



ACE: ALL-ROUND CREATOR AND EDITOR

Following Instructions via Diffusion Transformer

Zhen Han*, Zeyinzi Jiang*, Yulin Pan*, Jingfeng Zhang*, Chaojie Mao*,
Chenwei Xie, Yu Liu, Jingren Zhou
Tongyi Lab, Alibaba Group



Contents

1. ACE: Making Editing as Simple as Generation
2. Unified Multi-Task Generation / Editing Framework
3. Data Structuring and Quality Tuning
4. Model Performance and Applications
5. Follow-up Work Related to ACE



ACE: Making Editing as Simple as Generation

ACE provides a more convenient and versatile way of editing with instructions



Editing With Comfyui workflows




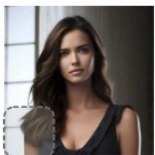
















Editing With ACE



ACE: Making Editing as Simple as Generation

- ✓ Visual generation and editing can be categorized into 9 types.
- ✓ The various tasks can be defined through a unified format.

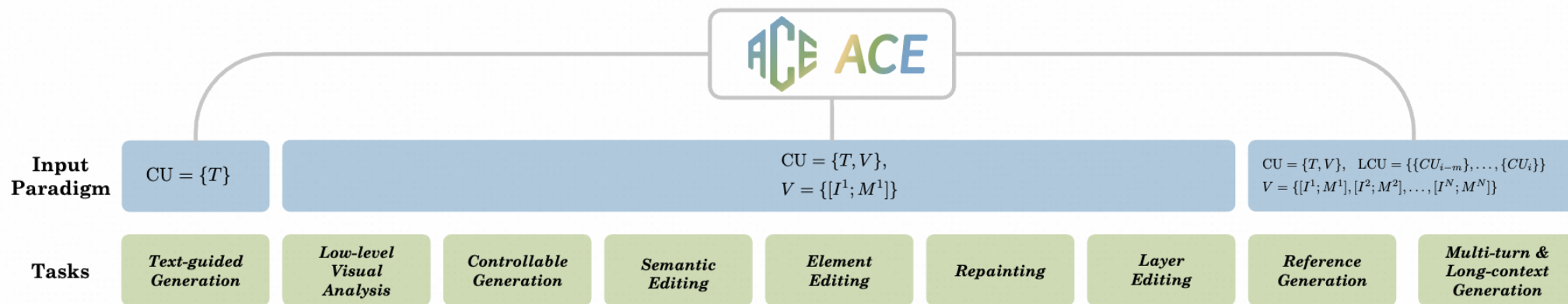
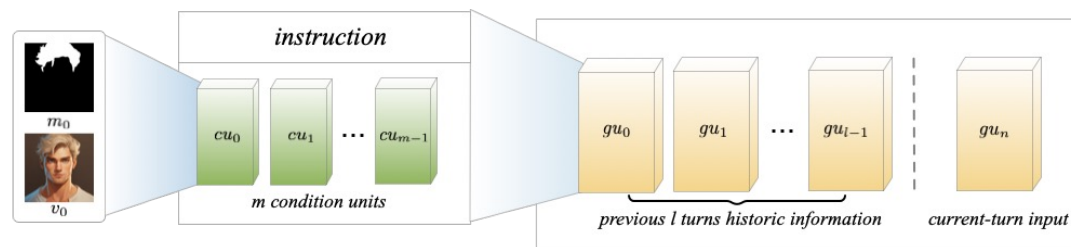
Tasks	Text-guided Generation	Low-level Visual Analysis	Controllable Generation	Semantic Editing	Element Editing	Repainting	Layer Editing	Reference Generation	Multi-turn & Long-context Generation
Source Image (V)									
Instruction (T)	A girl with blue hair, glasses, anime style 2D artwork.	Proceed with the depth extraction of {image}.	Please utilize the contour {image} to develop a restoration image.	Convert {image} characters into different style, ensuring the facial characteristics unchanged.	Incorporate the text "who" into {image}.	Repaint the parts of {image} identified by mask with "a fluffy white cat"	To fuse {image} with {image1}	{image}, {image1}, {image2}, a girl with two pigtails is standing against a light blue background.	<instruction1~3>, a black and white cartoon character sits at a table with bread and two cups of coffee, with Shanghai landmarks.
Target Image									

ACE: Making Editing as Simple as Generation

Unified Input Paradigm (LCU)

