Generative Visual Chain-of-Thought for Image Editing

Zijin Yin[1,2,3†] Tiankai Hang[3] Yiji Cheng[3] Shiyi Zhang[3] Runze He[3] Yu Xu[3]
Chunyu Wang[3‡] Bing Li[4] Zheng Chang[1] Kongming Liang[1,2§] Qinglin Lu[3] Zhanyu Ma[1,2]
[1]Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance
[3]Tencent Hunyuan [4]King Abdullah University of Science and Technology

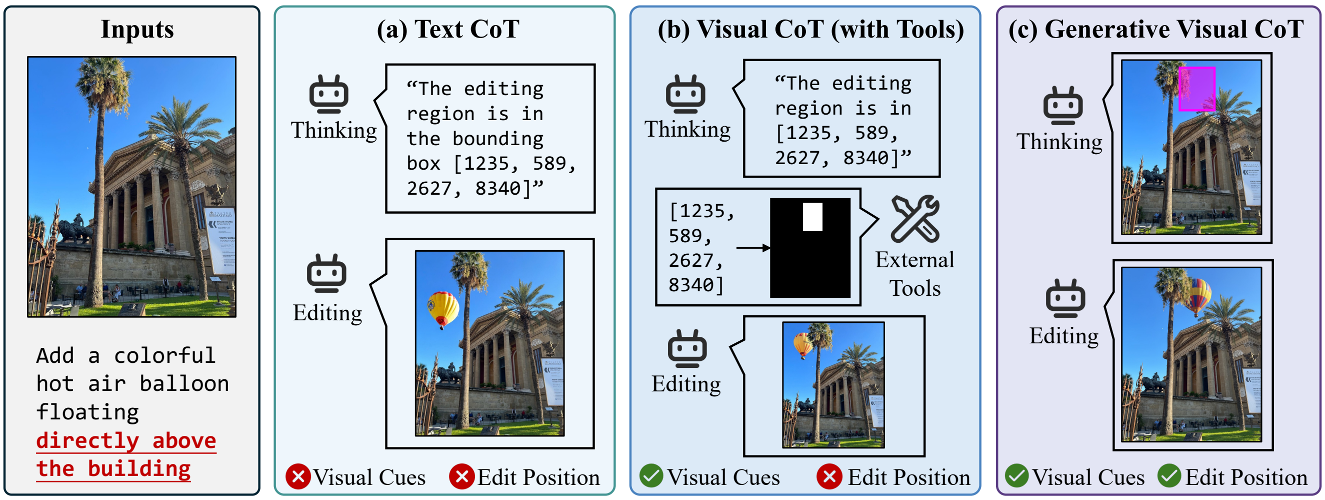


Figure 1. **Generative Visual Chain-of-Thought (GVCoT).** A comparison of three reasoning paradigms: (a) **Text CoT**, which reasons purely within the text space; (b) **Visual CoT (with Tools)**, which leverages external tools to highlight target regions; and (c) **Our GVCoT**, which performs native visual reasoning via a generative diffusion process within a unified space.

Abstract

Existing image editing methods struggle to perceive where to edit, especially under complex scenes and nuanced spatial instructions. To address this issue, we propose Generative Visual Chain-of-Thought (GVCoT), a unified framework that performs native visual reasoning by first generating spatial cues to localize the target region and then executing the edit. Unlike prior text-only CoT or tool-dependent visual CoT paradigms, GVCoT jointly optimizes visual tokens generated during the reasoning and editing phases in an end-to-end manner. This way fosters the emergence of innate spatial reasoning ability and enables more effective utilization of visual-domain cues. The main challenge of training GCVoT lies in the scarcity of large-scale editing data with precise edit region annotations; to this end, we construct GVCoT-Edit-Instruct, a dataset of 1.8M high-quality samples spanning 19 tasks. We adopt a progressive training strategy: supervised fine-tuning to build foundational localization ability in reasoning trace before final editing, followed by reinforcement learning to further improve reasoning and editing quality. Finally, we introduce SREdit-Bench, a new benchmark designed to comprehensively stress-test models under sophisticated scenes and fine-grained referring expressions. Experiments demonstrate that GVCoT consistently outperforms state-of-the-art models on SREdit-Bench and ImgEdit. We hope our GV-CoT will inspire future research toward interpretable and precise image editing.

1. Introduction

Recent advances in large-scale datasets and training have enabled significant progress in instruction-guided image editing, through both unified understanding-generation models [9, 32, 49, 58] and diffusion-based approaches [4, 28, 36, 51, 55]. However, these methods still struggle to localize intended edit regions reliably under complex scenarios, such as tasks involving intricate spatial relations, images with multiple entities, and finely nuanced instructions.

Several studies [14, 50, 64] have shown that inference-

† Work done during internship at Tencent Hunyuan.
‡ Project leader.
§ Corresponding author. liangkongming@bupt.edu.cn

time scaling, such as Chain-of-Thought (CoT) [54], improves performance on complex tasks. Motivated by this, GoT-R1 [10, 11] adopts such a strategy into image editing, *i.e.*, predicting target location coordinates within the textual CoT, as illustrated in Fig. 1 (a). However, it remains a linguistic proxy and therefore does not fully leverage spatial information within the visual domain. Cognitive science suggests an alternative view: visual reasoning is an inherently modality-specific capacity [29]. A skilled artist "paints twice", first imagining in the mind, then drawing on the canvas. This raises a new question: *Can integrating reasoning through visual intermediates improve image editing more effectively than solely using textual reasoning results?*

To investigate this question, we conduct a preliminary study comparing two methods of providing spatial cues: (1) bounding-box coordinates in text modality, and (2) bounding-box masks in visual modality. As shown in Fig. 2, visual modality cues yield superior instruction adherence and better background preservation. These findings establish that *visual-level spatial cues are more effective than text-level cues for image editing.* One straightforward approach to incorporate such cues is through an agentic pipeline that integrates external visual aids (such as cropping, zooming, or tool-generated masks) into reasoning traces [31, 47, 68], as illustrated in Fig. 1 (b). However, this paradigm is fundamentally limited by the expressiveness of external tools. Since the reasoning remains text-driven, the model cannot develop *innate visual reasoning* capabilities.

