Step1X-Edit: A Practical Framework for General Image Editing

Step1X-Image Team
StepFun
https://github.com/stepfun-ai/Step1X-Edit

Abstract

In recent years, image editing technology has witnessed remarkable and rapid development. The recent unveiling of cutting-edge multimodal models such as GPT-4o and Gemini2 Flash has introduced highly promising image editing capabilities. These models demonstrate an impressive aptitude for fulfilling a vast majority of user-driven editing requirements, marking a significant advancement in the field of image manipulation. However, there is still a large gap between the open-source algorithm with these closed-source models. Thus, in this paper, we aim to release a state-of-the-art image editing model, called Step1X-Edit, which can provide comparable performance against the closed-source models like GPT-4o and Gemini2 Flash. More specifically, we adopt the Multimodal LLM to process the reference image and the user's editing instruction. A latent embedding has been extracted and integrated with a diffusion image decoder to obtain the target image. To train the model, we build a data generation pipeline to produce a high-quality dataset. For evaluation, we develop the GEdit-Bench, a novel benchmark rooted in real-world user instructions. Experimental results on GEdit-Bench demonstrate that Step1X-Edit outperforms existing open-source baselines by a substantial margin and approaches the performance of leading proprietary models, thereby making significant contributions to the field of image editing.

1 Introduction

Image editing with natural language instructions has become an increasingly important task in vision-language research. It offers intuitive interaction for end users while posing unique technical challenges: understanding nuanced semantics, precisely localizing regions to edit, and preserving image fidelity. While diffusion models have dramatically improved image generation quality, the existing design by integrating text encoder, e.g., CLIP [33] and T5 [34], with diffusion transformer often struggles with following editing instruction to maintain alignment between input image and edit instruction, especially when edit instructions are subtle or compositional.

Recent advances in proprietary multimodal foundation models, such as GPT-4o [29], Gemini2 Flash [14], and SeedEdit/Doubao [41], have pushed the frontier of instruction-based image editing. These systems leverage large-scale vision-language modeling capabilities to perform high-fidelity edits across diverse scenarios. However, their closed nature limits reproducibility and transparency. In parallel, open-source efforts like OmniGen [51] and ACE++ [27] aim to replicate similar capabilities but still fall short in terms of overall generalization, edit accuracy, and the quality of generated images.

In this work, we aim to narrow the performance gap between open-source and closed-source editing systems, while also pushing the boundary of practical and user-grounded editing evaluation. Although researchers have open-sourced editing datasets like AnyEdit [55] and OmniEdit [49], we argue that the quality and diversity of these datasets are not good enough to obtain comparable performance against the close-source algorithms like GPT-4o. Thus, to target the image edit problem, we first try to build a large-scale high quality dataset for training. More specifically, we identify 11 major

Figure 1: **Overview of Step1X-Edit**. Step1X-Edit is an open-source general editing model that achieves proprietary-level performance with comprehensive editing capabilities.

editing task categories based on the commonly used editing instructions. Guided by this taxonomy, we develop a scalable and flexible data pipeline to generate over 1 million high-quality training data. These image-instruction pairs encompass a broad spectrum of editing operations, including object manipulation, attribute modification, layout adjustment, and stylization, ensuring comprehensive coverage of real-world editing scenarios.

Building on this dataset, we propose, **Step1X-Edit**, a unified image editing model that combines the strong semantic reasoning of Multimedia Large Language Model (MLLM), e.g., Qwen-VL [3], with a DiT-style diffusion architecture. The reference image and editing prompts can be processed by the MLLM to generate a target image latent condition which will be integrated with the diffusion model to obtain the output image. Our approach maintains a good balance between reference image reconstruction and editing prompt following. To train the model, we start from a text-to-image model to retain aesthetic quality and visual consistency, which can be easily replace by the existing text-to-image models like SD3 [11] and FLUX[6]. To evaluate the existing editing models, we introduce a new benchmark named **GEdit-Bench**. By carefully collecting the images and editing prompts, GEdit-Bench ensures both real-world editing requirements and the diversity of the editing prompts. The experiments on GEdit-Bench validate that **Step1X-Edit** outperforms existing open-

source baselines with a large-margin and approaches the performance of leading proprietary models, *e.g.*, GPT-4o.

In summary, there will be three contributions of our work:

- We will open-source our **Step1X-Edit** model, to reduce the performance gap between open-source and closed-source image editing systems and boost further research in the field of image editing.
- A data generation pipeline is designed to produce high-quality image editing data. It ensures that the dataset is diverse, representative, and of sufficient quality to support the development of effective image editing models. The availability of such a pipeline provides a valuable resource for researchers and developers working on similar projects.
- A new benchmark, named **GEdit-Bench**, grounded in real-world usages is developed to support more authentic and comprehensive evaluation. This benchmark, which is carefully curated to reflect actual user editing needs and a wide range of editing scenarios, enables more authentic and comprehensive evaluations of image editing models.

2 Related Work

2.1 Controllable Image Generation and Editing

Autoregressive (AR) models have been actively studied for controllable image generation and editing by modeling images as sequences of discrete tokens. Works such as ControlAR [23], ControlVAR [21], and CAR [54] incorporate spatial and pixel-level guidance—like edges, segmentation masks, or depth maps—into the decoding process, enabling more structured and localized control. Training-Free VAR [48] enables editing without inversion through token-level distribution caching and adaptive masking, while M2M [39], MSGNet [9], and MVG [37] extend AR-based generation to multi-object, temporal, and multimodal scenarios. However, due to reliance on discrete visual tokens and sequence length limits, AR models often struggle to produce high-resolution and photorealistic results, especially in complex scenes.