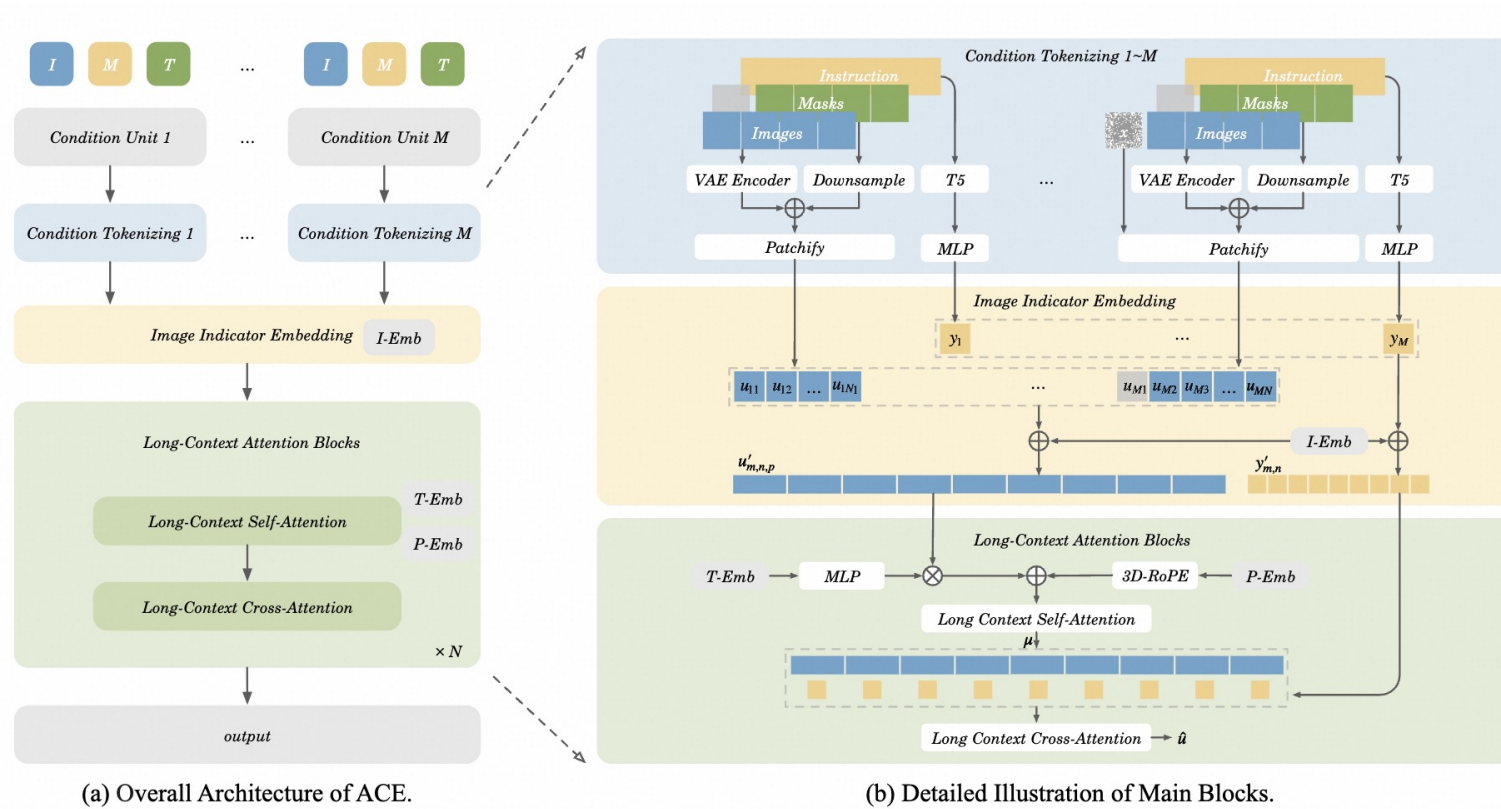
T: represents the instruction text,

V: represents the visual condition unit, consisting of the image I and the mask M



Unified Multi-Task Generation / Editing Framework

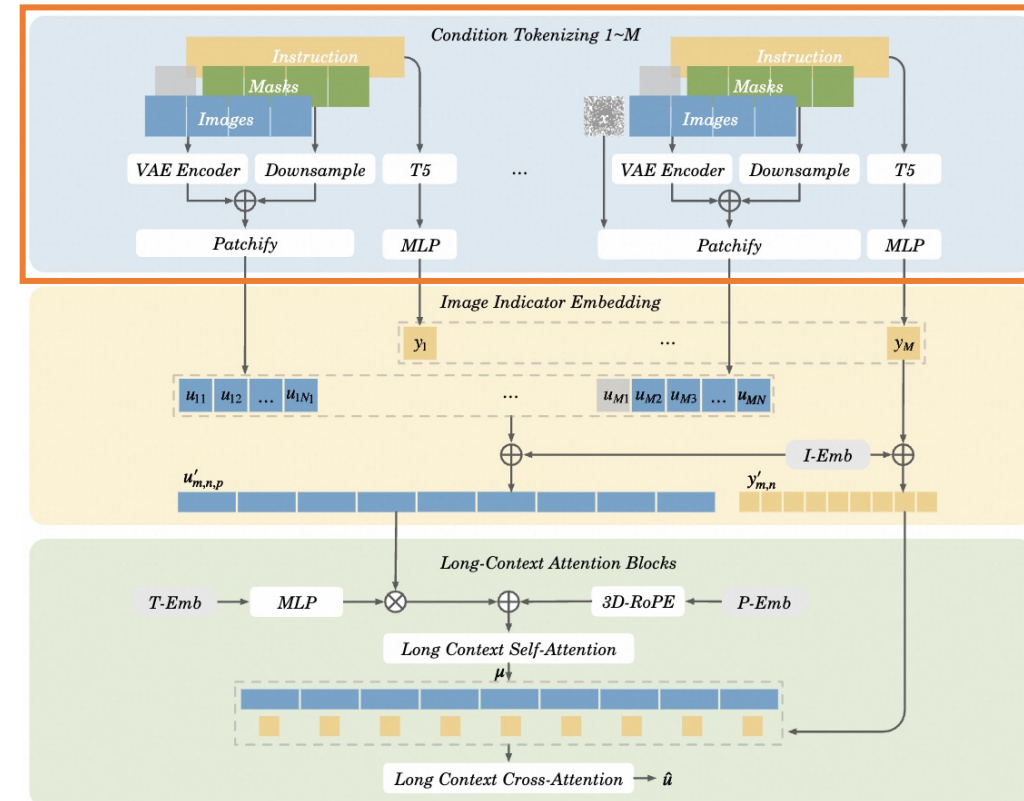
✓ Multimodal Editing and Generation Framework Based on Dit(0.6B)



Unified Multi-Task Generation / Editing Framework

✓ Multimodal Editing and Generation Framework Based on Dit

Condition Tokenizing



Unified Multi-Task Generation / Editing Framework

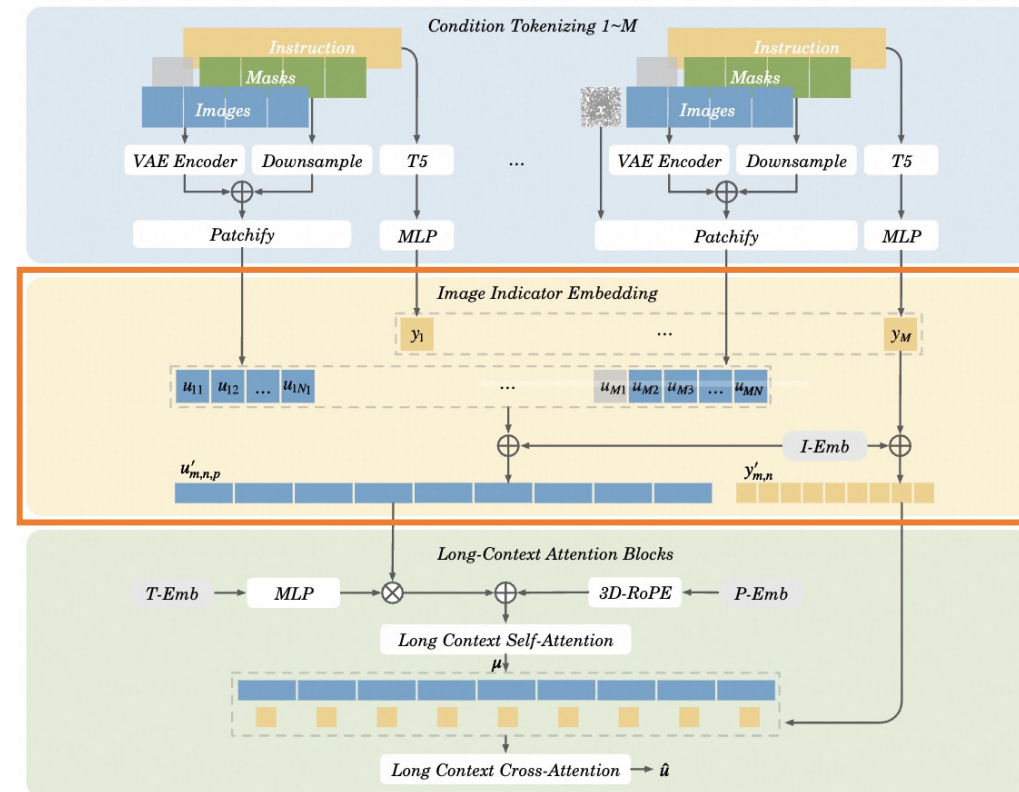
✓ Multimodal Editing and Generation Framework Based on Dit