In this paper, we propose **G**enerative **V**isual **C**hain-**o**f-**T**hought (**GVCoT**), a novel framework that enables a unified model to *generate visual spatial cues* as intermediate reasoning steps during image editing (see Fig. 1 (c)). Specifically, the process begins by identifying the editing region by drawing masks onto the input image, which corresponds to the visual thought, followed by the image editing step. The main advantage is that, by directly supervising the visual tokens generated during the reasoning process with a diffusion loss [18, 45], GVCoT integrates reasoning and editing into a unified end-to-end learning framework, thereby facilitating a more stable and effective emergence of intrinsic visual reasoning ability.

The key challenge in enabling GVCoT is the scarcity of image editing datasets with accurate edit region annotations. To overcome this, we develop a scalable multi-stage pipeline that automatically generates high-quality bounding boxes and segmentation masks for edited regions across diverse editing tasks. We utilize this pipeline to construct **GVCoT-Edit-Instruct**, a large-scale dataset containing 1.8 million high-quality training samples. In particular, we adopt a progressive training recipe that combines supervised fine-tuning (SFT) and reinforcement learning (RL). The first phase focuses on equipping the model with foundational capabilities of drawing masks onto original images

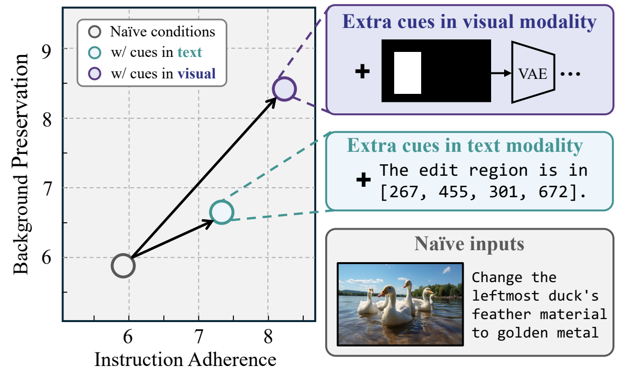


Figure 2. **Comparing spatial cue representation for image editing on ImgEdit [61].** We study two ways of injecting spatial information: (1) **text modality** uses bounding box coordinates, and (2) **visual modality** providing a binary mask. Providing spatial information in the visual modality yields a greater improvement in both instruction adherence and background preservation.

and producing structured visual reasoning chains before the image editing process. The second phase boosts both intermediate localization accuracy and final editing fidelity using Group Relative Policy Optimization (GRPO) [35].

While existing benchmarks such as ImgEdit [61] and GEdit-Bench [36] primarily focus on object-salient scenes, they fall short in evaluating a model's true spatial reasoning ability under complex editing scenarios. To address this, we introduce **SREdit-Bench**, a new benchmark comprising 590 carefully curated samples covering (1) non-object-salient and multiple entities scenes, and (2) fine-grained referring expressions in instructions. We evaluate 16 representative editing models and observe considerable performance gaps, highlighting the challenges of spatially grounded reasoning in image editing. We hope SpaEdit-Bench can serve as a new testbed for future research.

Our main contributions are summarized as follows:

- We introduce GVCoT, a new image editing paradigm that integrates reasoning via visual intermediates, outperforming state-of-the-art approaches.
- We develop a scalable curation pipeline and construct GVCoT-Edit-Instruct, a large-scale dataset comprising 1.8M high-quality pairs with region annotations.
- We propose a unified end-to-end training recipe that leverages progressive supervised fine-tuning and reinforcement learning with multi-dimensional rewards.
- We introduce SREdit-Bench, a new benchmark that assesses models' visual reasoning ability in image editing. Experiments demonstrate the superiority of our method.

2. Related Work

Instruction-Guided Image Editing. Diffusion models [18, 33, 45] have revolutionized visual content creation and manipulation. Early training-free works [5, 17, 39] modify content through latent inversion and attention-based con-

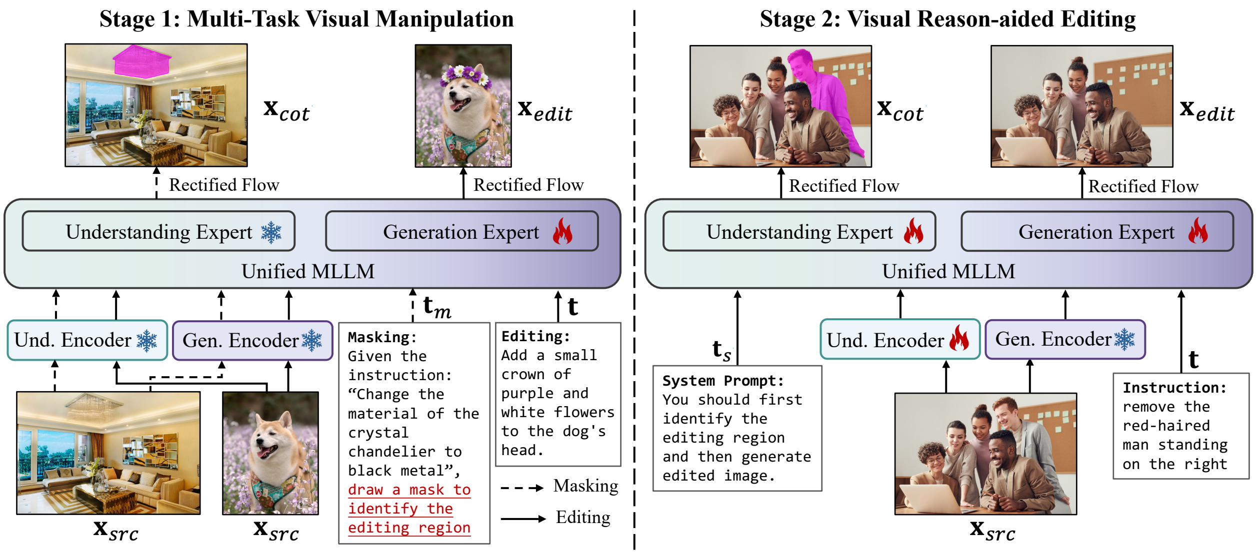


Figure 3. **Supervised Fine-Tuning of our GVCoT training recipe.** Stage 1: Multi-Task Visual Manipulation, where the model's generation expert is trained in a multi-task setup to inject the newly masking skill. Stage 2: Visual Reason-aided Editing, where the entire model is trained to generate a faithful and interpretable visual reasoning image and then an edited image within a single sequence.

trols. Training-based approaches [4, 13, 28, 36, 52, 55, 59, 65] have shown strong capability by constructing high-quality training pairs. To handle more complex and compositional editing tasks, several approaches [24, 26, 38] employ an agentic scheme, where an MLLM first plans the instruction and then drives the diffusion process to execute sub-tasks. Additionally, several benchmarks [23, 48, 60] evaluate model performance on complex tasks. CompBench [23] features scenes that require sophisticated spatial and contextual reasoning, and Complex-Edit [60] progressively tests models by increasing instruction complexity.

