RAA: Achieving Interactive Remove/Add Anything via Fully Synthetic Data

Delong Liu[1*], Haotian Hou[2,3*], Zhaohui Hou[3*], Shihao Han[3], Zhiyuan Huang[3], Mingjie Zhan[3], Fei Su[1,4,5], Zhicheng Zhao[1,4,5†]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Beihang University
[3]SenseTime
[4]Beijing Key Laboratory of Network System and Network Culture, China
[5]Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism, Beijing, China
{liudelong, zhaozc, sufei}@bupt.edu.cn; hthou@buaa.edu.cn;
{houzhaohui, hanshihao, huangzhiyuan, zhanmingjie}@sensetime.com

Abstract

Precise and controllable image editing, especially object removal and insertion, represents one of the most common demands in image manipulation. However, existing methods suffer from severe limitations. Mask-based inpainting often introduces visual artifacts and semantic inconsistencies, while instruction-based approaches lack accurate spatial control and tend to unintentionally modify background regions. To address these issues, we propose two key contributions. First, we develop a fully automated and self-improving pipeline for synthetic data generation. This pipeline utilizes a Large Language Model (LLM) to generate diverse prompts, a Diffusion Transformer (DiT) fine-tuned evolutionarily to synthesize high-quality images, and a Multimodal LLM (MLLM) combined with open-set object detector for automated quality control and annotation. This process produces the Remove/Add Dataset (RAD), consisting of over 514,510 high-quality image pairs, each richly annotated with bounding boxes, segmentation masks, and a variety of editing instructions. Second, based on RAD, we introduce Remove/Add Anything (RAA), a novel editing framework with precise spatial control. Built upon a diffusion-based inpainting model, RAA achieves high editing accuracy by conditioning on both textual instructions and an explicitly defined region of interest (ROI), enabling efficient fine-tuning while maintaining global visual coherence. Extensive experiments demonstrate that RAA significantly outperforms existing open-source methods on both addition and removal tasks, and even slightly surpasses costly proprietary models.

Code — https://github.com/Delong-liu-bupt/RAA

1 Introduction

Recent advancements in generative models have enabled the synthesis of images with unprecedented detail and realism. While *de novo* image generation achieves remarkable success, precise and controllable image editing remains challenging. A fundamental requirement in this context is the seamless insertion or removal of objects within an existing image.

Current methodologies for object editing, however, exhibit notable limitations. Mask-based inpainting approaches rely on binary masks to guide image modifications. There have been powerful diffusion models such as Stable Diffusion Inpainting (Rombach et al. 2022) and FLUX.1-Fill (Labs 2024). Despite their advancements, these methods are heavily dependent on mask quality and often struggle with accurately reconstructing elements outside the masked region (Song et al. 2023; Suvorov et al. 2022), leading to visual artifacts and semantic inconsistencies that disrupt the natural coherence of the edited image.

Instruction-based editing models, such as InstructPix2Pix (Brooks, Holynski, and Efros 2023), Smartedit (Huang et al. 2024), and ICEdit (Zhang et al. 2025), support natural language commands as a more user-friendly approach. Nevertheless, language ambiguity frequently results in inadequate spatial localization, forcing users into repeated prompt refinements without guaranteeing precise control (Gal et al. 2022). Moreover, these models may inadvertently cause "style bleed", where edits unintentionally affect backgrounds or adjacent elements, thereby compromising overall image fidelity (Cao et al. 2023).

A core bottleneck for robust image editing is the scarcity of high-quality training data, typically structured as {source image, edit instruction, target image} triplets. Although certain studies have established pipelines to extract realistic frames from videos (Sagong et al. 2022; Wei et al. 2025), these approaches are both time-consuming and limited in scope. Consequently, many leading methods resort to synthetic data generation (Zhang et al. 2023; Zhao et al. 2024; Zhang et al. 2024). However, synthesized target images may suffer from low fidelity or unnatural artifacts, and the corresponding edit instructions often misalign with visual changes, introducing significant noise into the training process. Thus, a robust pipeline ensuring synthetic data quality and stability is essential.