Condition Tokenizing

$$y'_{m,n} = y_m + I\text{-Emb}_{m,n},$$

$$u'_{m,n,p} = u_{m,n,p} + I\text{-Emb}_{m,n}.$$

Image Indicator Embedding



Unified Multi-Task Generation / Editing Framework

✓ Multimodal Editing and Generation Framework Based on Dit

Condition Tokenizing

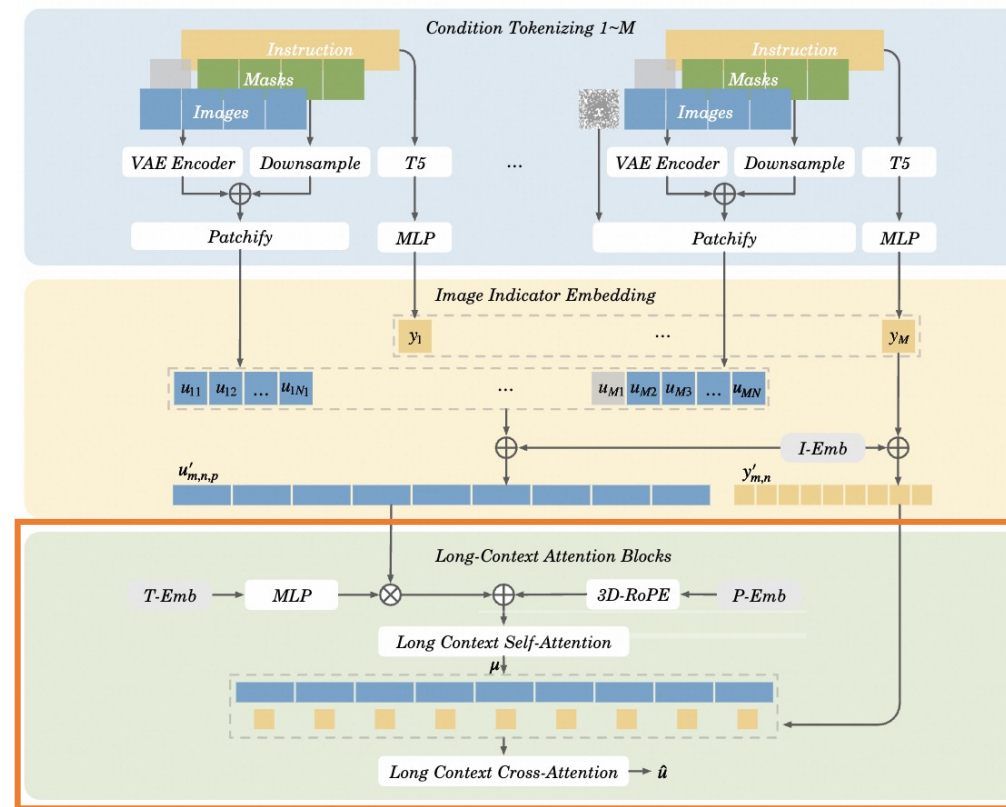
$$y'_{m,n} = y_m + I\text{-Emb}_{m,n},$$

$$u'_{m,n,p} = u_{m,n,p} + I\text{-Emb}_{m,n}.$$

Image Indicator Embedding

$$\hat{u}_{m,n} = \text{Attn}(\mu_{m,n}, y'_{m,n}).$$

Long-Context Attention Blocks

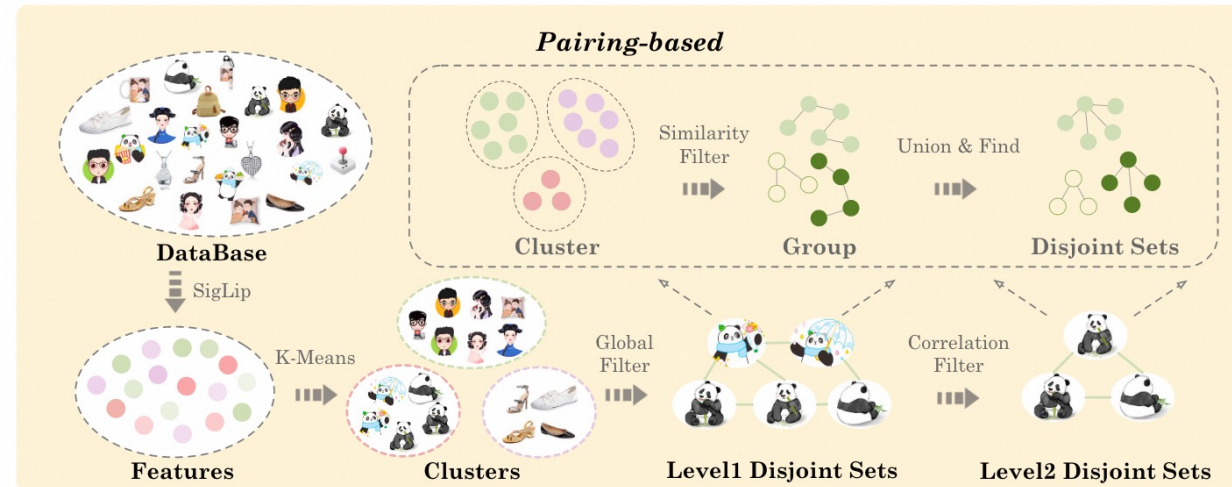
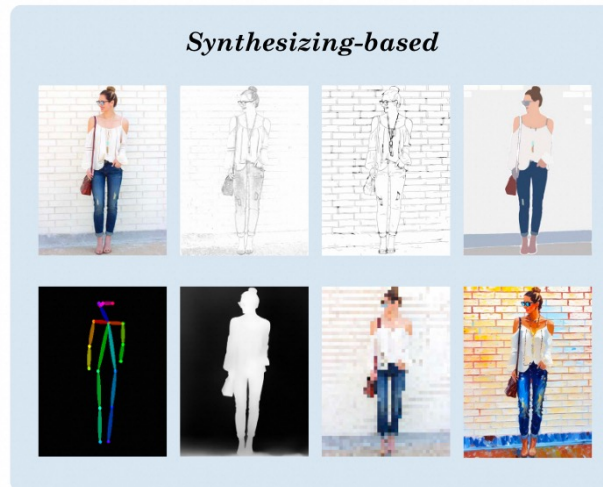


Data Structuring and Quality Tuning

Scaling Data Construction and Annotation Process

- ✓ Generate corresponding editing task data pairs based on existing generative models.
- ✓ Construct data pairs with similar attributes based on feature clustering, typically referencing image-related tasks.

