X2Edit: Revisiting Arbitrary-Instruction Image Editing Through Self-Constructed Data and Task-Aware Representation Learning

Jian Ma[1*†], **Xujie Zhu**[2†‡], **Zihao Pan**[2‡], **Qirong Peng**[1], **Xu Guo**[3‡], **Chen Chen**[1*], **Haonan Lu**[1],
[1]OPPO AI Center
[2]Sun Yat-sen University
[3]Tsinghua University
majian2@oppo.com, zhuxj6@mail2.sysu.edu.cn, panzh33@mail2.sysu.edu.cn, pengjirong@oppo.com,
guo-x24@mails.tsinghua.edu.cn, chenchen4@oppo.com, luhaonan@oppo.com

Abstract

Existing open-source datasets for arbitrary-instruction image editing remain suboptimal, while a plug-and-play editing module compatible with community-prevalent generative models is notably absent. In this paper, we first introduce the X2Edit Dataset, a comprehensive dataset covering 14 diverse editing tasks, including subject-driven generation. We utilize the industry-leading unified image generation models and expert models to construct the data. Meanwhile, we design reasonable editing instructions with the VLM and implement various scoring mechanisms to filter the data. As a result, we construct 3.7 million high-quality data with balanced categories. Second, to better integrate seamlessly with community image generation models, we design task-aware MoE-LoRA training based on FLUX.1, with only 8% of the parameters of the full model. To further improve the final performance, we utilize the internal representations of the diffusion model and define positive/negative samples based on image editing types to introduce contrastive learning. Extensive experiments demonstrate that the model's editing performance is competitive among many excellent models. Additionally, the constructed dataset exhibits substantial advantages over existing open-source datasets.

Code — https://github.com/OPPO-Mente-Lab/X2Edit

Datasets — https://huggingface.co/datasets/OPPOer/X2Edit-Dataset

Introduction

Instruction-based image editing has witnessed explosive growth recently, evolving from single-task frameworks to more flexible models capable of open-vocabulary and free-form editing. Open-source models(Wang et al. 2025b; OpenAI et al. 2024; Team, Anil et al. 2025) continue to lag behind their closed-source counterparts. Despite a significant increase in the release of editing datasets, the challenge of constructing a well-balanced, high-quality dataset that encompasses a wide range of editing tasks remains unresolved.

Existing editing datasets have three main limitations: **cumbersome construction processes**, **poor data quality**, and **limited support for complex editing tasks**. (1) Editing datasets such as AnyEdit(Yu et al. 2025a) and ImgEdit(Ye et al. 2025) require different data construction processes to be designed for each type of editing task, which not only consume substantial human effort but is also difficult to scale flexibly. (2) Existing open-source datasets perform poorly in terms of editing accuracy and data balance. Datasets like AnyEdit and SEED-Data-Edit(Ge et al. 2024) suffer from low image quality and imbalanced data across different editing tasks. HQ-Edit(Hui et al. 2024) and OmniEdit(Wei et al. 2025) extensively use synthetic images as source images in hopes of improving quality, but this creates deviations from real data distributions. Moreover, existing workflows that leverage tools such as Flux-Fill(Labs 2024) and SD 3.5(Esser et al. 2024) ControlNet necessitate instruction conversion and the acquisition of additional prompt information, both of which can readily result in deviations from the intended edits. (3) High-quality open-source data for difficult editing tasks involving complex reasoning, camera movement, style transfer, etc., is extremely scarce due to the difficulty of construction.

We propose an automated dataset construction pipeline and a large-scale training dataset called **X2Edit Dataset**. We uniformly use Vision-Language Model(VLM) to generate quantity-balanced instructions for different editing tasks based on source images and contextual examples. Subsequently, we leverage SOTA open-source and closed-source models to create editing pairs according to task-specific workflows. Finally, we introduce a comprehensive filtering mechanism, including pre-filtering of source images and further screening after comprehensive scoring, strictly ensuring data quality. Through this pipeline, we ultimately obtain a category-balanced, high-quality dataset of **3.7M** scale, which demonstrates strong competitiveness compared to existing open-source datasets in terms of data scale, number of supported tasks, and data quality.

Parallel to this, the rapid emergence of arbitrary-instruction image editing methods has dramatically advanced text-guided visual manipulation. However, these gains are often offset by prohibitive training costs. Step1X-Edit(Liu et al. 2025) and Kontext(Labs et al. 2025), fine-

*Corresponding author.
†Co-first authors.
‡The author did his work during internship at OPPO AI Center.