Multimodal Reasoning. The multimodal large language models [2, 6, 9, 22, 32, 46, 58] has unlocked powerful multimodal reasoning capabilities. Prior works [15, 20] employ text CoT to enhance visual perception [3], mathematical reasoning [20], and visual generation [10, 11, 16]. Unlikely, visual CoT integrates visual aids directly into the reasoning process. One approach uses external tools, *e.g.* drawing auxiliary lines [19], zooming in [47, 68], style transfer [34], and sub-region highlighting [12]. Another approach explores intrinsic visual CoT, where models generate visual thoughts natively [7, 8, 30, 31, 44]. Despite the promise, this approach is largely unexplored in image editing. Concurrently, MURE [69] employs native interleaved CoT for image editing. However, it does not evaluate its spatial reasoning ability under complex tasks.

3. Method

3.1. GVCoT Formulation

Different from existing methods relying on textual intermediate reasoning results, our proposed GVCoT first infers an intermediate visual Chain-of-Thought (CoT) image and subsequently generates the final edited result. Formally, given an input image $\mathbf{x}_{src} \in \mathbb{R}^{H\times W\times 3}$ and the editing instruction $\mathbf{t}$, the goal is to generate: (1) a visual thought map $\mathbf{x}_{cot} \in \mathbb{R}^{H\times W\times 3}$ that explicitly highlights editing regions, and (2) a final edited image $\mathbf{x}_{edit} \in \mathbb{R}^{H\times W\times 3}$. The overall process can be expressed as:

$$\mathbf{x}_{cot} = f_\theta(\mathbf{x}_{src}, \mathbf{t}), \quad \mathbf{x}_{edit} = f_\theta(\mathbf{x}_{src}, \mathbf{t}, \mathbf{x}_{cot}), \qquad (1)$$

where f_θ denotes the unified model.

3.2. GVCoT Training Recipe

We implement our GVCoT framework on Bagel [9], a unified model that has two distinct experts, an understanding expert and a generation expert. To stably internalize and improve the new visual reasoning skills without disrupting the model's original capability, we employ a two-phase training recipe: (1) Progressive Supervised Fine-tuning and (2) Reinforcement-based Refining.

Progressive Supervised Fine Tuning. The first phase aims to endow the model with the fundamental capability to generate an accurate and interpretable visual reasoning image $\mathbf{x}_{cot}$ before editing. We design a progressive strategy containing two steps, as shown in Fig. 3.

Step 1: Multi-Task Visual Manipulation. This stage injects explicit spatial localization capability into the generation expert. To prevent catastrophic forgetting of prior editing skills, we adopt a multi-task objective: (1) masking: generating an image $\mathbf{x}_{cot}$ (which draws masks on the $\mathbf{x}_{src}$) based on source image $\mathbf{x}_{src}$ and masking instruction $\mathbf{t}_m$; (2) editing: predicting an edited image $\mathbf{x}_{edit}$ conditioned on $\mathbf{x}_{src}$ and edit instruction T. All images provided in the question are encoded into clean VAE and ViT tokens, serving as visual context. To preserve the model's inherent rea-

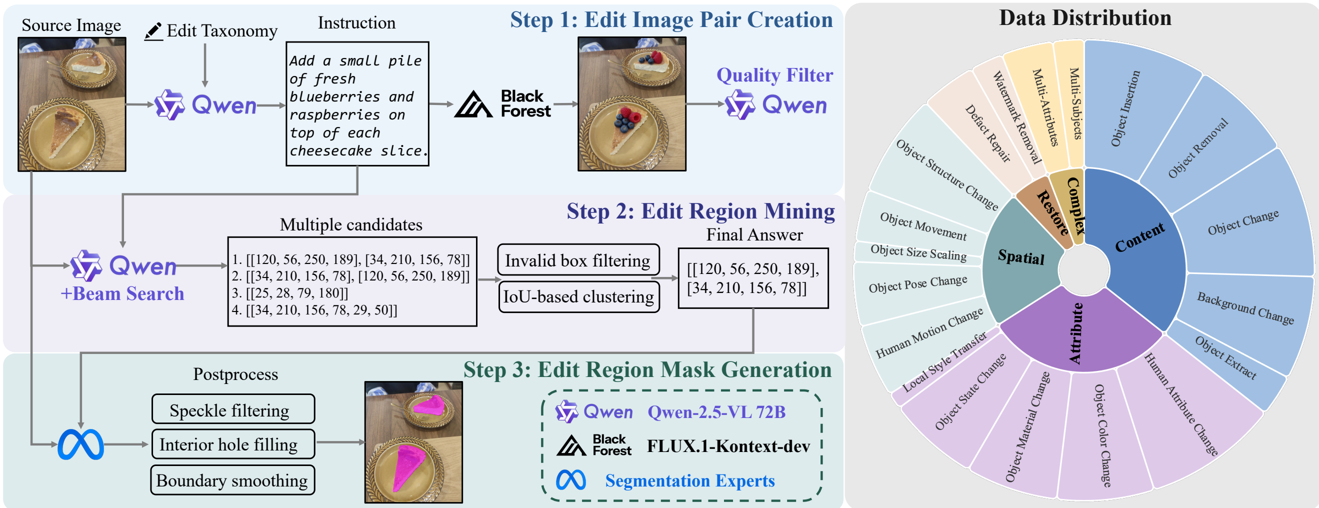


Figure 4. **GVCoT-Edit-Instruct Data Pipeline.** Left: We design a scalable multi-stage data pipeline to curate high-quality samples with faithful editing region annotations, *i.e.*, bounding boxes and masks. Right: The distribution of GVCoT-Edit-Instruct spanning 19 tasks.

soning abilities, we freeze the entire understanding expert and only train the generation expert as follows:

$$\lambda_m \mathcal{L}(\mathbf{x}^*_{cot}, f_\theta(\mathbf{x}_{src}, \mathbf{t}_m)) + \lambda_e \mathcal{L}(\mathbf{x}^*_{edit}, f_\theta(\mathbf{x}_{src}, \mathbf{t})) \quad (2)$$

where $\mathbf{x}^*_{cot}$ and $\mathbf{x}^*_{edit}$ indicates the ground-truth of visual reasoning and edit image, $\mathcal{L}$ is the flow matching loss [33, 37], and λ_m and λ_e are weights of two tasks to balance training dynamics.