Diffusion models, by contrast, have become the dominant approach for high-fidelity image synthesis, with strong capabilities in photorealism, structural consistency, and diversity. Beginning with DDPM [16] and DDIM [42], and further advanced by Latent Diffusion [38, 32, 62], these models operate in latent spaces for improved scalability. ControlNet [58] and T2I-Adapter [28] inject spatial or task-specific control into the generation process, while InstructPix2Pix [8], Pix2Pix-Zero [31], and UniDiffuser [4] support instruction-based or zero-shot editing without retraining. Despite these strengths, diffusion models often depend on static prompts or predefined conditions and lack the capacity for multi-turn reasoning or flexible language alignment, limiting their use in open-ended editing scenarios.

These limitations have led to growing interest in unified image editing frameworks that combine the symbolic control of AR models with the generative fidelity of diffusion. Such models aim to tightly couple instruction understanding, spatial reasoning, and photorealistic synthesis within a single architecture, offering more flexible, general, and user-controllable editing capabilities. We discuss this direction further in the next section on Unified Image Editing Models.

2.2 Unified Instruction-based Image Editing Models

Unified image editing models aim to bridge semantic instruction understanding with precise visual manipulation in a single, coherent framework. Early approaches often rely on modular designs, where MLLMs generate textual prompts or spatial instructions to guide diffusion models, as seen in Prompt-to-Prompt [15], InstructEdit [45], and BrushEdit [22]. InstructPix2Pix [8] trains a conditional diffusion model using synthetic instruction-image pairs, while MagicBrush [57] improves real-world usability through high-quality human annotations. To tackle data scarcity and alignment issues, SeedEdit [41] distills text-to-image models into consistent editing models via causal diffusion and iterative self-alignment. Recent works such as SmartEdit [17], X2I [26], and RPG [53] enhance interaction between language and vision by integrating multimodal attention and reasoning strategies. AnyEdit [55] introduces task-aware routing and visual prompt projection within a unified diffusion model, while OmniGen [51] adopts a single transformer backbone to jointly encode text and images,

removing reliance on external encoders. More generally, models such as Gemini [14] and GPT-4o [29] demonstrate strong visual fluency through joint vision-language training, showing promising capabilities in understanding and generating consistent, context-aware images. Collectively, these developments reflect a shift from loosely coupled systems toward tightly integrated, instruction-driven editing frameworks.

However, existing approaches still face key limitations. Most methods are task-specific and lack general-purpose editability. They typically do not support incremental editing, fine-grained region correspondence, or instruction feedback refinement. Moreover, architectural coupling remains shallow in many designs, failing to unify instruction understanding and generation into a cohesive framework. These challenges motivate **Step1X-Edit**, which tightly integrates MLLM-based multimodal reasoning with diffusion-based controllable synthesis, enabling scalable, interactive, and instruction-faithful image editing across diverse editing goals.

3 Step1X-Edit

3.1 Data Creation

3.1.1 Data Pipeline

In the existing literature, current image editing datasets are constrained either by the scale or the quality of the collected data. To address this gap, this report endeavors to assemble a large-scale, high-quality dataset specifically tailored for image editing tasks.

We initiate the dataset collection process by web crawling a diverse set of image editing examples from the Internet. Through in-depth analysis of these examples, we systematically categorize the image editing problem into 11 distinct categories, which has been partly referenced by [55, 30]. These categories are designed to comprehensively encompass the vast majority of image editing requirements in practice. An overview of these 11 categories, along with the detailed data collection pipeline, is illustrated in Fig. 3.

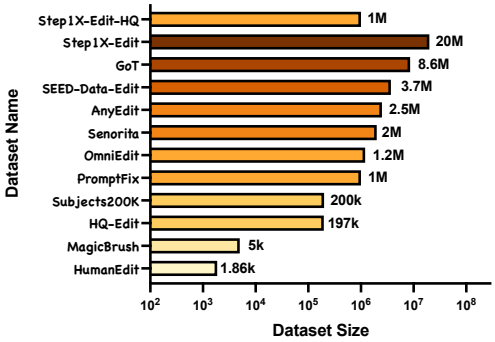


Figure 2: **Data Volume Comparison**.

To collect a large-scale high-quality triplets consisting of a source image, an editing instruction, and a target image, we designed a sophisticated data pipeline, which enabled us to generate over 20 million instruction-images triplets. Following rigorous filtering using both Multimodal LLMs, e.g. step-1o [43], and human annotators, we retained more than 1 million high-quality triplets. In Fig. 2, we present a side-by-side comparison of all existing editing datasets [12, 13, 64, 49, 56, 52, 18, 57, 2, 55]. Our Step1X-Edit dataset surpasses all others in scale. Even after a rigorous filtering process (with a retention ratio of 20:1), the **Step1X-Edit-HQ** subset remains on par with other datasets in terms of absolute magnitude. The full data collection pipeline for each subtask is outlined below.

Subject Addition & Removal: For subject-add and subject-remove tasks, we begin by annotating our proprietary dataset using Florence-2 [50], which supports diverse semantic granularities, spatial hierarchies, and annotation types such as object detection and classification. We then apply SAM-2 [36] for segmentation and use ObjectRemovalAlpha [24] to perform inpainting. Editing instructions are generated using a combination of Step-1o model [43] and GPT-4o, followed by manual review to ensure data validity.

Subject Replacement & Background Change: This category shares similar preprocessing steps with subject-add/remove, including Florence-2 [50] annotation and SAM-2 [36] segmentation. However, for these tasks, we utilize Qwen2.5-VL [3] and the Recognize-Anything Model [61] to identify target objects or keywords, followed by Flux-Fill [7] for content-aware inpainting. The instructions are automatically generated by Step-1o and the triplets are human-verified.

Color Alteration & Material Modification: After detecting objects in the image, we employ Zeodepth [5] for depth estimation to understand object geometry. Based on the identified target transformation (e.g., change of color or material), we use ControlNet [59] with diffusion model [1] to generate new images that preserve object identity while altering appearance attributes such as texture or color.

Text Modification: For text-editing tasks, we differentiate between valid and invalid text edits. We use PPOCR [10], which focuses on recognizing correct characters, alongside the Step-1o model to distinguish correct and incorrect regions of text. Based on this classification, we generate corresponding editing instructions. All outputs are finalized via human post-processing (e.g., manual retouching of text).