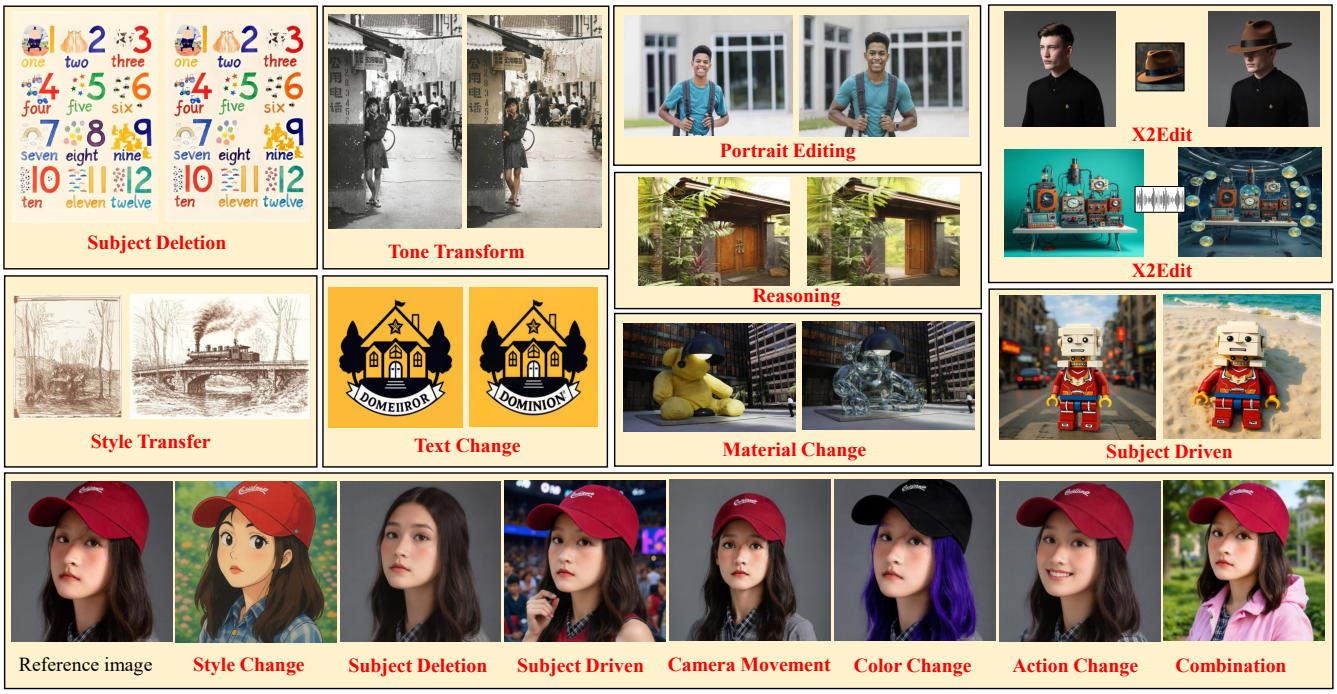

Figure 1: The X2Edit image generation results span 14 diverse editing types. In each unit, the left image serves as the reference. The central modality in the top-right unit is the input to X2I and can be leveraged by other modalities to assist in image editing.

tune the entire 12B-parameter DiT(Peebles and Xie 2023) backbone, whereas unified methods such as Bagel(Deng et al. 2025) and OmniGen2(Wu et al. 2025a) develop cross-modal understanding and generation capabilities from scratch on massive multimodal corpora. HiDream-E(Cai et al. 2025) follows the same recipe of end-to-end pre-training with a 17B-parameter model on ultra-large-scale data. ICEdit(Zhang et al. 2025b) inserts Mixture-of-Experts(MoE)(Shazeer et al. 2017) layers into LoRA modules based on FLUX-Fill. Despite effectively reducing training costs, it is limited by transferability and still falls short of full-model methods in terms of editing fidelity.

In this paper, we present **X2Edit**. Specifically, we first learn a task embedding matrix whose entries are injected into the MoE gating network to guide expert selection. We also analogize different editing tasks within the same batch to negative samples and the same editing task to positive samples in contrastive learning. This division can promote the mapping of different editing tasks to distinct, separable locations in the hidden space projection, enabling the model to learn discriminative features and avoid feature collapse. In addition, it can also ensure that the same editing task is mapped to similar encodings in the hidden space projection. Subsequently, through the constraint of the contrastive regularization loss, X2Edit achieves further enhancement in overall performance. Our contributions are three-fold:

- We construct X2Edit Dataset and GEdit-Bench++.
- To the best of our knowledge, X2Edit is an early attempt to explore the use of contrastive learning in arbitrary-instruction image editing.
- Extensive evaluations demonstrate that X2Edit Dataset surpasses existing open-source datasets across multiple objective metrics, while X2Edit rivals current SOTA editing methods in both automatic and human assessments.

Related Work

Datasets for Image Editing

Datasets	#Size	#Types	Res.(px)	Complex Tasks
AnyEdit	2.5M	25	512	✓
HQ-Edit	197K	6	$\geq$ 768	×
UltraEdit	4M	9	512	×
SEED-Data-Edit	3.7M	6	768	×
ImgEdit	1.2M	13	$\geq$ 1280	×
OmniEdit	5.2M/1.2M	7	$\geq$ 512	×
X2Edit(512)	2M	14	512	✓
X2Edit(1024)	1.7M	14	~1024	✓

Table 1: Comparison of existing image editing datasets.

We compare with current representative open-source arbitrary-instruction image editing datasets in Table 1. Represented by HQ-Edit, AnyEdit and OmniEdit, most existing datasets use automated pipelines to scale up as much as possible, while SEED-Data-Edit and UltraEdit(Zhao et al. 2024) add some manual quality control. The source images of HQ-Edit mostly come from synthetic data, causing the dataset to deviate from real-world images. Existing datasets also face problems of few editing categories and data imbalance. Although AnyEdit and ImgEdit enrich editing tasks, the overall pipeline suffers from critical flaws

including excessive complexity, poor replicability and low data quality. Moreover, existing image editing datasets often do not include some complex editing tasks such as reasoning, subject-driven generation and style transfer. While ensuring high-quality data construction, we streamline the pipeline as much as possible, using GPT-4o, BAGEL, and Kontext to supplement data for complex editing tasks.

Models for Image Editing

Arbitrary-instruction image editing is advancing along two primary paradigms. The first is full-parameter training, which encompasses unified vision editing models and unified understanding-generation models. In the former camp, Kontext and Hidream-E train large-scale networks on vast datasets, whereas Step1X-Edit performs full parameter fine-tuning on the entire DiT framework. In the latter camp, Bagel and OmniGen2 pursue a unified understanding-generation model via joint pre-training on massive data, yet incur prohibitive compute and training costs. For AnyEdit(Yu et al. 2025a), the first training stage involves pre-training the UNet backbone of the diffusion process, while the second stage separately trains IP-Adapter as task-specific expert with MoE layers. The second is parameter-efficient fine-tuning to drastically cut training costs. ICEdit integrates MoE layers within LoRA modules for a lightweight approach. UniControl(Qin et al. 2023) introduces a task-aware HyperNet to modulate the diffusion models. Conversely, while contrastive learning has proven effective for discriminative tasks, its application in generative models remains a nascent yet promising direction. REPA(Yu et al. 2025b) leverages external alignment for semantic inheritance at higher costs, whereas Dispersive Loss(Wang and He 2025) optimizes internal representations without external dependencies, both enhancing text-to-image generation.