To address these challenges, we propose a self-improving data curation pipeline, integrating a Large Language Model (LLM) to automatically generate high-quality prompts, em-

[*]These authors contributed equally.
[†]Corresponding author

Figure 1: Comparative visualization of different image editing methods for object removal and addition. (a) Image editing results produced by our proposed RAA framework. (b) Results from various advanced methods across different categories on the same examples. Although some methods, such as GPT-4o, produce plausible and natural edits, they generally show weaker preservation of other elements in the original image.

ploying the advanced generative image model FLUX.1 (Labs et al. 2025b) to synthesize target image pairs, and incorporating a Multimodal Large Language Model (MLLM) combined with Structural Similarity Index (SSIM) filtering for semantic quality control. Additionally, an open-set object detector generates precise bounding-box annotations. This self-improvement mechanism leverages generated high-quality image pairs for iterative training via Low-Rank Adaptation (LoRA), progressively enhancing the generative model's performance. Using this automated approach, we create Remove/Add Dataset (RAD), a large-scale, meticulously curated dataset containing over **514,510** interactive editing image pairs, each annotated with bounding boxes, segmentation masks, and multiple semantic editing instructions specifically designed for interactive object insertion and removal tasks.

Building upon this dataset, we present Remove/Add Anything (RAA), a spatially aware editing framework. Leveraging the FLUX.1-Fill (Labs 2024) model, RAA innovatively redefines the masking input by introducing regions of interest (ROIs) in the form of Bounding Box (BBox), alongside conventional repaint and preservation areas. By generating the edited image holistically, RAA addresses the visual artifacts and semantic inconsistencies often observed in mask-based methods. The introduction of ROIs enables interactive spatial localization that instruction-based methods lack. As demonstrated in Figure 1, our approach significantly surpasses existing open-source methods across relevant benchmarks and achieves performance comparable to leading proprietary closed-source systems. Our main contributions include four aspects:

- A self-improving synthetic data pipeline that evolutionarily generates and refines high-quality training data without reliance on real-world datasets;
- A large-scale dataset of over 514,510 high-quality image pairs, each annotated with regions of interest and editing instructions.
- A novel spatially-aware editing framework, RAA, enabling interactive object removal and insertion with high parameter efficiency.
- State-of-the-art performance on benchmarks for interactive object insertion and removal.

2 Related Work

2.1 Mask-based Inpainting

Given an explicit binary mask, early methods like DeepFill v2 (Yu et al. 2019) used gated convolutions to synthesize missing regions, but produced blurry textures for large or irregular holes. Transformer variants (Cao, Dong, and Fu 2023; Li et al. 2022) expanded the receptive field with self-attention modules, yielding sharper details. Diffusion inpainters (Rombach et al. 2022; Saharia et al. 2022) treat the known region as a hard constraint during the reverse process, achieving impressive texture harmonization. Despite their progress, undesirable artifacts remain to occur even with small or no mask errors (Song et al. 2023; Suvorov et al. 2022). Toolchains like Inpaint Anything (Yu et al. 2023) and ObjectStitch (Song et al. 2023) automate or refine masks, easing but not eliminating this dependency. For instance, when removing a boat from a lake scene, an automatically generated mask may successfully erase the boat

itself but fail to account for its reflection in the water, resulting in a semantically inconsistent final image.

2.2 Generative Models for Image Editing

Early research on generative adversarial network (GAN)-based editing introduced latent-space manipulation and inversion techniques, enabling semantically meaningful yet resolution-limited edits on real images (Richardson et al. 2021; Tov et al. 2021; Alaluf et al. 2022). A second wave of diffusion-based editors greatly improved fidelity while adding support for natural-language or example-based control. From unconditional inpainting methods (Saharia et al. 2022; Meng et al. 2021) to text-conditioned and exemplar-guided approaches (Couairon et al. 2022; Yang et al. 2023; Avrahami, Lischinski, and Fried 2022). Based on large-scale triplet datasets, recent work has focused on instruction-aligned generators (Huang et al. 2024; Sheynin et al. 2024; Zhang et al. 2025; Liu et al. 2025). InstructPix2Pix utilizes synthetic triplets to train a diffusion network capable of free-form natural-language edits in seconds (Brooks, Holynski, and Efros 2023). UltraEdit enhances its capability with an automatically generated image editing dataset based on real-image editing samples with diverse instructions. (Zhao et al. 2024). While user-friendly, these methods can drift semantically or bleed style to background regions in the absence of explicit spatial constraints (Cao et al. 2023; Brooks, Holynski, and Efros 2023).