Step 2: Visual Reason-aided Editing. This stage aims to endow the model with reasoning-aware editing competence. The model is required to generate an intermediate visual reasoning image, and then the final edit step-by-step within a single sequence. Thus, the loss is:

$$\mathcal{L}(\mathbf{x}^*_{cot}, f_\theta(\mathbf{t}_s, \mathbf{x}_{src}, \mathbf{t})) + \mathcal{L}(\mathbf{x}^*_{edit}, f_\theta(\mathbf{t}_s, \mathbf{x}_{src}, \mathbf{t}, \mathbf{x}^*_{cot})) \quad (3)$$

where $\mathbf{t}_s$ is a predefined system text prompt. Unlike the first stage, all model components except the VAE encoder are unfrozen and trained jointly. Please refer to our Supplementary Material for more implementation details.

Reinforcement-based Refining. Then we aim to further refine the model's grounding accuracy and overall instruction following through reinforcement learning, *i.e.*, Flow-GRPO [35]. However, jointly optimizing visual reasoning and final editing quality in a unified multi-task framework makes optimization unstable goal confusion. Thus, we adopt a progressive strategy, optimizing each generation step separately with tailored rewards.

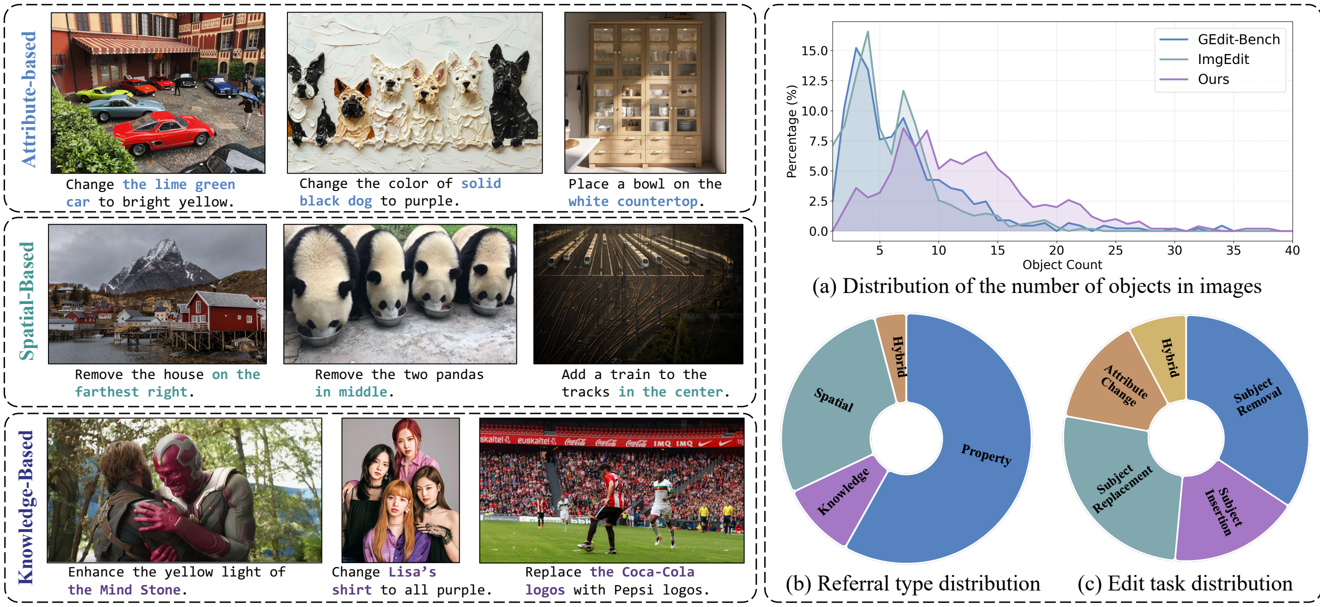
Step 1: Visual Reasoning with Verified Rewards. Low-quality visual reasoning may deteriorate the final result. We enhance the localization accuracy of the model's visual thoughts using two verified rewards. (1) Format Reward, which ensures the model follows a consistent reasoning–editing sequence rather than skipping or merging them. We train a binary classifier to distinguish whether an output image belongs to the visual thought stage or the editing stage. (2) IoU Reward, which measures the IoU between the ground-truth edit region mask and the predicted one. We extract the predicted mask by computing the pixel-wise difference between $\mathbf{x}_{src}$ and $\mathbf{x}_{cot}$.

Step 2: Editing with MLLM-as-a-Judge. Even when using teacher-forcing visual thought to guide edits, the final results can still be inaccurate. To address this, we employ two rewards: (1) CoT-Edit Consistency Reward, which encourages the model to faithfully translate the teacher-forcing visual thought into accurate edits. (2) Image Quality Reward, which improves visual realism. Both rewards are quantified by MLLM-as-a-judge, leveraging the Qwen2.5-VL-72B [2] to generate a score. More details on reward designs are provided in the supplementary.

3.3. GVCoT-Edit-Instruct Data Pipeline

The major challenge is the lack of large-scale image editing training data with corresponding editing region annotations. Thus, we design a scalable data construction pipeline and use it to create GVCoT-Edit-Instruct, comprising 1.8 million high-quality samples (see Fig. 4). Each sample consists of a quadruple: a source image, an edit instruction, edit region annotations, and the target image. The pipeline consists of three main steps, described below.