X2Edit Dataset

We propose **X2Edit Dataset**, a large-scale, high-quality, and well-balanced dataset tailored for arbitrary-instruction image editing, with its construction process depicted in Figure 2. Additional details can be found in the appendix D.

Edit Type Definition

To meet diverse user/setting image editing needs, we categorize 14 editing tasks covering local, global, complex editing and subject-driven generation. Local editing modifies specific regions without changing other areas; global editing impacts the entire image. Notably, complex tasks demand models' full understanding of instructions and source images. We also introduce subject-driven generation, rarely featured in existing open-source datasets.

Automatic Dataset Pipeline

Source Image Preparation. We select the majority of the source images from COYO-700M(Byeon et al. 2022), Wukong(Gu et al. 2022) and LAION(Schuhmann et al. 2022), aiming to thoroughly encompass a diverse range of real-world image inputs. In order to fulfill the need for high-quality reference image in subject-driven generation, we also employ an internal query to generate source images using Shuttle-3-Diffusion(Liu 2024). Rigorous filtering criterias are applied to samples that advance to the subsequent phase of the data construction pipeline. Specifically, we only retain images with high aesthetic score and minimum side lengths greater than 512 pixels. Furthermore, we filter the internal query for subject-driven generation with Qwen3(Yang, Li et al. 2025) to ensure the presence of keywords related to the subject in the foreground. Additionally, the data we construct encompasses a range of aspect ratios between 512 and 2048 to enhance diversity.

Diverse Editing Instruction Generation. Since image captions cannot capture rich visual detail information, we employ Qwen2.5-VL-7B(Bai et al. 2025) to directly generate diverse editing instructions based on source images, which differs from existing LLM-based methods. We provide source images, task definitions, and contextual examples, and use meticulously crafted prompts to guide the VLM in formulating editing instructions directly from the images. To minimize the risk that the VLM generates unfeasible instructions due to hallucinations, we incorporate a self-reflection mechanism that enables the VLM to verify the validity of its generated instructions. Additionally, we devise a load balancing strategy to ensure a balanced distribution of editing instructions across various tasks.

Edited Image Construction. We select open-source and closed-source models to create a variety of data construction pipelines tailored to the specifics of different tasks. For instance, in tasks such as subject addition and deletion, which demand high consistency in non-target areas, we employ RAM++(Huang et al. 2023) and SAM2(Ravi et al. 2024) on the original images and subsequently apply LaMa(Suvorov et al. 2021) to generate edited images. Our experiments indicate that this methodology results in superior data quality for these tasks compared to models like GPT-4o, Kontext, etc. Details of this comparison are provided in the appendix D. For more general editing tasks, we utilize Step1X-Edit to generate the needed data. In cases where Step1X-Edit performs poorly, such as style modification, we incorporate additional data using OmniConsistency(Song, Liu, and Shou 2025), TextFlux(Xie et al. 2025), and Kontext. For the challenging construction of complex reasoning and camera movement tasks, we employ GPT-4o and BAGEL. Furthermore, we leverage GPT-4o and Kontext to work with 1024-resolution images with varying aspect ratios to generate data that satisfies requirements for high fidelity.

Post Quality Enhancement. In order to guarantee the quality of the data, we implement an extensive framework for data quality evaluation and filtering. We specifically derive aesthetic score, as well as LIQE(Zhang et al. 2023) and CLIPIQA(Wang, Chan, and Loy 2022) scores for all the generated images, thus facilitating a assessment of image quality and the exclusion of samples that fall below predetermined standards. To ensure the precision of editing, we employ ImgEdit-Judge(Ye et al. 2025) and Qwen2.5-VL-72B to assess and filter images according to editing instructions, source images, and edited images. For subject-driven generation, we utilize CLIP(Hessel et al. 2021) and DINO(Caron

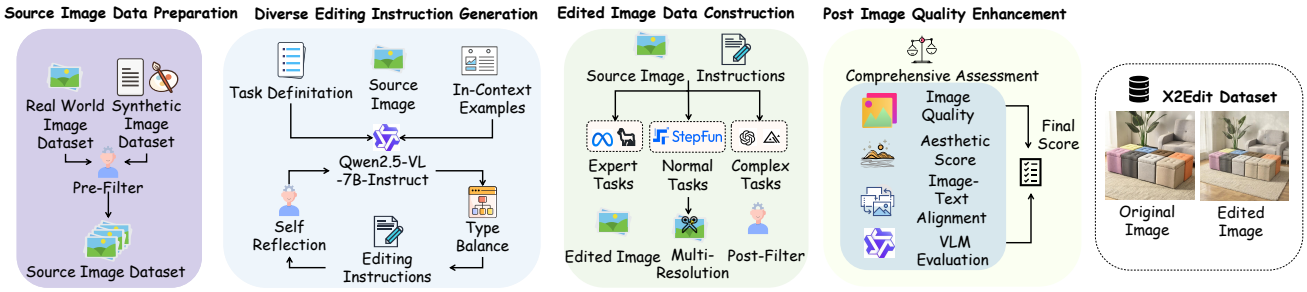

Figure 2: The comprehensive construction pipeline of X2Edit Dataset. We divide the pipeline into four stages: (1) Sampling from real-world datasets and synthesizing source images using our internal query dataset; (2) Generating diverse editing instructions using a VLM based on the source images; (3) Generating edited images using task-specific workflows according to the editing instructions; (4) Conducting comprehensive evaluation and filtering of all generated data to ensure quality.

et al. 2021) to assess subject consistency. For style transfer, Qwen2.5-VL-7B evaluates the stylistic match between source and generated images, and applies filters accordingly.