3 Method

An overview of our proposed method is illustrated in Figure 2. Our approach comprises a three-stage dataset synthesis pipeline (Figure 2a), designed to automatically generate high-quality diptych editing pairs, and the RAA framework for object removal and insertion (Figure 2b). In the subsequent sections, we detail each component respectively.

3.1 Dataset Synthesis Pipeline

Our method addresses two distinct editing scenarios: **Instruction-based editing**, where users specify via verbal instructions what should be added or removed in an image, and **Region-specific editing**, in which users define a particular region to control the insertion or removal operation. Following this rationale, we construct our dataset $\mathcal{D}$ consisting of tuples:

$$\mathcal{D}_{\text{tgt}} = \left\{ \left(I_-,\, I_+,\, B,\, P_{+/-}\right) \right\},$$

where I_- represents the image without the target object, I_+ represents the image containing the target object, B denotes the target bounding box for ROIs, and $P_{+/-}$ indicates the corresponding add or remove prompt. In the following subsections, we elaborate on the three-stage process for collecting the RAD dataset.

Stage I: Diversified Prompt Generation. We leverage a powerful LLM, Qwen3 (Yang et al. 2025), to synthesize diverse natural-language editing instructions. Initially, we construct a foundational corpus consisting of paired scene descriptions. Given a comprehensive set of human-created object categories $\mathcal{C}_0$, the LLM expands these manually provided categories, forming a richer and more diverse corpus $\mathcal{C}$. Subsequently, we employ the LLM to produce a collection of natural-language instructions:

$$\mathcal{D}_0 = \left\{ prefix,\, (d^{(i)},\, d_+^{(i)}) \right\}_{i=1}^{K},$$

in which a fixed instruction prefix guides the generation in form of a diptych, with $d^{(i)}$ depicting the base scene without the target object, and $d_+^{(i)}$ describing exclusively the added object, drawn from corpus $\mathcal{C}$.

This initial corpus $\mathcal{D}_0$ undergoes two parallel processing steps. First, for each description pair (d, d_+), we extract the central edited subject into concise labels $s \in \mathcal{S}$, formatted for precise localization by grounding models such as Grounded-SAM. Concurrently, for each distilled subject label s, we generate two sets of prompts: $\mathcal{P}_+(s)$ consisting of addition instructions (e.g., "Add a navy-blue shorts") and $\mathcal{P}_-(s)$ consisting of removal instructions (e.g., "Remove the shorts"). During training, editing prompts are randomly sampled from the respective prompt set to supervise object insertion or deletion, ensuring broad linguistic coverage and creative variety.

Stage II: Diptych Pair Synthesis. In this stage, we utilize FLUX.1-dev (Labs et al. 2025b) to synthesize the required diptych images. Empirically, we observe a higher success rate when generating diptychs where the left image contains the base scene and the right image includes the added object, compared to reversing this arrangement. Nevertheless, in most cases, obtaining results that fully satisfy the desired criteria remains challenging, as illustrated by the visual comparisons provided in Figure 3. Consequently, we employ LoRA to fine-tune FLUX.1-dev specifically for our current task scenario, thereby stabilizing the synthesis process and producing high-quality training data. Starting from the initial model $\mathcal{M}^{(0)}$, we evolutionarily enhance its capability to generate diptych-style images. Given the corpus $\mathcal{D}_0$, we prompt model $\mathcal{M}^{(0)}$ to produce synthetic composite diptych images:

$$\widetilde{\mathcal{D}} = \left\{ I_D^{(i)} \right\}_{i=1}^{K}.$$

where each diptych image $I_D^{(i)}$ consists of horizontally concatenated pairs $[I_-^{(i)} | I_+^{(i)}]$. The sub-images $I_-^{(i)}$ and $I_+^{(i)}$ represent the original image and the edited version with the target subject $s^{(i)}$ inserted, respectively. These composite images are then split into separate pairs for subsequent tasks.

