UltraEdit: Instruction-based Fine-Grained ImageEditing at Scale

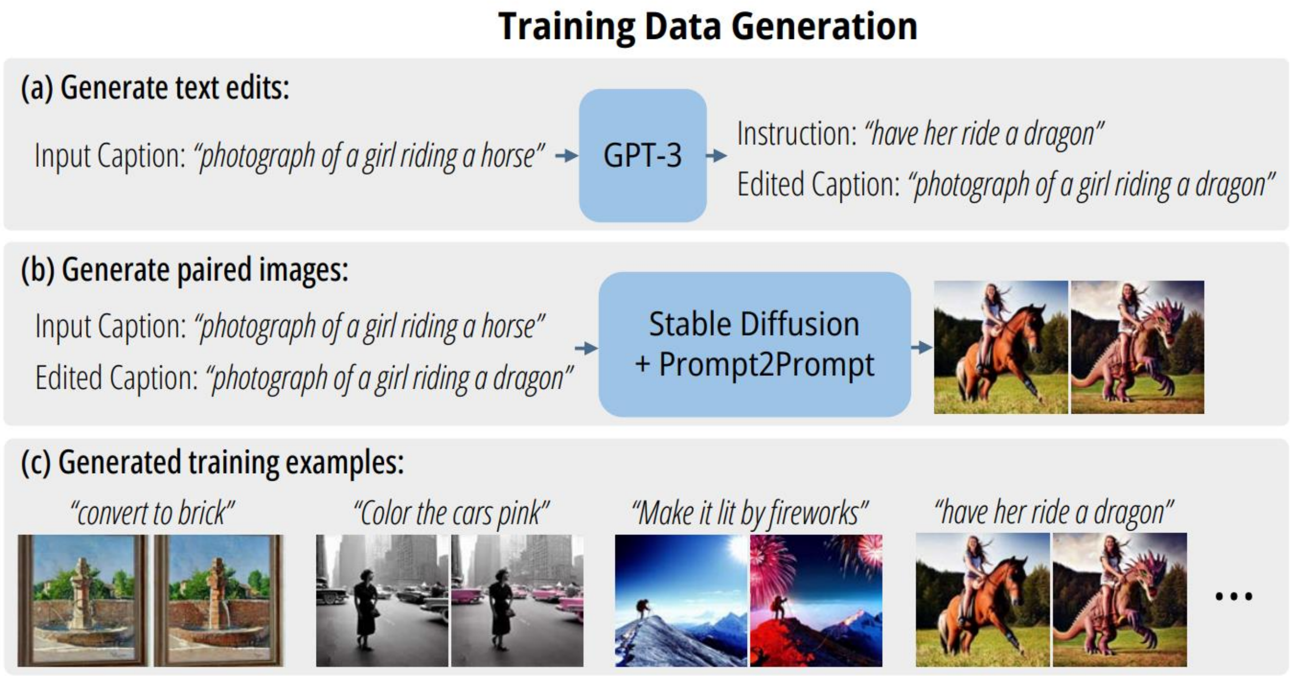
Haozhe zhao

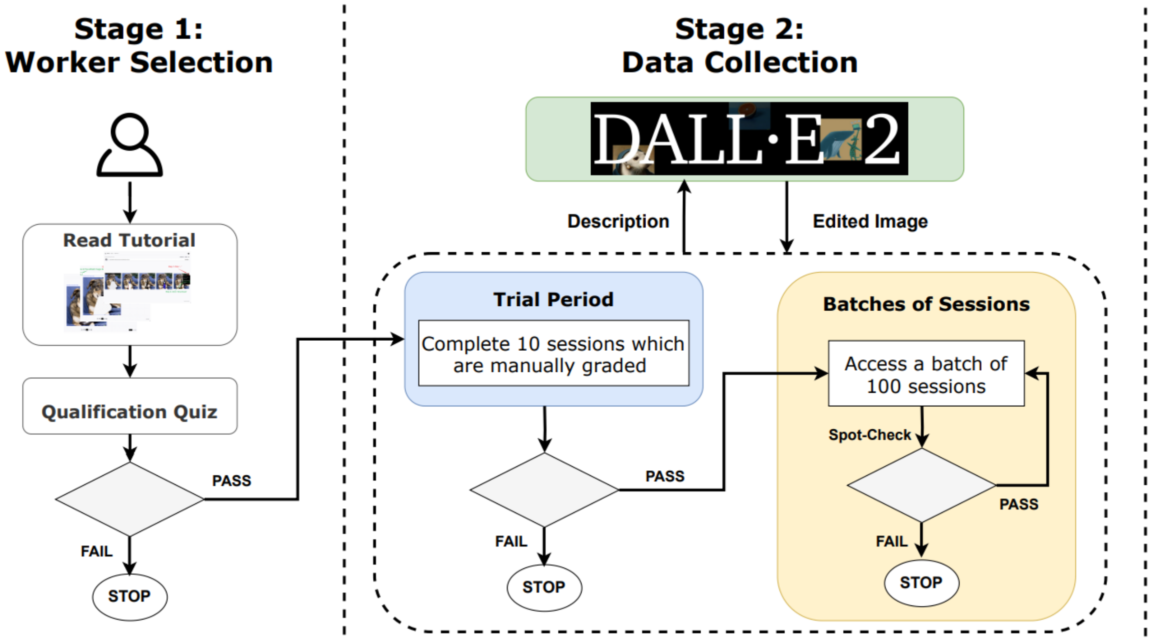
Sept 18, 2024

[UltraEdit (ultra-editing.github.io)](ultra-editing.github.io)

Large Scale Instruction-based Image Editing Dataset

Related Work: Existing Image Editing Data





InstructPix2Pix

MagicBrush

Drawbacks in existing image editing datasets

1. Limited instruction diversity

Datasets	Real Image Based	Automatic Generated	Editing Region	#Edits	#Editing Types	Source Example	Instruction	Target Example
EditBench [57]	✓	✗	✓	240	1		*an amber vase* with a narrow lip and a wide base	
MagicBrush [59]	✓	✗	✓	10,388	5		*replace the dove* with an owl.	
HQ-Edit [22]	✗	✓	✗	197,350	6		remove the chisel.	
InstructPix2Pix [10]	✗	✓	✗	313,010	4		*make it a* stone bridge	
ULTRAEDIT	✓	✓	✓	**4,108,262**	**9+**		*Change the hat* into a crown.	

Drawbacks in existing image editing datasets

2. Implicit biases in images

Moon bridge, Taiwan

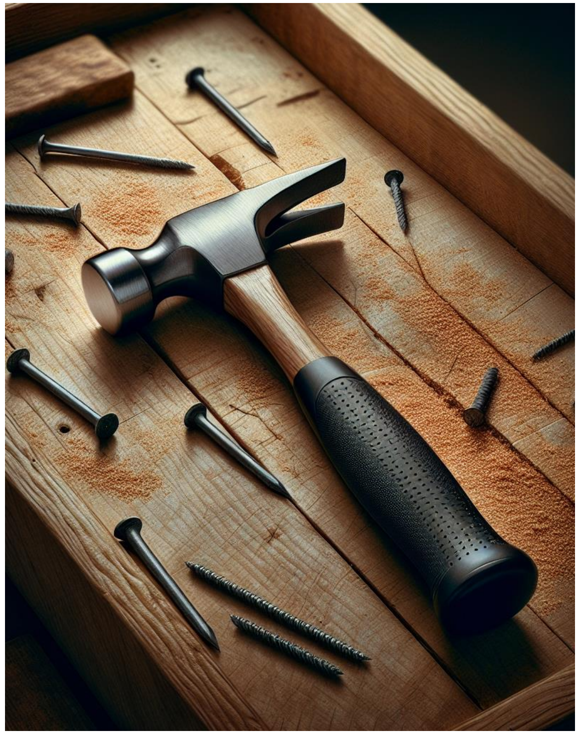
Make it a stone bridge

Stone bridge, Taiwan

Drawbacks in existing image editing datasets

Using advanced model still facing image biases

remove the chisel.