The X2Edit Dataset comprises 3.7 million pairs of high-quality image across 14 categories. It outperforms existing datasets in terms of task diversity, scale, and resolution, as shown in Table 1. Notably, it includes 342K reasoning editing samples, 94K camera movement samples, and 460K subject-driven generation samples, addressing the scarcity of such data in current open-source datasets. To evaluate editing precision and data quality, we randomly select 1k samples from each dataset and perform comprehensive evaluations using models such as Qwen2.5-VL-72B, ImgEdit-Judge, and GPT-4o. The results indicate that the X2Edit Dataset is competitive against other open-source image editing datasets in terms of VLM evaluations, aesthetics, and overall image quality. Appendix provides details on these task categories and their statistical data.

Methodology

Model Overview

As shown in Figure 3, we adopt X2I(Ma et al. 2025) as our backbone. The injection of reference image information draws on classic strategies(Tan et al. 2025a; Zhang et al. 2025a; Tan et al. 2025b; Wang et al. 2025a; Mou et al. 2025; Ma et al. 2023) of concatenating noise. During training, the AlignNet branch within X2I, the task-embedding matrix, and MoE-LoRA parameters are simultaneously updated. We apply task-aware contrastive regularization to the intermediate features of all MMDiT blocks, enforcing a structured hidden space. During inference, we optionally deploy Qwen3 to predict the editing task. We can further leverage X2I's multimodal inputs to enrich the textual editing instructions as shown in the top-right unit of Figure 1.

Task-Aware MoE-LoRA

Image editing tasks span from low-level manipulations to high-level semantic edits. Task heterogeneity can lead to parameter inefficiency when using a single model, as shared parameters internalize interference-prone representations, causing suboptimal specialization and increased parameter numbers. Task-aware MoE addresses this by activating sparse expert sub-networks based on the editing task, allocating capacity precisely where needed.

MoE consists of N_e expert networks, a shared expert network, and a gating network. The gating network outputs an N_e-dimensional vector representing the scores of experts, which are subsequently converted into normalized weights via a softmax function. We select the top-K experts with the highest weights for token processing, ensuring computational efficiency through sparse activation. The shared expert processes all tokens to balance knowledge distribution and eliminate redundant storage of shared representations.

We define $h^l \in \mathbb{R}^{b\times n\times d}$ as the intermediate representations at the l-th layer of the MMDiT. Let $t_{\text{emb}} \in \mathbb{R}^{N_t\times c}$ represents task embeddings and $y \in \mathbb{R}^b$ (where $y_i \in 1, 2, \ldots, N_t$) indicate the task type corresponding to the b samples of h^l, where N_t denotes the number of editing task types. We first extract the task-specific embedding corresponding to h^l from t_{emb}, then reshape and expand it to enable channel-wise concatenation with h^l:

$$t^h_{\text{emb}} = \text{Expand}(\text{Reshape}(t_{\text{emb}}[y], (b, 1, c))). \tag{1}$$

Subsequently, we process the concatenation of t^h_{emb} and h^l through the gating network and the softmax function to obtain the weights of N_e experts:

$$s_i = \text{Softmax}_i\left(\text{Gate}(\text{Concat}(h^l, t^h_{\text{emb}}))\right). \tag{2}$$

Finally, we select the top-K experts with the highest weights and aggregate the outputs of these experts with the corresponding gating weights through weighted summation to yield the MoE outputs q^l_{moe}, k^l_{moe}, and v^l_{moe}:

$$x^l_{\text{moe}} = \sum_{i=1}^{N_e}\left(g_i \text{Expert}^i_x\left(h^l\right)\right) + \text{SharedExpert}_x\left(h^l\right), \text{ for } x \in \{q, k, v\}, \tag{3}$$

$$g_i = \begin{cases} s_i, & s_i \in \text{Topk}\left(\{s_j | 1 \leqslant j \leqslant N_e\}, K\right), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

we add q^l_{moe}, k^l_{moe}, and v^l_{moe} to the corresponding outputs from the linear projector of MMDiT, respectively, and then

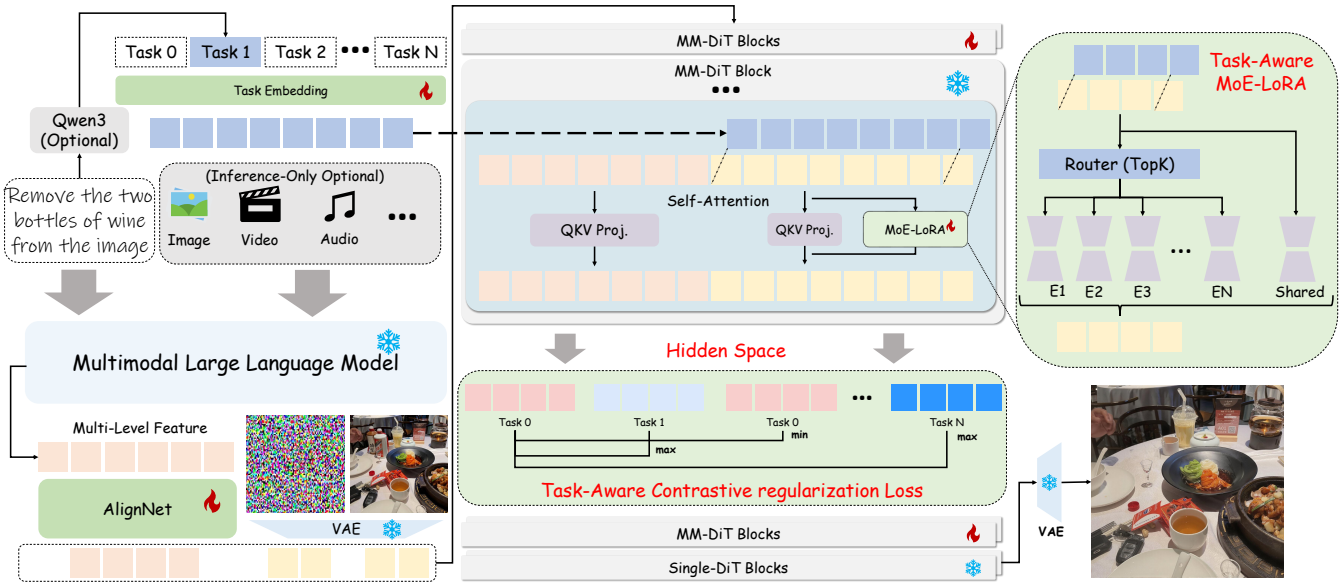

Figure 3: X2Edit consists of an MLLM for editing instruction understanding, a DiT fine-tuned based on FLUX.1, an optional intent perception model, and task embeddings. We introduce a task-aware MoE-LoRA structure and task-aware contrastive learning into the DiT to enhance the unified editing model's ability to perceive different editing tasks.