To prevent model collapse resulting from training on unfiltered synthetic outputs, we employ a MLLM to implement a multi-criteria evaluation function $\Phi : \widetilde{\mathcal{D}} \to \mathbb{R}$, defined as:

$$\Phi(x) = \mathrm{Acc}(x) + \mathrm{Nat}(x) + \mathrm{Aes}(x) + \mathrm{Ali}(x),$$

where the terms Acc, Nat, Aes, and Ali measure editing accuracy, naturalness, aesthetic quality, and textual alignment, respectively. We then select a high-quality subset $\mathcal{D}_1 = \{x \in \widetilde{\mathcal{D}} : \Phi(x) \geq \tau\}$, and fine-tune the model $\mathcal{M}^{(0)}$ on $\mathcal{D}_1$ to produce an updated model $\mathcal{M}^{(1)}$. By iteratively repeating this generate-filter-fine-tune cycle for T rounds, we progressively enhance both the diptych image synthesis quality and the success rate of images passing the evaluation criteria.

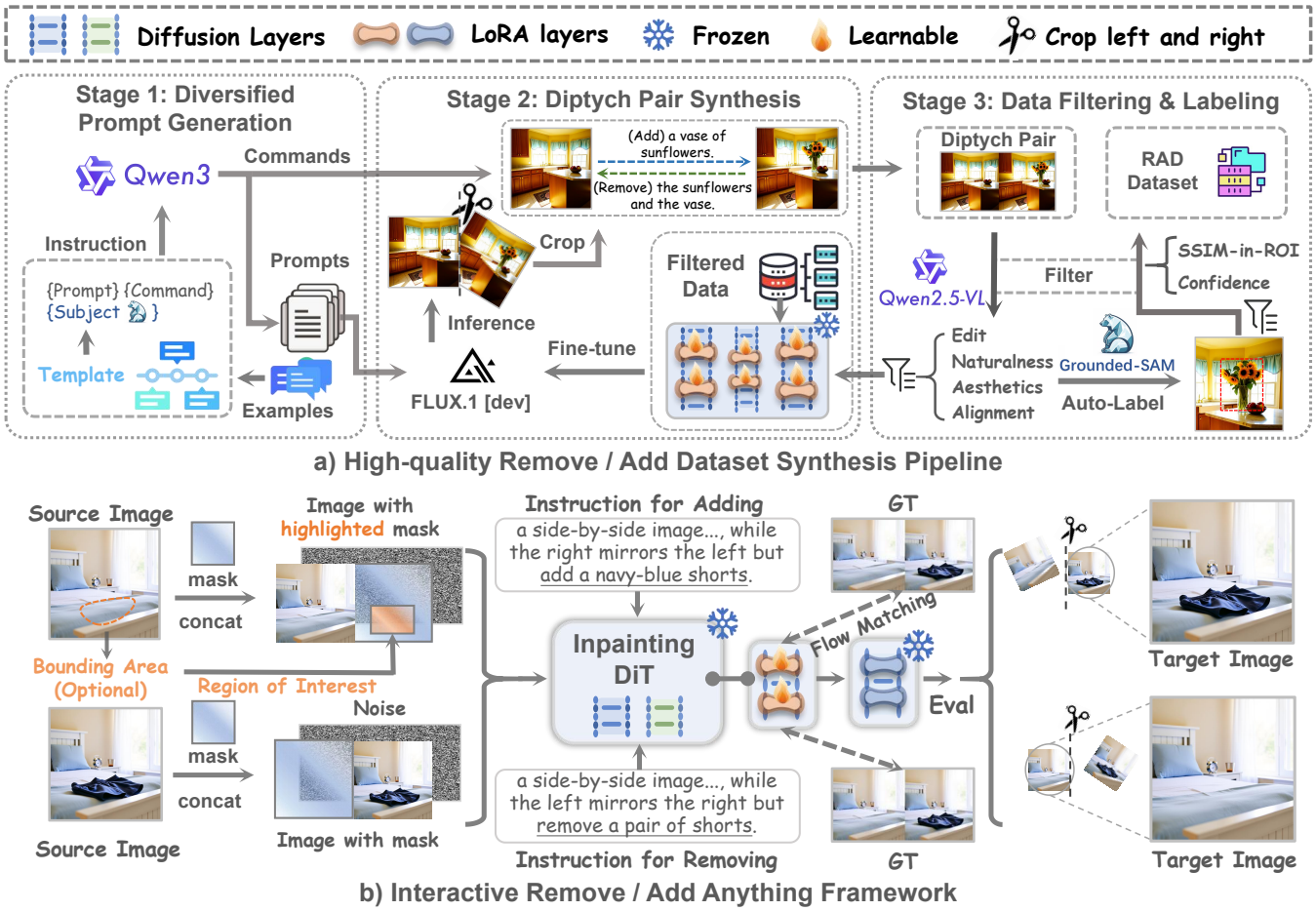

Figure 2: Overview of our proposed method. (a) Pipeline for synthesizing the high-quality Remove/Add Dataset, consisting of three stages: 1) diverse prompt generation via Qwen3, 2) diptych image synthesis with a recursively fine-tuned diffusion model, and 3) automated filtering and labeling using MLLM and Grounded-SAM. (b) Interactive RAA framework: given a source image, an optional ROI, and a text instruction, RAA generates the corresponding target image.

Stage III: Data Filtering and Labeling. In the final stage, we obtain the evaluated and cropped diptych image pairs along with distilled subject labels from Stage II, denoted as $\mathcal{D}_2 = \{(I_-^{(i)}, I_+^{(i)}, s^{(i)})\}_{i=1}^{N}$. We first employ Grounded-SAM (Ren et al. 2024) for automatic subject localization, generating segmentation masks and bounding boxes for each image pair based on the corresponding subject label $s^{(i)}$. Each identified bounding box is accompanied by a confidence score $c^{(i)} \in [0, 1]$.

We then apply a two-step filter to ensure data quality. First, we discard any bounding boxes with confidence below the threshold τ_{conf}. Next, to remove high-confidence boxes that capture unchanged backgrounds, we compute the SSIM score $\alpha^{(i)}$ between each pair's cropped regions and keep only those with $\alpha^{(i)} < \tau_{\text{ssim}}$. The remaining samples form the final RAD, which combines the quality and diversity needed for downstream editing.