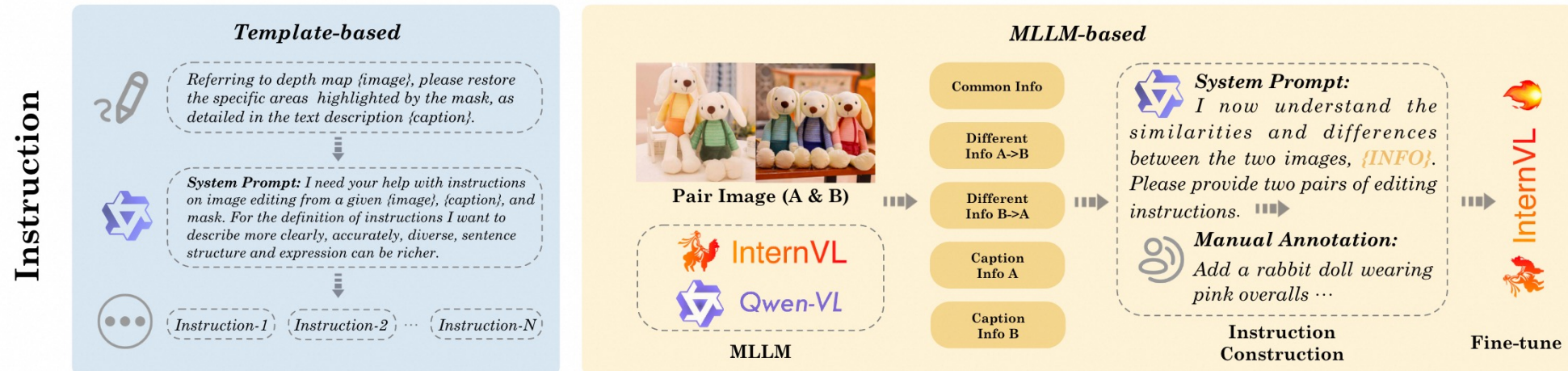
Pair Data



Data Structuring and Quality Tuning

Instruction Labeling With Template-based and MLLM-based Methods

- ✓ Template-based method constructs instruction templates for specific vision tasks by leveraging human knowledge priors.
- ✓ The Instruction Captioner finetuned with curated instructions datasets to generate unique instructions for each given editing pair.



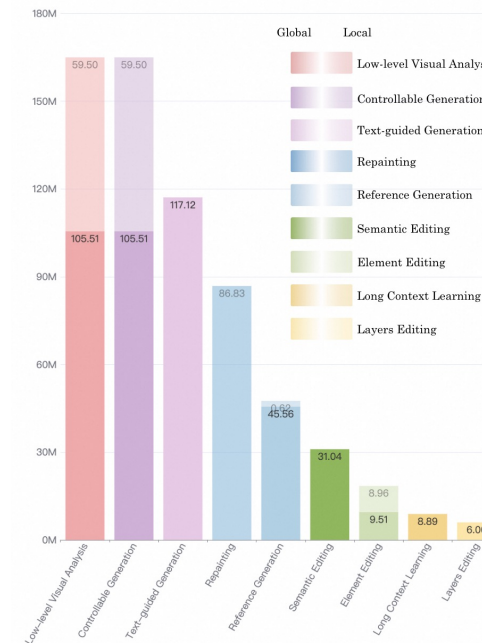
Data Structuring and Benchmark

ACE Dataset Distribution

- ✓ Covers 37 subtasks under 8 fundamental tasks.



a. The distribution of all tasks in the dataset



b. The data scale of basic tasks in the dataset

Data Structuring and Benchmark

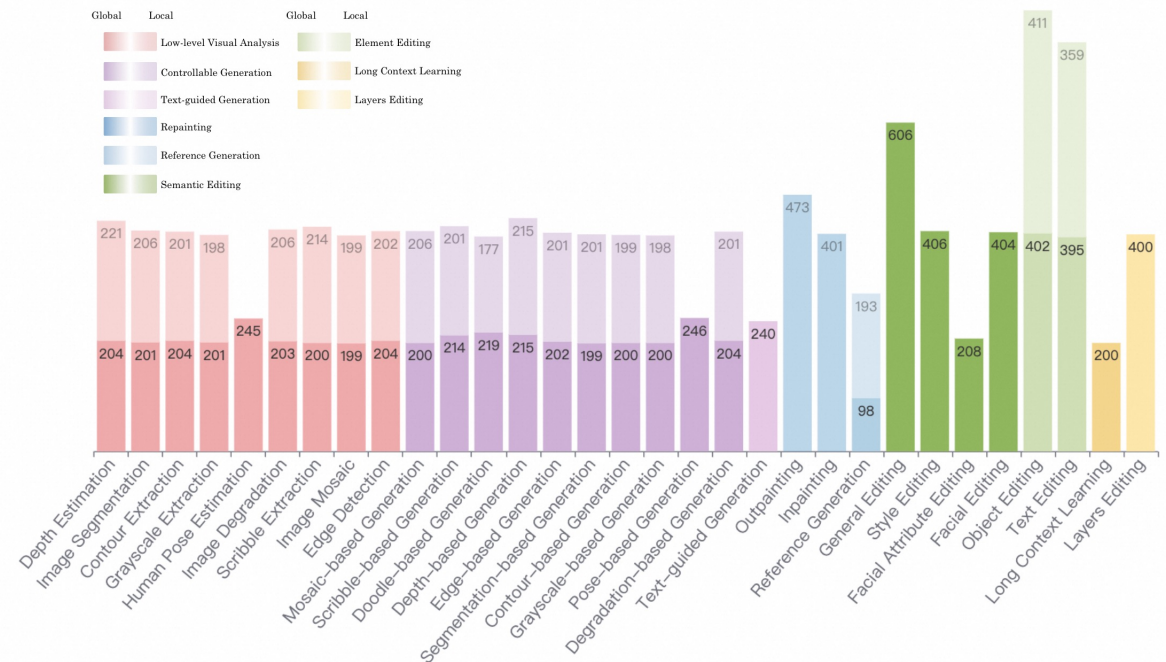
ACE Benchmark Distribution

(1) Total 12k: includes 6k real images and 6k generated images.

(2) Evaluation for tasks such as local and global editing, reference generation, multi-round editing, and composite generation.

(3) Total 31 subtasks: it is currently the largest benchmark in the field with the most comprehensive coverage of tasks.