execute the self-attention mechanism.

$$h^l_{\text{moe}} = \text{Attention}(q^l_{\text{moe}} + q^l, k^l_{\text{moe}} + k^l, v^l_{\text{moe}} + v^l), \quad (5)$$

where q^l, k^l, and v^l denote the vectors projected from h^l through the linear projector of MMDiT.

Task-aware Contrastive Learning

Current diffusion models primarily rely on regression objective for training, generally lacking explicit regularization of internal representations. To address this, we introduce a task-aware contrastive regularization term during diffusion model training. This explicitly structures the hidden space by dispersing different representations while collapsing similar representations into compact regions, thereby enhancing inter-class separability and intra-class compactness.

Unlike Dispersive Loss which employs a contrastive loss without positives, we adopt a standard contrastive loss. For samples from the same editing task, we enforce region-consistent representations in the hidden space, while promoting maximal dispersion across different tasks to ensure inter-class separability. For the construction of positives and negatives, whereas conventional contrastive learning(Chen et al. 2020) constructs positives through self-supervision and treats all other images in the batch as negatives, we leverage task labels to construct semantically meaningful samples: all intra-task samples within a batch form positives, while inter-task pairs automatically serve as negatives.

We first perform L2 normalization on h^l to increase training stability and linear separability in the hidden space, and calculate its corresponding distance matrix $D \in \mathbb{R}^{b\times b}$:

$$h^l_{\text{norm}} = \text{Normalize}(\text{Reshape}(h^l, (b, n \times d))), \quad (6)$$

$$D_{ij} = \|h^l_{\text{norm}}(i) - h^l_{\text{norm}}(j)\|_2^2, \quad (7)$$

where D_{ij} denotes the Squared Euclidean Distance between the representations of the i-th sample and j-th sample. We then calculate the task mask $M \in \mathbb{R}^{b\times b}$ and apply the InfoNCE(van den Oord, Li, and Vinyals 2019) loss:

$$M_{ij} = \begin{cases} 1 & \text{if } y_i = y_j \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

$$\mathcal{L}_{\text{task}} = -\frac{1}{b}\sum_{i=1}^{N} \log\left(\frac{\sum_{j=1}^{N} \exp(-\frac{D_{ij}}{\tau}) \cdot M_{ij}}{\sum_{k=1}^{N} \exp(-\frac{D_{ik}}{\tau})}\right), \quad (9)$$

where τ is the temperature hyperparameter that controls the model's discrimination against negatives.

In practice, we add this term to the original diffusion-based objectives $\mathcal{L}_{\text{diff}}$, the final objective becomes:

$$\mathcal{L} := \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{diff}}, \quad (10)$$

where λ is a hyperparameter that controls the tradeoff between denoising and contrastive learning.

Experiment

Implementation

Our model configuration uses a LoRA rank of 64, sets the number of expert networks N_e to 12, the number of editing task types N_t to 15(includes an "other" editing task), the number of activated experts K to 2, the regularization parameter λ to 0.2 and the temperature parameter τ to 0.5. We train the X2Edit model using 48 H20 GPUs. We conduct 16k training steps at 512-resolution with a micro-batch size of 12, followed by 5k training steps at 1024-resolution with a micro-batch size of 4 on the same GPUs. The random seed for inference on all evaluation is set to 1.

Benchmark Suite

Arbitrary-Instruction Image Editing. Evaluation datasets include GEdit-Bench++, ImgEdit-Bench, AnyEdit-Test, and KontextBench(Labs et al. 2025), each of which includes multiple types of editing tasks. GEdit-Bench++ extends GEdit-Bench by incorporating reasoning and camera movement, expanding the total task count from 11 to 13. For the construction of evaluation data for reasoning and camera movement, we manually select 50 real-world images and generate appropriate instructions in both Chinese and English based on the content of the images. **Comparative methods** include closed-source models SeedEdit(Wang et al. 2025b) and GPT-4o, as well as open-source models Step1X-Edit, Kontext, Bagel, OmniGen2, ICEdit, Hidream-E, and AnyEdit. All models are evaluated using the same prompts and images. **Evaluation metrics** employ VIEScore(Ku et al. 2023) and ImgEdit-Judge Score. VIEScore includes three metrics: Semantic Consistency(SC), Perceptual Quality(PQ), and Overall Score(O). SC measures the alignment of the generated image with prompt, while PQ assesses the visual authenticity and naturalness of the generated image and $\mathrm{O} = \sqrt{\mathrm{SC} \times \mathrm{PQ}}$. Similar to Step1X-Edit, we use GPT-4o and Qwen2.5-VL-72B to automatically calculate the VIEScore. ImgEdit-Judge Score assesses the degree of instruction following and absence of unintended changes in the generated images.

Subject-Driven Generation. Evaluation dataset employs classical DreamBench(Ruiz et al. 2023). **Comparative methods** include recent methods such as X2I, UNO(Wu et al. 2025b), Kontext, Bagel, OmniGen2, OmniGen(Xiao et al. 2024), ACE++(Mao et al. 2025) and Previous methods. **Evaluation metrics** include the average DINO and CLIP-I scores between source images and generated images, as well as the CLIP-T score between generated images and prompts.

