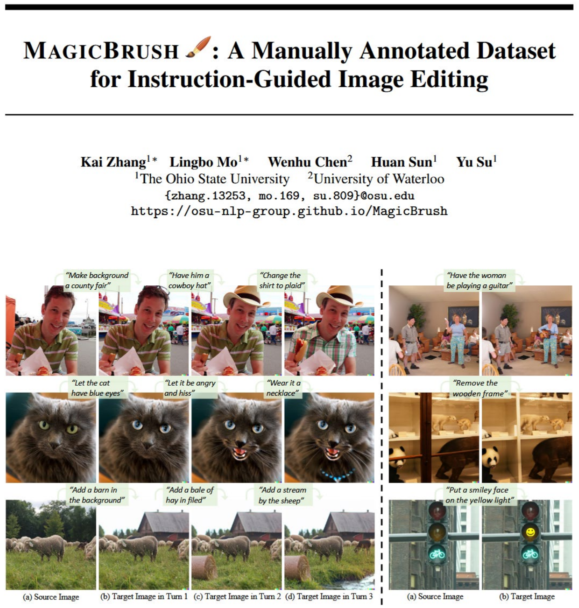
(b) Finetuning Strategy for ICEdit

Training Data (Randomly Selected)

Table 1: Dataset Statistics by Task Type

Task Type	Removal	Addition	Swap	Attribute Mod.	Style	Total
Count	13,272	11,938	5,823	11,484	10,530	53,047

OmniEdit



Magicbrush

求是創新

Experiment

Test results in paper

Table 1: **Quantitative results on Emu Test set (§4.1).** Following [4, 3], we compute CLIP-I and DINO scores between the source and edited image, while CLIP-out measures the distance between output caption and edited image. We also employ GPT-4o to evaluate the edited results. The Train. Pa. means parameters finetuned for the editing task. * indicates methods that rely on output captions.

