RCEdit-500K: Reference Completion for Image-Conditioned Image Editing

Jingxu Zhang[1,2], Daneul Kim[3], Yueming Pan[2,4], Dong Chen[2], Kai Qiu[2], Yang Liu[2], Yifan Yang[2], Qi Dai[2], Xiaoyan Sun[1*], and Chong Luo[1,2*]

[1] University of Science and Technology of China, Hefei, China
zjxjz@mail.ustc.edu.cn, sunxiaoyan@ustc.edu.cn
[2] Microsoft Research Asia, Beijing, China
{doch, Kai.Qiu, yangliu, yifanyang, Qi.Dai, chong.luo}@microsoft.com
[3] Seoul National University, Seoul, South Korea
carpedkm@snu.ac.kr
[4] Xi'an Jiaotong University, Xi'an, China
pym13474072916@stu.xjtu.edu.cn

Abstract. Image-conditioned image editing (ICIE) guides edits with a reference image to convey visual attributes—such as style, tone, or object identity—that are difficult to specify through text alone. Despite growing practical demand, no large-scale unified ICIE dataset currently exists in the open-source ecosystem; existing efforts cover only a narrow subset of edit types at small scale, typically constructed via costly forward synthesis. We reformulate ICIE data construction as a *reference-completion* problem: high-quality text-conditioned image editing (TCIE) datasets already supply the input image, instruction, and edited target, and can be augmented with aligned reference images through type-specific synthesis and lightweight instruction adaptation. Building on this insight, we propose a scalable pipeline equipped with weak-instruction augmentation and five-dimensional VLM-based post-filtering, and use it to construct **RCEdit-500K**—the first large-scale unified open ICIE dataset comprising 477K quadruplets across six edit categories (add, remove, replace, background, style, alter) with both concrete and abstract reference types. Training on RCEdit-500K consistently improves reference-guided editing: LoRA adaptation on diffusion models yields up to +1.22 average gain, and an autoregressive model without native editing ability acquires competitive ICIE performance, demonstrating that data availability is the primary bottleneck for open-source ICIE.

Keywords: data pipeline · dataset · image-conditioned image editing

1 Introduction

Text-conditioned image editing (TCIE) enables image manipulation through natural-language instructions and has become a widely adopted paradigm. [10,

* Corresponding authors

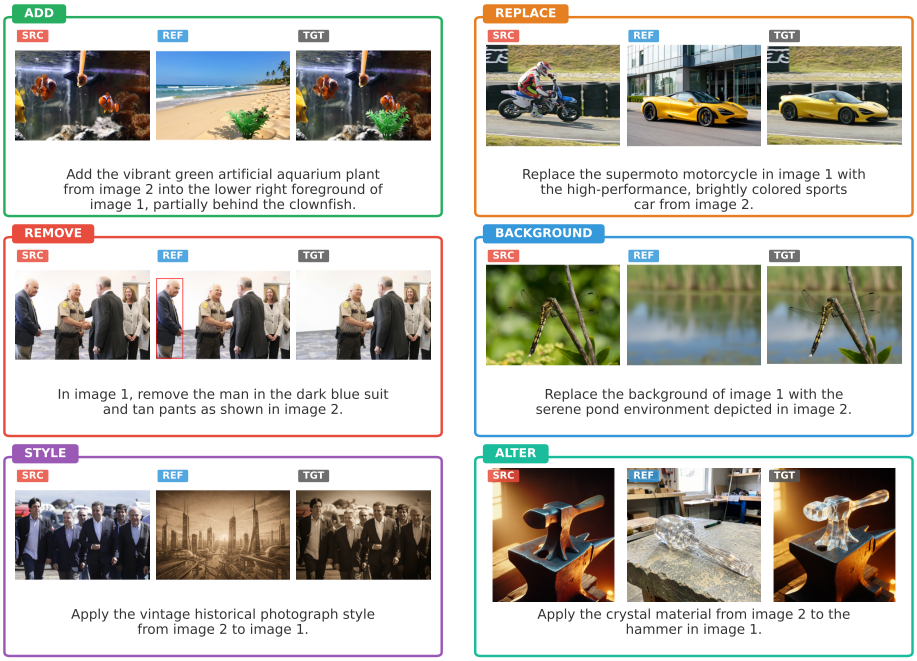

Fig. 1: Examples of six image-conditioned image editing (ICIE) types from RCEdit-500K. Each column shows an input image (left) paired with a reference image (right) and the corresponding edited output. Our dataset enables diverse editing capabilities including object addition, replacement, removal, attribute alteration, style transfer, and background replacement.

11, 30, 36] However, text alone provides a limited channel for specifying many real-world editing objectives. Global visual attributes — such as photographic tone, lighting distributions, material appearance, and stylistic aesthetics — are difficult to describe precisely in language and are often communicated through visual references. This motivates image-conditioned image editing (ICIE), where edits are guided jointly by an input image, a reference image, and a textual instruction.

Despite its practical relevance and demonstrated capabilities in closed-source systems, ICIE remains comparatively underdeveloped in the open-source ecosystem, **with no open-source unified ICIE dataset available**. Existing public datasets [1, 19, 27, 36, 37] and models [10, 11, 30] largely focus on TCIE or on multi-image conditioning for subject-identity preservation [17, 31], while reference-guided editing for distribution-level visual objectives — such as style alignment, tone harmonization, or aesthetic consistency — remains less well supported. As a result, open-source ICIE systems often struggle on these tasks.