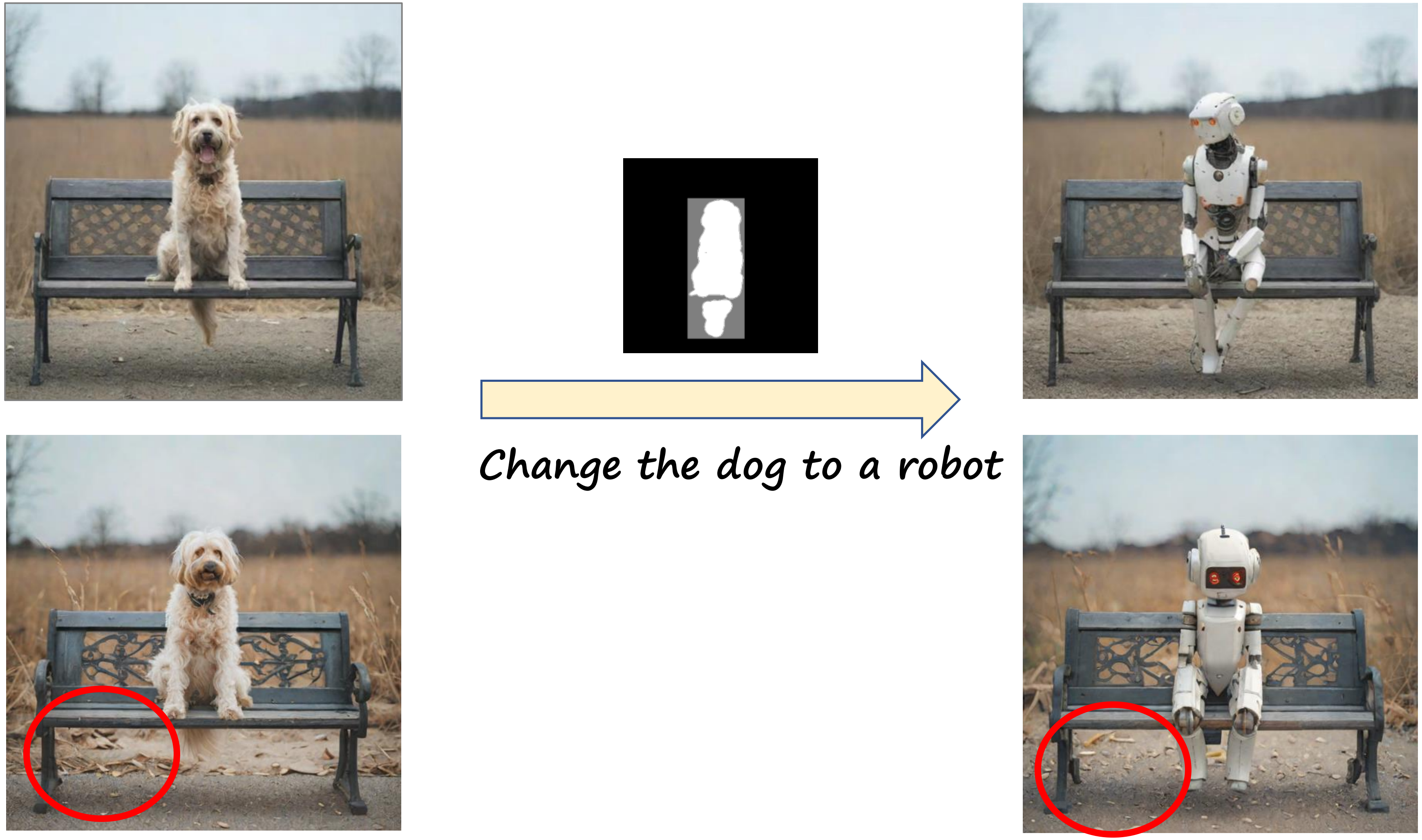


"A close-up of a hammer with a black grip resting on a wooden workbench, surrounded by nails, screws, sawdust, and a chisel with a wooden handle, evoking a scene of detailed craftsmanship."

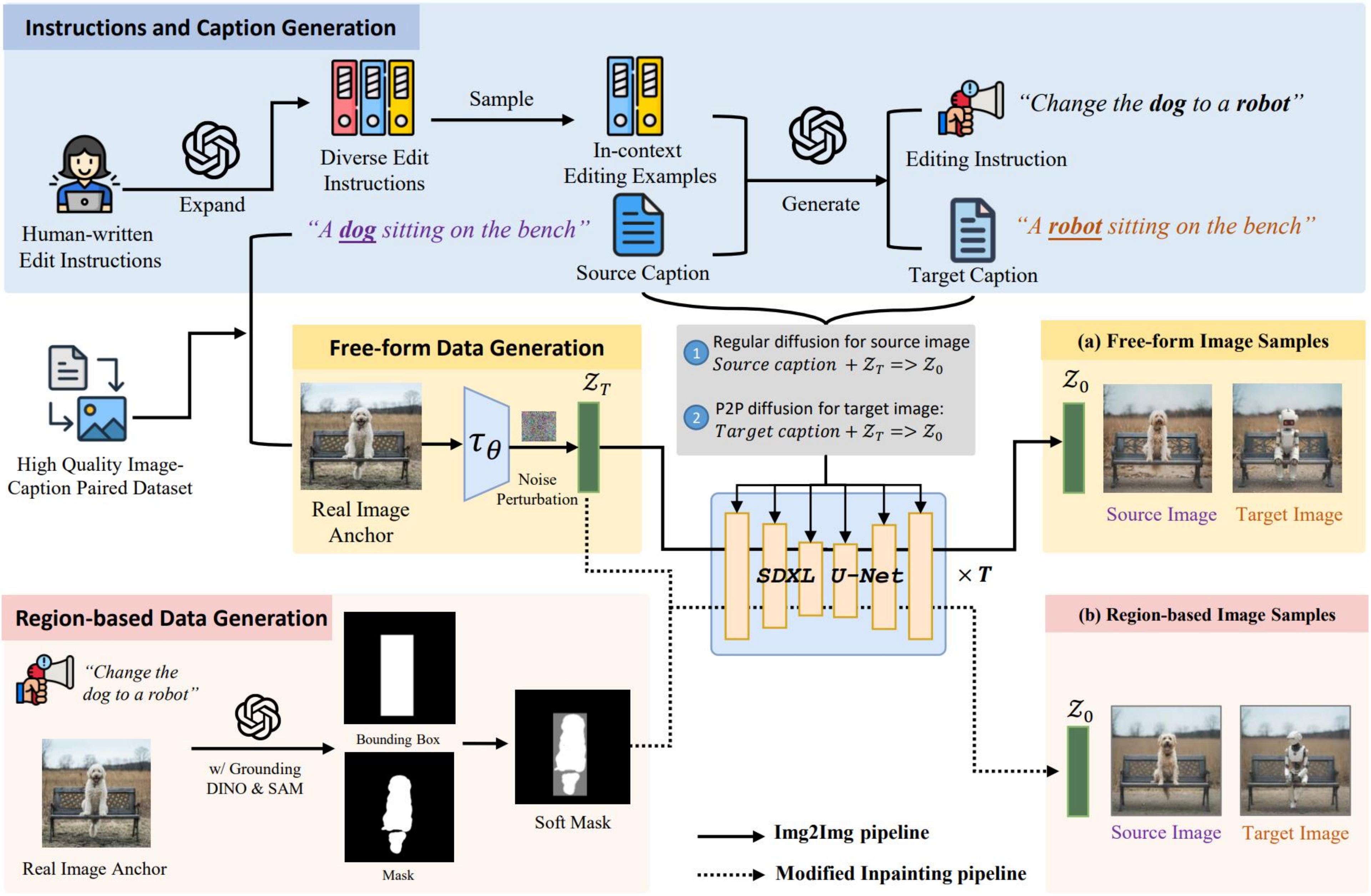
A close-up of a hammer with a black grip on a wooden workbench, surrounded by scattered nails, screws, and sawdust, evoking a scene of craftsmanship.