3.2 RAA Framework

With the high-quality RAD dataset, we fine-tune a pre-trained inpainting DiT model by concatenating "before" and "after" images into a unified side-by-side input, always placing the image with inserted object (I_+) on the right. Formally, each input becomes

$$I_D^{(i)} = \left[I_-^{(i)} \mid I_+^{(i)}\right] \in \mathbb{R}^{H \times 2W}.$$

Next, we generate a binary mask $M \in \mathbb{R}^{H \times W}$ by rasterizing the target bounding box B, and subsequently encode a highlight map H within this masked region using a distinctive pixel-value template.

During training, we employ FLUX.1-Fill (Labs et al. 2025b) as our base model. For object-addition, we mask the right panel and feed the concatenated input into the fill module; for object-removal, we instead mask the left panel. In both cases, the unmasked half provides contextual guidance. When a bounding box B is provided as input, the newly encoded highlight map H does not alter the repaint region of the image; instead, it serves as additional mask informa-

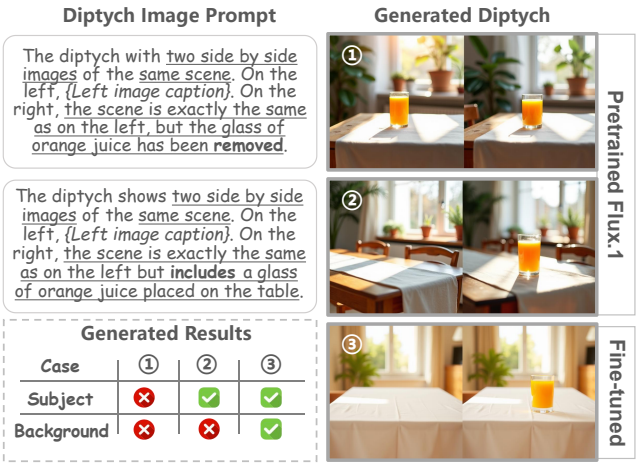

Figure 3: Generated diptychs using different prompts and models (pre-trained and fine-tuned), along with corresponding evaluations of the results.

Figure 4: Visualization samples from RAD. Each image pair is accompanied by three prompt variants (short to long) and corresponding editing bounding boxes.

tion specifying the ROI. To promote mutual improvement between object addition and removal tasks, we use identical ground-truth images for both tasks, differing only in the side of the diptych panel masked during training. To achieve precise and efficient editing without modifying the entire transformer, we introduce LoRA layers into each attention block. Training is guided by a flow-matching objective function, and only these lightweight LoRA parameters are fine-tuned, yielding a computationally efficient model capable of producing highly localized and aesthetically coherent edits.