To understand this gap, it is important to examine how ICIE training data is typically constructed. A common approach is forward synthesis [34, 36, 37]:

jointly generating input images, reference images, editing instructions, and target outputs using generative models, with at most one image sampled from real data. In practice, this process is expensive and introduces several problems, including model bias, distributional artifacts, and degradation of semantic and aesthetic realism. Crucially, these effects undermine the very qualities — diversity, realism, and data quality — that curated editing datasets aim to preserve.

In contrast, existing high-quality TCIE datasets [19,27,36,37] already provide three of the four elements required for ICIE: an input image, an editing instruction, and a target edited image. The missing component is a compatible reference image. This observation reframes ICIE dataset construction as a reference-completion problem rather than a forward-synthesis task: instead of generating entire editing instances from scratch, one can complete existing TCIE triplets by synthesizing aligned reference images while preserving the original input–target relationship and the edit distribution.

Based on this insight, we introduce a scalable and efficient reference-completion pipeline that converts text-guided image editing datasets into image-conditioned ones. The pipeline performs minimal instruction adaptation and synthesizes reference images that are consistent with the target edit, avoiding joint forward generation and retaining the semantic coverage, aesthetic quality, and realistic edit distributions of the source TCIE corpora. We apply a multi-aspect post-filtering strategy to further enhance the data quality.

Using this pipeline, we construct RCEdit-500K, the first large-scale unified open dataset for image-conditioned image editing containing approximately 500,000 aligned quadruplets: input image, reference image, editing instruction, and target image. The dataset preserves the diversity and realism of curated TCIE sources while enabling structured reference conditioning at scale. **Both the dataset and pipeline will be open-sourced to foster community advancement.**

We validate the effectiveness of RCEdit-500K across multiple model families. With lightweight LoRA adaptation, diffusion-based editors consistently improve on reference-guided editing tasks. Moreover, an autoregressive image-generation model — without explicit editing mechanisms — acquires competitive reference-guided editing behavior when trained on the proposed dataset. These results indicate that training-data construction plays a central role for ICIE and that reference completion provides an effective, scalable foundation for advancing open-source reference-guided image editing. Our contributions are summarized as follows:

- We introduce RCEdit-500K, the first large-scale unified open-source dataset dedicated to image-conditioned image editing, comprising 500K high-quality quadruplets.
- We propose a scalable and efficient data conversion pipeline that transforms existing text-driven image editing data into image-conditioned image editing data, preserving quality strengths from the source corpora while substantially reducing synthetic overhead.

– We provide comprehensive empirical validation across both diffusion-based and autoregressive model families, demonstrating consistent ICIE improvements with RCEdit-500K.

2 Related Work

2.1 Text-Conditioned Image Editing

Text-conditioned image editing (TCIE) has progressed rapidly, with both closed-source models [5,9,24] and open-source models [10,11,30] achieving strong editing quality. Correspondingly, TCIE datasets have grown in scale and diversity, while maintaining high quality. Early datasets [6,38] rely on human annotation to ensure quality but remain small in scale. To obtain paired data at larger scale, tool-based automatic pipelines have become the dominant construction paradigm [36,37,40], often combined with VLM-based post-filtering to mitigate errors introduced by intermediate models [36]. More recently, pipelines that leverage closed-source editing models for data synthesis [1,19,27] have shown that higher-precision data can yield competitive performance with considerably smaller data scale.

2.2 Multi-Image Combination

Multi-image combination, also referred to as multi-image personalization, has attracted growing attention, in part driven by the strong capabilities demonstrated by Nano Banana [5]. A number of recent methods [4, 29, 31, 33] extend single-subject personalization to multi-subject settings. However, these approaches primarily focus on combining concrete subjects and lack support for abstract visual attributes such as style or tone. USO [32] and DreamO [17] address this limitation by generating training pairs that share the same style via style-transfer models. DreamOmni2 [34] further broadens multi-image combination to cover both concrete objects and abstract attributes across a wide distribution. While these methods share the multi-image conditioning paradigm with ICIE, they target image generation rather than editing: the input image is not an explicit anchor whose structure must be preserved.

2.3 Image-Conditioned Image Editing

The need to specify editing operations with precise visual details motivates image-conditioned image editing (ICIE). Several TCIE datasets include a "visual editing" subset as part of their coverage, which shares the same concept with ICIE. ImgEdit [36] produces visual editing data via forward synthesis using an in-context editing model. AnyEdit [37] primarily focuses on structural conditions via ControlNet [39], with only a small portion dedicated to visual reference editing through trained identity and detail extractors. In both cases the ICIE portion remains small ($\sim$10K samples) and restricted to concrete-object addition/replacement or material transfer. DreamOmni2 [34] extends this by training

a concept-extraction model that handles both concrete objects and abstract attributes, enabling forward synthesis of ICIE data across more edit types and concept distributions.

Although several tasks related to ICIE—such as virtual try-on [7,16,26], image structure control [3,25], and interleaved generation [13,31,35]—have individually seen notable progress, each addresses only a specific subtask of ICIE. None provides a unified framework that covers the full spectrum of reference-guided editing types defined in our work. In contrast, this paper focuses on constructing a large-scale dataset that spans all six ICIE edit categories, enabling unified training for reference-guided editing.

3 Dataset

As discussed in Sec. 1, the absence of large-scale unified training data is a key bottleneck for open-source ICIE. In this section, we address this gap by constructing **RCEdit-500K**, shown in Fig. 1. We begin by formalizing the ICIE task and defining a taxonomy of edit types (Sec. 3.1). We then describe our reference-completion pipeline, which converts existing TCIE triplets into ICIE quadruplets by synthesizing aligned reference images and adapting instructions (Sec. 3.2). Finally, we report dataset statistics (Sec. 3.3).