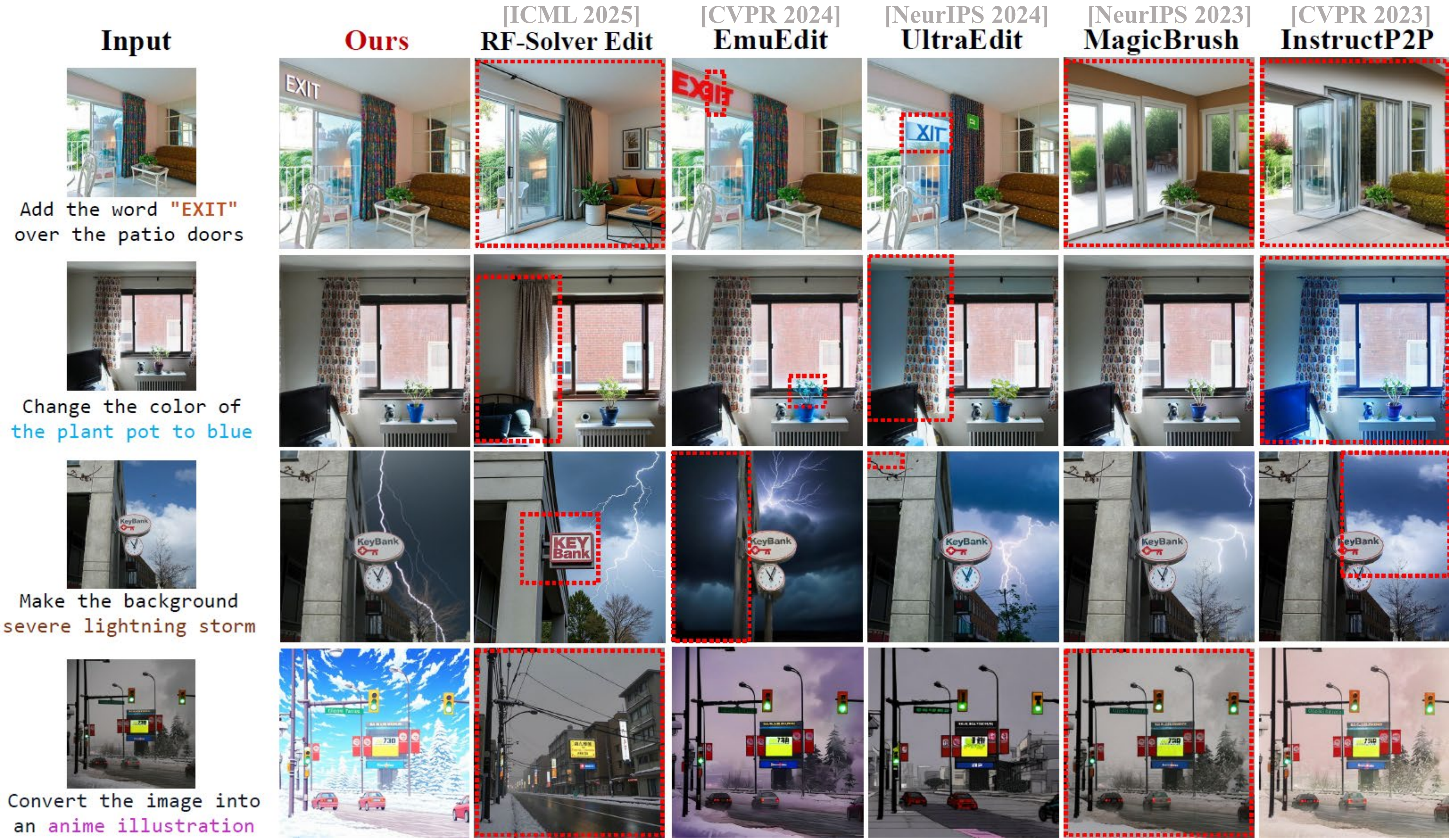
Methods	Base Model	Train. Pa.	Data Usage	CLIP-I ↑	CLIP-Out ↑	DINO ↑	GPT ↑
InstructP2P [CVPR23]	SD 1.5	0.9B	0.45M	0.856	0.292	0.773	0.36
MagicBrush [NeurIPS23]	SD 1.5	0.9B	0.47M	0.877	0.298	0.807	0.48
EmuEdit [CVPR24]	Close Source	2.8B	10M	0.877	0.306	0.844	**0.72**
UltraEdit [NeurIPS24]	SD 3	2.5B	3M	0.880	0.304	0.847	0.54
FluxEdit [huggingface]	Flux.1 dev	12B	1.2M	0.852	0.282	0.760	0.22
FLUX.1 Fill [huggingface]	Flux.1 Fill	-	-	0.794	0.273	0.659	0.24
RF-Solver Edit* [ICML25]	Flux.1 dev	-	-	0.797	**0.309**	0.683	0.32
ACE++ [arXiv25]	Flux.1 Fill	12B	54M	0.791	0.280	0.687	0.24
ICEdit (ours)	Flux.1 Fill	0.2B	0.05M	**0.907**	0.305	**0.866**	0.68

Table 2: **Quantitative results on MagicBrush test set.** Following [4], all metrics are calculated between the edited image and GT edited image provided by MagicBrush [2].

Methods	$L1$ ↓	CLIP-I ↑	DINO ↑
InstructP2P	0.114	0.851	0.744
MagicBrush	0.074	0.908	0.847
UltraEdit	0.066	0.904	0.852
FluxEdit	0.114	0.779	0.663
FLUX.1 Fill	0.192	0.795	0.669
RF-Solver Edit*	0.112	0.766	0.675
ACE++	0.195	0.741	0.591
ICEdit (ours)	**0.060**	**0.928**	**0.853**

Results

Our method achieves higher editing accuracy and better preservation of non-editing regions compared to existing approaches.



Test time scaling

- During inference, we find that initial noise significantly shapes editing outcomes, with some inputs producing results better aligned with human preferences.