Edit Image Pair Creation. We begin by constructing the source images, instructions, and edited images. We collect 5.6 million images with at least $1K$ resolution from public datasets and websites, ensuring broad coverage of humans, objects, and scenes. We define a comprehensive edit taxonomy that spans diverse, real-world editing intents. Since our focus is on localized reasoning and editing, we exclude global edits such as style transfer and viewpoint change. Guided by this taxonomy, Qwen2.5-VL [2] produces con-



(a) Distribution of the number of objects in images

(b) Referral type distribution (c) Edit task distribution

Figure 5. **Illustration of the SREdit-Bench**. Left: We provide challenging scenarios featuring complex scenes and fine-grained referring expressions. Right: (a) We quantify scene complexity by counting editable objects and regions. Results show that SpaEdit-Bench concentrates on more sophisticated scenes than ImgEdit [61] and GEdit-Bench [36]. (b) Referral type distribution. (c) Edit tasks distribution.

cise, natural user-style instructions, and FLUX.1 Kontext [Dev] [28] synthesizes edited images. At last, an MLLM-based verifier filters out low-quality samples by measuring image naturalness and edit faithfulness.

Edit Region Mining. Then we mine the editing regions' annotations. Previous attempts [22, 32] compute pixel differences between source and target images to acquire a regional mask. While this method is effective for rigid edits (*e.g.*, color shift), it fails for flexible edits such as object motion or structural changes. We instead propose a more robust localization strategy. Qwen2.5-VL [2] predicts bounding box coordinates for the intended edit regions; multiple candidates are generated via beam search to minimize hallucinations. We filter out invalid boxes (*e.g*, out-of-bounds, zero area, extreme aspect ratios) and perform IoU-based clustering to remove outliers. The averaged coordinates of the valid boxes are taken as the final results.

Edit Region Mask Generation. At last, we generate a precise mask for each mined edit region. For object insertion, we directly use a box mask since the object boundary is unknown before editing. For modification and removal tasks, we leverage segmentation experts, *i.e.*, SAM2 [43] and BiRefNet [67], to produce instance masks. Finally, a post-process is applied to fill the interior hole, remove exterior speckle, and smooth the boundaries.

3.4. SREdit-Bench

Existing benchmarks such as ImgEdit [61] and GEdit-Bench [36] under-represent spatially complex editing scenarios, *e.g.*, multiple similar editable entities, non-object-salient scenes, and tasks that demand fine-grained object referral. Some prior works [23, 48, 60] consider multi-region editing and complex scenes; however, they do not target evaluating spatial reasoning ability and often rely on low-resolution images ($<1024\times1024$). To fill this gap, we introduce **SREdit-Bench**, a new benchmark focused on editing scenarios that require spatial reasoning.

Benchmark Construction. We curate a diverse set of high-quality source images ($>1024\times1024$) from the Internet (e.g., Unsplash), and remove similar scenes to maximize diversity. To comprehensively evaluate models' spatial reasoning in editing, we have two critical designs as shown in Fig. 5. The first is sophisticated scenes, including multiple entities and non-object-centric images. The second is fine-grained target referral in instruction, including three modes: (1) *spatial*: explicit location or relational cues; (2) *property*: appearance or attribute-based descriptions; and (3) *knowledge*: implicit, context-dependent cues that require background or commonsense knowledge.

Evaluation Protocol. We use GPT4.1 [40] as an automated judge for consistent, scalable evaluation. Following VIEScore [27], we report three metrics: (1) SC (Semantic Consistency) — how well the edited result follows the instruction; (2) PQ (Perceptual Quality) — image naturalness and artifact presence; (3) O (Overall) — the geometric mean of SC and PQ, averaged across all samples.

4. Experiments

4.1. Main Results

Comparison with general image editing methods. We first compare our Bagel-GVCoT against 17 prominent gen-

Table 1. **Quantitative results on SREdit-Bench.** SC_g, PQ_g, and O_g indicate scores on Semantic Consistency, Perceptual Quality, and Overall. The best and second-best results are highlighted in bold and underlined, respectively.

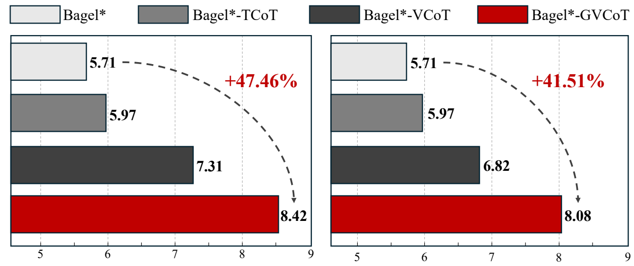
Model	SREdit-Bench ↑		
	SC_g	PQ_g	O_g
Product-level models			
GPT Image 1 [High] [41]	9.02	8.42	8.56
FLUX.1 Kontext [Pro]	8.69	8.40	8.41
Qwen-Image [55]	8.57	8.43	8.32
Generation Only			
Instruct-Pix2Pix [4]	2.10	6.40	2.58
MagicBrush [63]	2.99	5.91	3.42
ICEdit [65]	4.41	7.71	4.81
OmniGen [57]	7.05	7.38	6.68
FLUX.1 Kontext [Dev] [28]	7.52	8.17	7.27
Step1X-Edit [36]	6.96	7.87	6.98
Unified Understanding and Generation			
GoT [11]	5.78	7.59	5.83
Bagel [9]	8.02	7.90	7.75
Bagel-think [9]	<u>8.13</u>	8.01	<u>7.82</u>
UniWorld-v1 [32]	5.92	<u>8.47</u>	6.02
OmniGen2 [56]	6.52	7.54	6.44
Ming-UniVision [22]	6.05	6.85	5.96
Bagel-GVCoT (Ours)	**8.87**	**8.76**	**8.53**
Δ Over Base Model	+0.85	+0.86	+0.78

Table 2. **Quantitative results in HumanEdit.** Our Bagel-GVCoT allowing native visual reasoning outperforms previous approaches that rely on additional mask guidance.