Benchmark	Real Image?	Generated Image?	Multi-turn?	Regional?	Tasks	Data Scale
MagicBrush	Y	N	Y	Y	-	1588
Emu Edit	Y	N	N	N	8	3589
ACE	Y	Y	Y	Y	31	12000



Model Performance and Applications

Quantitative Analysis

Settings	Methods	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
Single-turn	<i>Global Description-guided</i>					
	SD-SDEdit (Meng et al., 2021)	0.1014	0.0278	0.8526	0.7726	0.2777
	Null Text Inversion (Mokady et al., 2022)	0.0749	0.0197	0.8827	0.8206	0.2737
	GLIDE (Nichol et al., 2022)	3.4973	115.8347	0.9487	0.9206	0.2249
	Blended Diffusion (Avrahami et al., 2022)	3.5631	119.2813	0.9291	0.8644	0.2622
	ACE (Ours)	0.0505	0.0160	0.9436	0.9184	0.2833
	<i>Instruction-guided</i>					
	HIVE (Zhang et al., 2024)	0.1092	0.0380	0.8519	0.7500	-
	InstructPix2Pix (Brooks et al., 2023)	0.1122	0.0371	0.8524	0.7428	0.2764
	MagicBrush (Zhang et al., 2023a)	0.0625	0.0203	0.9332	0.8987	0.2781
Multi-turn	UltraEdit (Zhao et al., 2024)	0.0575	0.0172	0.9307	0.8982	-
	ACE (Ours)	0.0507	0.0165	0.9453	0.9215	0.2841
	<i>Global Description-guided</i>					
	SD-SDEdit (Meng et al., 2021)	0.1616	0.0602	0.7933	0.6212	0.2694
	Null Text Inversion (Mokady et al., 2022)	0.1057	0.0335	0.8468	0.7529	0.2710
	GLIDE (Nichol et al., 2022)	11.7487	1079.5997	0.9094	0.8494	0.2252
	Blended Diffusion (Avrahami et al., 2022)	14.5439	1510.2271	0.8782	0.7690	0.2619
	ACE (Ours)	0.0778	0.0290	0.9124	0.8611	0.2843
	ACE (Ours w/ LC)	0.0768	0.0285	0.9136	0.8635	0.2819
	<i>Instruction-guided</i>					
	HIVE (Zhang et al., 2024)	0.1521	0.0557	0.8004	0.6463	0.2673
	InstructPix2Pix (Brooks et al., 2023)	0.1584	0.0598	0.7924	0.6177	0.2726
	MagicBrush (Zhang et al., 2023a)	0.0964	0.0353	0.8924	0.8273	0.2754
	UltraEdit (Zhao et al., 2024)	0.0745	0.0236	0.9045	0.8505	-
	ACE (Ours)	0.0773	0.0293	0.9128	0.8661	0.2855
	ACE (Ours w/ LC)	0.0761	0.0284	0.9140	0.8668	0.2809

Comparison on the MagicBrush Benchmark

Method	CLIPdir↑	CLIPout↑	L1↓	CLIPimg↑	DINO↑
InstructPix2Pix (Brooks et al., 2023)	0.0739	0.2681	0.1240	0.8508	0.7647
MagicBrush (Zhang et al., 2023a)	0.0831	0.2701	0.0995	0.8664	0.7927
Emu Edit (Sheynin et al., 2024)	0.1073	0.2791	0.0893	0.8743	0.8398
UltraEdit (Zhao et al., 2024)	0.0888	0.2783	0.0532	0.8814	0.8524
CosXL (StabilityAI, 2024)	0.0901	0.2775	0.0940	0.8686	0.8340
ACE (Ours)	0.0855	0.2746	0.0761	0.8952	0.8620

Comparison on the EmuEdit Benchmark

Method	Face Similarity	Effective Score
InstantID [†] Wang et al. (2024b)	84.08	0.96
CosXL StabilityAI (2024)	66.49	0.37
UltraEdit Zhao et al. (2024)	62.91	0.16
IP-Adapter Ye et al. (2023)	66.51	0.31
FaceChain Liu et al. (2023b)	65.46	0.42
ACE (Ours)	70.07	0.67

Quantitative Evaluation of Portrait Preservation

Method	Edit Distance	Sentence Accuracy
UDiffText (Zhao & Liar, 2024)	0.6827	0.4110
AnyText (Tuo et al., 2023)	0.6035	0.3313
ACE (Ours)	0.8211	0.5767

Quantitative Evaluation of Text Editing

Model Performance and Applications

Human Study

(1) Results generated by different methods are ranked manually.

(2) Five designers serve as annotators, ensuring that each sample is evaluated by at least three individuals.

(3) Among the 21 benchmarkable tasks, a win rate of 15 out of 21 has been achieved.

	Txt2img ●	Canny ●	Depth ●	Scribble ●	Pose ●	Face ●	Style ●	General ●	Add Text ●	Rm Text ●	Add Obj. ●	Rm Obj. ●	Inpaint ●	Repainting ●
<i>Global Editing</i>														
SD1.5 (A., 2022a)	3.3/2.2	-	-	-	-	-	-	-	-	-	-	-	-	-
SDXL (StabilityAI, 2022)	4.1/2.8	-	-	-	-	-	-	-	-	-	-	-	-	-
CtrlNet (Zhang et al., 2023b)	-	2.5/2.0	3.8/2.4	1.9/2.0	2.9/1.9	-	-	-	-	-	-	-	-	-
StyleBooth (Han et al., 2024)	-	-	-	-	-	-	3.3/2.6	-	-	-	-	-	-	-
IP-Adapter (Ye et al., 2023)	-	-	-	-	-	2.0/2.2	-	1.7/2.5	-	-	-	-	-	-
InstantID (Wang et al., 2024b)	-	-	-	-	-	2.5/2.7	-	-	-	-	-	-	-	-
FaceChain (Liu et al., 2023b)	-	-	-	-	-	2.0/3.0	-	-	-	-	-	-	-	-
SDEdit (Meng et al., 2021)	-	1.4/1.9	1.3/1.8	1.1/1.6	1.2/1.4	1.3/2.1	1.1/1.7	1.5/2.1	1.1/2.2	1.1/1.7	1.5/2.1	1.1/2.0	-	-
IP2P (Brooks et al., 2023)	-	1.9/2.0	1.7/2.0	1.5/2.3	1.4/1.4	2.3/2.4	2.4/2.5	2.2/2.4	1.1/2.6	1.3/2.6	2.0/2.4	1.5/2.4	-	-
MB (Zhang et al., 2023a)	-	1.3/1.8	1.3/1.7	1.3/1.9	1.1/1.3	2.4/2.3	1.4/2.0	2.2/2.3	1.5/2.4	2.2/2.5	3.1/2.2	2.1/2.4	-	-
SEED-X (Ge et al., 2024b)	-	1.6/2.1	1.7/2.0	1.7/2.2	1.5/1.5	2.0/2.7	2.2/2.5	2.1/2.7	1.3/2.6	2.1/2.6	1.9/2.6	2.5/2.4	-	-
CosXL (StabilityAI, 2024)	-	4.1/2.9	4.1/2.8	2.6/2.9	3.7/2.1	2.9/3.1	3.2/3.0	3.2/2.9	1.4/2.7	1.0/2.9	2.8/2.5	1.1/3.1	-	-
UltraEdit (Zhao et al., 2024)	-	1.7/2.2	1.2/1.8	1.3/2.3	1.1/1.3	2.3/2.5	2.1/2.4	2.6/2.5	1.7/2.6	1.1/2.7	2.7/2.3	1.5/2.6	-	-
ACE (Ours)	3.7/2.5	4.6/2.7	4.5/2.8	4.8/2.9	4.1/2.3	2.8/2.8	2.4/2.6	2.1/2.5	2.8/2.7	4.4/2.9	2.6/2.4	3.9/2.5	-	-
<i>Local Editing</i>														
LaMa (Suvorov et al., 2021)	-	-	-	-	-	-	-	-	-	3.6/2.8	-	4.5/2.8	1.6/2.3	3.0/2.4
SDInpaint (A., 2022b)	-	-	-	-	-	-	-	-	-	2.6/2.6	1.6/2.7	2.2/2.5	3.6/2.6	-
CtrlNet (Zhang et al., 2023b)	-	-	-	-	-	-	-	-	-	2.9/2.7	1.9/2.5	2.6/2.2	3.0/2.1	3.2/2.1
AnyText (Tuo et al., 2023)	-	-	-	-	-	-	-	-	3.5/2.7	-	-	-	-	-
UDiffText (Zhao & Lian, 2024)	-	-	-	-	-	-	-	-	3.6/2.7	-	-	-	-	-
UltraEdit (Zhao et al., 2024)	-	1.4/1.9	1.2/1.8	1.2/2.0	-	-	-	-	1.1/2.8	1.2/2.9	2.9/2.5	1.4/2.5	1.1/1.7	1.1/2.1
ACE (Ours)	-	4.8/2.6	4.3/2.5	4.8/2.6	-	-	-	-	4.5/2.9	4.5/2.9	3.7/2.5	4.3/2.5	4.4/2.7	4.6/2.8