Motion Change: To handle motion-related transformations, we leverage videos from Koala-36M [46], extracting frame pairs as input. We use BiRefNet [63] and RAFT [44] for foreground-background separation and optical flow estimation. Specifically, we compute the mean of the foreground flow norm and the norm of the background flow mean, ensuring robustness in selecting pairs where only the foreground exhibits motion. Finally, GPT-4o is used to annotate the change in motion between frames as editing instructions.

Portrait Editing and Beautification: Data are collected and created by two major sources: **(a)** Beautification pairs from public sources. Faces are detected and passed through Step-1o to assess layout and background consistency. **(b)** Beautification of the human editor, we invite the human editor to conduct beatification on collected data. All data are manually validated.

Style Transfer: We handle stylization in two directions depending on the target visual domain: For styles such as Ghibli, ink painting, or 3D anime style, generating photorealistic images from stylized inputs yields better alignment. We extract edges from stylized images and generate realistic outputs using controlled diffusion model [59, 1]. Conversely, for styles like oil painting or pixel art, we begin with realistic images and generate stylized outputs using the same edge-to-image pipeline.

Tone Transformation: This category focuses on global tonal adjustments, including color grading, dehazing, deraining, and seasonal transformations. These changes are largely driven by algorithmic tools and automated filters to simulate realistic environmental changes.

3.1.2 Caption Strategy

To obtain high-quality and fine-grained editing instruction–image pairs, we adopt the following annotation strategies:

Redundancy-Enhanced Annotation: Given the well-known limitations of Vision-Language Models (VLMs)—such as vague background descriptions and susceptibility to hallucinations—we employ a multi-round annotation strategy. Specifically, the annotation results from a previous round are fed into the next round as contextual input. This recursive refinement strengthens semantic consistency across annotations and significantly mitigates hallucination-related issues. Deterministic information is reinforced through repeated confirmations, ensuring higher reliability of the final annotation.

Stylized Annotation via Contextual Examples: During the captioning process, we provide annotators (or models) with a large set of style-aligned examples as contextual references. These examples guide the tone, structure, and granularity of the captions, ensuring a consistent and stylized annotation format throughout the dataset.

Cost-Efficient Pipeline: To control annotation costs while maintaining quality, we first use GPT-4o to perform the aforementioned annotation procedures. The annotated results are then used to fine-tune our in-house Step1o model, which is employed to scale up annotation for larger datasets in a more cost-effective manner.

Bilingual Annotation (Chinese-English): All our annotations are conducted bilingually, in both Chinese and English. This not only enhances accessibility and usability across different linguistic communities but also lays the groundwork for multilingual model training and evaluation.

3.2 Our Method

As illustrated in Fig. 4, our algorithm involves three key components: a Multimedia Large Language Model (MLLM), a connector module, and a Diffusion in Transformer (DiT). The input editing