- In instruction-based editing, we observe that **success in instruction alignment often become evident in few inference steps,** we can evaluate edit success with only a few steps

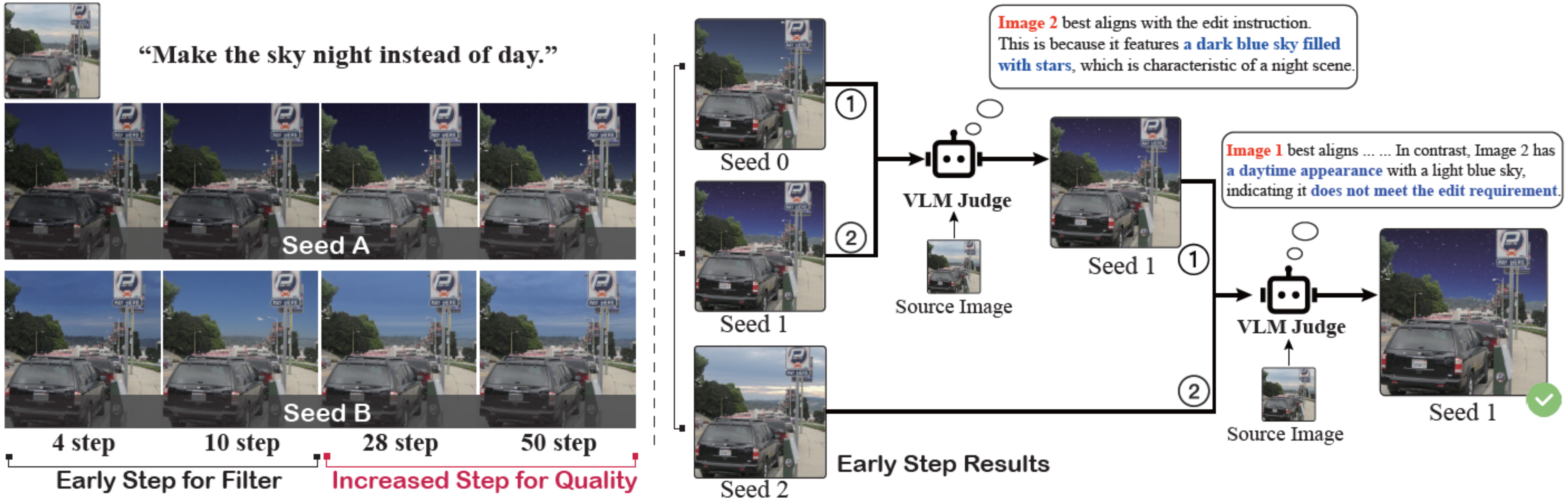
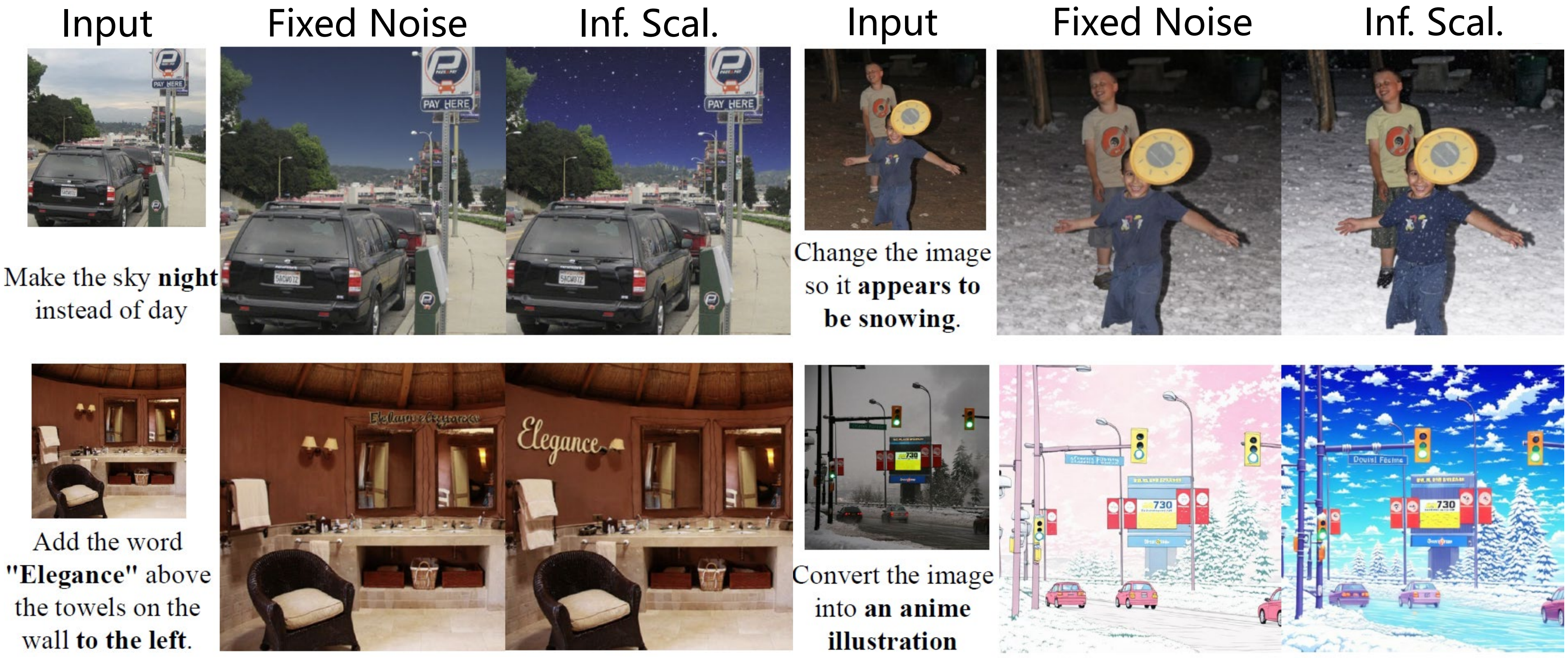