Model Performance and Applications

Visual Comparison

1. ControlNet

2. IP2P

3. MagicBrush

4. CosXL

5. SEED-X

6. UltraEdit

7. StyleBooth

8. SDEdit

9. LoRA

10. SD-Inpaint

11. LaMa

12. IP-Adapter

13. InstantID

14. FaceChain

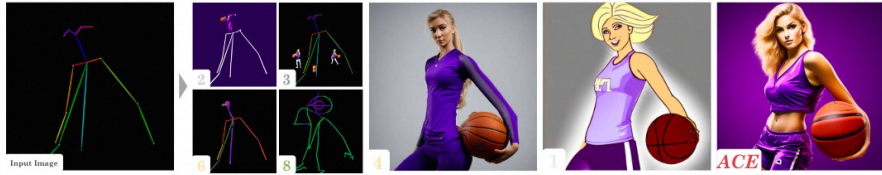
15. AnyText

16. UDiTText


ACE

Controllable Generation

Instruction: Here is a posture *[image]*, I need a color picture based on this image according to the brief: A blonde female with a purple basketball uniform and a basketball in hand.




Instruction: create an equivalent image for the provided edge *[image]* to represent "flower on background, oil painting, Italya, pink, 3D"




Face Editing

Instruction: Maintain facial consistency of the man in *[image]*, A man exploring a local market filled with spices and textiles, looking fascinated, dressed in light, summery attire, absorbing the vibrant culture.




Instruction: Change the head color of the woman in *[image]* to light hair with a blue gradient. Change dress for white scarf, and red dress.




Style Editing

Instruction: Convert *[image]* into vibrant pop art style



Instruction: Transform *[image]* into cartoon style



Model Performance and Applications

Visual Comparison

	1. ControlNet	2. IP2P	3. MagicBrush	4. CosXL	5. SEED-X	6. UltraEdit	7. StyleBooth	8. SDEdit	9. LoRA	10. SD-Inpaint	11. LaMa	12. IP-Adapter	13. InstantID	14. FaceChain	15. AnyText	16. UDiffText	ACE
General Editing	<p>Instruction: Let the girl face the camera</p>																
Text Remove	<p>Instruction: On {image}, obliterate the text found in the mask area</p>																
Text Render	<p>Instruction: Embed the text 'shop' at the place highlighted by mask in the {image}</p>																
	<p>Instruction: Delete all text in {image}</p>																
	<p>Instruction: Add a dark green word 'cat' over the cat in {image}</p>																

Model Performance and Applications

Visual Comparison

1. ControlNet

2. IP2P

3. MagicBrush

4. CosXL

5. SEED-X

6. UltraEdit

7. StyleBooth

8. SDEdit

9. LoRA

10. SD-Inpaint

11. LaMa

12. IP-Adapter

13. InstantID

14. FaceChain

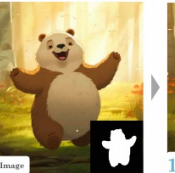





15. AnyText

16. UDiffText



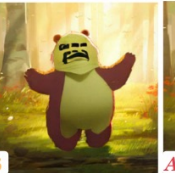
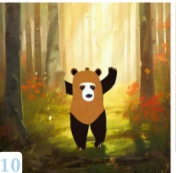
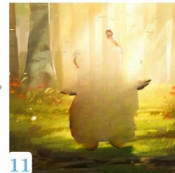
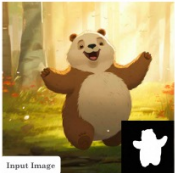
ACE

Object Removal

Instruction: Remove the dragonfly in {image}.









Instruction: Remove the bear in mask of {image}.



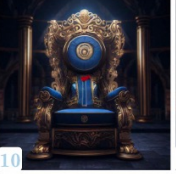
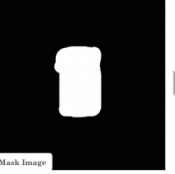



Object Addition

Instruction: I want to add a lemon character on the left of {image} next to the middle lemon.








Instruction: Add a dog sit on the chair on the mask of {image}.


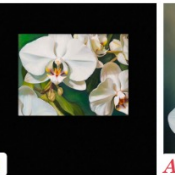
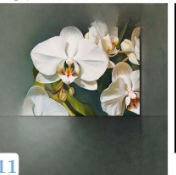




Repaint

Instruction: Using the text description "The village hut is surrounded by endless fields. The fields are full of golden rapeseed flowers, occasionally mixed with pink cherry blossoms, blue sky and white clouds, clear sky", repaint the black regions indicated by the mask in the {image}.



Instruction: According to "oil painting of 3 white orchid flowers on a thin branch with green orchid leaves in background", please expand the content in mask of {image}.



