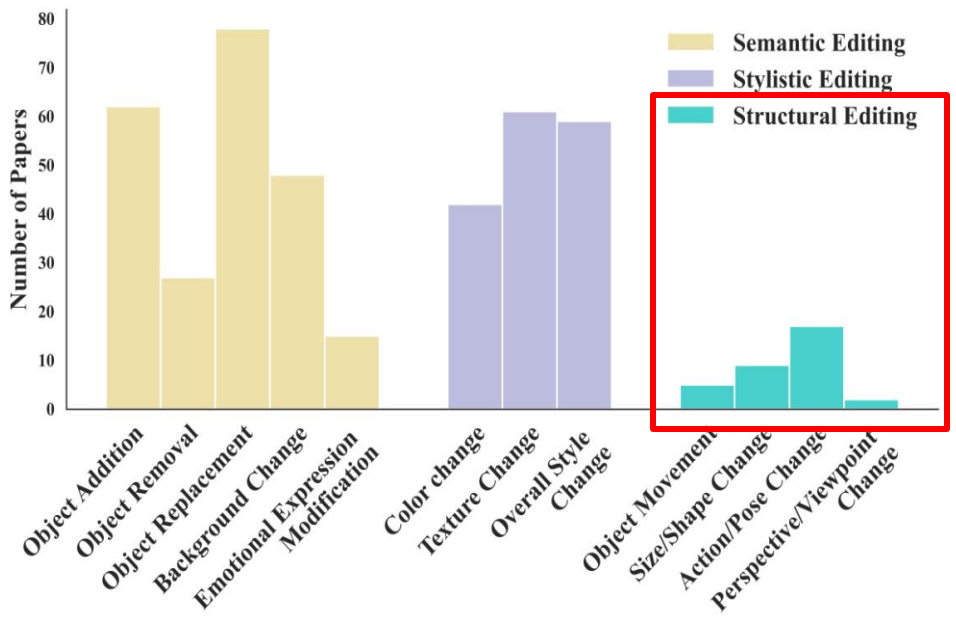
Training-Free Geometric Image Editing on Diffusion Models

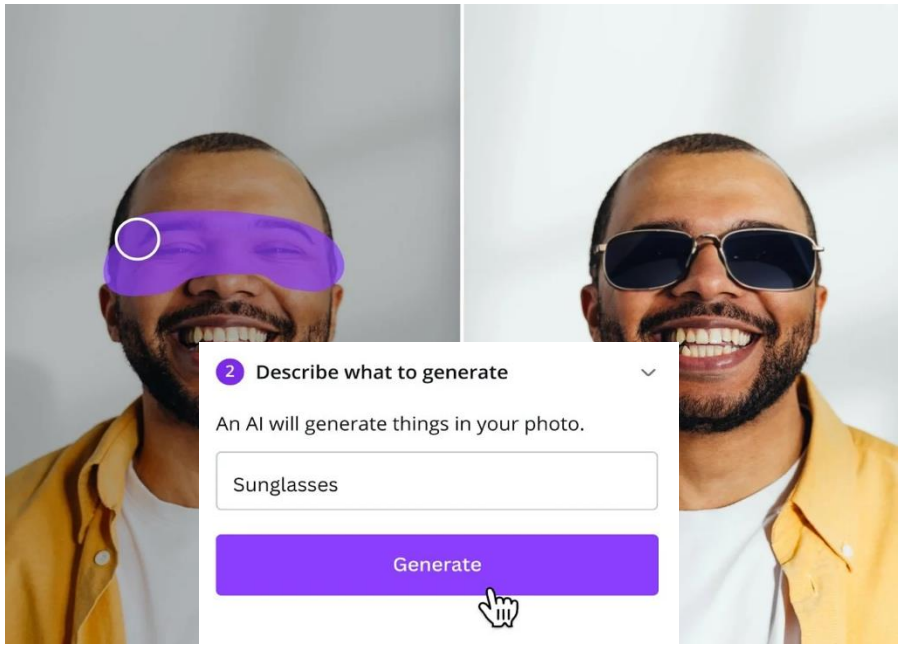
Hanshen Zhu[1*] Zhen Zhu[2*] Kaile Zhang[1] Yiming Gong[2] Yuliang Liu[1] Xiang Bai[1†]
[1] Huazhong University of Science and Technology [2] University of Illinois at Urbana-Champaign

Background

1. Image generation models achieve photorealistic quality
2. Limited controllability remains a core bottleneck
3. Geometric Image editing remains underexplored



* Diffusion Model-Based Image Editing: A Survey

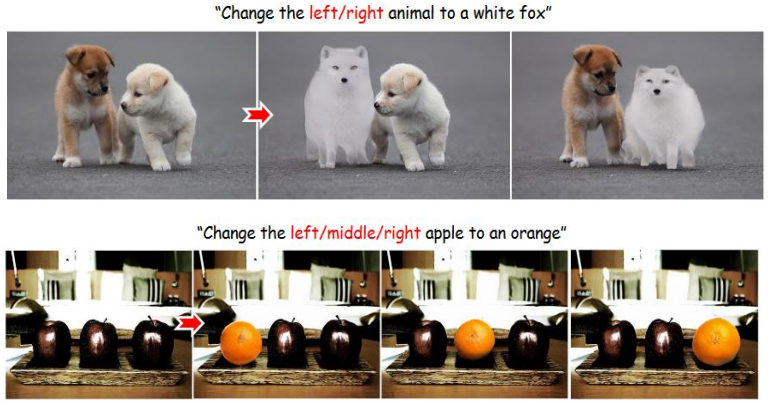


可控图像编辑示例 * Canva可画

Task Definition

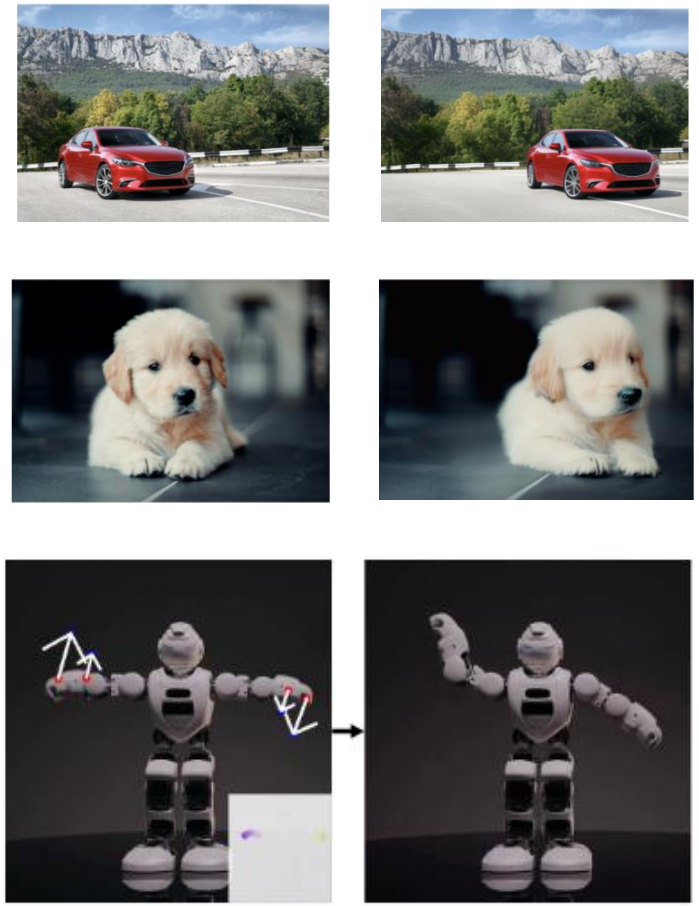
- **Repositioning, reorienting, or reshaping an object within an image while preserving overall scene coherence.**