3.1 Task definition

We formulate **image-conditioned image editing** (ICIE) as the task of producing a target image that applies a desired edit to a given **input** (anchor) image under guidance from both a natural-language instruction and a separate **reference** image. Concretely, each ICIE example is represented as a quadruplet

$$\langle \texttt{input\_image}, \texttt{reference\_image}, \texttt{instruction}, \texttt{target\_image} \rangle,$$

where the `input_image` is the anchor whose structural content and unedited regions should be preserved, the `reference_image` supplies visual guidance (a concrete object such as a red handbag or a golden retriever, or an abstract attribute such as style or tone), the `instruction` describes the desired edit in natural language, and the `target_image` is the desired edited output.

ICIE differs from common multi-image generation or multi-reference datasets in that the images do not share equal semantic priority: the input image is explicitly designated as the anchor and must retain the majority of its original structure (including spatial arrangement and any regions outside the edit), while the reference image functions purely as a conditioning signal.

Following prior work [36], we categorize ICIE edits into six types: **add**, **remove**, **replace**, **background**, **style**, and **alter**. We adopt this taxonomy because it cleanly separates semantic edits (add/replace/remove) from distributional or appearance-level edits (background/style/alter), which is useful when designing evaluation protocols and conditioning mechanisms. We intentionally

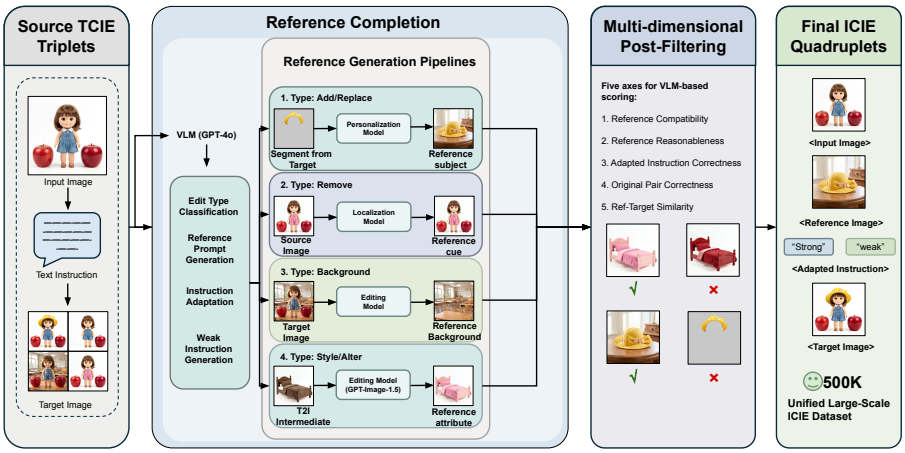

Fig. 2: Data curation pipeline to construct RCEdit-500K.

exclude a separate **motion** category from our ICIE definition since it is hard to describe with single reference (see supplementary for taxonomy details).

The **remove** category merits special attention. In scenes containing many instances of the same object class, purely textual instructions (e.g., "remove the person") are often ambiguous. For ICIE we therefore recommend supplementing the instruction with an explicit spatial cue derived from the input image. Acceptable forms of such cues include a binary mask, a tight bounding box, or a cropped exemplar of the object to be removed; each of these options disambiguates which instance should be edited while remaining compatible with image-conditioned modeling. For example, a removal request may be specified as "remove the object indicated by the bounding box" or by providing a cropped patch of the object to be excised.

3.2 Data Curation Pipeline

Source Data Preparation. To the best of our knowledge, no publicly available unified dataset currently exists for ICIE, making it essential to construct one. A key observation is that existing high-quality TCIE datasets already contain three of the four elements required for ICIE: ⟨`input_image`, `instruction`, `target_image`⟩; only the reference image is missing. As discussed in Sec. 3.1, every ICIE edit type corresponds to a valid TCIE edit, meaning that ICIE dataset construction can be naturally cast as a **reference-completion** problem rather than a full forward-synthesis task.

We select source data from two large-scale, high-quality TCIE corpora: the complete single-turn split of Pico-Banana-400K [19] and a 500K-sample subset of GPT-IMAGE-EDIT-1.5M [27]. Together, these two sources provide broad

coverage of semantic and appearance-level edits while preserving high-quality edit precision.

Reference Completion. The data construction pipeline is shown in Fig. 2. For each source triplet, we first employ a vision-language model (VLM; specifically GPT-4o [9]) to (i) classify the edit into one of the six defined types, (ii) generate several prompts for reference-image synthesis, and (iii) rewrite the original text instruction so that it explicitly references the supplementary image. Triplets whose edits fall outside the defined taxonomy (*e.g.* motion editing) are discarded at this stage.

In addition to the adapted instruction, we ask the VLM to produce a **weak** textual instruction in which the reference-specific description is deliberately abstracted. For example, "*Add the cute orange cat from image 2 to image 1*" is weakened to "*Add the animal from image 2 to image 1*," removing fine-grained appearance details that could be resolved from text alone. By mixing weak instructions into training, we encourage the model to attend more heavily to the reference image as a visual conditioning signal, rather than relying solely on its pre-existing text-conditioned editing capability.

We then design four dedicated pipelines to synthesize reference images according to edit type:

1. **Add / Replace.** The target subject is extracted from the target image using a segmentation model [14, 21, 22]. A reference image is then generated either by re-synthesizing the subject via an image personalization model [10] or by compositing the segmented subject onto a novel background.
2. **Remove.** The subject to be removed is localized in the input image with the same segmentation model [14, 21, 22]. A reference cue is produced by overlaying a tight bounding box on the input image or by cropping the target region, providing an unambiguous spatial indicator of which instance to remove.
3. **Background.** A TCIE model is applied to inpaint all foreground objects while retaining the desired background, yielding a clean background reference that captures the target scene layout and appearance.
4. **Style / Alter.** An intermediate image is first generated by a text-to-image (T2I) model from a descriptive prompt. This intermediate is then transformed to exhibit the required abstract attribute (*e.g.* artistic style, color palette, texture) using GPT-Image-1.5.