Drawbacks in existing image editing datasets

3. Missing of region-based editing



Dataset Formation

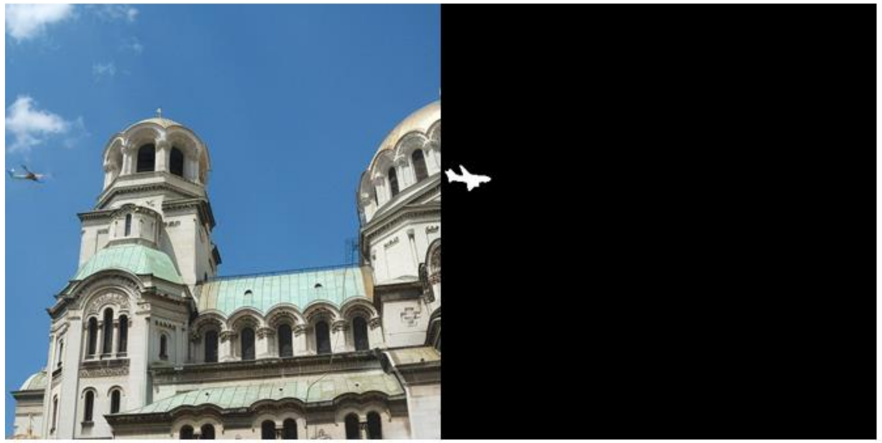


Region-based image generation

Mask Segmentation

(a) overly large mask

(c) fragmented mask



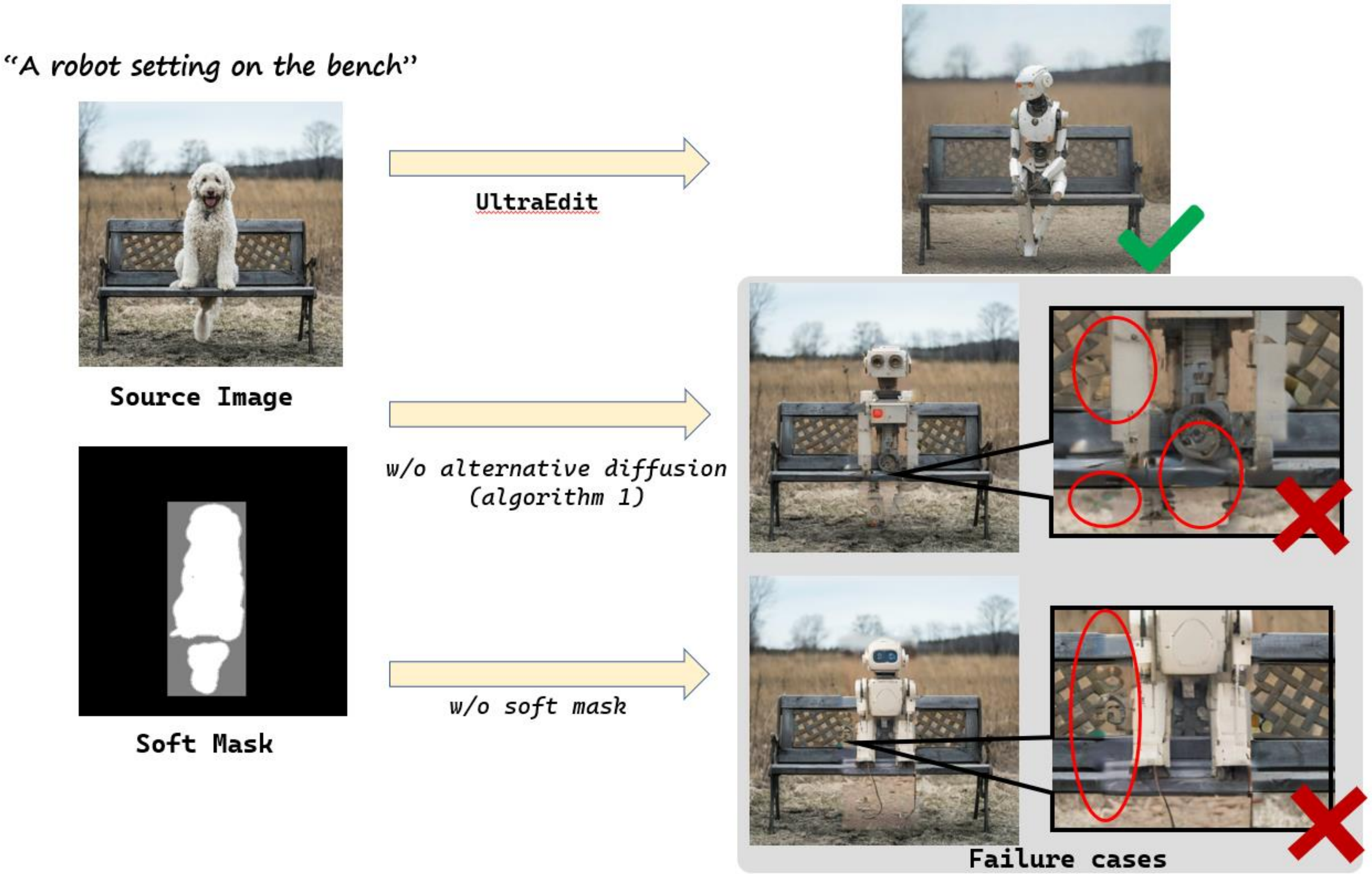
(b) overly small mask

(d) fine-grained mask

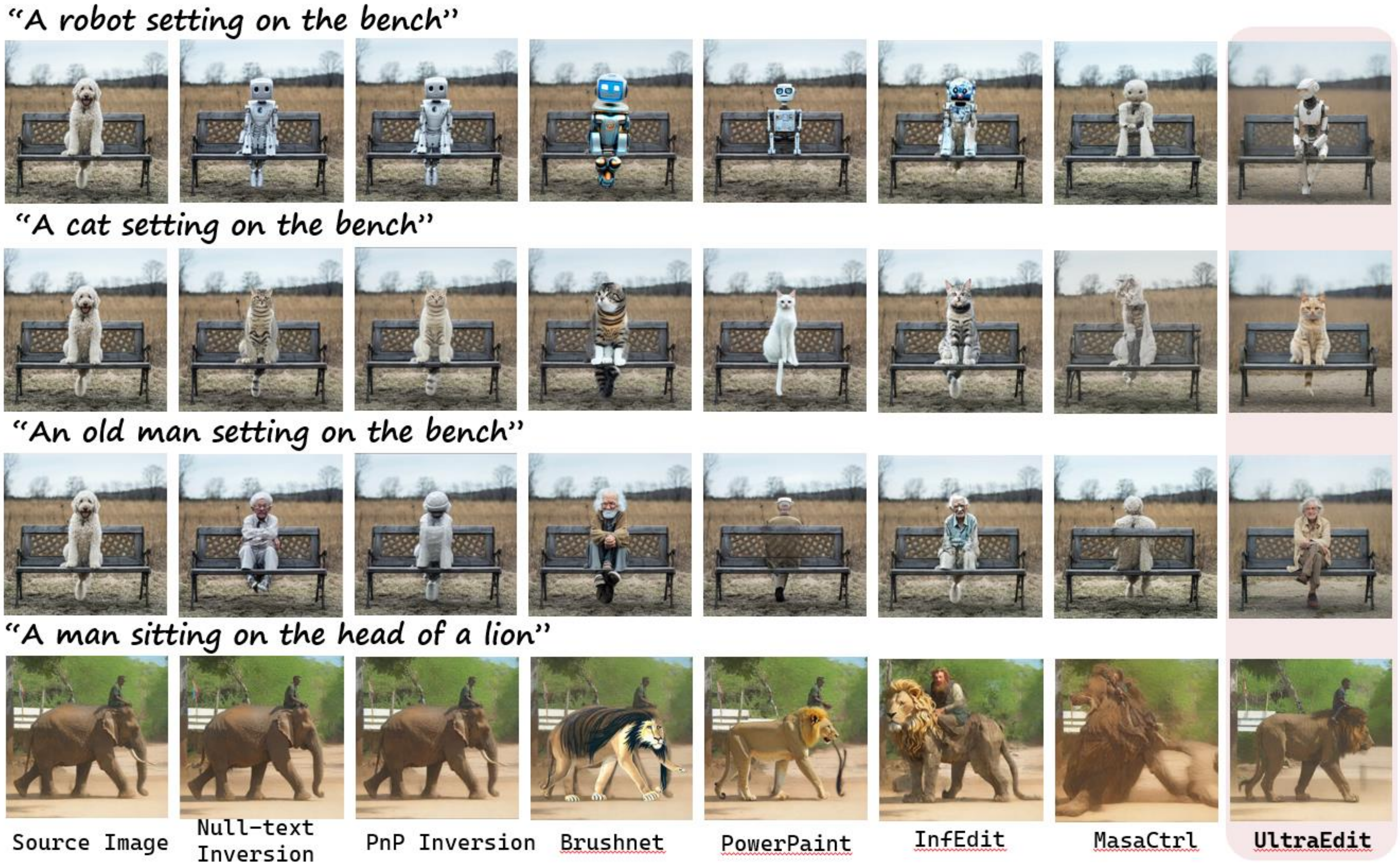
Region-based image generation

$$z_{t-1} = \begin{cases} (1 - M_s) \cdot z_T + M_s \cdot DM(z_t) & \text{if } t \mod 2 == 0 \\ DM(z_t) & \text{otherwise} \end{cases}$$

Usage of the soft mask



Comparison with other generation methods



Characteristics and Statistics

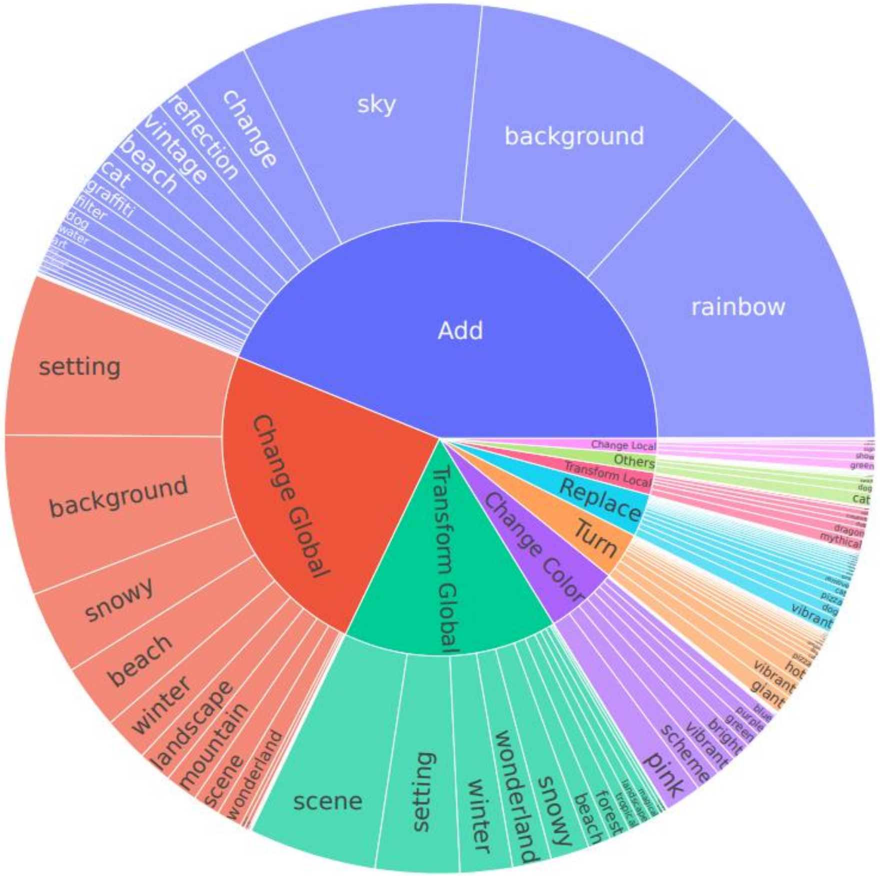


Figure 3: Distribution of edit types and keywords in the instructions of ULTRAEDIT. The inner ring illustrates the various types of edit instructions, while the outer ring presents the frequency of instruction keywords. This visualization highlights the rich diversity found within our instructions.

Table 2: Editing Instruction Types in ULTRAEDIT.

Type	Description
Add	Inserting a new object or texture at a specific location in the image.
Change Global	Modifying the entire image to achieve a clear and noticeable effect.
Change Local	Altering a specific object or texture, affecting only a portion of the image.
Change Color	Adjusting the color within the image.
Transform Global	Smoothly transforming images into a different setting, scene, or style.
Transform Local	Modifying part of image features while preserving its overall structure.
Replace	Substituting existing objects in the image with those specified in the instructions.
Turn	Implicitly changing objects, background, or texture, often without a specific target.
Others	Miscellaneous editing types such as text edits and altering quantities.