Semantic Editing



Stylistic Editing

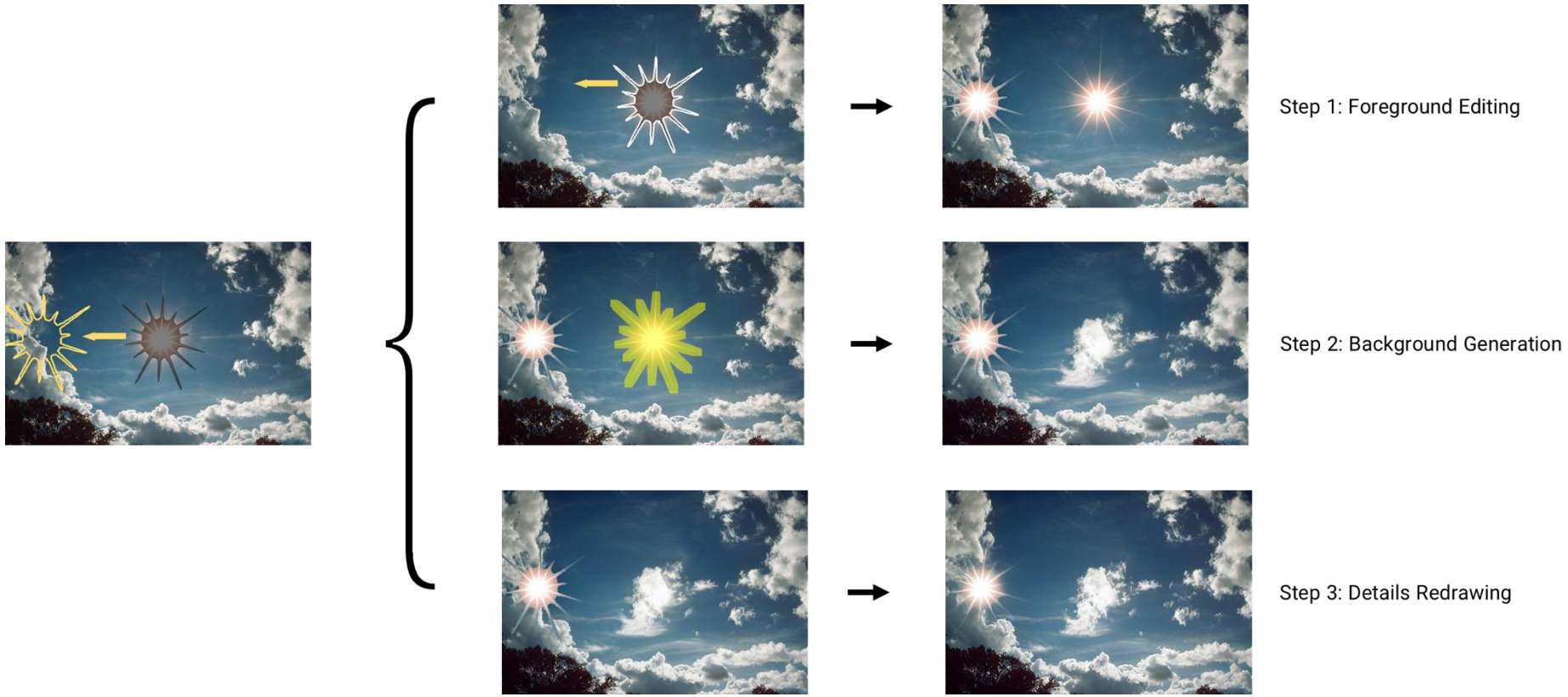
Geometric Editing



Challenges

A complex task involving multiple subtasks:

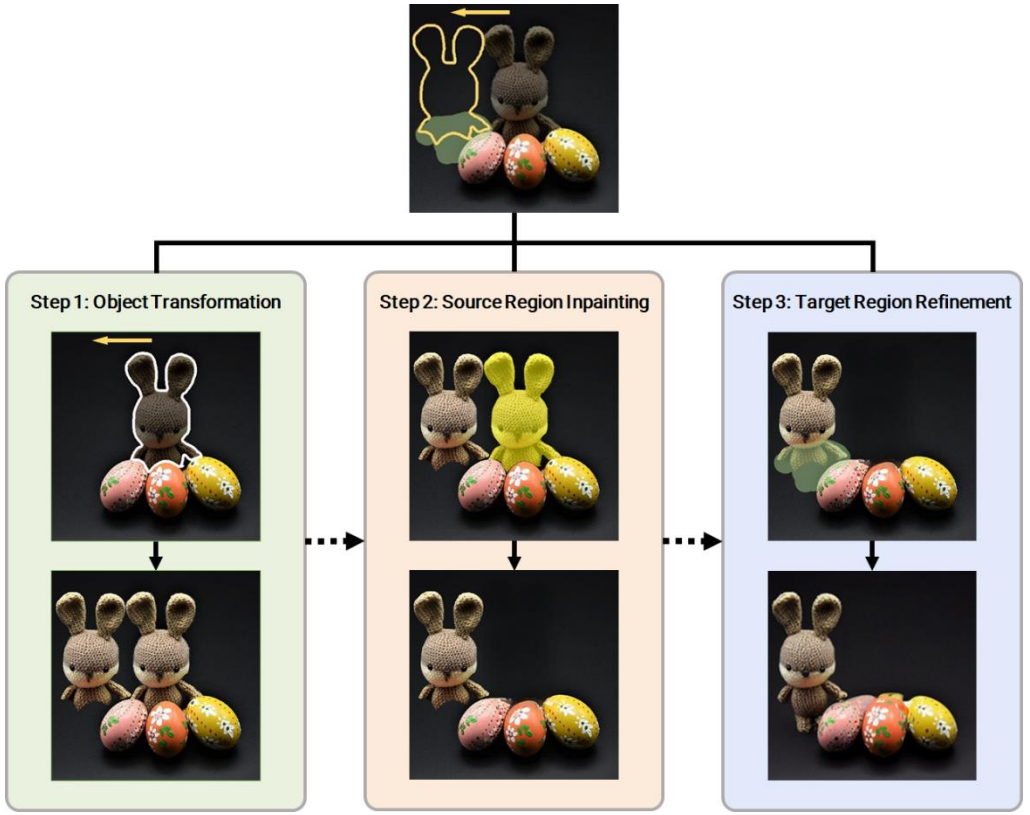
- Precise copying and transformation of target objects.
- High-quality inpainting of the source region to avoid artifacts.
- Seamless blending of transformed objects with the background.



Challenges

A complex task involving multiple subtasks:

- Precise copying and transformation of target objects.
- High-quality inpainting of the source region to avoid artifacts.
- **More difficult with structure completion demand.**

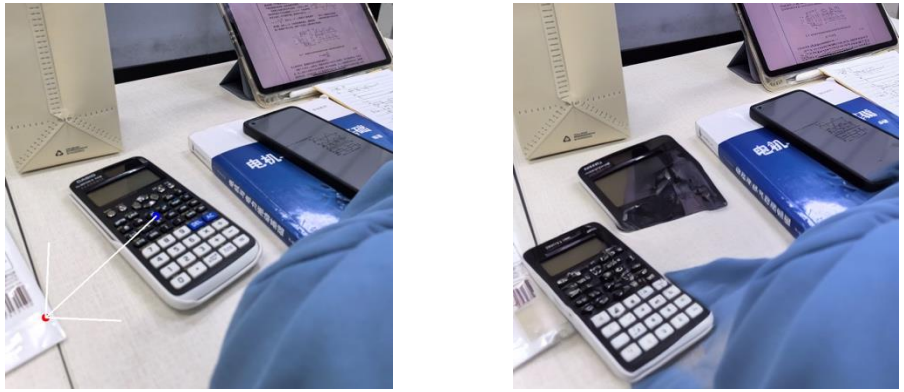


Limitations on existing methods

Recent methods typically address these goals with a single, unified objective. Balancing multiple subtasks in one optimization framework leads to:

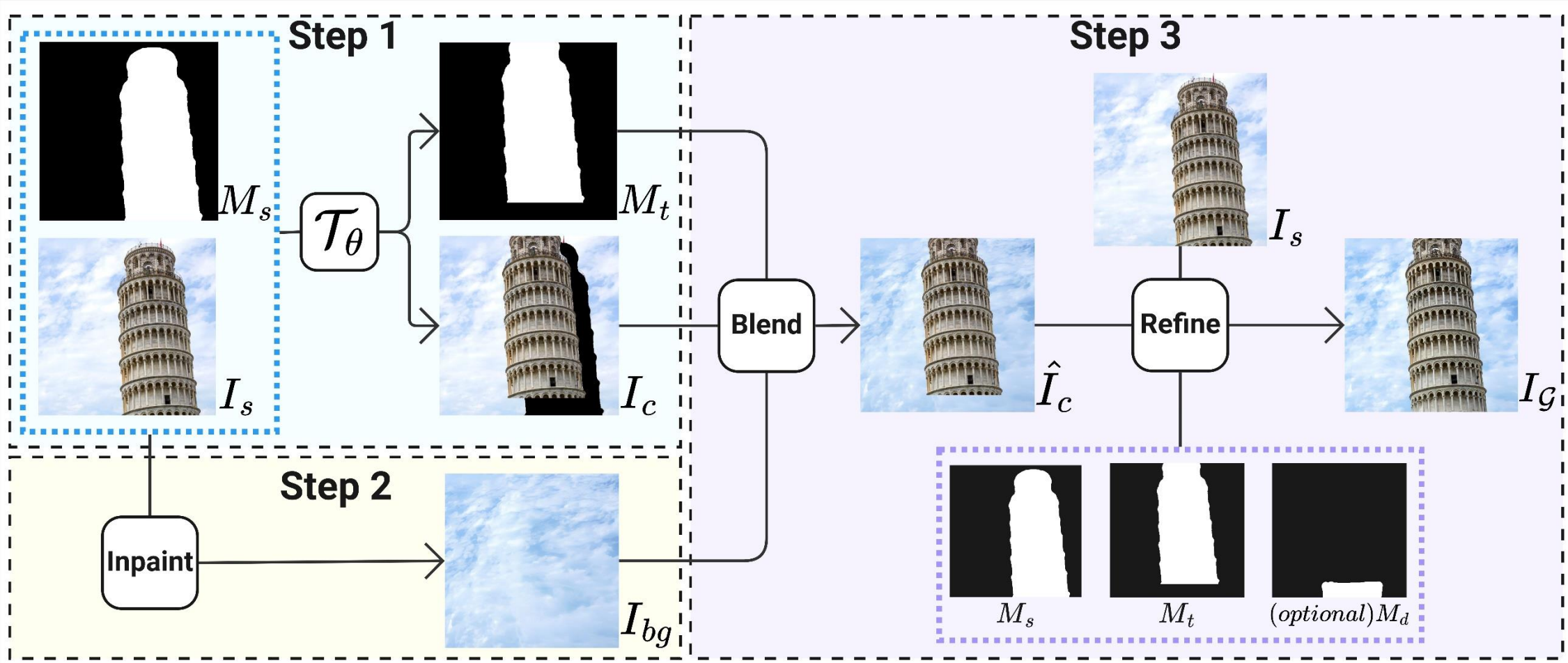
- Struggle in preserving details and avoiding artifacts.
- Primarily restricted to 2D or minor 3D adjustments.
- Cannot naturally achieve structural completion

Self-Guidance(NIPS2023)



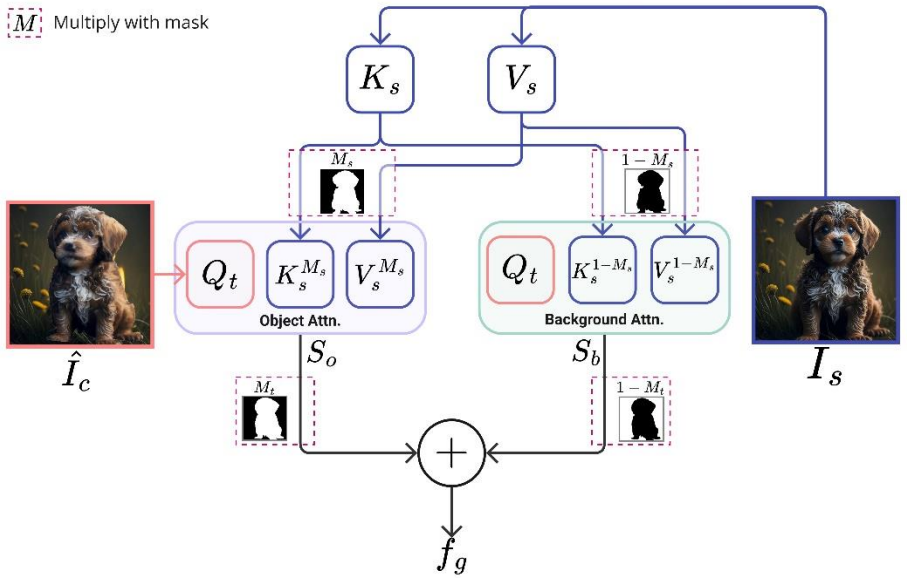
DragonDiffusion(ICLR2024)

Method

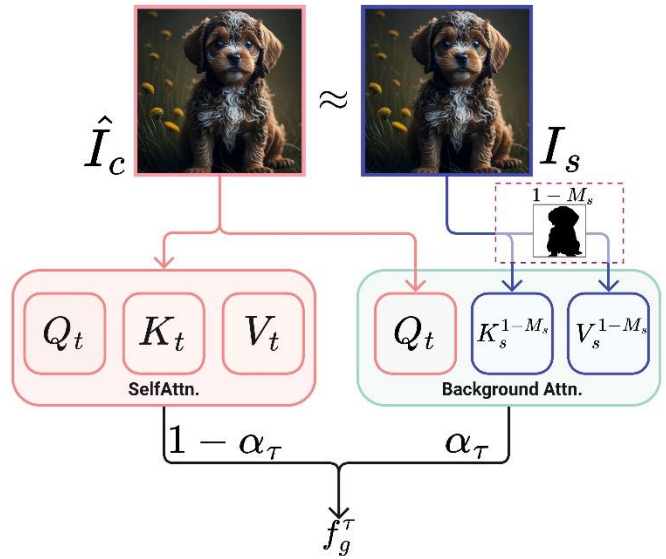


A training-**free** image re**fine**ment approach, **FreeFine**

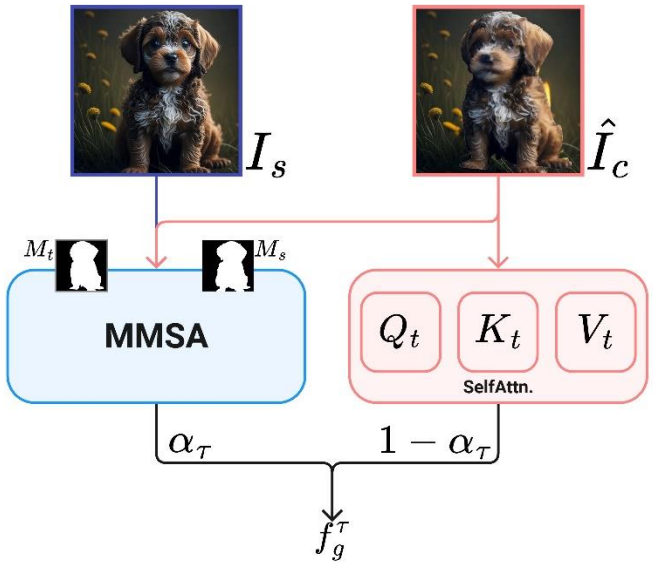
Method



(a) Mask-guided Mutual Self Attention (**MMSA**)



(b) Temporal Contextual Attention (**Step 2**)



(c) Temporal Contextual Attention (**Step 3**)

Figure 3. Comparison of Context Aggregation Methods. This figure illustrates different approaches for context alignment in image editing tasks: (a) MMSA [3] replaces Key-Value (KV) pairs and enforces explicit feature interaction between regions. (b) TCA (Step2) for source region inpainting, and (c) TCA (Step3) for target region refinement, which smoothly transition from MMSA to full self-attention.

1. Temporal Contextual Attention

Method

2.Local Perturbation

$$x_{t-1} = \begin{cases} \text{DDPM}(x_t), & \text{if } x \in \mathcal{M}, \\ \text{DDIM}(x_t), & \text{otherwise.} \end{cases}$$

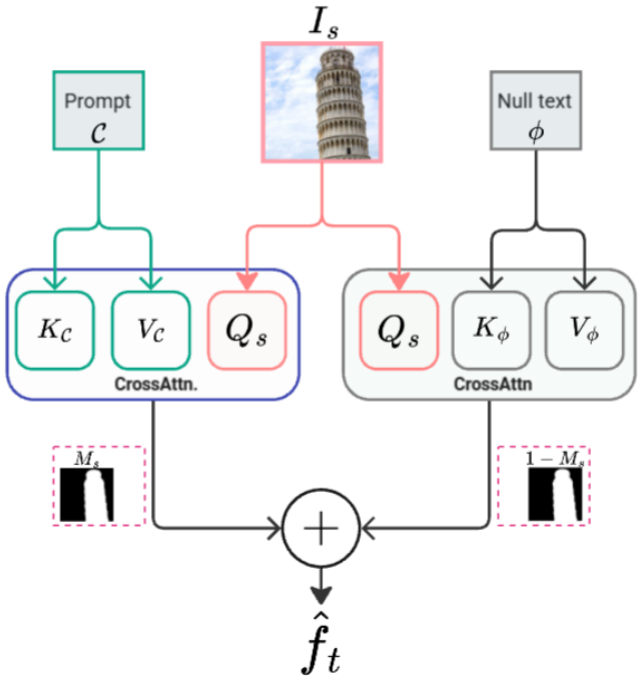
$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}, \quad (1)$$

$$\tilde{x}_{t-1} = \sqrt{\alpha_{t-1}} \cdot \tilde{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \cdot \epsilon, \quad (2)$$

$$\sigma_t = \begin{cases} \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)}, & \text{if } \mathcal{M} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This allows LP to selectively apply DDPM updates [7] within the mask and DDIM updates [21] elsewhere, balancing flexibility and control.