We use Grounded-SAM-2 [14, 21, 22] for localization and segmentation, and Flux-Klein-9B [10] for personalization, T2I generation, and TCIE. For style and alter edits, open-source models do not yet achieve sufficient quality for abstract attribute transfer; we therefore employ GPT-Image-1.5 for these categories to ensure data fidelity. **The choices of our toolkit enables data harvesting with low-capacity machines like V100.** Full VLM prompts and synthesis details are in the supplementary.

Post-Filtering. After reference completion, we apply a multi-dimensional quality filter to remove low-quality samples. Prior image edit data-construction pipelines [36, 37] typically rely on CLIP [20] similarity or a holistic VLM score,

Table 1: Post-filter validation. Cohen's κ: agreement beyond chance (1=perfect, 0=random).

Setting	**Comparison**	Agreement↑	κ↑
Cross-VLM (N=600)	GPT-4o vs. Claude 4.6	72.9%	0.46
	GPT-4o vs. Gemini 3.1	78.6%	0.57
Human (10 ann., N=60)	Mean human–human	73.1%	0.38
	Mean GPT-4o–human	72.2%	0.44
GPT-4o Precision / Recall / F1		96.7 / 76.3 / 85.3%	

which conflates multiple quality axes and leads to incomplete quality assessment. Similarly, existing benchmark evaluation protocols [18] focus primarily on the fidelity of the target image, without verifying consistency across all four elements of the quadruplet.

To address these limitations, we propose a **five-dimensional VLM-based filtering** strategy that independently scores each candidate quadruplet on: (i) **reference compatibility for original TCIE pair**, (ii) **reference image reasonableness**, (iii) **adapted instruction correctness**, (iv) **original pair correctness**, and (v) **reference–target similarity**. Detailed scoring prompts and threshold selections are provided in the supplementary material.

Each dimension captures a distinct failure mode in reference-conditioned editing. A sample is retained only if **every** dimension exceeds a predefined threshold; a low score on any single axis suffices for rejection. This strict, axis-wise filtering ensures that the final dataset maintains high consistency across all components of the quadruplet.

To validate the reliability of VLM-based filtering, we conduct cross-VLM and human-agreement studies (Tab. 1). Re-running the 5-axis filter on 600 stratified samples with Claude Opus 4.6 and Gemini 3.1 Pro yields cross-VLM Cohen's κ that meets or exceeds human inter-annotator κ. A 10-annotator human study on 60 samples confirms that GPT-4o–human agreement exceeds mean human–human agreement, with high precision and moderate recall, confirming a conservative filter that rarely retains flawed samples.

Pipeline Comparison. Fig. 3 compares our reference-completion pipeline with two alternative paradigms: forward synthesis and random combination. Forward synthesis constructs all four quadruplet components sequentially, requiring 2–3 image-model calls per sample; this compounds errors and limits scalability. Random combination pairs each input with a randomly sampled reference, achieving high volume scalability but suffering from low semantic compatibility between input and reference, leading to high failure rates.

Our reference-completion pipeline inherits the triplets from curated TCIE corpora, which makes it efficient and ensure high data success rate and reasonableness. The dataset scale can be enriched by incorporating more TCIE data. Moreover, while both forward synthesis and random combination are constrained by pre-defined keyword pools or image pools and can only benefit from basic

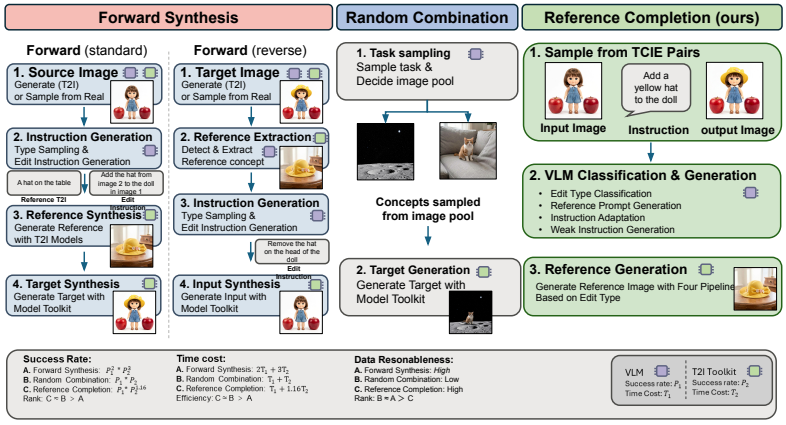

Fig. 3: Comparison of three ICIE data construction paradigms. Forward synthesis generates all quadruplet components from a single starting image; random combination samples references from an image pool; our reference completion inherits curated TCIE triplets and synthesizes only the missing reference image.

Table 2: Comparison of image-conditioned image-editing datasets. RCEdit-500K is larger in scale and provides a unified taxonomy covering all edit types.

	Data Size	Resolution	Concrete	Abstract	Type					
					Add	Remove	Replace	Background	Style	Alter
ImgEdit [36]	60K	1024	✓	×	✓	×	✓	×	×	×
AnyEdit [37]	40K	512	✓	✓	✓	×	✓	×	×	✓
RCEdit-500K (Ours)	477K	1024	✓	✓	✓	✓	✓	✓	✓	✓

model improvements, our pipeline can further enhance data quality by adopting stronger model toolkits and leveraging better TCIE datasets as they become available.