ACE

Model Performance and Applications

Visual Comparison

1. ControlNet

2. IP2P

3. MagicBrush

4. CosXL

5. SEED-X

6. UltraEdit

7. StyleBooth

8. SDEdit

9. LoRA

10. SD-Inpaint

11. LaMa

12. IP-Adapter

13. InstantID

14. FaceChain

15. AnyText

16. UDiffText

ACE

Text-guided & Lowlevel

Instruction: An adorable, anthropomorphic rabbit dressed for adventure in a lush forest setting. The rabbit is the focal point, positioned slightly off-center to the left. It has large, expressive blue eyes that convey a sense of wonder and

Reference Generation

Instruction: {image}, {image1}, {image2}, {image3}, old man in a trendy yellow dunejacket. capture photography, Graphics

Reference Editing

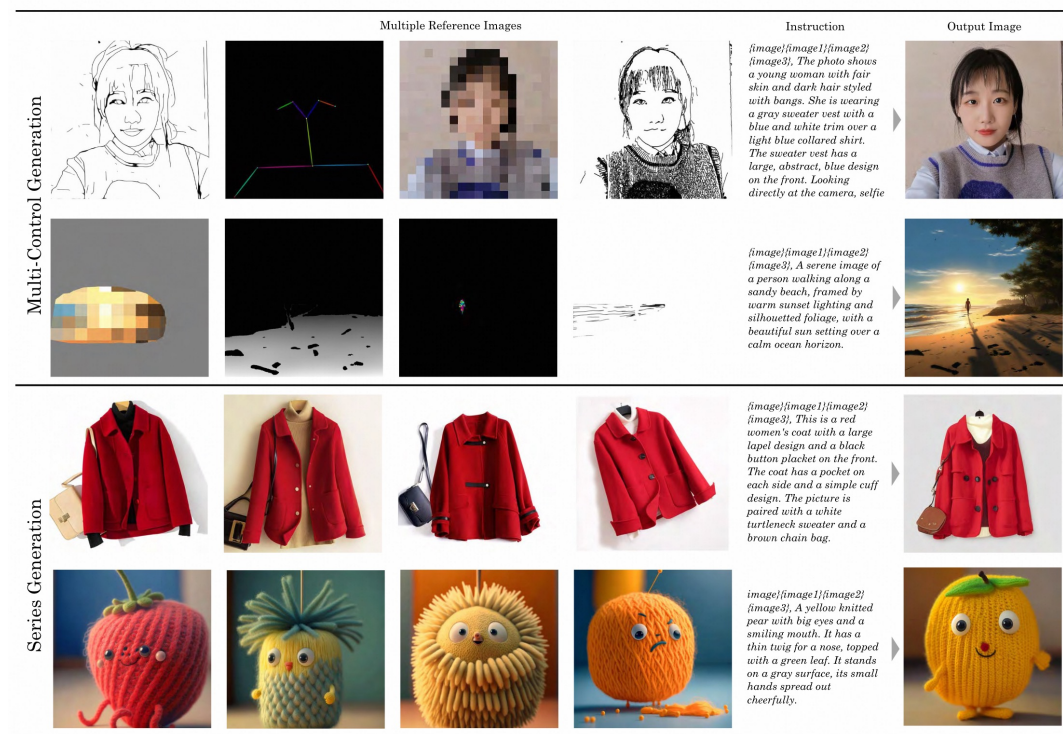
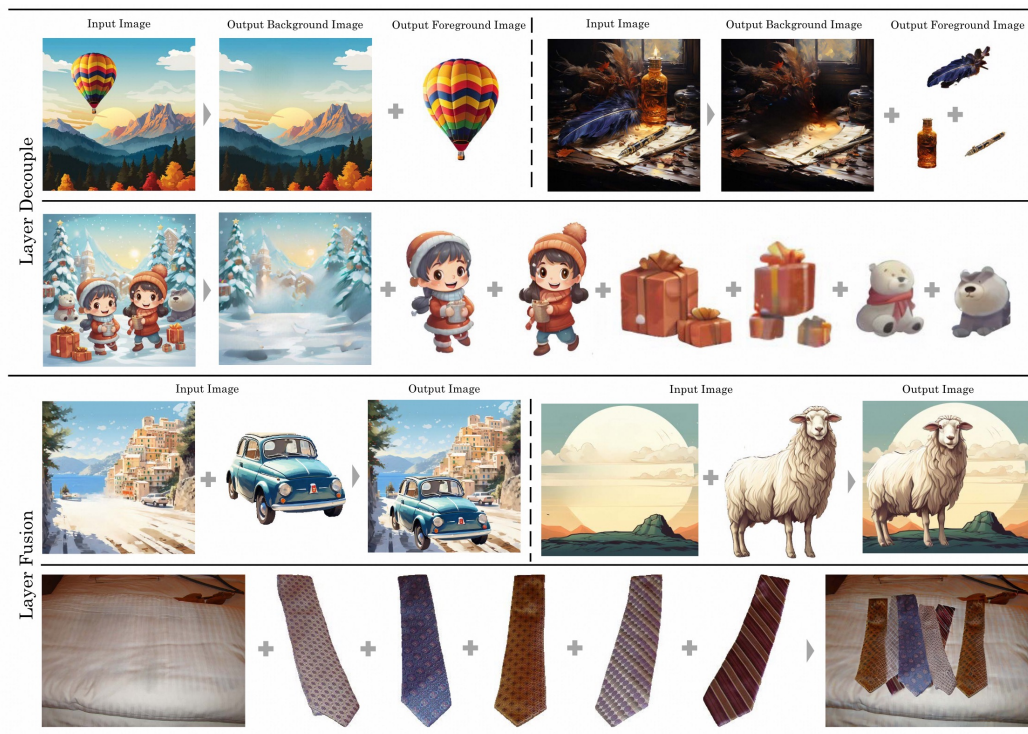
Instruction: Replace the existing face in {image} by accurately integrating the face from {image1} in the mask zone.