3. Content-specified Generation



$$\tilde{f}_t = \text{CrossAttn}(Q_t, K_\mathcal{C}, V_\mathcal{C}) \cdot \mathcal{M}_1 + \text{CrossAttn}(Q_t, K_\varnothing, V_\varnothing) \cdot (1 - \mathcal{M}_1), \quad (2)$$

$$\hat{\epsilon}_\theta(x_t, \mathcal{C}) = \epsilon_\theta(x_t, \varnothing) + w\left[\epsilon_\theta(x_t, \mathcal{C}) - \epsilon_\theta(x_t, \varnothing)\right] \cdot \mathcal{M}_2, \quad (3)$$

Results

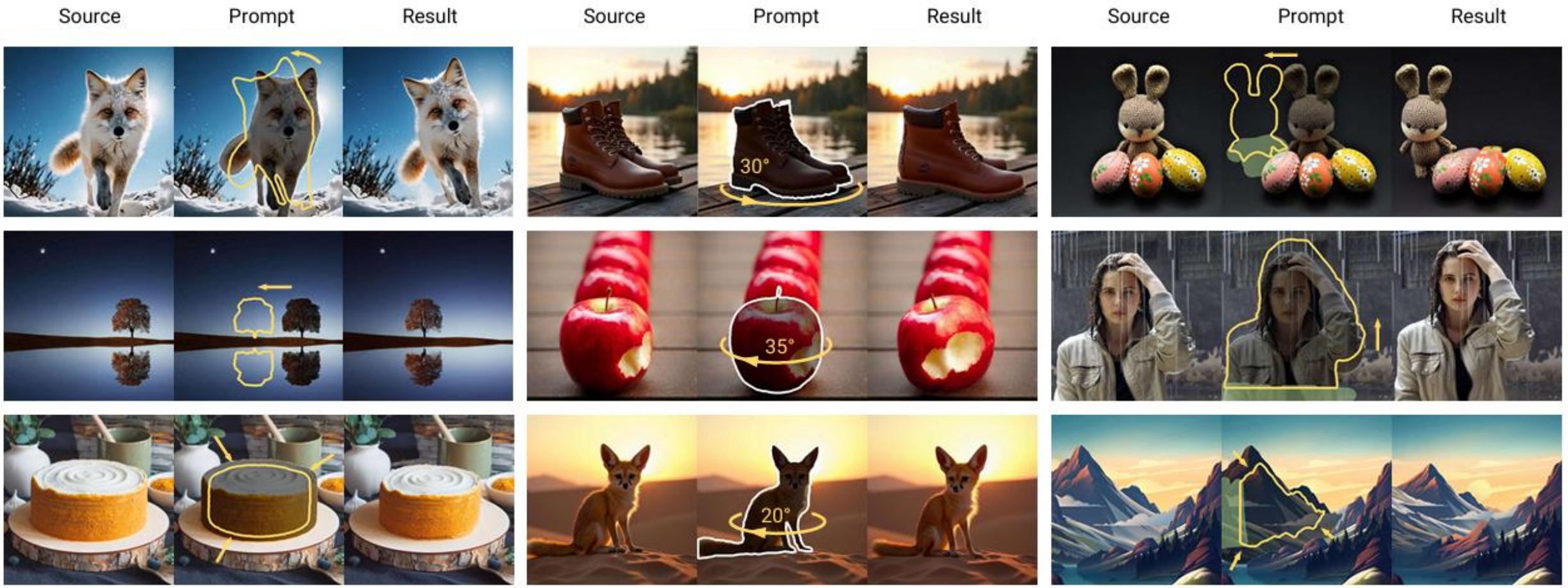
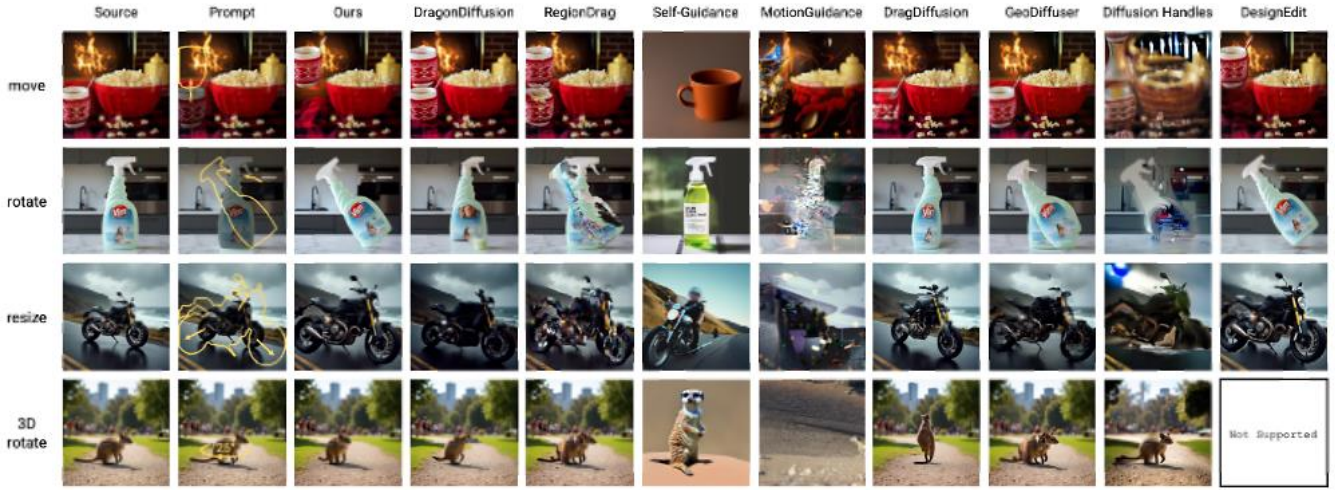
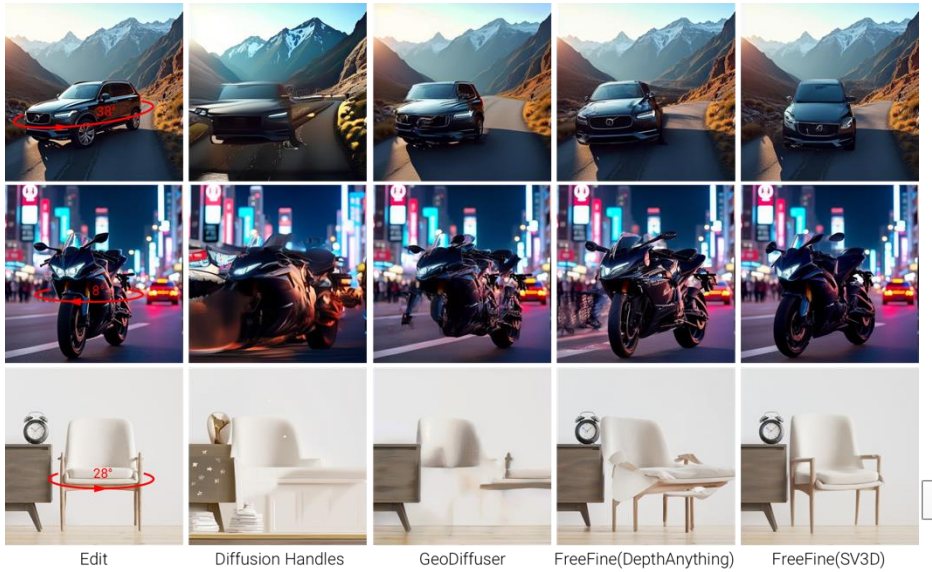