As shown in Tab. 2, although some TCIE datasets [36, 37] include a "visual editing" subset that overlaps with ICIE, their coverage is limited to concrete-object edits (with only a small material-transfer subset in AnyEdit [37]) and remains at a small scale. In contrast, RCEdit-500K provides 477K samples at 1024 resolution, covering all six edit categories with both concrete and abstract reference types.

3.3 Data Statistics

Fig. 4 and Fig. 5 summarize the statistics of RCEdit-500K. We use LAION-Aesthetics-Predictor-V1 [23] to estimate aesthetic scores, and employ patch-wise SSIM [28] to quantify edit-region size: for each image pair, we compute per-patch SSIM between input and target and apply k-means [15] clustering to separate high- and low-similarity patches, with the low-similarity cluster indicating the edited region.

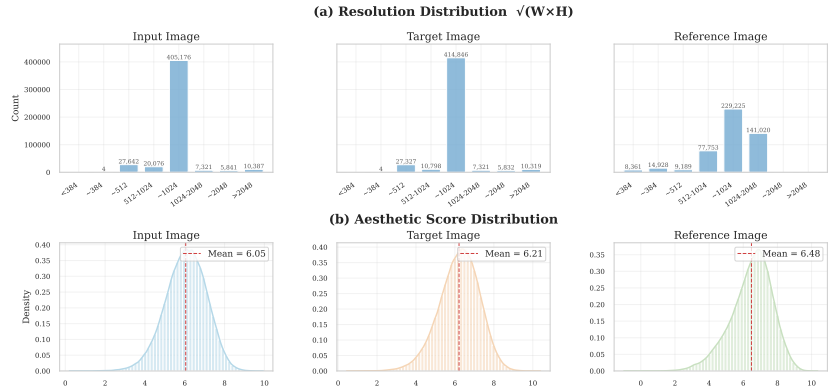

Fig. 4: Data statistics of RCEdit-500K: (a) resolution distribution and (b) aesthetic score distribution.

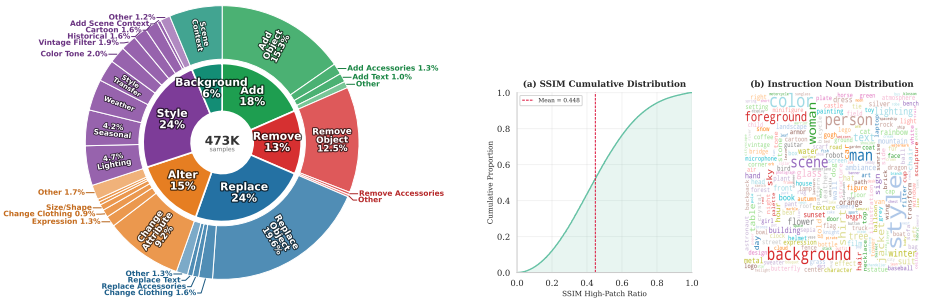

Fig. 5: Left: Edit-type diversity of RCEdit-500K—6 coarse categories and 35 fine-grained types classified by GPT-4o (validated accuracy 92.7%); no coarse type exceeds 24%. Right: Patch-wise SSIM distribution and word cloud; the SSIM patch ratio measures the fraction of patches with high similarity between input and target images.

As shown in Fig. 4, the resolution distributions are well-balanced and consistent with those of recent TCIE datasets [1, 36]. The majority of images achieve aesthetic scores above 4, with a mean of approximately 6, indicating high visual quality. Fig. 5 (right) shows that the edit-region ratio is neither too small nor too large for most samples, providing a healthy distribution for model training. The word cloud further illustrates the lexical diversity of instructions in RCEdit-500K.

Among the six edit categories, add and replace together account for the largest share, reflecting the dominance of subject-centric edits in the source TCIE corpora. Style and alter edits, while smaller in absolute count, still comprise a substantial portion, ensuring adequate coverage of abstract-attribute conditioning—a capability largely absent from prior ICIE datasets (Tab. 2). As shown in Fig. 5 (left), 35 fine-grained sub-types are identified by GPT-4o classification, with no single coarse type exceeding 24% of the total, confirming balanced coverage

across edit semantics. The post-filtering step discards approximately 12.5% of candidates, with the highest rejection rates observed for style and alter types, where synthesized references are most prone to semantic mismatch. This confirms that our multi-dimensional filtering is especially valuable for the more challenging abstract-attribute edits.

4 Experiments

We validate the effectiveness of RCEdit-500K by fine-tuning multiple models and evaluating their ICIE performance. Sec. 4.1 describes the training configurations, Sec. 4.2 introduces the evaluation benchmark, Sec. 4.3 presents quantitative and qualitative results, and Sec. 4.4 ablates the contributions of post-filtering and weak-instruction.

4.1 Experimental Settings

We fine-tune two diffusion models with LoRA [8]: Flux-Kontext [11] and Qwen-Image-Edit-2509 [30]. These two models are chosen to examine whether our dataset benefits models at different capability levels—Flux-Kontext has weak baseline ICIE performance, whereas Qwen-Image-Edit-2509 already exhibits reasonable editing ability. Both LoRA rank and alpha are set to 256. For the autoregressive family, we select Janus-Pro [2] and perform full fine-tuning. During inference on ICIE tasks, we concatenate image tokens from both the generation encoder and the understanding encoder to preserve editing fidelity. Further implementation details are provided in the supplementary material.