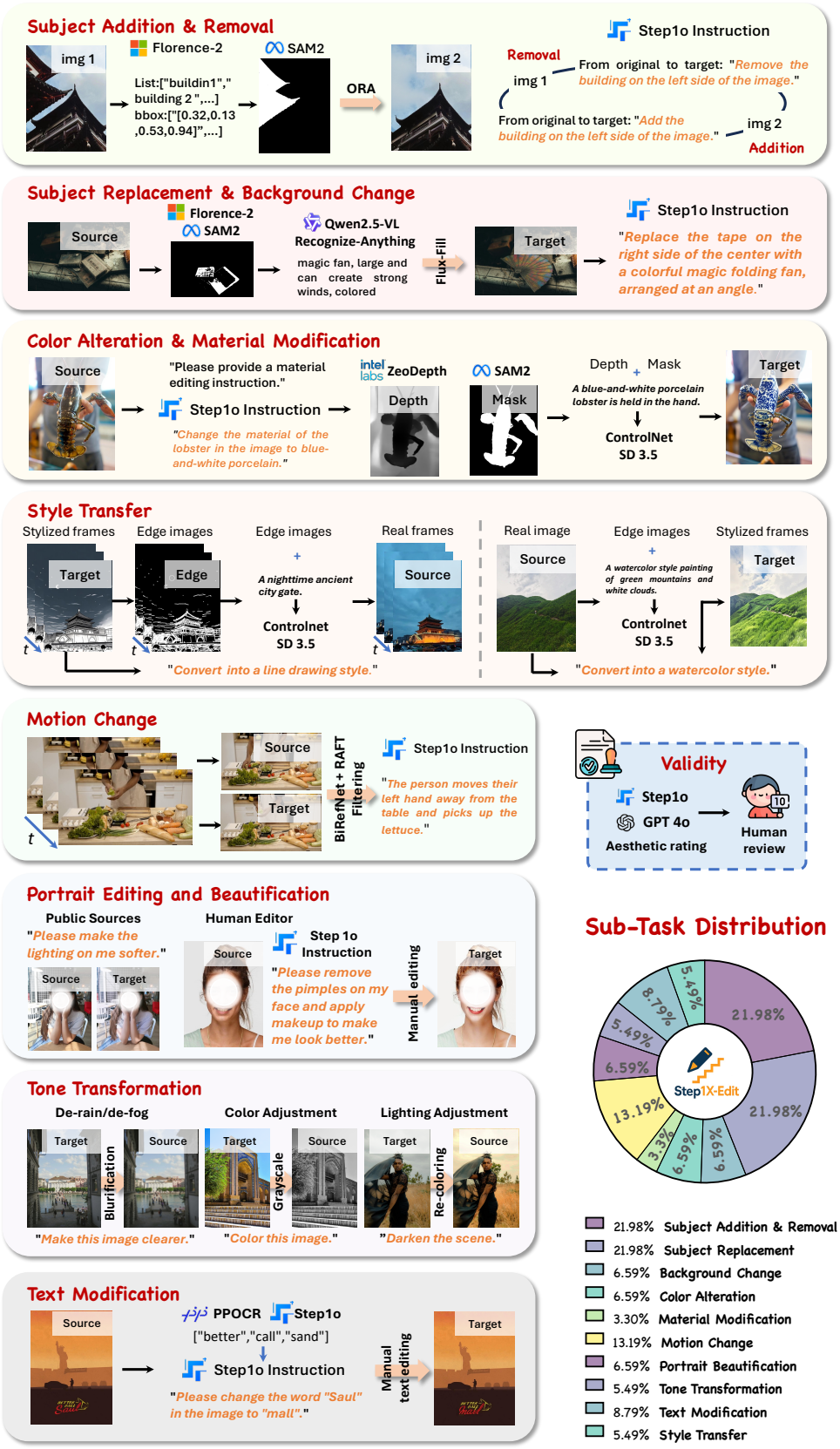


Figure 3: **Data Construction Pipeline and Sub-Task Distribution**.

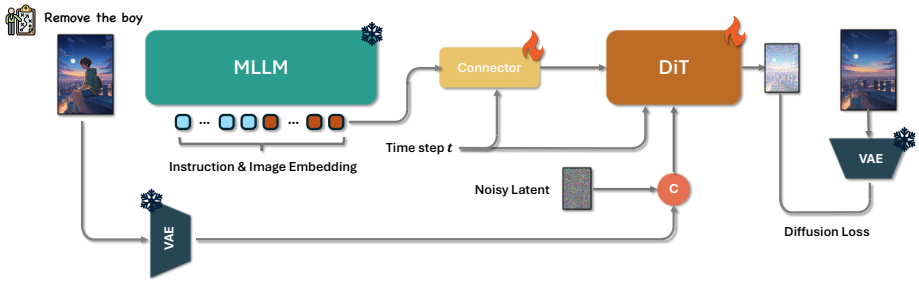


Figure 4: **Framework of Step1X-Edit**. Step1X-Edit leverages the image understanding capabilities of MLLMs to parse editing instructions and generate editing tokens, which are then decoded into images using a DiT-based network.

instruction, accompanied by the reference image, is first introduced to the MLLM; e.g., Qwen-VL [3] (hereafter abbreviated as Qwen). In conjunction with a system prefix, these inputs are jointly processed through a single forward pass of the MLLM, enabling the model to capture the semantic relationships between the instruction and the visual content. To isolate and emphasize the semantic elements relevant to the editing task, we selectively discard the token embeddings associated with the prefix. This filtering process retains only the token embeddings that directly align with the edit instruction, ensuring that subsequent processing focuses precisely on the user-specified editing requirements.

The extracted embeddings are then fed into a lightweight connector module, such as the token refiner [25, 19]. This module restructures the embeddings into a more compact textual feature representation. The refined feature subsequently substitutes the text embedding that was initially generated by the T5 [35] encoder in the downstream DiT network, e.g., FLUX [6]. Furthermore, we calculate the mean of all output embeddings from Qwen. This mean value is then projected through a linear layer, resulting in the generation of a global visual guidance vector. By doing so, the image editing network can leverage Qwen's enhanced semantic comprehension capabilities, enabling more accurate and context-aware editing operations.

To effectively train the Token Refiner and enable rich cross-modal conditioning, we draw inspiration from the token concatenation mechanism introduced in FLUX-Fill [7]. The key idea is to enhance the model's ability to reason over contrastive visual contexts. During training, both a target image and a reference image are fed into the system. The target image is first encoded by a VAE encoder, followed by the addition of Gaussian noise to promote generalization. The resulting latent is then linearly projected into an image token representation. In contrast, the reference image is encoded without noise and projected similarly. These two sets of image tokens are concatenated along the token length dimension, forming a fused feature of doubled token length, which is used as the final visual input.

The model is trained in a joint learning setup, where the connector and the downstream DiT are optimized simultaneously. We initialize both components with pretrained weights from our in-house Qwen and DiT text-to-image model, enabling better convergence and performance transfer. The learning rate is set to $1e^{-5}$, balancing training stability and convergence speed.

By combining structured language guidance, token-level visual conditioning, and strong pretrained backbones within a unified framework, our method significantly boosts the system's capability to perform high-fidelity, semantically aligned image edits across a diverse range of user instructions.

4 Benchmark and Evaluation

4.1 GEdit(Genuine Edit)-Bench

To evaluate the performance of the image editing models, we collect a new benchmark called **GEdit(Genuine Edit)-Bench**. The main motivation of the benchmark is to collect the real-world

Benchmarks	Size	Real Image	Genuine Instruction	Human Filtering	#Sub-tasks	Public Availability
EditBench [47]	240	✓	✗	✗	1	✓
EmuEdit [40]	3,055	✓	✗	✗	7	✓
HIVE [60]	1,000	✓	✗	✓	1	✓
HQ-Eidt [18]	1,640	✗	✗	✗	7	✓
MagicBrush [57]	1,053	✓	✗	✓	7	✓
AnyEdit [39]	1,250	✓	✗	✗	25	✓
ICE-Bench [30]	6,538	✓	✗	✓	31	✗
GEdit-Bench(Ours)	606	✓	✓	✓	11	✓

Table 1: **Key Attributes of Open-source Edit Benchmarks**. The reliance of existing open-source benchmarks on synthetic user inputs and minimal human involvement highlights the necessity of our proposed GEdit-Bench.

user editing instances in order to evaluate how the existing editing algorithms can be suffice for the practical editing instructions. More specifically, we collect more than 1K user editing instances from the Internet, e.g., reddit, and manually split these editing instructions into the 11 categories. To keep the diversity of the benchmark, we filter those editing instructions with similar purpose. Finally, we obtain 606 testing examples whose reference images are from the real-world cases which make it more genuine for the applications. Based on **GEdit-Bench**, we evaluate the existing open-source image editing algorithms like ACE++ and AnyEdit, as well as the closed-source algorithms like GPT-4o and Gemini2 Flash.