Method	CLIP-I↑	LPIPS↓	PSNR↑	SSIM↑
SmartEdit [21]	0.8841	0.2915	17.1728	0.6828
BrushNet [25]	0.8986	0.1830	19.2172	0.7877
MagicQuill [38]	<u>0.9381</u>	<u>0.1162</u>	22.2380	**0.8981**
MIND-Edit [53]	0.9310	0.1245	<u>22.2714</u>	0.8517
Bagel [9]	0.9124	0.1721	22.1843	0.8640
Bagel-GVCoT (Ours)	**0.9451**	**0.1066**	**23.0161**	<u>0.8943</u>
Δ Over Base Model	+0.0327	-0.0655	+0.8341	+0.0303

eral image editing algorithms, including top-performing product-level models FLUX.1 Kontext Pro [28], Qwen-Image [55], GPT Image 1 [41], and powerful open-source methods, including the pure generation models [4, 36, 55, 57, 62, 63, 65, 66] and unified models [6, 22, 32, 56].

Tab. 1 shows results on our SREdit-Bench. Bagel-GVCoT achieves the highest overall performance under challenging image-editing scenarios with multiple objects and complex spatial relations. It attains an overall score of 8.53, surpassing the diffusion specialists Qwen-Image [55]. Moreover, our Bagel-GVCoT outperforms text-CoT-based reasoning models, including GoT [11] and Bagel-think [9], demonstrating the superiority of incorporating explicit visual reasoning into the editing process.



(a) Mask-Form (b) Zoom-In-Form

Figure 6. **Quantitative comparison of different visual reasoning paradigms on SREdit-Bench.** The performance is measured by O_g. Under both visual cue forms, our method consistently surpasses the text CoT and Visual CoT with considerable margins.

We also evaluate Bagel-GVCoT across various sub-tasks on ImgEdit [61]. As listed in Tab. 3, our method achieves the best overall performance among open-source models, second only to Qwen-Image [55]. Although the model attains a relatively lower score (3.83) on the style transfer task, this represents a reasonable trade-off—our framework is specifically optimized for precise spatial reasoning and localized editing rather than global stylistic manipulation. We provide more results in the supplementary.

Comparison with mask-based image editing methods. Since our method generates intermediate spatial cues to guide the subsequent editing process, we also compare it with mask-based editing models [21, 25, 38, 42, 53] that rely on input masks as external spatial guidance. We follow the evaluation setup of HumanEdit [1] in [53] for a fair comparison. Tab. 2 demonstrates that our Bagel-GVCoT outperforms all mask-based counterparts. As the base model Bagel [9] fails to surpass these methods, it further validates the strength of our visual reasoning paradigm.

4.2. Ablation Studies

Comparison of Visual Reasoning Paradigms. We compare two distinct visual reasoning paradigms: Visual CoT (VCoT), which relies on external tools, and our Generative Visual CoT (GVCoT), which produces reasoning cues in an end-to-end generative manner. To comprehensively evaluate, we design two forms of visual thought that reflect how spatial reasoning is represented and applied during editing:

- **Mask-Form**: In VCoT, the model first predicts bounding box coordinates and then employs SAM2 [43] to generate a segmentation mask, which is fused with the input image. In contrast, GVCoT directly generates the mask onto the image through a generative process.
- **Zoom-In-Form**: VCoT predicts the bounding box and then crops the corresponding region from the input image, whereas GVCoT generates a zoomed-in sub-image.

We include control group: Bagel*, (the base model directly fine-tuned on our dataset), and Bagel with text CoT (Bagel*-TCoT), which only generates textual coordinates for fair

Table 3. **Quantitative results on ImgEdit.** We use GPT-4.1 to evaluate all metrics. "Overall" is calculated by averaging all scores across tasks. The best and second-best results in open-sourced models are highlighted in **bold** and <u>underlined</u>, respectively.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
Product-level models										
FLUX.1 Kontext [Pro] [28]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT Image 1 [High] [41]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Qwen-Image [55]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Generation Only										
MagicBrush [63]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [4]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [62]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [66]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen [57]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit [65]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [36]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
FLUX.1 Kontext [Dev] [28]	**4.12**	<u>3.80</u>	2.04	<u>4.22</u>	3.09	3.97	4.51	**3.35**	4.25	<u>3.71</u>
Unified Understanding and Generation										
GoT [11]	3.74	3.06	1.33	2.72	2.46	2.33	3.45	1.77	2.50	2.65
Bagel [9]	3.56	3.31	1.70	3.3	2.62	3.24	4.49	2.38	4.17	3.20
Bagel-think [9]	3.65	3.53	2.03	3.60	3.03	3.45	4.43	2.59	4.22	3.39
UniWorld-V1 [32]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	<u>2.96</u>	2.74	3.26
OmniGen2 [56]	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	3.44
Ming-UniVision [22]	3.55	3.14	1.52	3.25	<u>3.29</u>	2.77	3.99	2.74	3.91	3.06
BLIP3o-NEXT [6]	4.00	3.78	<u>2.39</u>	4.05	2.61	**4.30**	<u>4.64</u>	2.67	4.13	3.62
Bagel-GVCoT (Ours)	<u>4.02</u>	**4.07**	**2.92**	**4.23**	**3.74**	<u>4.16</u>	3.83	2.82	<u>4.48</u>	**3.82**
Δ Over Base Model	+0.46	+0.76	+1.22	+0.93	+1.12	+0.92	-0.66	+0.44	+0.31	+0.62

Table 4. **Comparative results of GVCoT and Visual CoT using SREdit-Bench.** TF-thought denotes skipping the reasoning step by using teacher-forcing visual thought.

	Model	IoU↑	SC_g↑	PQ_g↑	O_g↑
	Bagel*	–	6.21	6.09	5.71
	TCoT	–	6.54	6.22	5.97
Mask	VCoT	0.68	7.55	7.92	7.31
Mask	**GVCoT**	0.60	8.53	8.62	8.42
Mask	VCoT w/ TF-thought	–	8.03	7.96	7.75
Mask	**GVCoT w/ TF-thought**	–	8.79	8.95	8.72
Zoom-In	VCoT	–	7.05	7.14	6.82
Zoom-In	**GVCoT**	–	8.12	8.30	8.08
Zoom-In	VCoT w/ TF-thought	–	7.78	7.67	7.54
Zoom-In	**GVCoT w/ TF-thought**	–	8.39	8.58	8.33

comparison. The comparative results under two visual cue forms are illustrated in Fig. 6, showing that Bagel*-GVCoT consistently outperforms Bagel*-VCoT and Bagel*-TCoT with considerable margins. Our GVCoT paradigm are more effective than one rely on external tools.