At inference, we maintain the same diptych arrangement, masking either the left or right panel appropriately to ensure consistency with the training distribution. This design enables precise, interactive, and instruction-guided image editing.

4 Experiments

4.1 Remove/Add Dataset

Our model is trained on the RAD, a large-scale, fully synthetic dataset constructed using the pipeline described in Section 3.1. Through a rigorous multi-stage synthesis and filtering process, we obtain a comprehensive collection comprising **514,510** high-quality editing samples. Each RAD sample is structured as a tuple $\mathcal{D}_{\text{tgt}} = \{(I_-, I_+, B, P_{+/-})\}$ to facilitate diverse editing tasks. The diptych images (I_-, I_+) are synthesized by our evolutionarily fine-tuned DiT model, where I_- represents the base image and I_+ includes the target object. The textual editing instructions $(P_{+/-})$ for object addition and removal are generated by LLM based on diptych prompts. To ensure linguistic robustness, we generate three types of instructions per action, varying in length and complexity, as illustrated in Figure 4. Each edited object is precisely annotated with a bounding box B, automatically produced by Grounded-SAM (Ren et al. 2024) and refined via confidence-score filtering and SSIM validation. The corresponding object segmentation is also retained as auxiliary information. This spatial annotation enables our RAA model to perform edits based on bounding boxes, text prompts, or a combination of both. Being fully synthetic, RAD encompasses a broad variety of objects, scenes, and styles that are difficult to obtain from real-world data, while fully avoiding privacy and copyright concerns. Consequently, RAD provides a strong and versatile foundation for training our RAA editing framework.

4.2 Experimental Settings

Evaluation Benchmarks and Metrics. We select the widely used instruction-based editing test sets MagicBrush (Zhang et al. 2023) and Emu Edit (Sheynin et al. 2024), specifically filtering samples involving object addition and removal tasks to serve as benchmarks for evaluating model capabilities. For the MagicBrush benchmark, we directly compare the generated images with the ground truth edited images (GT), measuring pixel-level differences using the L1 metric, and assessing global image preservation using CLIP-I (Radford et al. 2021) and DINO (Oquab et al. 2023). For Emu Edit, due to the absence of ground truth images, we follow previous approaches (Zhao et al. 2024; Sheynin et al. 2024) by evaluating content preservation through computing the L1, CLIP-I, and DINO scores between generated and source images. Emu Edit provides captions describing the desired edited outcomes; thus, we introduce an additional text-image semantic similarity metric, CLIP-T, to measure editing success, denoted as CLIP-Out, reflecting the alignment between the edited image and the provided caption. However, all aforementioned metrics suffer from the issue that high scores can be obtained even if the image is not edited at all. Therefore, we additionally introduce GPT-4o to rigorously evaluate the final edited images for both editing success and naturalness, with the corresponding pass rate recorded as the GPT score.

Figure 5: Qualitative comparison on challenging samples from Emu Edit and MagicBrush test sets. Markers at the top-right corner indicate editing quality: green (satisfactory), blue (acceptable), red (failed). Artifacts or poor consistency regions are highlighted with red boxes. Bounding boxes (BBox) added by RAA are indicated by orange dashed rectangles.

	Method	CLIP-I ↑	DINO ↑	L1 ↓	GPT ↑
Instruction	InstructPix2Pix	0.855	0.804	0.094	0.060
	MagicBrush	0.910	0.900	0.061	0.124
	UltraEdit	0.914	0.901	0.056	0.341
	FluxEdit	0.893	0.863	0.067	0.112
	ICEdit	0.924	0.918	0.057	0.393
	Flux-Kontext	0.931	0.935	0.045	0.682
Mask	UltraEdit	0.947	0.954	**0.035**	0.423
	Flux-Fill	**0.949**	**0.957**	0.042	0.498
Ours	**RAA**	0.931	0.939	0.049	0.701
	RAA (+ BBox)	0.942	0.950	0.041	**0.794**

Table 1: Performance on the MagicBrush benchmark. Best results are in **bold**, second-best results are underlined

Method	CLIP-I ↑	DINO ↑	L1 ↓	CLIP-out ↑	GPT↑
InstructPix2Pix	0.856	0.809	0.097	0.266	0.108
MagicBrush	0.925	0.914	0.063	0.272	0.254
UltraEdit	0.923	0.920	0.042	0.278	0.273
FluxEdit	0.875	0.843	0.070	0.259	0.177
ICEdit	0.931	0.936	0.046	0.276	0.526
Flux-Kontext	0.938	0.945	0.040	0.283	0.621
RAA (Ours)	**0.940**	**0.948**	**0.039**	**0.291**	**0.665**

Table 2: Performance on the Emu Edit benchmark.