In order to assess the performance of image editing models comprehensively, we have curated a novel benchmark named **GEdit-Bench**. The core impetus behind developing this benchmark lies in gathering real-world user image editing instances. This enables us to thoroughly examine whether existing editing algorithms can effectively meet practical user editing requirements. Specifically, we have collected over 1,000 user editing examples from platforms like Reddit across the Internet. Subsequently, these editing instructions have been carefully categorized into 11 distinct groups manually. To ensure the diversity of the proposed benchmark, we systematically filtered out instructions with overlapping intents. Finally, we obtain a set of 606 testing samples, all featuring reference images sourced from authentic real-life scenarios. This characteristic renders the benchmark highly applicable for practical use cases.

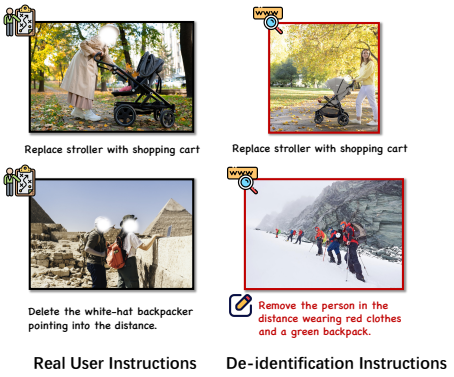


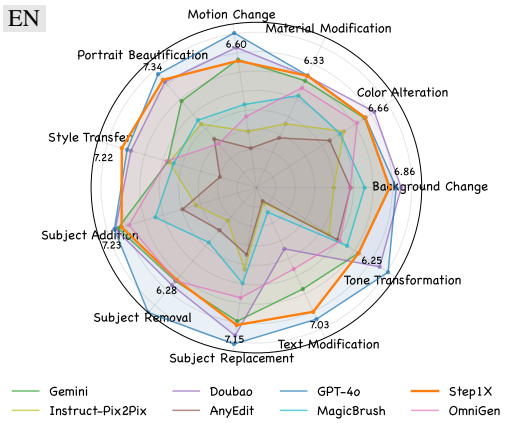
Figure 5: **De-Identification Process**.

To safeguard privacy, a comprehensive de-identification protocol was meticulously implemented for all user-uploaded images prior to their utilization within the benchmarking framework as shown in Fig. 5. For each individual original image, a multi-faceted reverse image search strategy was employed, spanning across multiple public search engines. This process aimed to identify publicly accessible alternative images that demonstrated both visual similarity and semantic consistency with the original one, thereby aligning seamlessly with the corresponding editing instructions. In instances where public image alternatives could not be procured through this search methodology, a systematic approach to modifying the editing instructions was adopted. These modifications were carefully calibrated to maintain the highest degree of fidelity between the anonymized image-instruction example and the original user intents. This approach not only ensures the ethical integrity of the benchmark dataset but also preserves the essential characteristics required for accurate and meaningful evaluation of image editing models.

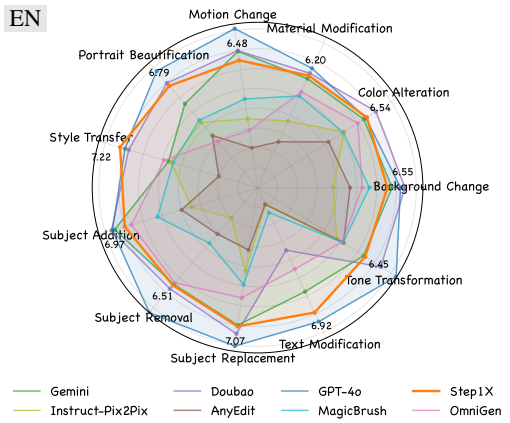
4.2 Experimental Results

4.2.1 Evaluation on GEdit-Bench

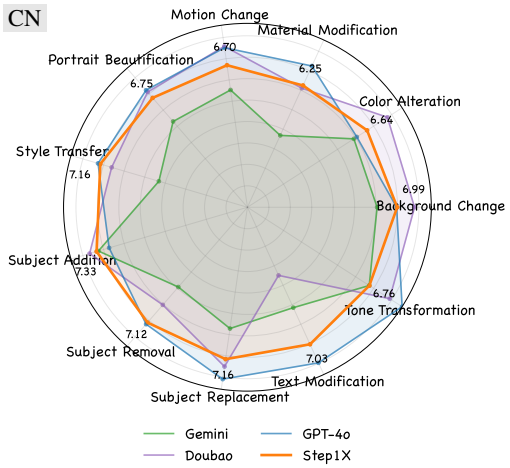
Based on the **GEdit-Bench**, we evaluated a diverse range of image editing algorithms, covering state-of-the-art open-source solutions such as Instruct-Pix2Pix [8], MagicBrush [57], AnyEdit [55],



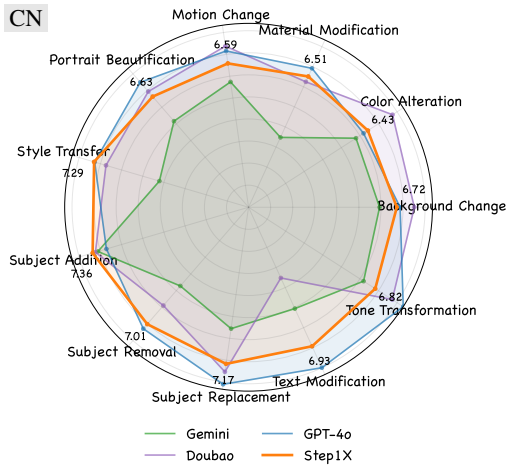
(a) VIEScore for the Intersection-subset.



(b) VIEScore for the Full set.



(c) VIEScore for the Intersection-subset.



(d) VIEScore for the Full set.