Table 3: Quantitative evaluation for ULTRAEDIT.

Metric	Free-form.	Region-based.
CLIPimg	0.8427	0.8813
SSIM	0.6401	0.7413
DINOv2	0.7231	0.7688
CLIPin	0.2834	0.2848
CLIPout	0.3049	0.2848
CLIPdir	0.2950	0.3052

4,108,262 image editing data (757,879 unique edits):

free-form image editing: 4,000,083 samples

region-based editing: 108,179 samples

Experiments on the MagicBrush benchmark

Settings	Methods	L1↓	L2↓	CLIP-I↑	DINO↑
	Global Description-guided				
	SD-SDEdit	0.1014	0.0278	0.8526	0.7726
	Null Text Inversion	0.0749	0.0197	0.8827	0.8206
	GLIDE	3.4973	115.8347	0.9487	0.9206
	Blended Diffusion	3.5631	119.2813	0.9291	0.8644
Single-turn	***Instruction-guided***				
	HIVE	0.1092	0.0380	0.8519	0.7500
	InstructPix2Pix (IP2P)	0.1141	0.0371	0.8512	0.7437
	IP2P w/ MagicBrush	0.0625	0.0203	**0.9332**	**0.8987**
	Ours, trained w/o region data	0.0689	0.0201	0.8986	0.8477
	Ours, eval w/o region	0.0614	0.0181	0.9197	0.8804
	Ours, eval w/ region	**0.0575**	0.0172	0.9307	0.8982
	Global Description-guided				
	SD-SDEdit	0.1616	0.0602	0.7933	0.6212
	Null Text Inversion	0.1057	0.0335	0.8468	0.7529
	GLIDE	11.7487	1079.5997	0.9094	0.8494
	Blended Diffusion	14.5439	1510.2271	0.8782	0.7690
Multi-turn	***Instruction-guided***				
	HIVE	0.1521	0.0557	0.8004	0.6463
	InstructPix2Pix (IP2P)	0.1345	0.0460	0.8304	0.7018
	IP2P w/ MagicBrush	0.0964	0.0353	0.8924	0.8273
	Ours, trained w/o region data	0.0883	0.0276	0.8685	0.7922
	Ours, eval w/o region	0.0780	0.0246	0.8954	0.8322
	Ours, eval w/ region	**0.0745**	**0.0236**	**0.9045**	**0.8505**

Trained on the same amount of data, ours already attains significant improvement over the baseline, confirming the advantages brought by our dataset to general image editing

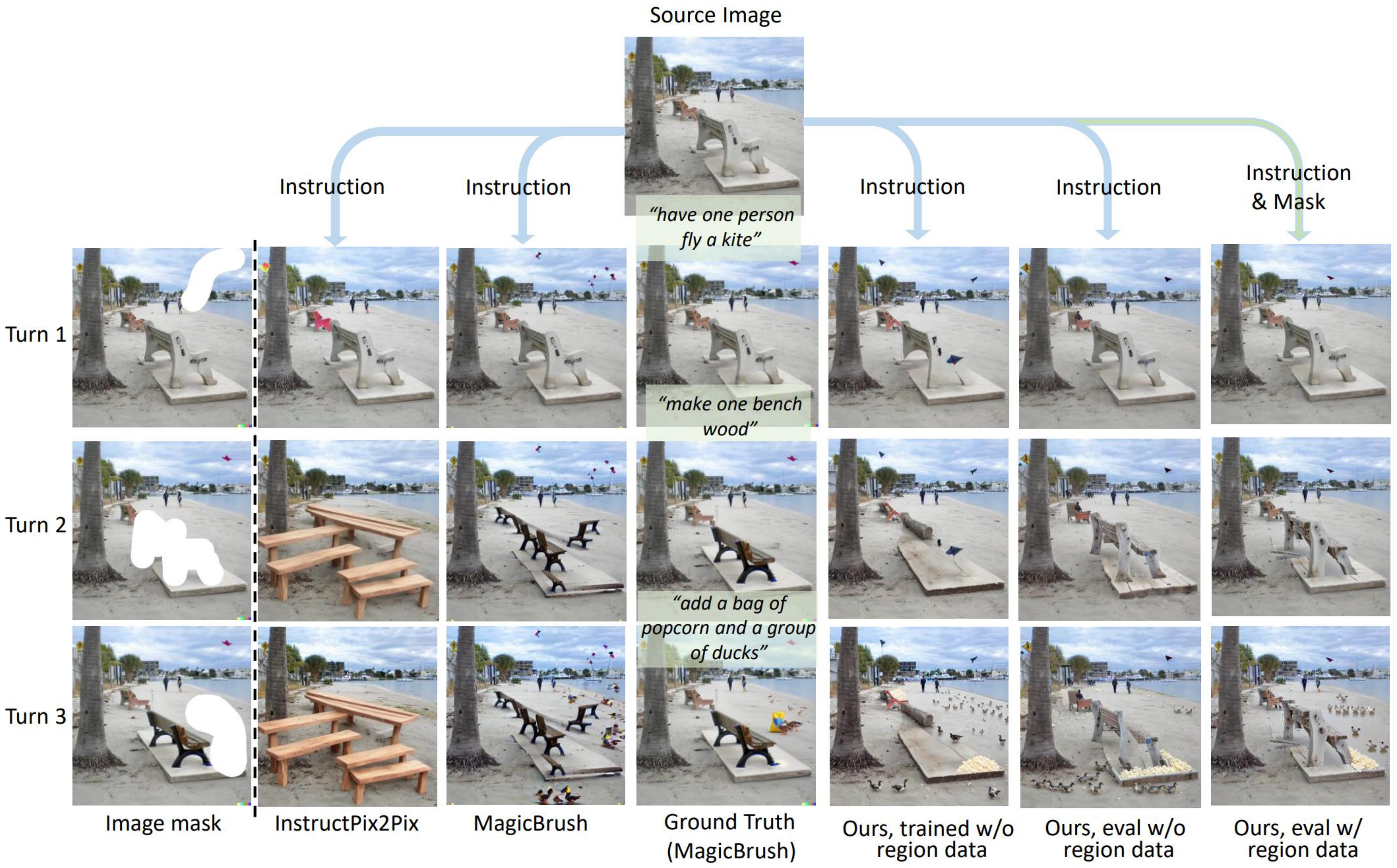
Experiments on the MagicBrush benchmark

Settings	Methods	L1↓	L2↓	CLIP-I↑	DINO↑
Single-turn	***Global Description-guided***				
	SD-SDEdit	0.1014	0.0278	0.8526	0.7726
	Null Text Inversion	0.0749	0.0197	0.8827	0.8206
	GLIDE	3.4973	115.8347	0.9487	0.9206
	Blended Diffusion	3.5631	119.2813	0.9291	0.8644
	Instruction-guided				
	HIVE	0.1092	0.0380	0.8519	0.7500
	InstructPix2Pix (IP2P)	0.1141	0.0371	0.8512	0.7437
	IP2P w/ MagicBrush	0.0625	0.0203	**0.9332**	**0.8987**
	Ours, trained w/o region data	0.0689	0.0201	0.8986	0.8477
	Ours, eval w/o region	0.0614	0.0181	0.9197	0.8804
	Ours, eval w/ region	**0.0575**	0.0172	0.9307	0.8982
Multi-turn	***Global Description-guided***				
	SD-SDEdit	0.1616	0.0602	0.7933	0.6212
	Null Text Inversion	0.1057	0.0335	0.8468	0.7529
	GLIDE	11.7487	1079.5997	0.9094	0.8494
	Blended Diffusion	14.5439	1510.2271	0.8782	0.7690
	Instruction-guided				
	HIVE	0.1521	0.0557	0.8004	0.6463
	InstructPix2Pix (IP2P)	0.1345	0.0460	0.8304	0.7018
	IP2P w/ MagicBrush	0.0964	0.0353	0.8924	0.8273
	Ours, trained w/o region data	0.0883	0.0276	0.8685	0.7922
	Ours, eval w/o region	0.0780	0.0246	0.8954	0.8322
	Ours, eval w/ region	**0.0745**	**0.0236**	**0.9045**	**0.8505**