Figure 6: **Illustration of Inference-Time Scaling Strategy (§3.3).** The upper rows demonstrate that edit success can be assessed within a few initial steps. These early results are used to filter the optimal initial noise with VLM judges.

求是創新

Results

The proposed inference-time scaling strategy can quickly filter the best editing candidates at the inference stage, improving editing quality and stability.



Results

ZHEJIANG UNIVERSITY

More harmonious editing results

Commercial
SeedEdit(Doubao)

Ours

Results

Multi-task and ID consistent editing

In the evening, the **fire** burning.

Replace the building to **Eiffel tower**.

Remove the **dog**.

Original Image

Holding a cup of tea, eye **closed**.

Wears a **diamond** earring, and a **golden ruby** crown.

Make her hair **dark green** and her clothes **checked**.

Girl is on the beach, **colorful cloud** in the sky.

What if it looks like **watercolor** painting?

Results

Multi-turn edits

Original Image | Change the background to **Hawaii scenery**. | Dress in **Aloha Shirt, Hawaiian shorts** and **surf on board**. | Replace the boy with **SpongeBob** and make it **a comic book photo**. | Add the **text "ICCV2025"** on the bottom in **bold white color**.

Image-to-image translation

input

output

Some Limitations

ZHEJIANG UNIVERSITY

➢ Object Movement: Instructions requiring spatial relocation (e.g., "move the chair to the corner") may fail due to insufficient exposure to motion oriented data in general editing datasets.

➢ Semantic Understanding Limitations: While T5 demonstrates strong text encoding capabilities, its semantic understanding remains constrained, particularly in resolving polysemous terms

Figure 7: Some failure cases of our methods, such as object movement, semantic ambiguity.

Thanks!