Figure 6: **VIEScore of Each Sub-task in GEdit-Bench. All the results are evaluated by GPT-4o**.

and OmniGen [51], as well as proprietary algorithms like GPT-4o [29][1], doubao [41][2], and Gemini2 Flash [14][3]. Following VIEScore [20], we adopt three metrics: SQ (Semantic Consistency), PQ (Perceptual Quality), and O (Overall Score). SQ assesses the degree to which the edited results conform to the given editing instruction, with a score ranging from 0 to 10. PQ evaluates the naturalness of the image and the presence of artifacts, also using a scoring scale that ranges from 0 to 10. The overall score is calculated based on these evaluations. To perform the automatic evaluation for VIEScore, we adopt the state-of-art MLLM model GPT-4.1[4]. Also, the evaluation based on the open-source model Qwen2.5-VL-72B [3] is included for reproduction. To comprehensively assess model capabilities on different languages, each image in our benchmark is paired with one English (EN) and one Chinese (CN) instruction. For EN instructions (GEdit-Bench-EN), both closed and open-source models are evaluated. For CN instructions (GEdit-Bench-CN), only those models which supports Chinese prompts, i.e., the close-source systems, are tested. During the evaluation process, we find that close-source image editing system such as GPT-4o may refuse certain instructions due to safety policies. To address this issue, we report two scores for two testing sets: (1) Intersection Subset — the subset of images whose results can be successfully returned from all the tested models, and (2) Full set — all the testing samples from **GEdit-Bench**. For the full set of results, we will

[1]The results are obtained based on ChatGPT APP in April 2025.
[2]The results are obtained based on doubao APP in April 2025.
[3]The results are obtained in April 2025.
[4]API access as of April 2025

Model	GEdit-Bench-EN (Intersection subset) ↑						GEdit-Bench-EN (Full set) ↑					
	G_SC	G_PQ	G_O	Q_SC	Q_PQ	Q_O	G_SC	G_PQ	G_O	Q_SC	Q_PQ	Q_O
Instruct-Pix2Pix [8]	3.473	5.601	3.631	4.836	6.948	4.655	3.575	5.491	3.684	4.772	6.870	4.576
MagicBrush [57]	4.646	5.800	4.578	5.806	7.162	5.632	4.677	5.656	4.518	5.733	7.066	5.536
AnyEdit [55]	3.177	5.856	3.231	3.583	6.751	3.498	3.178	5.820	3.212	3.438	6.729	3.361
OmniGen [51]	6.070	5.885	5.162	7.022	6.853	6.565	5.963	5.888	5.061	6.900	6.781	6.413
Step1X-Edit	**7.183**	**6.818**	**6.813**	**7.380**	**7.229**	**7.161**	**7.091**	**6.763**	**6.701**	**7.332**	**7.204**	**7.104**
Gemini [14]	6.697	6.638	6.322	7.276	7.306	6.978	6.732	6.606	6.315	7.287	7.315	6.982
Doubao [41]	7.004	7.215	6.828	7.417	7.635	7.273	6.916	7.188	6.754	7.382	7.639	7.241
GPT-4o [29]	7.844	7.592	7.517	7.873	7.690	7.694	7.850	7.620	7.534	7.826	7.689	7.646

Table 2: **Quantitative evaluation on GEdit-Bench-EN.** All metrics are reported as higher-is-better (↑). The Intersection subset reflects the subset of prompts where all methods return valid responses with a total of 434 instances; the Full set includes all the 606 instances. G_SC, G_PQ, and G_O refer to the metrics evaluated by GPT-4.1, while Q_SC, Q_PQ, and Q_O refer to the metrics evaluated by Qwen2.5-VL-75B.

Model	GEdit-Bench-CN (Intersection subset) ↑						GEdit-Bench-CN (Full set) ↑					
	G_SC	G_PQ	G_O	Q_SC	Q_PQ	Q_O	G_SC	G_PQ	G_O	Q_SC	Q_PQ	Q_O
Gemini [14]	5.580	6.757	5.505	5.658	7.362	5.522	5.427	6.767	5.360	5.617	7.360	5.485
Doubao [41]	7.076	7.315	6.869	7.122	7.669	7.062	6.984	7.273	6.772	7.116	7.667	7.063
GPT-4o [29]	7.722	7.590	7.353	7.771	7.625	7.572	7.673	7.559	7.302	7.698	7.634	7.529
Step1X-Edit	7.250	6.855	6.898	7.347	7.327	7.232	7.204	6.869	6.861	7.282	7.303	7.161

Table 3: **Quantitative evaluation on GEdit-Bench-CN.** All metrics are reported as higher-is-better (↑). The Intersection subset reflects the subset of prompts where all methods return valid responses with a total of 422 instances; the Full set includes all the 606 instances. G_SC, G_PQ, and G_O refer to the metrics evaluated by GPT-4.1, while Q_SC, Q_PQ, and Q_O refer to the metrics evaluated by Qwen2.5-VL-75B.

calculate the average scores only for the cases where the models successfully generate and return the target image. For each evaluated model, instances where no result image is returned due to reasons such as safety concerns will be excluded from the averaging process.

Fig. 6 demonstrates the groundbreaking capacity of Step1X-Edit, which outperforms open-source counterparts across 11 distinct evaluation axes. When compared to closed models, it surpasses Gemini2 Flash [14] and even beats GPT-4o [29] in axes such as style change and color alteration. As detailed in Tab. 2, Step1X-Edit significantly outperforms the existing open-source models like OmniGen [51] and has comparable results against the closed model like Gemini2 Flash and Doubao. Furthermore, as shown in Tab. 3, Step1X-Edit demonstrates consistent performance, even surpassing Gemini2 and doubao when handling the Chinese editing instructions in GEdit-Bench-CN benchmark. These results highlight the outstanding performance of our model across all dimensions with a unified architecture, eliminating the requirement for masks during the editing process. Fig. 7 and Fig. 8 provide illustrative examples for English instructions and Chinese instructions, respectively.