In-depth Analysis of GVCoT and VCoT. To understand why our GVCoT outperforms VCoT, we conduct an in-depth analysis, as shown in Tab. 4. First, we measure the localization accuracy of visual thoughts by computing IoU with ground-truth masks. The two paradigms exhibit similar localization ability (0.68 for VCoT vs. 0.66 for GVCoT). Second, to isolate the impact of visual thought

Table 5. **Ablation results on progressive SFT.**

Step 1	Step 2	IoU↑	SC_g↑	PQ_g↑	O_g↑
✓		0.64	7.92	8.01	7.75
	✓	0.61	8.21	8.33	8.10
✓	✓	**0.66**	**8.53**	**8.62**	**8.42**

accuracy to edit results, we utilize teacher-forcing visual thoughts (denoted as TF-thought) for two paradigms. Results show GVCoT achieves a significantly higher O_g score (8.72) compared to VCoT (7.75). GVCoT demonstrates stronger thought–edit consistency and the trend holds across both Mask and Zoom-In cue forms. Given the similar IoU but large performance gap in editing quality, the advantage of GVCoT clearly lies not in localization precision, but in how effectively it leverages spatial information.

Effectiveness of supervised fine-tuning designs. Tab. 5 shows the ablation results of our two-stage SFT. Step 1 primarily enhances localization capability, Step 2 improves thought-editing consistency and editing quality. Combining both stages leads to the best overall performance.

Effectiveness of reinforcement learning designs. In Tab. 6, we perform ablation studies on our progressive reinforcement learning with multi-reward designs. Firstly, the incorporation of RL boosts performance (O_g from 8.42 to 8.53) and enhances spatial accuracy (IoU from 0.60 to 0.67). Secondly, the multi-stage setup is effective, as removing Stage 2 substantially degrades results. Thirdly, two rewards for

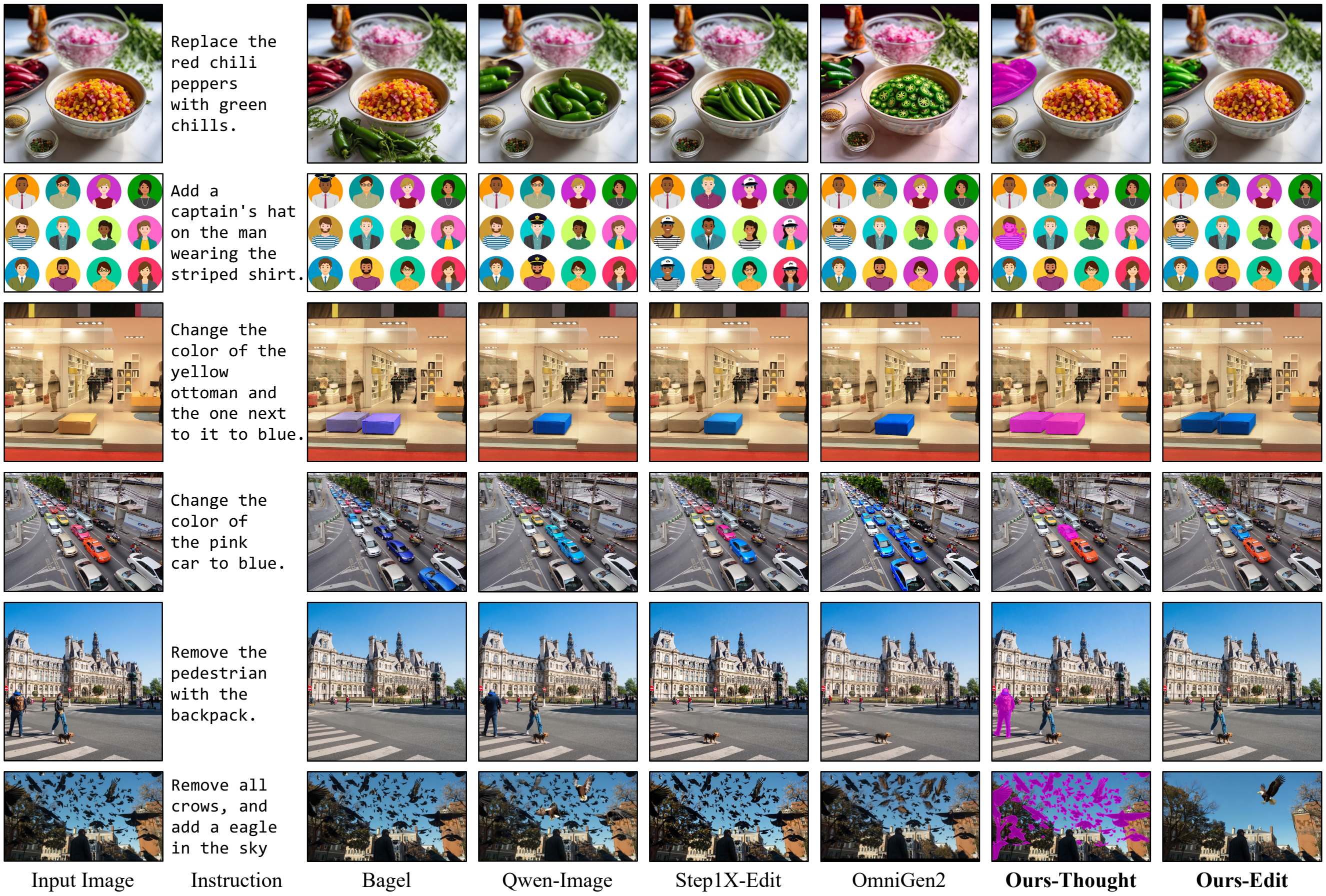


Figure 7. **Qualitative comparison in SREdit-Bench.** Our method demonstrates superior spatial reasoning and instruction adherence compared to existing open-source models, especially when handling complex, multi-object editing tasks.

improving the localization accuracy of visual thought prove vital, with their removal drops the IoU score, e.g., from 0.67 to 0.50 and 0.62, respectively. Finally, both the CoT-Edit consistency reward and the Image quality reward contribute positively to ensuring both fidelity and the faithful translation of the visual thought into the final edit.