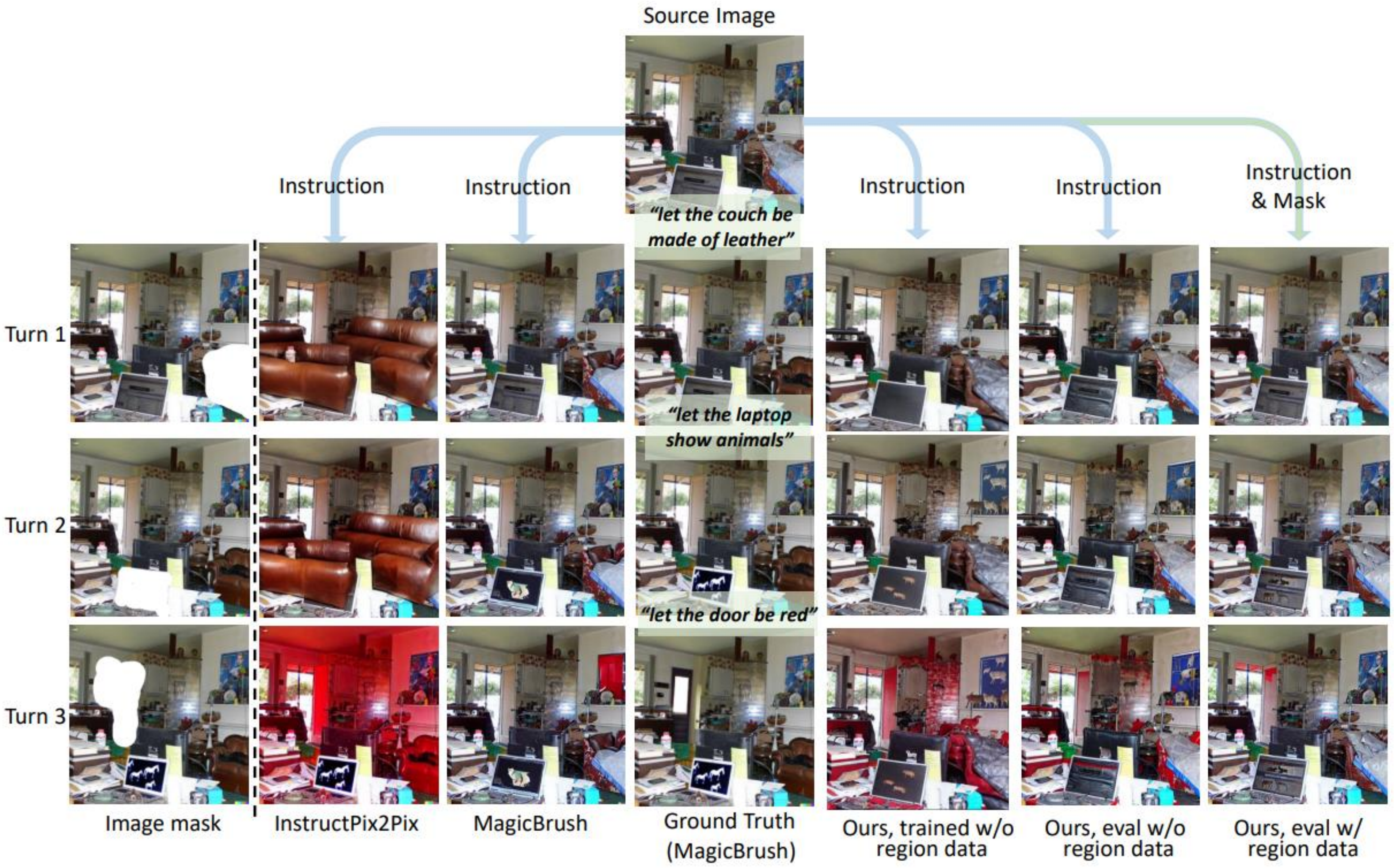
Incorporating region-based editing data during training, and evaluate on the same setting without editing region input, the general editing performance can be boosted considerably.