4.3 Quantitative Results

Effectiveness of RAA. We conduct a comparative evaluation of our RAA framework against several advanced editing models (Zhang et al. 2025; Brooks, Holynski, and Efros 2023; Zhang et al. 2023; Zhao et al. 2024; Labs et al. 2025b) on the MagicBrush (Zhang et al. 2023) and Emu Edit (Sheynin et al. 2024) benchmarks. As shown in Tables 1 and 2, RAA achieves the best performance using only instruction-based editing, even slightly outperforming FLUX.1-Kontext (Labs et al. 2025a), which is trained with extensive data and computational resources. On MagicBrush, which provides masks for editing regions, we feed the bounding rectangles of these masks as additional input to RAA, further enhancing its performance and demonstrating the effectiveness of our ROI-based training strategy. For methods that support mask input, they do not apply edits beyond the masked region, thereby achieving the highest performance in metrics related to preserving unchanged image areas. However, this also leads to possible artifacts or inconsistencies in the final results, causing these methods to fall behind in terms of editing success rate as judged by GPT-based evaluation. On Emu Edit, although ROI annotations are absent, RAA still achieves the best performance in both original image detail preservation and editing success rate with instruction-only input.

Iteration	Pass-Rate ↑	CLIP-T ↑	GPT ↑	$\overline{\Phi}(x)$ ↑
0	0.056	0.289	0.093	28.94
1	0.526	0.302	0.643	32.17
2	0.699	0.313	0.765	32.75
3	**0.773**	**0.318**	**0.829**	**32.92**

Table 3: Image quality metrics at each iteration of the multi-round self-improving strategy. Pass-Rate denotes the percentage of samples passing the MLLM-based scoring criteria at each iteration, while $\overline{\Phi}(x)$ indicates the average score.

Effectiveness of the Self-Improving Strategy. To verify the effectiveness of the self-improving generation model used during the construction process of RAD, we evaluate the quality of generated images at each iteration, as shown

Method	CLIP-I ↑	DINO ↑	L1 ↓	GPT ↑
RAA (w/o BBox)	0.931	0.939	0.049	0.701
RAA (Cross-Attention)	0.935	0.945	0.047	0.734
RAA (LoRA rank=64)	<u>0.939</u>	<u>0.948</u>	**0.040**	<u>0.775</u>
RAA (LoRA rank=32)*	**0.942**	**0.950**	<u>0.041</u>	**0.794**

Table 4: Ablation results on the MagicBrush benchmark.

in Table 3. The results indicate that after each round of self-refinement, the alignment between the generated image pairs and their corresponding texts, the average scores from the MLLM, and the filtering pass rates all exhibit significant improvements. As iterations continue, the magnitude of these improvements gradually decreases. After three iterations, the vast majority of images pass the filtering criteria, and therefore further iterations are unnecessary. By comparing the GPT scores and pass rate metrics, we find a consistent trend between these two evaluations. Moreover, our filtering criteria based on MLLM are relatively stricter, confirming the effectiveness of our filtering approach.

4.4 Qualitative Results

Qualitative comparisons with related methods on challenging samples are presented in Figure 5. Under solely semantic guidance, the images edited by RAA not only maintain semantic correctness but also seamlessly blend with the lighting, perspective, and style of the scene, achieving results competitive with state-of-the-art models such as GPT-4o and FLUX.1-Kontext and significantly surpassing the editing outcomes of other models. As shown in the fourth row example, RAA perfectly removes the trash bin and seamlessly reconstructs the previously occluded area with contextually consistent content. Compared to the satisfactory results of FLUX.1-Kontext, it is evident that the grass restored by RAA appears more natural. Additionally, by introducing the ROI region through BBox, RAA further enhances editing flexibility, generating outcomes more aligned with user expectations. The incorporation of such reasonable prior knowledge significantly improves editing accuracy. For instance, in the third row of the figure, RAA (BBox) achieves the only satisfactory result; without the introduced BBox, the addition of the lamp may be inaccurately positioned.