Quantitative Results

Arbitrary-Instruction Image Editing. Table 2 shows the performance of different methods on four evaluation datasets. In the evaluation on GEdit-Bench++, X2Edit achieves scores of 8.313, 6.354, and 5.639 for the Chinese sub-dataset, and scores of 8.334, 6.158, and 5.55 for the English sub-dataset. These results indicate that X2Edit is on par with other mainstream open-source methods such as Bagel and Kontext, while surpassing Step1X-Edit and OmniGen2 across multiple metrics. Additionally, it significantly outperforms ICEdit, Hidream-E and AnyEdit across all metrics. X2Edit exhibits robust cross-lingual capability, appendix G demonstrates more language tests. On the ImgEdit-Bench and AnyEdit-Test benchmarks, X2Edit achieves performance comparable to that of GEdit-Bench++. It is close to Kontext and Bagel in most metrics, outperforms OmniGen2 and Step1X-Edit in the majority of indicators, and significantly surpasses ICEdit, Hidream-E, and AnyEdit across all metrics. The KontextBench is somewhat different from the other three benchmarks, especially under the evaluation systems of Qwen2.5-VL-72B and GPT-4o. X2Edit shows a significant gap compared to the two open-source methods, Kontext and Bagel. Nevertheless, X2Edit's support for style transfer and subject-driven generation also results in a significant performance gap compared to the other five comparative methods. KontextBench comprises 1,026 unique image-prompt pairs derived from 108 base images, including personal photos, CC-licensed art, public domain images, and AI-generated content. Compared with other benchmarks, it is relatively complex. Additional subjective comparisons are provided in the appendix H.

Subject-Driven Generation. X2Edit has a slight advantage in the DINO score on DreamBench, while it holds a middle position in the CLIP-I score. In terms of the CLIP-T metric, X2Edit slightly lags behind OmniControl, OmniGen2, and Bagel. However, it outperforms these three methods in the other two image fidelity metrics. Overall, X2Edit still demonstrates a relatively competitive result.

Summary. Overall, X2Edit holds its own among top-tier performers. While it still has room to close the gap with closed-source models, it shows a clear edge over fully trained models like Step1X-Edit, OmniGen2, and HiDream-E. It runs neck and neck with Bagel and Kontext, and it markedly surpasses lightweight fine-tuning models such as ICEdit; On DreamBench, X2Edit exhibits enhanced subject-driven generation surpassing most competitors. We additionally assess out-of-domain generalization across complicated editing tasks. Please see the appendix F for details.

Plug-and-Play

As shown in Table 4, we validate the plug-and-play capability of X2Edit on two types of modules with high frequency of use in the Flux.1 community. We provide more subjective results in the appendix E.

FLUX.1 Dev Variants. FLUX.1-Schnell and Shuttle-3-Diffusion are 4-step accelerated models. PixelWave can generate images of multiple artistic styles. FLUX.1-Krea-dev(Lee, Ebbecke et al. 2025) offers strong performance with highly distinctive aesthetics and exceptional realism. The experimental results indicate that through flexible plug-and-play capability, we can improve inference speed while maintaining comparable performance. **FLUX.1 Dev LoRA Ecosystem**. FLUX.1-Turbo-Alpha supports 8-step inference. FLUX.1-dev-LoRA-AntiBlur enhances depth of field and clarity, while FLUX-Midjourney-Mix2-LoRA emulates MidJourney v6's distinctive aesthetic. FLUX-Super-Realism-LoRA and FLUX-Chatgpt-Ghibli-LoRA specialize in stylization. The experimental results demonstrate that X2Edit seamlessly adapts to the FLUX.1 ecosystem.

Ablation Study

We design a comprehensive suite of ablation experiments on both English and Chinese sub-datasets of GEdit-Bench++ with ImgEdit-Judge, systematically dissecting the contributions of MoE routing, task-awareness(TA), expert cardinality, and the contrastive regularization. First, a vanilla single-rank LoRA baseline confirms its limited general-purpose editing capability. Incorporating task priors into our MoE-LoRA yields consistent positive gains on all metrics in Table 5. Notably, increasing the number of experts while proportionally reducing the LoRA rank, which preserves the total parameter budget, delivers further improvements.

Methods	Params	US	GEdit-Bench++_CN			GEdit-Bench++_EN			ImgEdit-Bench			AnyEdit-Test			KontextBench		
			IJ	Q_VIE	G_VIE	IJ	Q_VIE	G_VIE	IJ	Q_VIE	G_VIE	IJ	Q_VIE	G_VIE	IJ	Q_VIE	G_VIE
GPT-4o	-	2.911	9.062	7.706	7.943	9.003	7.684	7.848	8.202	7.634	7.328	8.460	7.414	6.836	8.358	7.281	7.078
Seededit	-	2.660	8.686	6.215	6.460	8.604	6.449	6.344	8.143	6.820	6.552	8.204	6.677	5.840	8.106	6.770	5.983
Kontext	12B	2.881	-	-	-	8.408	6.170	5.712	8.149	6.087	5.258	8.110	5.419	4.900	8.095	6.250	5.718
Bagel	7B+7B	2.632	8.461	6.649	5.627	8.326	6.748	5.722	7.925	6.748	6.022	7.960	6.292	5.451	7.929	6.586	5.880
Omnigen2	3B+4B	2.427	8.001	4.265	4.199	7.973	4.523	4.321	8.018	5.797	5.482	7.609	4.162	3.825	7.548	4.879	4.278
Step1X-Edit	12B	2.305	8.146	5.994	5.430	8.017	5.844	5.108	7.653	6.064	5.425	7.753	5.342	4.674	7.476	5.272	4.481
Hidream-E	17B	2.198	-	-	-	7.461	5.630	4.257	7.264	6.079	4.344	6.923	5.014	3.282	7.154	4.860	3.770
ICEdit	0.2B	2.036	-	-	-	7.203	4.984	4.109	7.615	5.443	5.228	7.498	4.510	3.782	7.192	4.390	3.720
AnyEdit	0.9B	-	-	-	-	6.841	2.608	2.242	6.784	3.991	3.448	7.078	2.884	2.528	6.452	1.959	1.760
X2Edit	0.9B	2.432	8.313	6.354	5.639	8.334	6.158	5.550	8.025	5.987	5.402	8.095	5.578	5.147	7.606	5.768	5.214

Table 2: Objective performance on four benchmark datasets. IJ, Q_VIE, and G_VIE refer to the overall score evaluated by ImgEdit-Judge, the VIEScore calculated by Qwen2.5-VL-72B and GPT-4o. US refer to the user study on GEdit-Bench++.