Experiments on the MagicBrush benchmark

Multi-step Image Editing



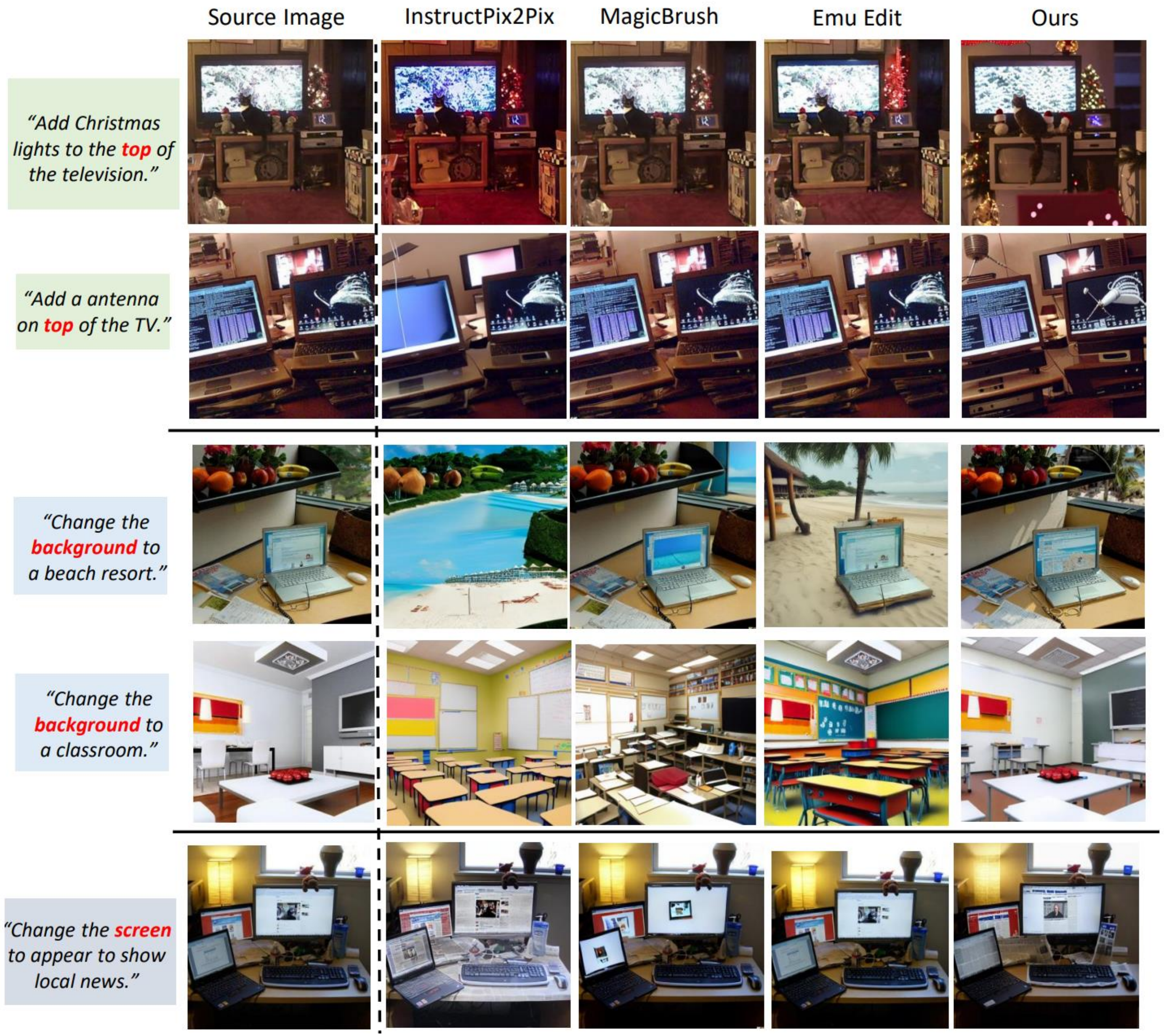
Multi-step Image Editing



Experiments on the EmuEdit benchmark

Method	CLIPdir↑	CLIPout↑	L1↓	CLIPimg↑	DINO↑
InstructPix2Pix (450K)	0.0784	0.2742	0.1213	0.8518	0.7656
MagicBrush (450+20K)	0.0658	0.2763	0.0652	0.9179	**0.8924**
Emu Edit(10M)	0.1066	**0.2843**	0.0895	0.8622	0.8358
Ours (450k, w/o region data)	0.0823	0.2778	0.0626	0.8617	0.8190
Ours (1M w/o region data)	0.0862	0.2804	**0.0515**	**0.8915**	0.8656
Ours (1.5M, w/o region data)	0.0952	0.2808	0.0600	0.8659	0.8243
Ours (2M, w/o region data)	0.0960	0.2811	0.0608	0.8689	0.8269
Ours (2.5M, w/o region data)	0.0997	0.2822	0.0854	0.8407	0.7814
Ours (3M, w/o region data)	**0.1076**	0.2832	0.0713	0.8446	0.7937