4.2 Evaluation Setups

Evaluation protocols for ICIE remain under-explored [12]. Traditional pairwise similarity metrics (*e.g.* CLIP [20] similarity, SSIM [28]) are inadequate because they do not account for multi-image conditioning. VLM-based evaluation is therefore more appropriate [36]. We adopt MultiBanana [18] as our benchmark, which uses a VLM to assess multi-reference text-to-image generation along five axes: Instruction Alignment (IA), Reference Consistency (RC), Background-Subject Match (BSM), Physical Realism (PR), and Visual Quality (VQ). We evaluate on its two-reference subset, which directly corresponds to the ICIE setting. The "makeup" category is excluded from our evaluation for privacy considerations. All VLM evaluations are conducted with GPT-4o.

4.3 Results

Quantitative Results. Tab. 3 reports the MultiBanana two-reference evaluation results. The closed-source GPT-Image-1.5 attains the highest overall score by a clear margin, underscoring the persistent gap between proprietary and open-source ICIE systems.

Table 3: Quantitative comparison on the MultiBanana two-reference subset. Please refer to Sec. 4.2 for metric full-names. All scores are rated by GPT-4o on a 1–10 scale.

Model	IA↑	RC↑	BSM↑	PR↑	VQ↑	Avg↑
Closed-source models						
Nano Banana [5]	4.59	4.00	5.13	5.72	5.44	4.67
GPT-Image-1.5	5.87	4.94	5.40	5.89	5.85	5.51
Open-source models						
Flux-Kontext [11]	2.94	2.80	2.39	2.94	2.81	2.82
Omnigen2 [31]	3.62	3.28	4.50	5.17	5.05	3.93
Qwen-Image-Edit-2509 [30]	3.60	4.06	4.83	5.61	5.39	4.31
Dreamomni2 [34]	4.45	3.61	5.07	5.83	5.80	4.54
Qwen-Image-Edit-2511 [30]	4.15	4.18	4.95	5.68	5.59	4.58
Flux-Klein-4B [10]	4.69	4.01	5.05	5.62	5.57	4.71
Flux-Klein-9B [10]	4.78	4.09	5.01	5.61	5.50	4.75
Janus-Pro	1.02	1.01	1.02	1.04	1.11	1.03
+ RCEdit-500K (Ours)	4.65	3.81	4.52	5.04	4.91	4.43
Flux-Kontext	2.94	2.80	2.39	2.94	2.81	2.82
+ RCEdit-500K (Ours)	3.52	3.42	4.76	5.50	5.27	4.04
Qwen-Image-Edit-2509	3.60	4.06	4.83	5.61	5.39	4.31
+ RCEdit-500K (Ours)	4.08	4.03	5.11	5.86	5.72	4.56

Fine-tuning on RCEdit-500K yields consistent gains across both diffusion models. For Flux-Kontext, LoRA adaptation raises the average by +1.22, demonstrating that RCEdit-500K can transform a model with weak ICIE capability into a competitive one through parameter-efficient finetuning. For Qwen-Image-Edit-2509, which already possesses reasonable editing ability, LoRA fine-tuning brings moderate but consistent improvement across all axes. The gain on Instruction Alignment further confirms that RCEdit-500K strengthens the model's capacity to follow reference-conditioned instructions. Notably, this lightweight adaptation achieves performance comparable to the newer Qwen-Image-Edit-2511, which was full-fine-tuned on large-scale data.

For the autoregressive model, Janus-Pro-ICIE, fully fine-tuned on our dataset, surpasses both Qwen-Image-Edit-2509 and Omnigen2. Its Instruction Alignment ranks third overall, revealing strong instruction-following potential in autoregressive architectures. This demonstrates that autoregressive models, despite not being originally designed for image editing, can acquire competitive reference-guided editing capabilities when trained on high-quality ICIE data—reinforcing that data availability is the primary bottleneck for ICIE.

Qualitative Results. Fig. 6 provides qualitative comparisons between each base model and its RCEdit-500K fine-tuned counterpart, spanning both concrete-object edits (add, replace) and abstract-attribute edits (style, alter). For Qwen-Image-Edit-2509, the original model tends to copy-paste the reference image, largely ignoring the instruction; after fine-tuning, instruction alignment improves substantially, consistent with the IA gain in Tab. 3. Flux-Kontext originally