Methods	Base	DINO	CLIP-I	CLIP-T
Textual Inversion		0.569	0.780	0.255
DreamBooth		0.668	0.803	0.305
BLIP-Diffusion		0.594	0.779	0.300
Subject-Diffusion	SD	0.711	0.787	0.293
IP-Adapter		0.667	0.813	0.289
KOSMOS-G		0.694	0.847	0.287
SuTI	Imagen	0.741	0.819	0.304
OmniGen	Phi-3	-	0.801	0.315
OmniGen2	Qwen-VL	0.807	0.814	0.332
Bagel	Qwen	0.739	0.756	0.332
UNIC-Adapter	SD3	0.816	0.841	0.306
IP-Adapter		0.768	0.803	0.322
OminiControl		0.740	0.768	0.329
Kontex		0.822	0.839	0.322
UNO	FLUX.1	0.764	0.806	0.319
ACE++		0.760	0.787	0.319
X2I		0.817	0.826	0.304
X2Edit	FLUX.1	0.822	0.826	0.326

Table 3: Performance of different methods on DreamBench.

Models	Steps	GC	GE
FLUX.1-Schnell	4	8.254	8.085
Shuttle-3-Diffusion	4	8.401	8.185
FLUX.1-Krea-dev	28	8.412	8.367
PixelWave	28	8.458	8.324
FLUX.1-Turbo-Alpha	8	8.258	8.154
FLUX.1-dev-LoRA-AntiBlur	28	8.414	8.104
FLUX-Midjourney-Mix2-LoRA	28	8.399	8.237
FLUX-Super-Realism-LoRA	28	8.247	8.115
FLUX-Chatgpt-Ghibli-LoRA	28	7.954	7.885
X2Edit	28	8.313	8.334

Table 4: X2Edit transfers seamlessly to Flux.1-based modules. GC and GE refer to the Chinese and English subset of GEdit-Bench++, we utilize ImgEdit-Judge for evaluation.

This evidences that expert granularity trumps capacity: task-specific low-rank sub-spaces retain ample discriminative power, and the routing diversity from more experts offsets any capacity loss from rank compression, thus affirming the "narrow-yet-numerous" expert approach for diverse editing tasks. Furthermore, we conduct three ablation studies specifically on the contrastive regularization loss. Firstly, we use

Methods	Loss	Layers	Experts	Rank	GC	GE
LoRA	-	-	-	512	7.834	7.649
MoE-LoRA	-	-	6	128	7.943	7.751
MoE-LoRA w/TA	-	-	6	128	8.087	7.985
MoE-LoRA w/TA	-	-	12	64	8.161	8.084
X2Edit	cosine	4	12	64	8.253	8.113
X2Edit	L2	4	12	64	8.289	8.120
X2Edit	L2	all	12	64	8.313	8.334

Table 5: Ablation results on GEdit-Bench++.

the conventional cosine similarity and select representations in the fourth layer of MMDiT. The experiments show that the inclusion of regularization loss further enhance the overall performance. Similar to dispersive loss, we explore the impact of different similarity metrics and the position of the regularization layer on performance. Ultimately, we choose squared L2 distance as the similarity metric and calculate the regularization loss using the representations in all 19 layers.

User Study

We recruited four participants to evaluate X2Edit and comparative methods on 1.3k editing pairs from GEdit-bench++, focusing on instruction following (model's instruction execution) and image fidelity (edited image quality/consistency). Scores (0=poor, 1=fair, 2=good) are summed for an overall score; Table 2 Column 3 shows the mean overall score across participants. We average results for models supporting both Chinese and English on GEdit-Bench++, excluding AnyEdit (subpar subjective performance, no human annotations). Our method demonstrates competitiveness; detailed chart experiments are in Appendix B.

Conclusion

In this paper, we construct and release X2Edit Dataset, a large-scale, high-quality 3.7M-sample corpus spanning 14 editing tasks, alongside X2Edit—a lightweight, plug-and-play editing model with a novel framework. Extensive experiments show the dataset outperforms existing open-source counterparts in quality and task diversity, while the model matches or outperforms state-of-the-art (SOTA) models on multiple benchmarks.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Cai, Q.; Chen, J.; Chen, Y.; Li, Y.; Long, F.; Pan, Y.; Qiu, Z.; Zhang, Y.; Gao, F.; Xu, P.; et al. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; Shi, G.; and Fan, H. 2025. Emerging Properties in Unified Multimodal Pretraining. arXiv:2505.14683.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206.

Ge, Y.; Zhao, S.; Li, C.; Ge, Y.; and Shan, Y. 2024. SEED-Data-Edit Technical Report: A Hybrid Dataset for Instructional Image Editing. arXiv:2405.04007.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Niu, M.; Xu, H.; Liang, X.; Zhang, W.; Jiang, X.; and Xu, C. 2022. Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework. arXiv:2202.06767.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Huang, X.; Huang, Y.-J.; Zhang, Y.; Tian, W.; Feng, R.; Zhang, Y.; Xie, Y.; Li, Y.; and Zhang, L. 2023. Open-Set Image Tagging with Multi-Grained Text Supervision. arXiv:2310.15200.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2024. HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing. *arXiv preprint arXiv:2404.09990*.