Instruction: Change the cloth in {image} to the one in {image1}

Instruction: Shape {image} with inspiration from {image1}

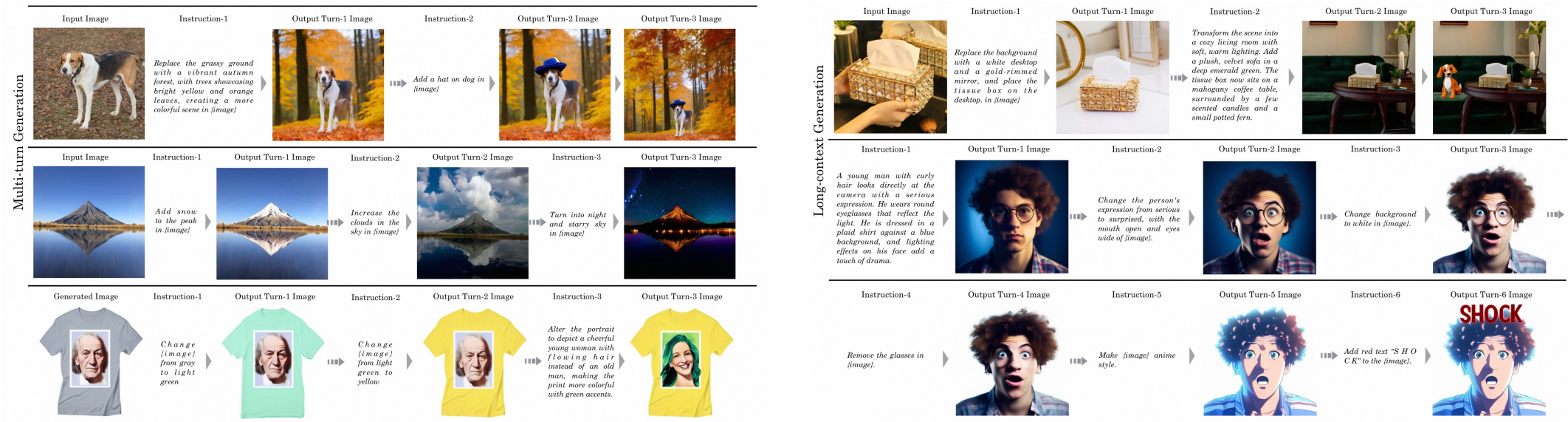
Model Performance and Applications

Visual Comparison



Model Performance and Applications

Visual Comparison

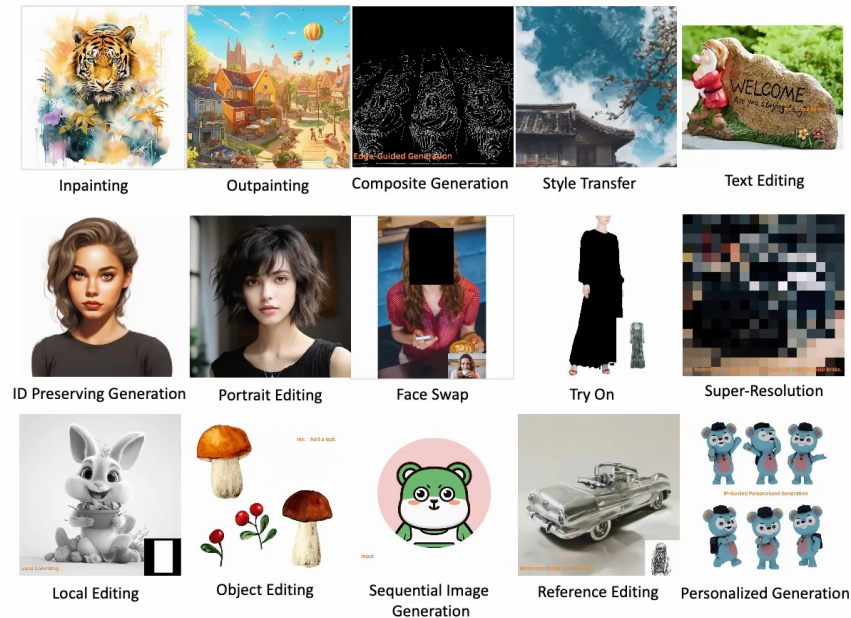


Multi-Round Editing and Identity Preservation Generation

Model Performance and Applications

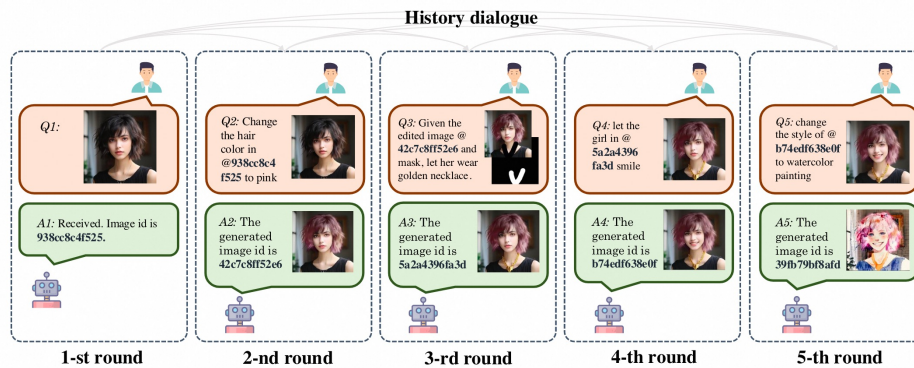
Applications: Basic Editing and Generation Tasks

- ✓ It can be used for 15 common image generation and editing tasks.
- ✓ Interactive generation and editing through instructions can significantly reduce deployment and usage costs.



Model Performance and Applications

Applications: Conversational Editing and Generation.



- ✓ In conversational mode, multi-round editing of images can be achieved through chat.

ACE Usage

Project Page: <https://ali-vilab.github.io/ace-page/>



<https://github.com/ali-vilab/ACE/>



<https://huggingface.co/spaces/scepter-studio/ACE-Chat>



<https://www.modelscope.cn/studios/iic/ACE-Chat>



Follow-up Works Related to ACE

ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling

- ✓ ACE++ is an upgrade Version based on post-training using FLUX-Dev.

Project Page: https://ali-vilab.github.io/ACE_plus_page/



https://github.com/ali-vilab/ACE_plus?tab=readme-ov-file



<https://huggingface.co/spaces/scepter-studio/ACE-Plus>



<https://www.modelscope.cn/studios/iic/ACE-Plus>



Follow-up Works Related to ACE

Wan2.1:

- ✓ The text data from ACE is used to train the Chinese and English text rendering capabilities of Wan 2.1.

Project Page: <https://wanxai.com/>

 Wan <https://wan.video>

 <https://github.com/Wan-Video>

 <https://huggingface.co/Wan-AI>

 <https://www.modelscope.cn/models/Wan-AI/>

Follow-up Works Related to ACE

ICEBench: A Unified and Comprehensive Benchmark for Image Creating and Editing

- ✓ ICEBench is a more refined definition and collection of ACEBench.

Project Page: <https://ali-vilab.github.io/ICE-Bench-Page/>



Follow-up Works Related to ACE

VideoACE: All-in-One Video Creation and Editing

- ✓ VideoACE is the adaptation of ACE for video tasks.

Project Page: <https://ali-vilab.github.io/VACE-Page/>



<https://github.com/ali-vilab/VACE>



Thanks

Zhen Han*, Zeyinzi Jiang*, Yulin Pan*, Jingfeng Zhang*, Chaojie Mao*,
Chenwei Xie, Yu Liu, Jingren Zhou

Tongyi Lab, Alibaba Group