4.5 Ablation Studies

Effectiveness of ROI Training Strategy. The key to enabling interactive editing is incorporating the ROI as a conditioning factor during model training. To demonstrate the effectiveness of the ROI training strategy within our proposed RAA framework, we compare it with a commonly used conditioning method, where the BBox information of the ROI region is projected into an embedding space through a simple learnable linear layer and subsequently fed into the DiT model along with the textual prompt for cross-attention operations. Performance results for this baseline are presented in Table 4, illustrating that the cross-attention approach can indeed leverage ROI information to enhance editing outcomes, although the performance improvement is

Figure 6: Editing results of different methods on images at varying scales. Red boxes indicate the input ROI regions.

notably inferior to that achieved by RAA. Moreover, we observe empirically that the cross-attention method tends only to bias the editing result toward the ROI direction without achieving precise spatial control. Additionally, it entirely lacks size generalization capabilities, as illustrated in Figure 6. After training on RAD images of uniform square size, our proposed RAA method demonstrates adaptability to interactive edits at various image scales, whereas the cross-attention method fails to retain spatial control. We also attempt training RAA using LoRA with a higher rank of 64; however, this increase in computational cost does not consistently improve performance, prompting us to select a LoRA rank of 32.

5 Conclusion

In this work, we address the longstanding challenge in generative image editing of simultaneously achieving precise spatial control and natural image quality. Our core contribution is the development of a scalable, automated pipeline that generates the RAD, a fully synthetic dataset containing over 514,510 pairs of high-quality images. Leveraging this robust dataset, our proposed RAA framework introduces a novel ROI conditioning mechanism, which provides flexible spatial control while ensuring high-fidelity image generation. Extensive experiments demonstrate that RAA significantly outperforms existing open-source methods and achieves competitive performance compared to proprietary commercial models. Ultimately, our research confirms that purely synthetic data pipelines can effectively train robust editing models. Future work could explore how to bridge the subtle domain gap between synthetic datasets and complex real-world styles, thereby further enhancing the model's ability to generate more realistic details.

Acknowledgments

This work was supported by the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing (GJJ-24-021) and the BUPT Innovation and Entrepreneurship Support Program (2025-YC-T043).

References

Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; and Bermano, A. 2022. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, 18511–18521.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Cao, C.; Dong, Q.; and Fu, Y. 2023. Zits++: Image inpainting by improving the incremental transformer on structural priors. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12667–12684.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22560–22570.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Huang, Y.; Xie, L.; Wang, X.; Yuan, Z.; Cun, X.; Ge, Y.; Zhou, J.; Dong, C.; Huang, R.; Zhang, R.; et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8362–8371.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux. Accessed: May 2025.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025a. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025b. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768.

Liu, D.; Lei, C.; Jiang, S.; Zhao, Z.; and Su, F. 2025. CE-LoRA: Consistent Person Synthesis by Exploring the Model's Spatial Consistency. In *2025 IEEE International Conference on Multimedia and Expo ICME*, 1–6. IEEE;. ISBN 979-8-3315-9495-4.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sagong, M.-C.; Yeo, Y.-J.; Jung, S.-W.; and Ko, S.-J. 2022. RORD: A Real-world Object Removal Dataset. In *BMVC*, 542.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Sheynin, S.; Polyak, A.; Singer, U.; Kirstain, Y.; Zohar, A.; Ashual, O.; Parikh, D.; and Taigman, Y. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8871–8879.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18310–18319.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14.

Wei, R.; Yin, Z.; Zhang, S.; Zhou, L.; Wang, X.; Ban, C.; Cao, T.; Sun, H.; He, Z.; Liang, K.; and Ma, Z. 2025. OmniEraser: Remove Objects and Their Effects in Images with Paired Video-Frame Data. arXiv:2501.07397.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18381–18391.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36: 31428–31449.

Zhang, S.; Yang, X.; Feng, Y.; Qin, C.; Chen, C.-C.; Yu, N.; Chen, Z.; Wang, H.; Savarese, S.; Ermon, S.; et al. 2024. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9026–9036.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*.

Zhao, H.; Ma, X. S.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37: 3058–3093.