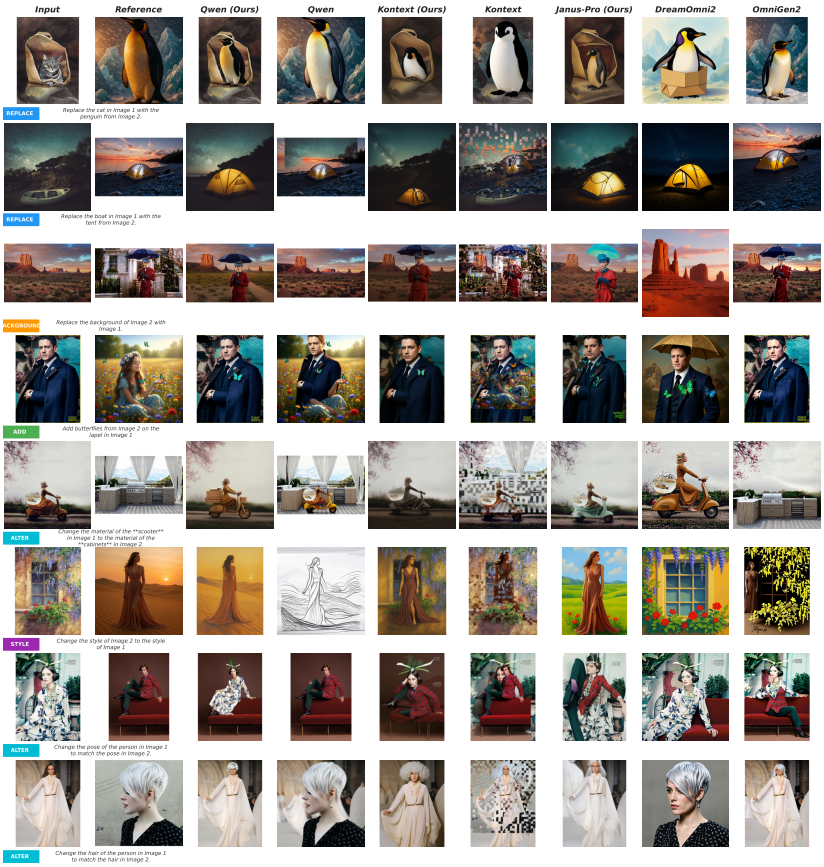

Fig. 6: Qualitative comparison of different models. Models labeled "Ours" are trained with RCEdit-500K, demonstrating consistent performance improvements across various tasks. "Qwen" refers to Qwen-Image-Edit-2509, and "Kontext" refers to Flux-Kontext.

suffers from visual blending between input and reference, producing incoherent outputs; this artifact is strongly reduced after fine-tuning. Janus-Pro trained on RCEdit-500K produces successful edits despite occasional artifacts from its limited base capability.

4.4 Ablation Study

We ablate two key designs in our data curation pipeline—post-filtering and weak-instruction augmentation (Tab. 4)—and compare against existing ICIE datasets at both matched and full scale to validate data quality and scalability (Tab. 5). All experiments use Flux-Kontext LoRA as the base configuration.

Data Post-Filtering. Removing the five-dimensional VLM-based post-filtering (Sec. 3.2) causes the largest overall degradation. The most pronounced declines

Table 4: Ablation study on post-filtering and weak-instruction augmentation evaluated on MultiBanana two-reference subset. All variants use Flux-Kontext with LoRA fine-tuning on RCEdit-500K.

Variant	IA↑	RC↑	BSM↑	PR↑	VQ↑	Avg↑
w/o post-filtering	3.60	3.30	3.62	4.20	4.04	3.62
w/o weak instruction	**3.67**	3.35	3.84	4.50	4.27	3.74
full	3.52	**3.42**	**4.76**	**5.50**	**5.27**	**4.04**

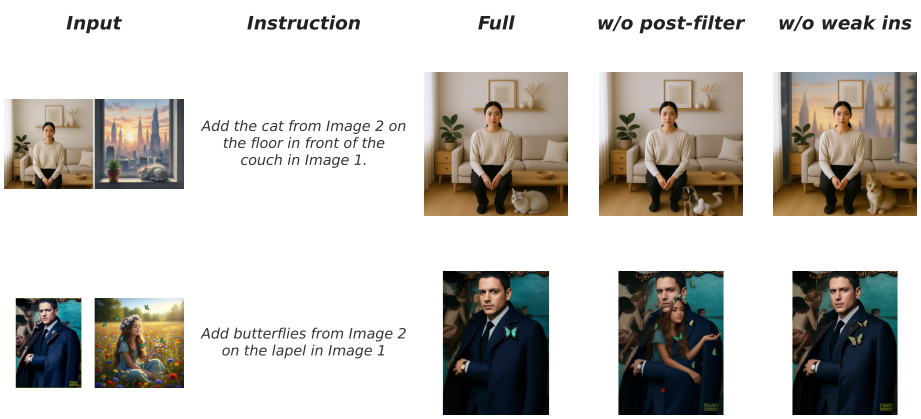

Fig. 7: Qualitative comparison of ablation study.

appear in BSM, PR, and VQ, indicating that unfiltered quadruplets introduce noise that severely impairs the model's ability to produce physically plausible and visually coherent outputs. In contrast, IA and RC are less affected, suggesting that these axes are more robust to data noise. As illustrated in Fig. 7, training without post-filtering leads to noticeably more artifacts in the generated images, validating that our multi-dimensional filtering strategy is critical for maintaining high data quality.

Weak Instruction. Removing weak-instruction augmentation also degrades overall performance, with the drops again concentrated on BSM, PR, and VQ. Without weak instructions, the model can partially resolve the edit from text alone, reducing its reliance on the reference image as a visual conditioning signal. This explains the slight IA increase alongside the RC decrease: the model follows the textual instruction but attends less to the reference's visual content. Qualitatively, Fig. 7 shows that although the model correctly performs the requested edit (*e.g.* adding a cat or butterfly), the generated subject differs noticeably from the one depicted in the reference image, confirming that the model falls back on its pre-existing TCIE capability. This validates our design motivation in Sec. 3.2:

Table 5: Ablation on training data. All models use Flux-Kontext LoRA. Baseline is without LoRA.

Training Data	IA↑	RC↑	BSM↑	PR↑	VQ↑	Avg↑
Baseline	2.94	2.80	2.39	2.94	2.81	2.82
ImgEdit	3.05	2.83	2.58	3.10	2.92	2.92
AnyEdit (w/o structure)	3.02	2.65	2.10	2.57	2.44	2.68
AnyEdit	2.17	1.87	4.55	5.15	4.93	2.97
RCEdit-60K (Ours)	**3.62**	3.28	3.78	4.45	4.16	3.68
RCEdit-500K (Ours)	3.52	**3.42**	**4.76**	**5.50**	**5.27**	**4.04**

deliberately abstracting reference-specific details in the instruction forces the model to ground its edits in the reference image, yielding stronger visual fidelity.