Ku, M.; Jiang, D.; Wei, C.; Yue, X.; and Chen, W. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

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

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

Lee, S.; Ebbecke, T.; et al. 2025. FLUX.1 Krea [dev]. https://github.com/krea-ai/flux-krea.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; Li, G.; Peng, Y.; Sun, Q.; Wu, J.; Cai, Y.; Ge, Z.; Ming, R.; Xia, L.; Zeng, X.; Zhu, Y.; Jiao, B.; Zhang, X.; Yu, G.; and Jiang, D. 2025. Step1X-Edit: A Practical Framework for General Image Editing. arXiv:2504.17761.

Liu, T. 2024. Shuttle-3-Diffusion. https://huggingface.co/shuttleai/shuttle-3-diffusion. Accessed: 2025-07-29.

Ma, J.; Peng, Q.; Guo, X.; Chen, C.; Lu, H.; and Yang, Z. 2025. X2I: Seamless Integration of Multimodal Understanding into Diffusion Transformer via Attention Distillation. arXiv:2503.06134.

Ma, J.; Zhao, M.; Chen, C.; Wang, R.; Niu, D.; Lu, H.; and Lin, X. 2023. GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation. arXiv:2303.17870.

Mao, C.; Zhang, J.; Pan, Y.; Jiang, Z.; Han, Z.; Liu, Y.; and Zhou, J. 2025. ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling. arXiv:2501.02487.

Mou, C.; Wu, Y.; Wu, W.; Guo, Z.; Zhang, P.; Cheng, Y.; Luo, Y.; Ding, F.; Zhang, S.; Li, X.; Li, M.; Liu, M.; Zhang, Y.; Wu, S.; Zhao, S.; Zhang, J.; He, Q.; and Wu, X. 2025. DreamO: A Unified Framework for Image Customization. arXiv:2504.16915.

OpenAI; :; Hurst, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; Ermon, S.; Fu, Y.; and Xu, R. 2023. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. arXiv:2305.11147.

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.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538.

Song, Y.; Liu, C.; and Shou, M. Z. 2025. OmniConsistency: Learning Style-Agnostic Consistency from Paired Stylization Data. arXiv:2505.18445.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2025a. OminiControl: Minimal and Universal Control for Diffusion Transformer. arXiv:2411.15098.

Tan, Z.; Xue, Q.; Yang, X.; Liu, S.; and Wang, X. 2025b. OminiControl2: Efficient Conditioning for Diffusion Transformers. arXiv:2503.08280.

Team, G.; Anil, R.; et al. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wang, H.; Peng, J.; He, Q.; Yang, H.; Jin, Y.; Wu, J.; Hu, X.; Pan, Y.; Gan, Z.; Chi, M.; Peng, B.; and Wang, Y. 2025a. UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer. arXiv:2503.09277.

Wang, J.; Chan, K. C. K.; and Loy, C. C. 2022. Exploring CLIP for Assessing the Look and Feel of Images. arXiv:2207.12396.

Wang, P.; Shi, Y.; Lian, X.; Zhai, Z.; Xia, X.; Xiao, X.; Huang, W.; and Yang, J. 2025b. SeedEdit 3.0: Fast and High-Quality Generative Image Editing. *arXiv preprint arXiv:2506.05083*.

Wang, R.; and He, K. 2025. Diffuse and Disperse: Image Generation with Representation Regularization. arXiv:2506.09027.

Wei, C.; Xiong, Z.; Ren, W.; Du, X.; Zhang, G.; and Chen, W. 2025. OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision. arXiv:2411.07199.

Wu, C.; Zheng, P.; Yan, R.; Xiao, S.; Luo, X.; Wang, Y.; Li, W.; Jiang, X.; Liu, Y.; Zhou, J.; Liu, Z.; Xia, Z.; Li, C.; Deng, H.; Wang, J.; Luo, K.; Zhang, B.; Lian, D.; Wang, X.; Wang, Z.; Huang, T.; and Liu, Z. 2025a. OmniGen2: Exploration to Advanced Multimodal Generation. arXiv:2506.18871.

Wu, S.; Huang, M.; Wu, W.; Cheng, Y.; Ding, F.; and He, Q. 2025b. Less-to-More Generalization: Unlocking More Controllability by In-Context Generation. arXiv:2504.02160.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2024. OmniGen: Unified Image Generation. arXiv:2409.11340.

Xie, Y.; Zhang, J.; Chen, P.; Wang, Z.; Wang, W.; Gao, L.; Li, P.; Sun, H.; Zhang, Q.; Qiao, Q.; Fan, J.; and Lian, Z. 2025. TextFlux: An OCR-Free DiT Model for High-Fidelity Multilingual Scene Text Synthesis. arXiv:2505.17778.

Yang, A.; Li, A.; et al. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Ye, Y.; He, X.; Li, Z.; Lin, B.; Yuan, S.; Yan, Z.; Hou, B.; and Yuan, L. 2025. ImgEdit: A Unified Image Editing Dataset and Benchmark. arXiv:2505.20275.

Yu, Q.; Chow, W.; Yue, Z.; Pan, K.; Wu, Y.; Wan, X.; Li, J.; Tang, S.; Zhang, H.; and Zhuang, Y. 2025a. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. arXiv:2411.15738.

Yu, S.; Kwak, S.; Jang, H.; Jeong, J.; Huang, J.; Shin, J.; and Xie, S. 2025b. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. arXiv:2410.06940.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. arXiv:2303.14968.

Zhang, Y.; Yuan, Y.; Song, Y.; Wang, H.; and Liu, J. 2025a. EasyControl: Adding Efficient and Flexible Control for Diffusion Transformer. arXiv:2503.07027.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025b. In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer. arXiv:2504.20690.

Zhao, H.; Ma, X.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. arXiv:2407.05282.