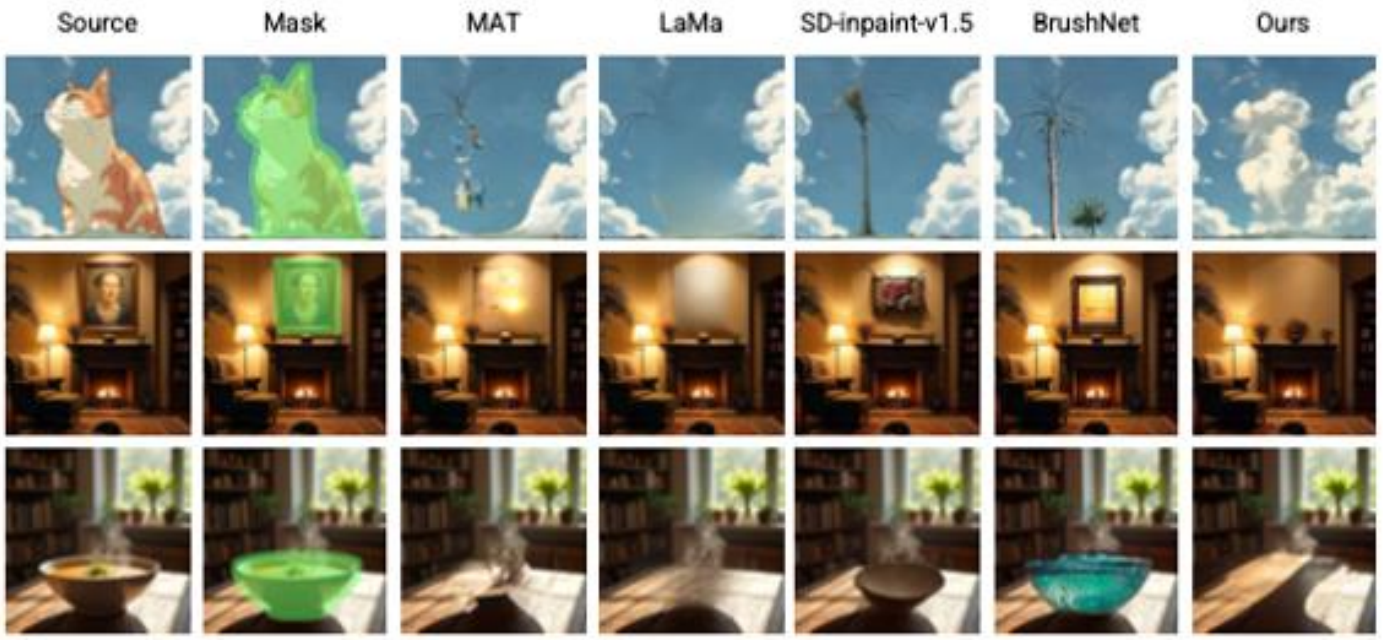
Figure 1. Given an image and an editing instruction, our method precisely performs geometric edits while maintaining high fidelity and avoiding artifacts. Besides, our training-free framework achieves impressive structural completion and background generation.

Results

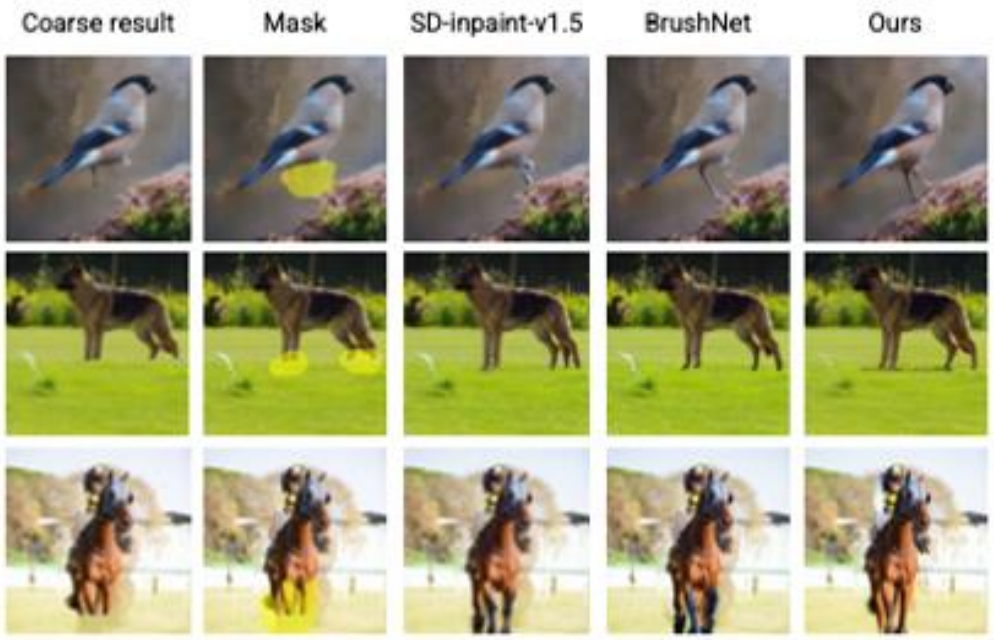
Methods	External Model	Editing Type	FID	DINOv2	KD	SUBC	BC	WE	MD
Self-Guidance [8]	SAM [26]	2D	49.15	647.56	0.438	0.575	0.759	0.268	116.89
RegionDrag [33]	SAM [26]		40.21	504.50	0.241	0.796	0.970	0.120	32.50
DragonDiffusion [39]	SAM [26]		37.09	507.67	0.144	0.840	0.968	0.158	32.36
MotionGuidance [10]	SAM [26], RAFT [59]		106.39	1189.23	3.871	0.521	0.736	0.186	90.03
DragDiffusion [50]	SAM [26]		36.58	455.68	**0.142**	0.758	0.966	0.199	41.31
Diffusion Handles [42]	SAM [26], LaMa [57], DepthAnything [64]		44.81	549.69	0.618	0.725	0.852	0.180	40.27
GeoDiffuser [49]	SAM [26], DepthAnything [64]		**33.89**	**437.75**	0.173	0.762	0.938	0.166	34.88
DesignEdit [21]	SAM [26]		35.22	480.91	0.179	0.874	0.959	0.098	10.15
FreeFine	SAM [26]		34.72	478.18	0.144	**0.907**	**0.971**	**0.055**	**9.25**
DragDiffusion [50]	SAM [26]	3D	157.42	**1867.02**	0.348	0.603	**0.958**	0.199	61.97
Diffusion Handles [42]	SAM [26], LaMa [57], DepthAnything [64]		156.90	1882.66	0.523	0.705	0.882	0.128	26.10
GeoDiffuser [49]	SAM [26], DepthAnything [64]		152.06	1894.26	0.351	0.749	0.941	0.097	34.34
FreeFine	SAM [26], DepthAnything [64]		**150.89**	1879.69	**0.310**	**0.786**	0.956	**0.079**	**20.32**
BrushNet [22]	SAM [26]	SC	186.93	**2516.52**	**0.971**	0.925	0.948	0.060	11.31
SD-inpainting [54]	SAM [26]		193.71	2556.44	1.047	0.913	0.928	0.064	14.43
FreeFine	SAM [26]		**184.84**	2526.38	0.982	**0.928**	**0.952**	**0.056**	**9.56**







(a) Source Region Inpainting



(b) Target Region Refinement

Results

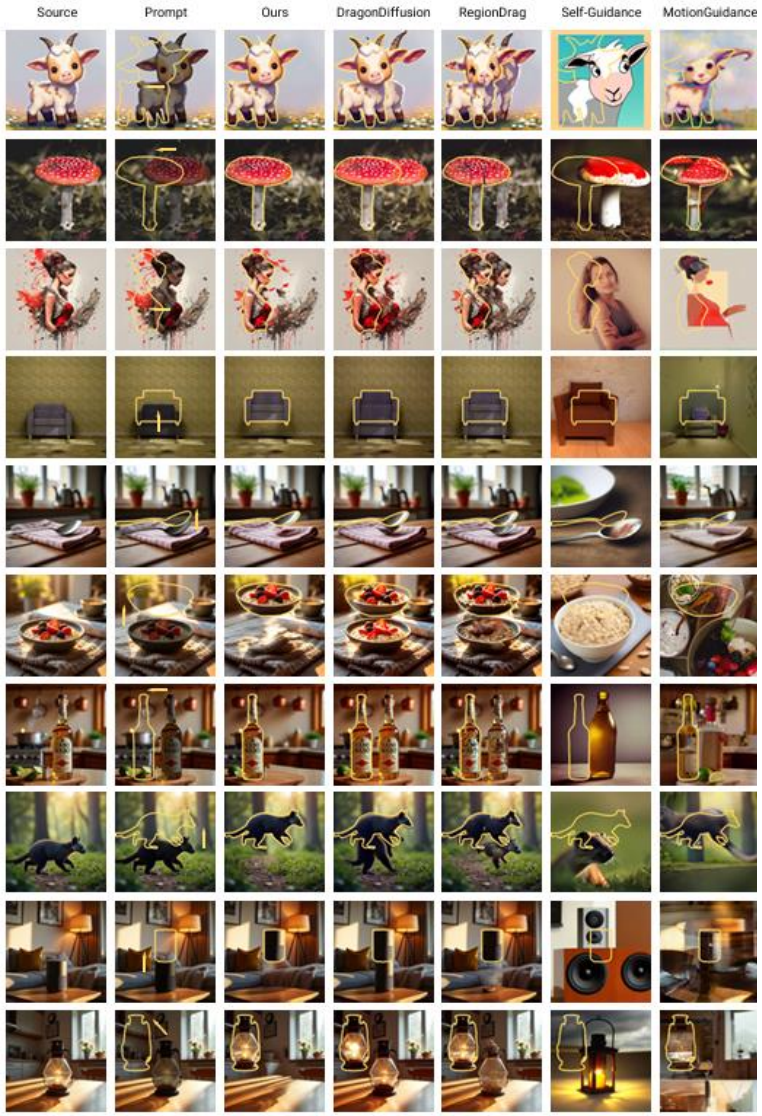


Figure 4. Qualitative comparison with state-of-the-art editing methods in moving operations