Experiments on the EmuEdit benchmark



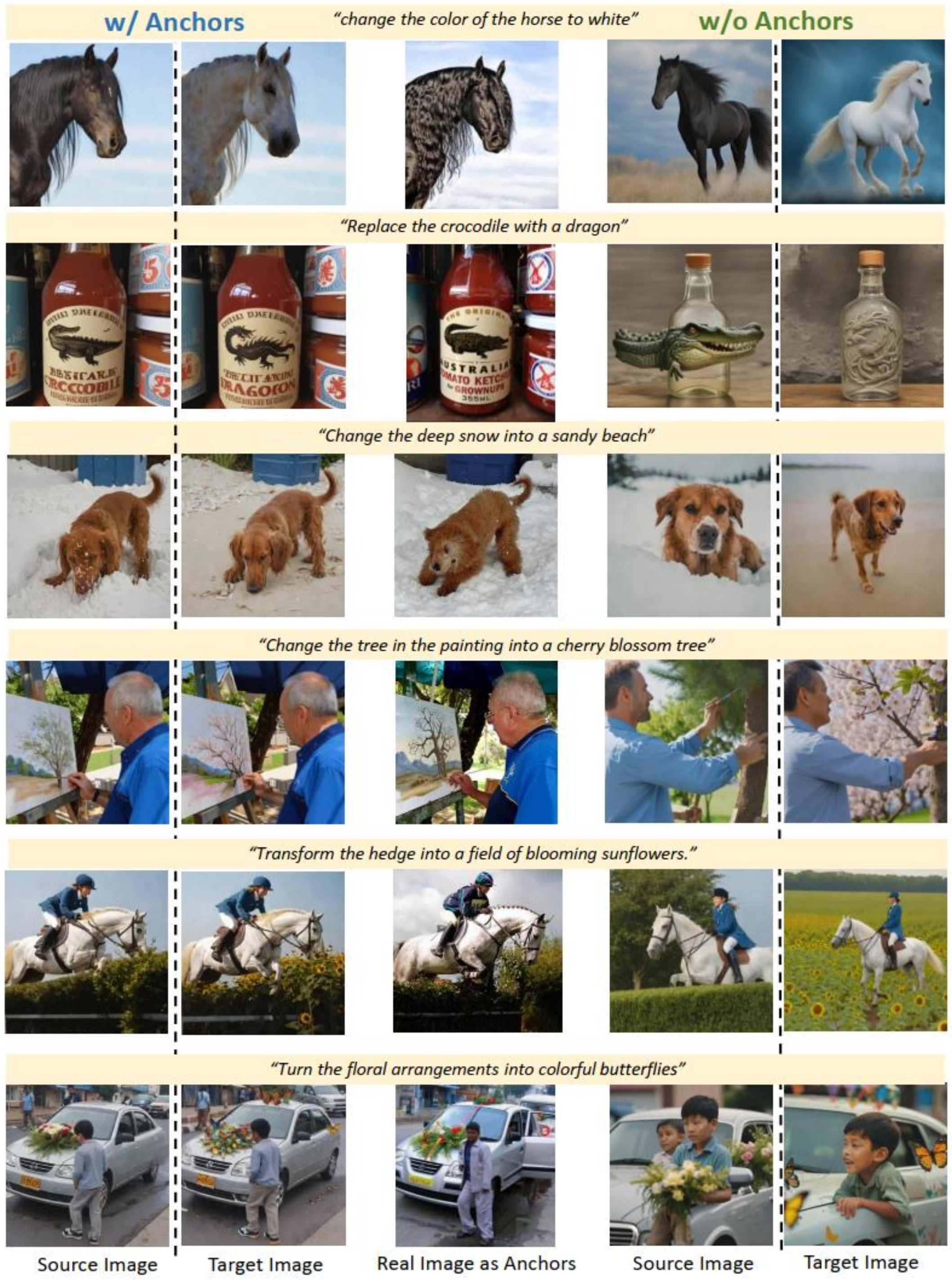
Insights and Analysis

Real Image Anchors for Generation

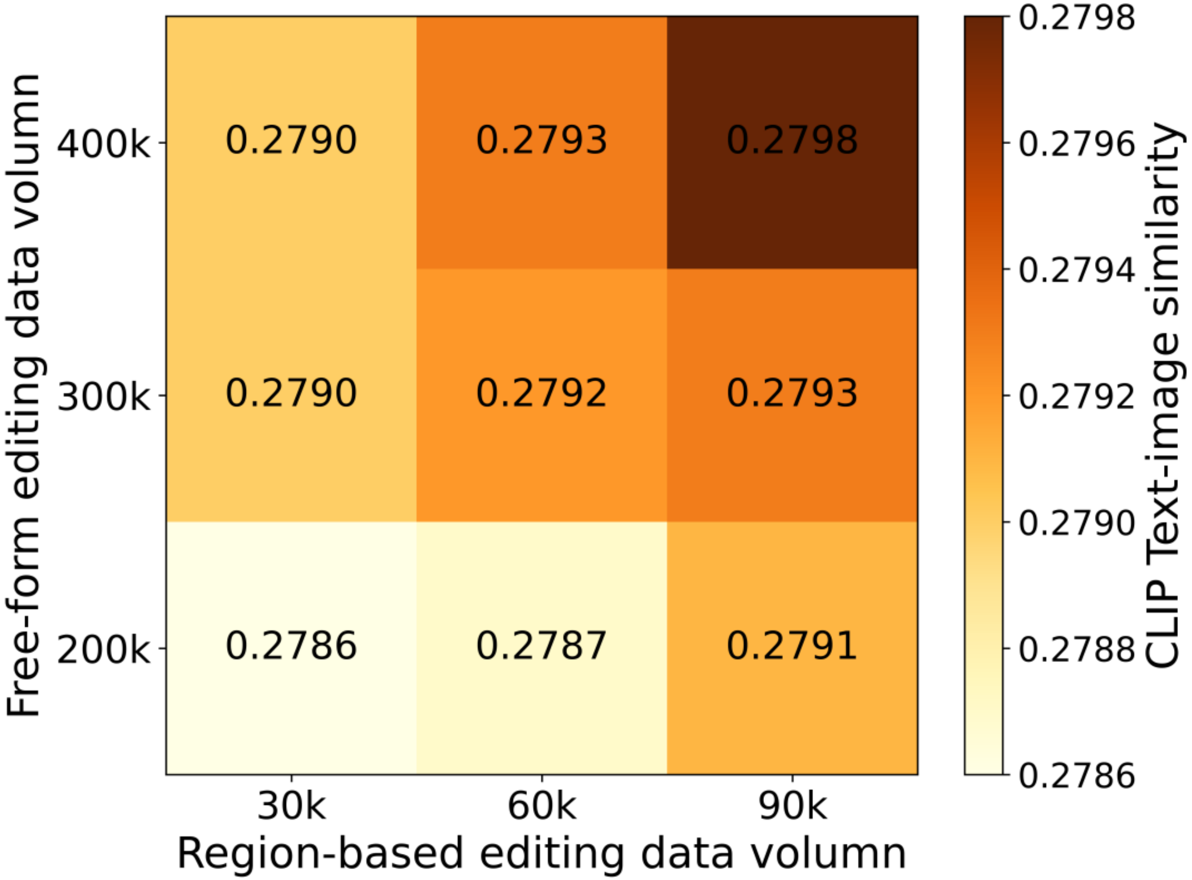
Data Type	Data Volume	CLIPdir↑	CLIPimg↑	CLIPout↑	L1↓	DINO↑
UltraEditing	450k	0.0823	0.8617	0.2778	0.0626	0.8190
	1M	0.0925	0.8696	0.2807	0.0599	0.8307
	1.5M	0.0952	0.8659	0.2808	0.0600	0.8243
w/o image anchor	450k	0.0728	0.8716	0.2796	0.0848	0.8154
	1M	0.0638	0.8837	0.2770	0.0674	0.8353
	1.5M	0.0720	0.8643	0.2781	0.0714	0.8105

(1) Dataset generated with real image anchors generally leads to better models.

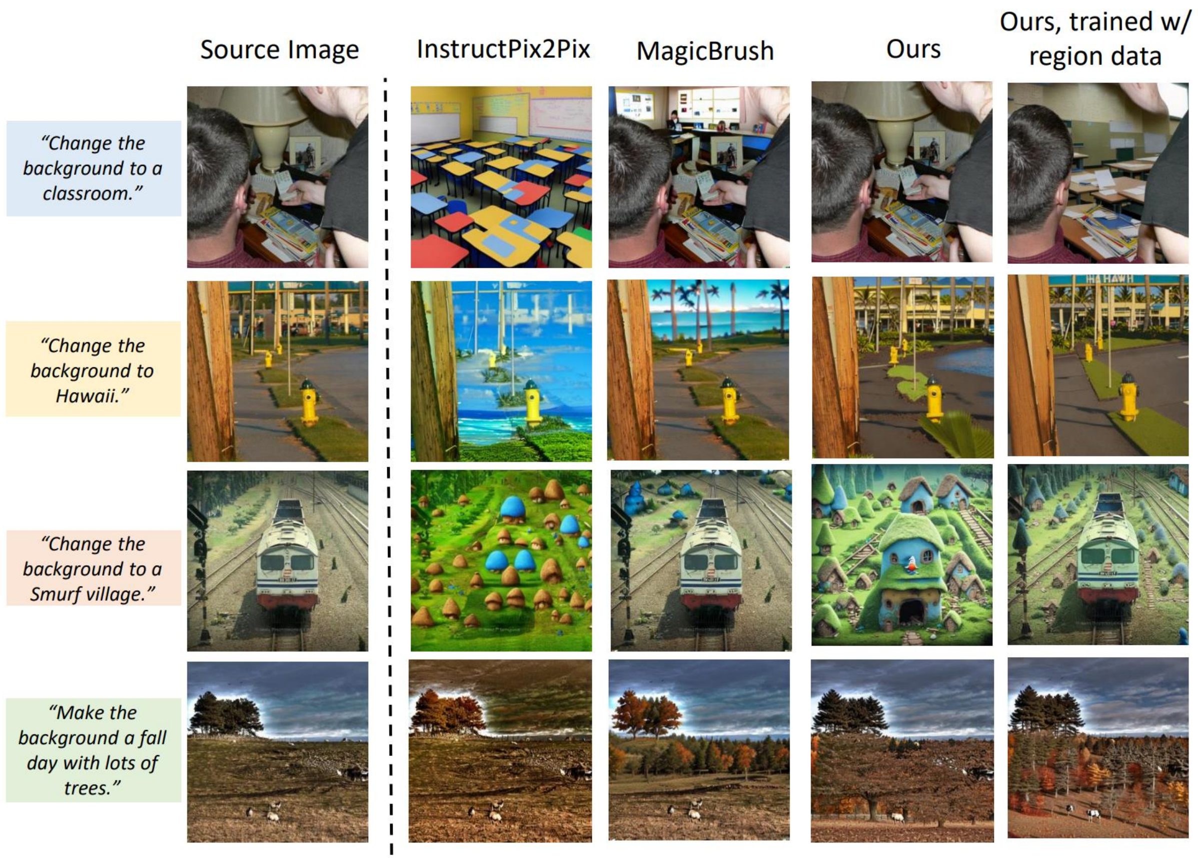
(2) The scaling effect only presents when real image anchors are adopted



Free-from vs. Region-based Editing.



Incorporating region-based editing data during model training can help with free-form editing tasks, model exhibits significantly more precise operations for background and localized edits



Conclusion

- We’ve presented **ULTRAEDIT**, a large-scale, high-quality dataset for instruction-based image editing.
- We **mitigate the issues** in existing editing datasets with a **systematic** approach for **automatic** data generation.
- Experiments on challenging benchmarks confirm the high quality of the dataset, as well as the effectiveness of training on our dataset.