4.2.2 User Study

To assess the subjective quality of image editing results, we conduct a comprehensive user preference study built upon the GEdit-Bench. A total of 55 participants are recruited to evaluate the outputs of four algorithms—Gemini2 Flash [14], Doubao[41], GPT-4o [29], and our method, Step1X-Edit. Each participant is presented with a series of test images and asked to rank the editing results generated by the four methods. This evaluation is performed in a blinded and subjective setting to minimize bias and ensure fairness.

Participants rate the outputs using a five-level quality scale, ranging from worst to excellent. To facilitate consistent comparison with quantitative evaluation metrics such as VIEScores, we map these qualitative ratings to numerical scores: worst = 2, poor = 4, fair = 6, good = 8, and excellent = 10. For each editing task, we compute the mean preference score across all participants. The overall performance of each method is then summarized by averaging the scores over all editing tasks.

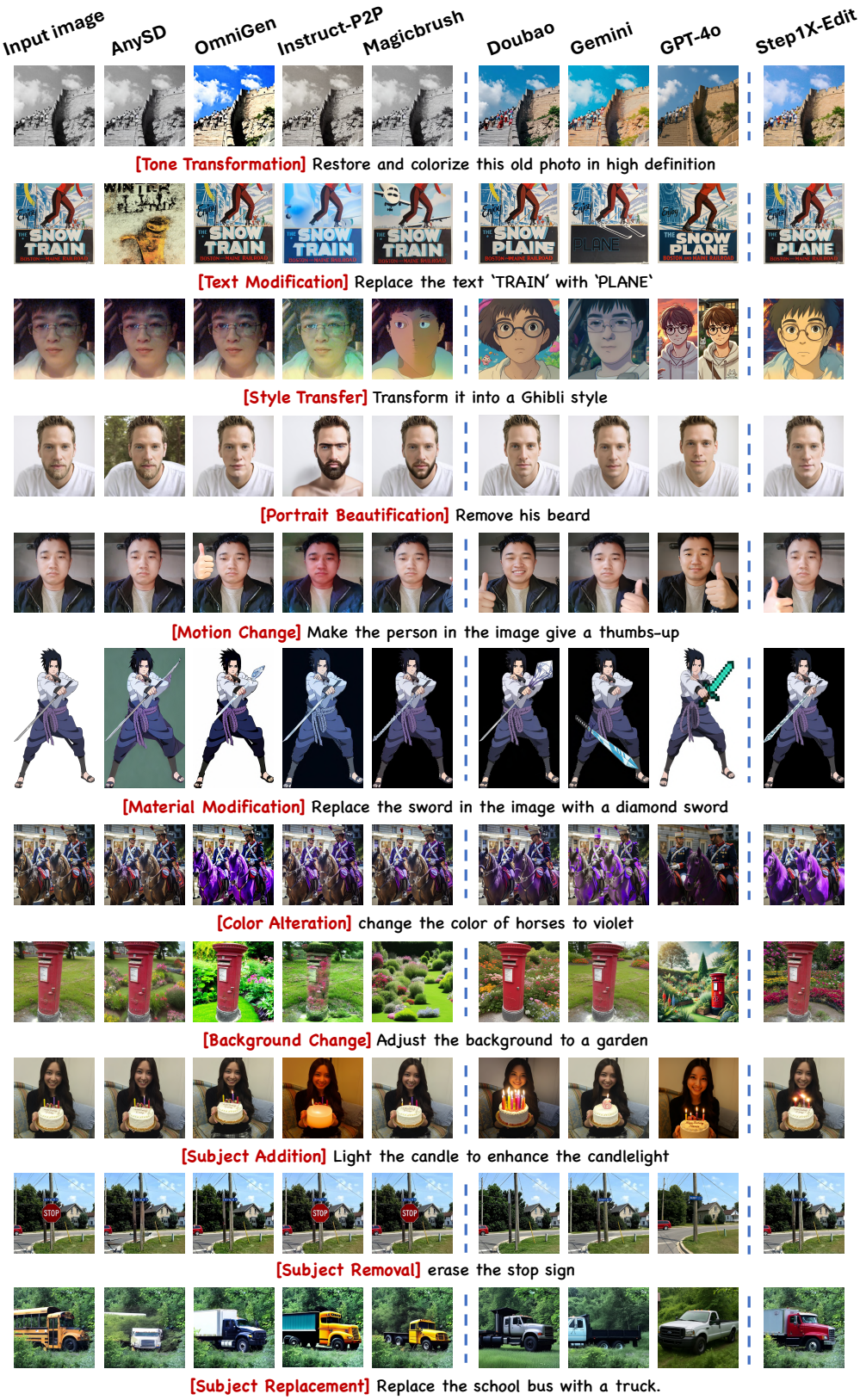


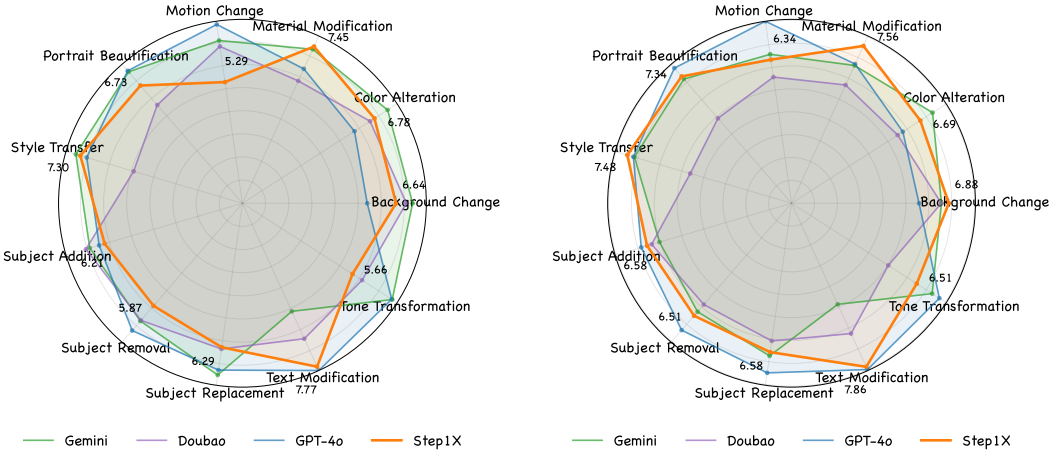
Figure 7: **A Comparative Illustration of Open-Source Approaches and Commercial systems for English Editing Instructions.**