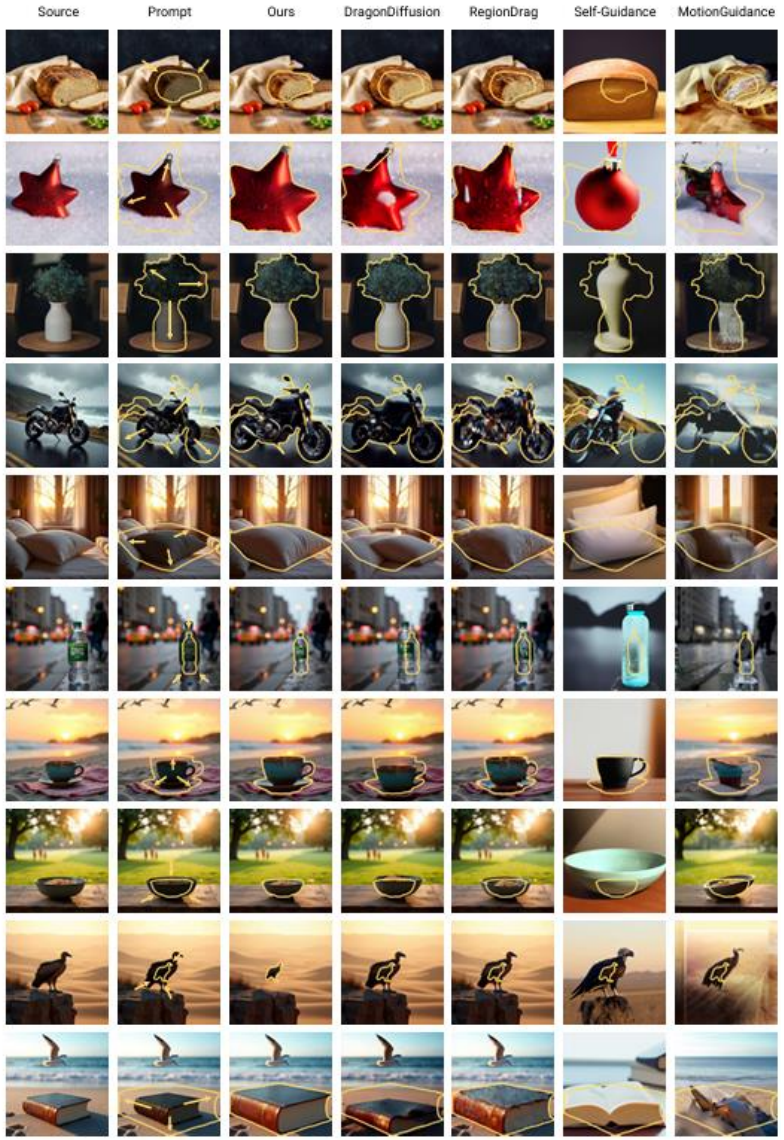


Figure 5. Qualitative comparison with state-of-the-art editing methods in scaling operation

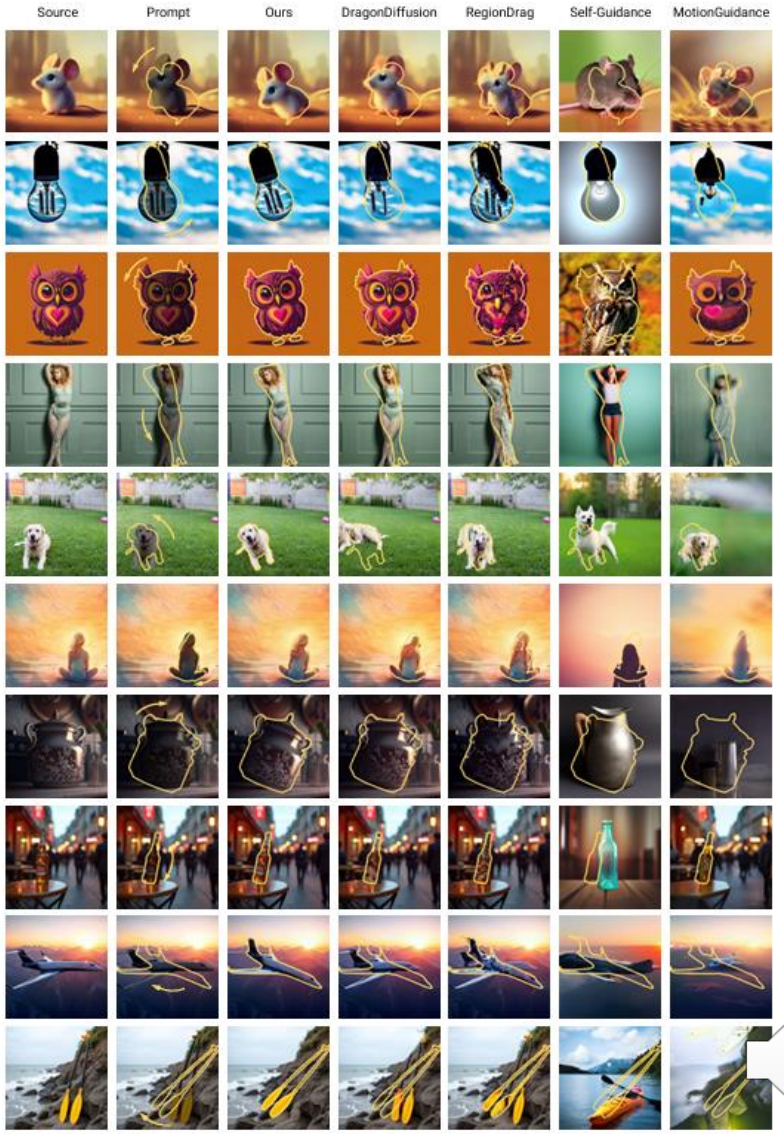
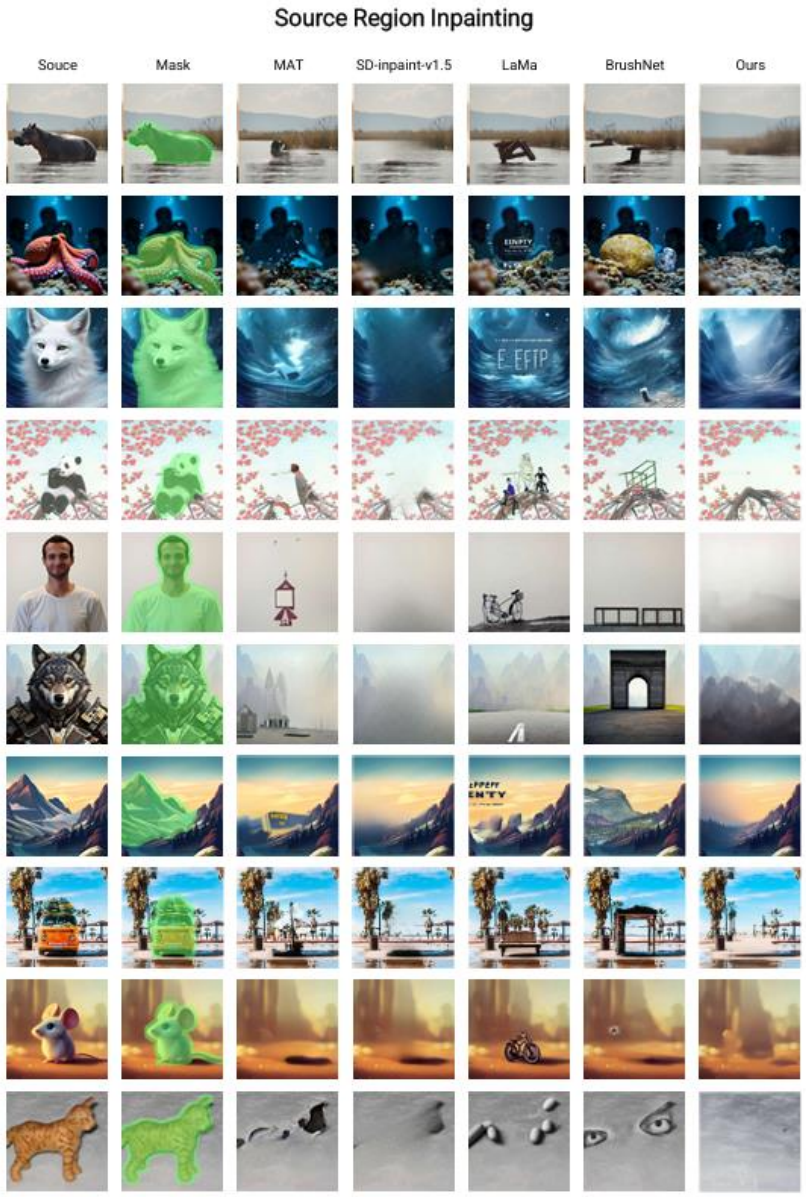