Comparison with Existing ICIE Datasets. We compare training on existing ICIE subsets—ImgEdit [36] and AnyEdit [37]—against RCEdit variants, including **RCEdit-60K** for scale-controlled comparison.

As shown in Tab. 5, existing ICIE subsets yield only marginal gains over the baseline. Notably, AnyEdit severely degrades IA and RC despite boosting BSM, PR, and VQ, suggesting the model defaults to unconditional generation rather than genuine reference-guided editing. The contrast with RCEdit is striking: at matched scale, RCEdit-60K outperforms both by a large margin while improving all five metrics, demonstrating that only high-quality ICIE data yields meaningful gains. Scaling to RCEdit-500K yields further improvement, confirming that our reference-completion pipeline scales favorably with the size of the underlying TCIE corpus. Per-type results are in the supplementary.

5 Conclusion

We introduce **RCEdit-500K**, the first large-scale unified open-source dataset for image-conditioned image editing (ICIE). By reframing ICIE data construction as a reference-completion problem, our pipeline converts existing TCIE triplets into ICIE quadruplets by synthesizing only the missing reference image and adapting instructions accordingly. Combined with five-dimensional VLM-based post-filtering and weak-instruction augmentation, this yields 477K high-quality quadruplets spanning six edit categories. Experiments across diffusion-based and autoregressive models demonstrate consistent improvements, with even an autoregressive model acquiring competitive ICIE capabilities—reinforcing that **data availability** is the primary bottleneck for ICIE. The dataset is publicly available at `https://huggingface.co/datasets/carpedkm/RCEdit-500K`.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62472399 and the Open Fund of APKL of BIIP, IAI, Hefei Comprehensive National Science Center under Grant 24YGXT003.

References

1. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. arXiv preprint arXiv:2506.18095 (2025)
2. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
3. Chen, Y., Ma, Z., Jia, G., Jiang, C., Li, J., Zhou, B.: Context-aware autoregressive models for multi-conditional image generation. arXiv preprint arXiv:2505.12274 (2025)
4. Cheng, Y., Wu, W., Wu, S., Huang, M., Ding, F., He, Q.: Umo: Scaling multi-identity consistency for image customization via matching reward. arXiv preprint arXiv:2509.06818 (2025)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Ge, Y., Zhao, S., Li, C., Ge, Y., Shan, Y.: Seed-data-edit technical report: A hybrid dataset for instructional image editing. arXiv preprint arXiv:2405.04007 (2024)
7. He, Z., Ning, Y., Qin, Y., Wang, G., Yang, S., Lin, L., Li, G.: Vton 360: High-fidelity virtual try-on from any viewing direction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26388–26398 (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
10. Labs, B.F.: FLUX.2: Frontier Visual Intelligence (2025)
11. 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., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025)
12. Li, R., Qu, L., Zhang, J., Gui, D., Xu, M., Zhang, X., Hu, H., Wang, W., Wang, J.: Genarena: How can we achieve human-aligned evaluation for visual generation tasks? arXiv preprint arXiv:2602.06013 (2026)
13. Li, S., Gu, J., Liu, K., Lin, Z., Wei, Z., Grover, A., Kuen, J.: Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. arXiv preprint arXiv:2509.19244 (2025)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
15. McQueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proc. of 5th Berkeley Symposium on Math. Stat. and Prob. pp. 281–297 (1967)
16. Mei, S., Ni, B.: Physdiff-vton: Cross-domain physics modeling and trajectory optimization for virtual try-on. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

17. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
18. Oshima, Y., Miyake, D., Matsutani, K., Iwasawa, Y., Suzuki, M., Matsuo, Y., Furuta, H.: Multibanana: A challenging benchmark for multi-reference text-to-image generation. arXiv preprint arXiv:2511.22989 (2025)
19. Qian, Y., Bocek-Rivele, E., Song, L., Tong, J., Yang, Y., Lu, J., Hu, W., Gan, Z.: Pico-banana-400k: A large-scale dataset for text-guided image editing. arXiv preprint arXiv:2510.19808 (2025)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024)
22. Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., Xiong, Y., Zhang, H., Li, F., Tang, P., Yu, K., Zhang, L.: Grounding dino 1.5: Advance the "edge" of open-set object detection (2024)
23. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
24. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
25. Sun, H., He, Q., Peng, J., Tang, P., Zhang, J., Zhu, J., Hu, X., Yan, S.: Tokenar: Multiple subject generation via autoregressive token-level enhancement. arXiv preprint arXiv:2510.16332 (2025)
26. Wan, S., Chen, J., Cai, Q., Pan, Y., Yao, T., Mei, T.: Vton-vllm: Aligning virtual try-on models with human preferences. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
27. Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., Xie, C.: Gpt-image-edit-1.5m: A million-scale, gpt-generated image dataset (2025)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
29. Wei, Y., Zhang, S., Yuan, H., Gong, B., Tang, L., Wang, X., Qiu, H., Li, H., Tan, S., Zhang, Y., et al.: Dreamrelation: Relation-centric video customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12381–12393 (2025)
30. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025)
31. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)

32. Wu, S., Huang, M., Cheng, Y., Wu, W., Tian, J., Luo, Y., Ding, F., He, Q.: Uso: Unified style and subject-driven generation via disentangled and reward learning. arXiv preprint arXiv:2508.18966 (2025)
33. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
34. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
35. Ye, M., Liu, J., Song, Y.: Loom: Diffusion-transformer for interleaved generation. arXiv preprint arXiv:2512.18254 (2025)
36. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
37. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
38. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
40. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)