Figure 8: **A Comparative Illustration of state-of-art algorithms for Chinese Editing Instructions.**

The results, presented in Tab. 4 and Fig. 9, highlight the effectiveness of Step1X-Edit. Notably, our method achieves comparable subjective quality to other state-of-the-art approaches, reinforcing its capability in producing visually pleasing and user-preferred edits. It is worth noting that Gemini2 Flash achieves an astonishingly high user preference score primarily attributed to its strong identity-preserving capabilities in the testing examples. This characteristic was more favored by the participants in the user study.

Model	Gemini [14]	Doubao [41]	GPT-4o [29]	Step1X-Edit
UP-IS (↑)	7.109	6.320	6.961	6.544
UP-Full (↑)	6.603	5.678	7.134	6.939

Table 4: **Overall user preference (UP) evaluation on GEdit-Bench.** UP-IS and UP-Full represent user preference score for Intersection subset (IS) and Full set (Full), respectively. All metrics are reported as higher-is-better (↑).



(a) User Preference score in the Intersection subset.

(b) User Preference score in the full subset.

Figure 9: **User Preference score of Each Sub-task in GEdit-Bench**.

5 Conclusion

In this report, we present a new general image editing algorithm called Step1X-Edit, which will be publicly released to foster further innovation and research within the image editing community. To train the model effectively, we propose a new data generation pipeline which can generate large-scale high-quality image editing triples, each consisting of a reference image, an editing instruction, and a corresponding target image. Based on the collected dataset, we train our Step1X-Edit model by seamlessly integrating powerful Multimedia Large Language Model with a diffusion-based image decoder. According to the evaluations on our collected GEdit-Bench, our proposed algorithm outperforms the existing open-source image editing algorithms with a substantial margin.

Contributors and Acknowledgments

We designate core contributors as those who have been involved in the development of Step1X-Edit throughout its entire process, while contributors are those who worked on the early versions or contributed part-time.

Core Contributors: Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Xianfang Zeng, Gang Yu.

Contributors: Honghao Fu, Ruoyu Wang, Yongliang Wu, Tianyu Wang, Haozhen Sun, Wen Sun, Bizhu Huang, Mei Chen, Kang An, Shuli Gao, Wei Ji, Tianhao You, Chunrui Han, Guopeng Li,

Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Daxin Jiang.

Corresponding Authors: Xianfang Zeng(zengxianfang@stepfun.com), Gang Yu (yugang@stepfun.com), Daxin Jiang (djiang@stepfun.com).

References

[1] Stability AI. Stable diffusion 3.5. `https://huggingface.co/stabilityai/stable-diffusion-3.5-large`, 2024. Accessed: 2025-04-17.

[2] Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shiliang Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 2023.

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[6] Black Forest Labs. Flux.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

[7] Black Forest Labs. Flux.1 fill [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2024. Accessed: 2025-04-19.

[8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2022.

[9] Bryan Cardenas, Devanshu Arya, and Deepak K Gupta. Generating annotated high-fidelity images containing multiple coherent objects. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 834–838. IEEE, 2021.

[10] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[12] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

[13] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

[14] Google Gemini2. Experiment with gemini 2.0 flash native image generation, 2025.

[15] Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ArXiv*, abs/2208.01626, 2022.

[16] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[17] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8362–8371, 2024.

[18] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

[19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[20] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.

[21] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024.

[22] Yaowei Li, Yuxuan Bian, Xu Ju, Zhaoyang Zhang, Ying Shan, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *ArXiv*, abs/2412.10316, 2024.

[23] Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024.

[24] lrzjason. Objectremovalalpha dataset. `https://huggingface.co/datasets/lrzjason/ObjectRemovalAlpha`, 2025. Accessed: 2025-04-19.

[25] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*, 2024.

[26] Jiancang Ma, Qirong Peng, Xu Guo, Chen Chen, H. Lu, and Zhenyu Yang. X2i: Seamless integration of multimodal understanding into diffusion transformer via attention distillation. *ArXiv*, abs/2503.06134, 2025.

[27] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025.

[28] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *ArXiv*, abs/2302.08453, 2023.

[29] OpenAI. Introducing 4o image generation, 2025.

[30] Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. *arXiv preprint arXiv:2503.14482*, 2025.

[31] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[32] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[37] Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv preprint arXiv:2406.05565*, 2024.

[38] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[39] Ying Shen, Yizhe Zhang, Shuangfei Zhai, Lifu Huang, Joshua M Susskind, and Jiatao Gu. Many-to-many image generation with auto-regressive diffusion models. *arXiv preprint arXiv:2404.03109*, 2024.

[40] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[41] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020.

[43] StepFun. step-1o-turbo-vision. `https://platform.stepfun.com/`, 2025.

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[45] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *ArXiv*, abs/2305.18047, 2023.

[46] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024.

[47] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023.

[48] Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. *arXiv preprint arXiv:2503.23897*, 2025.

[49] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. *arXiv preprint arXiv:2411.07199*, 2024.

[50] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[51] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

[52] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023.

[53] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *ArXiv*, abs/2401.11708, 2024.

[54] Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024.

[55] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

[56] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024.

[57] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023.

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[60] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024.

[61] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[62] Zibo Zhao, Wen Liu, Xin Chen, Xi Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *ArXiv*, abs/2306.17115, 2023.

[63] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

[64] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *arXiv preprint arXiv:2502.06734*, 2025.