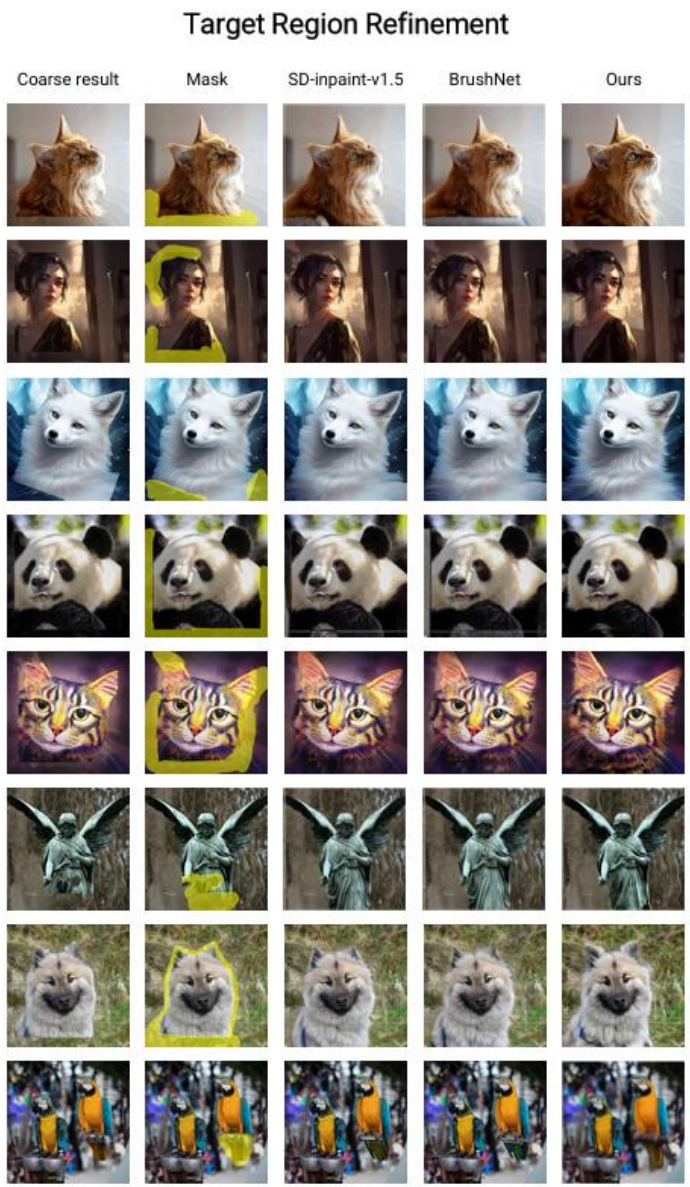
Figure 6. Qualitative comparison with state-of-the-art editing methods in rotation

Appendix



9

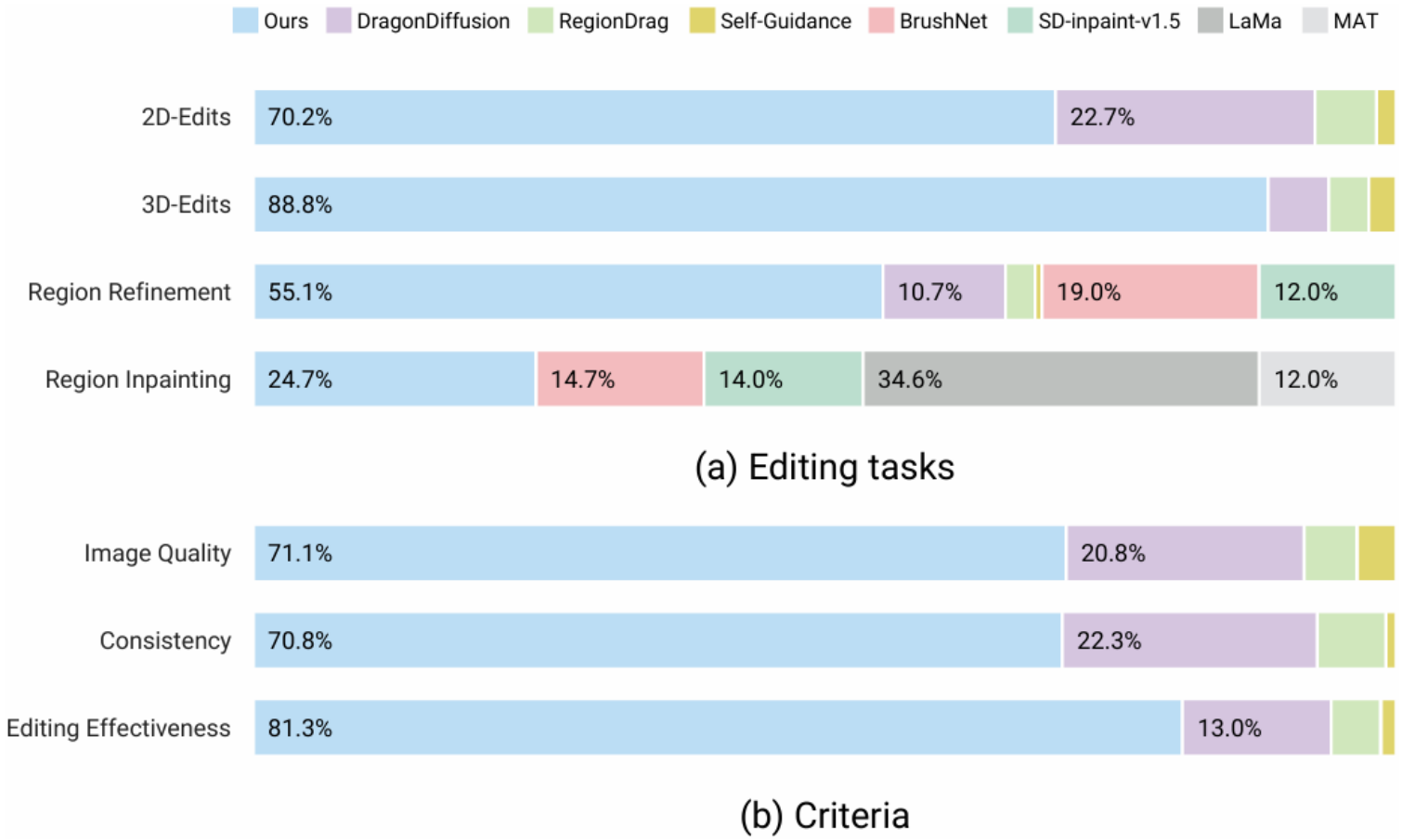
Figure 7. Qualitative comparison with state-of-the-art inpainting methods in source region inpainting



10

Figure 8. Qualitative comparison with state-of-the-art inpainting methods in target region refinement

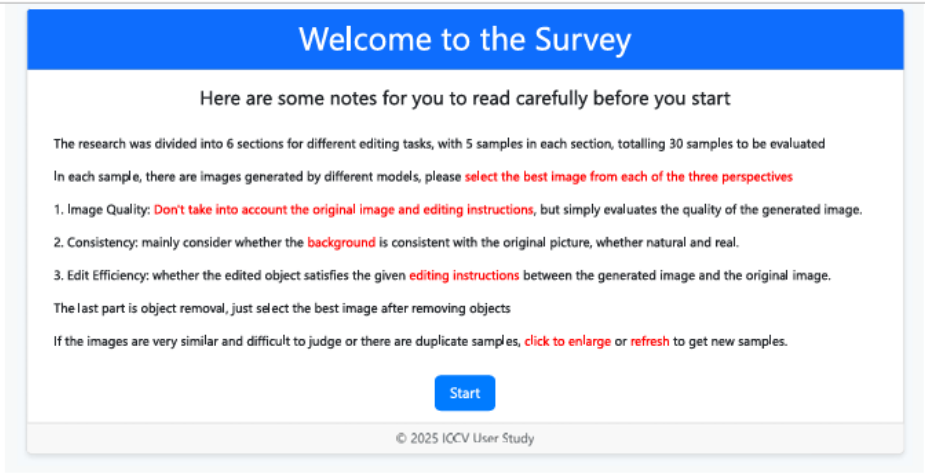
Results



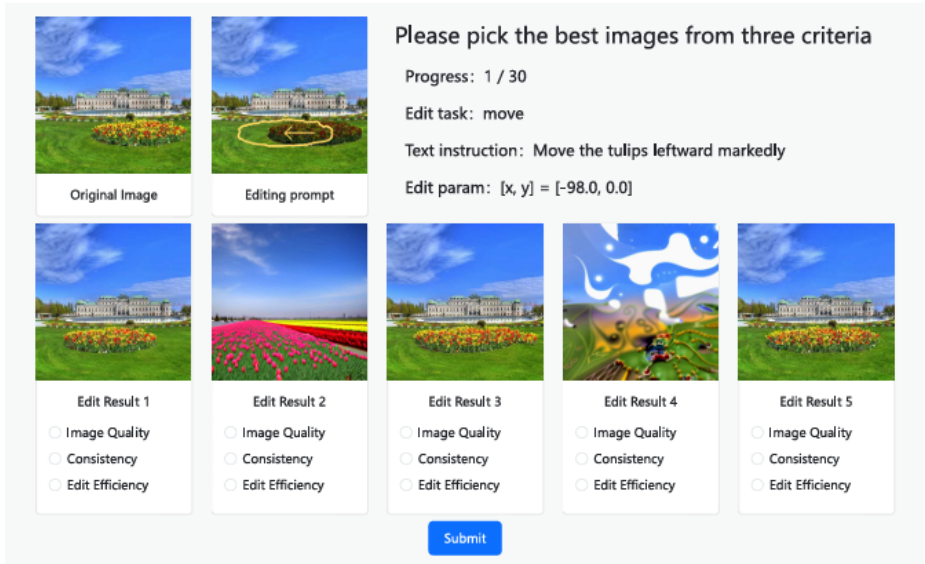
User Study. For a comprehensive quantitative evaluation, we conducted a user study to assess the perceptual quality and editing effectiveness of our method. We recruited 35 participants with diverse backgrounds in computer vision and collected 2,622 valid votes. Each participant was presented with 30 randomly selected editing samples from different tasks (2D-edits, 3D-edits, region refinement and region inpainting). Each sample contained the original im-

Figure 5. Visualization results of the user study. Participants preferred our edited images both in the different editing tasks and from three different criteria.

Results



(a) Home page



(b) Vote page

Figure 1. Screen shots of the website for user study.

Table 1. Voting statistics in the 2D-edits and 3-edits from different criteria, only the editing models are counted.

Method	Image Quality	Consistency	Editing Effectiveness	Total
Ours	**475**	**473**	**543**	**1491**
DragonDiffusion [15]	139	149	87	375
RegionDrag [14]	31	40	29	100
Self-Guidance [4]	23	6	9	38
MotionGuidance [5]	0	0	0	0
Total	668	668	668	2004

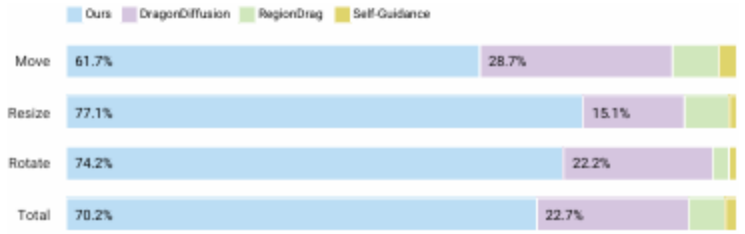


Figure 2. Visualization results of perceptual study in 2D-edits (Move, Rotate, Resize).

Table 2. Voting statistics in different editing tasks. The blank cells indicate that the model was not compared in the task.

Method	2D-Edits				3D-Edits	Region Refinement	Region Inpainting	Total
	Move	Resize	Rotate	Total				
Ours	**374**	**347**	**365**	**1086**	**405**	**258**	37	1786
DragonDiffusion[15]	174	68	109	351	24	50		425
RegionDrag[14]	42	30	12	84	16	12		112
Self-Guidance[4]	16	5	6	27	11	3		41
MotionGuidance[5]	0	0	0	0	0	0		0
BrushNet[9]						89	22	111
SD-inpaint-v1.5[18]						56	21	77
LaMa[23]							**52**	52
MAT[13]							18	18
Total	606	450	492	1548	456	468	150	2622

Ablations

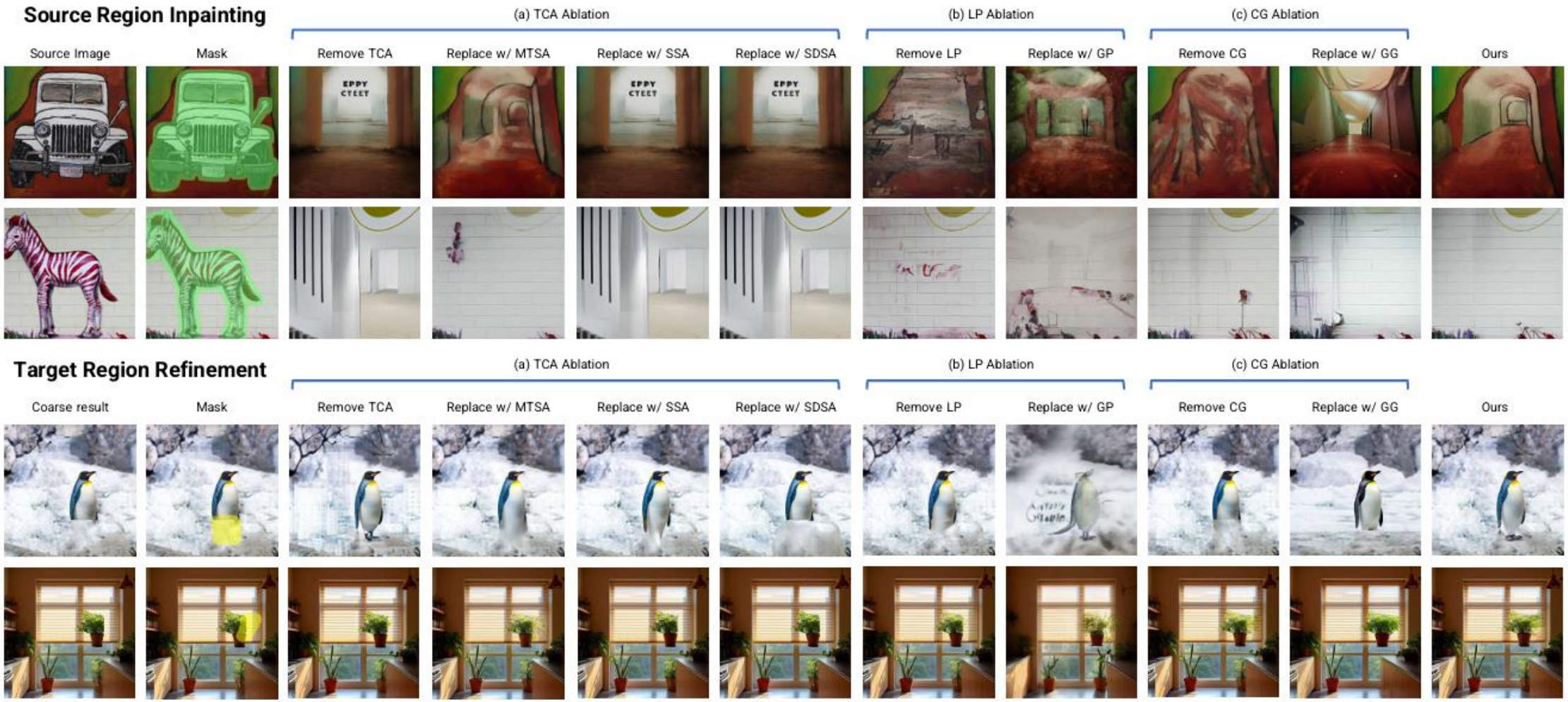


Figure 7. Ablation studies on the impact of removing individual components from **FreeFine** and different internal variations of each component while keeping other techniques applied at the same scale.

Thanks