5. Conclusion

We introduce the Generative Visual Chain-of-Thought (GVCoT) framework, designed to endow unified models with intrinsic spatial reasoning capabilities for image editing. Leveraging our curated large-scale dataset, GVCoT-Edit-Instruct, which contains 1.8 million high-quality editing images with detailed region annotations, we adopt a two-phase training recipe to develop Bagel-GVCoT. This enables the model to accurately ground instructions and effectively handle complex image editing scenarios, including sophisticated scenes, intricate spatial relationships, and fine-grained object referring. Results on our SREdit-Bench benchmark demonstrate that Bagel-GVCoT achieves a 47.46% relative improvement over the baseline. Crucially, we find that the generative visual reasoning way can more effectively exploit the spatial signals than the agentic one, which needs external tools or models to produce them. We hope this work will serve as a robust baseline for tackling complex and challenging image editing tasks.

Table 6. **Ablation studies on Reinforcement Learning (RL).** We analyze (1) multi-stage training strategy, (2) visual-thought reward design, and (3) editing reward design.

Setting	IoU↑	SC_g↑	PQ_g↑	O_g↑
SFT Only	0.60	8.53	8.62	8.42
SFT+RL	**0.67**	**8.57**	**8.76**	**8.53**
(1) Multi-stage RL training				
w/o Stage 2	0.67	8.32	8.47	8.25
(2) Visual thought reward design				
Full reward set	**0.67**	**8.32**	**8.47**	**8.25**
w/o IoU reward	0.50	8.32	8.39	8.13
w/o Format reward	0.62	8.13	8.24	8.19
(3) Editing reward design				
Full reward set	**0.67**	**8.57**	**8.76**	**8.53**
w/o CoT-Edit consistency	0.67	8.49	8.63	8.45
w/o Image quality reward	0.67	8.45	8.75	8.49

6. Acknowledgment

This work was supported by the National Nature Science Foundation of China (Grant 62476029, 62225601, U23B2052), funded by the Fundamental Research Funds for the Beijing University of Posts and Telecommunications under Grant 2025TSQY08, the Beijing Natural Science Foundation Project No. L242025, the BUPT Excellent Ph.D. Students Foundation No. CX20242081, and sponsored by Beijing Nova Program and the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

References

[1] 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. 6

[2] 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. 3, 4, 5

[3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 1, 3, 6, 7

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2

[6] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025. 3, 6, 7

[7] Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025. 3

[8] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025. 3

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 3, 6, 7

[10] Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025. 2, 3

[11] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025. 2, 3, 6, 7

[12] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025. 3

[13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12709–12720, 2024. 3

[14] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[16] Runze He, Yiji Cheng, Tiankai Hang, Zhimin Li, Yu Xu, Zijin Yin, Shiyi Zhang, Wenxun Dai, Penghui Du, Ao Ma, et al. Re-align: Structured reasoning-guided alignment for in-context image generation and editing. *arXiv preprint arXiv:2601.05124*, 2026. 3

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[19] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024. 3

[20] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 3

[21] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 6

[22] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin

Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025. 3, 5, 6, 7

[23] Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, et al. Compbench: Benchmarking complex instruction-guided image editing. *arXiv preprint arXiv:2505.12200*, 2025. 3, 5

[24] Qifei Jia, Yu Liu, Yajie Chai, Xintong Yao, Qiming Lu, Yasen Zhang, Runyu Shi, Ying Huang, and Guoquan Zhang. Lego-edit: A general image editing framework with model-level bricks and mllm builder. *arXiv preprint arXiv:2509.12883*, 2025. 3

[25] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 6

[26] Hyunseung Kim, Chiho Choi, Srikanth Malla, Sai Prahladh Padmanabhan, Saurabh Bagchi, and Joon Hee Choi. Camila: Context-aware masking for image editing with language alignment. *arXiv preprint arXiv:2509.19731*, 2025. 3

[27] 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. 5

[28] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 3, 5, 6, 7

[29] Lawrence, W., and Barsalou. Perceptual symbol systems. *Behavioral & Brain Sciences*, 1999. 2

[30] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025. 3

[31] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 2, 3

[32] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 1, 3, 5, 6, 7

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 4

[34] Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*, 2025. 3

[35] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 2, 4

[36] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 1, 2, 3, 5, 6, 7

[37] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4

[38] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13072–13082, 2025. 3, 6

[39] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[40] OpenAI. Gpt-4.1, 2025. 5

[41] OpenAI. Gpt-image-1, 2025. 6, 7

[42] Leigang Qu, Feng Cheng, Ziyan Yang, Qi Zhao, Shanchuan Lin, Yichun Shi, Yicong Li, Wenjie Wang, Tat-Seng Chua, and Lu Jiang. Vincie: Unlocking in-context image editing from video. *arXiv preprint arXiv:2506.10941*, 2025. 6

[43] 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. 5, 6

[44] Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025. 3

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[46] Yuxin Song, Wenkai Dong, Shizun Wang, Qi Zhang, Song Xue, Tao Yuan, Hu Yang, Haocheng Feng, Hang Zhou, Xinyan Xiao, et al. Query-kontext: An unified multimodal model for image generation and editing. *arXiv preprint arXiv:2509.26641*, 2025. 3

[47] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. 2, 3

[48] Chenglin Wang, Yucheng Zhou, Qianning Wang, Zhe Wang, and Kai Zhang. Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies. *arXiv preprint arXiv:2506.12830*, 2025. 3, 5

[49] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. 1

[50] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025. 1

[51] Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, et al. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025. 1

[52] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025. 3

[53] Shuyu Wang, Weiqi Li, Qian Wang, Shijie Zhao, and Jian Zhang. Mind-edit: Mllm insight-driven editing via language-vision projection. *arXiv preprint arXiv:2505.19149*, 2025. 6

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[55] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 1, 3, 6, 7

[56] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 6, 7

[57] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 6, 7

[58] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 1, 3

[59] Yu Xu, Hongbin Yan, Juan Cao, Yiji Cheng, Tiankai Hang, Runze He, Zijin Yin, Shiyi Zhang, Yuxin Zhang, Jintao Li, et al. Tag-moe: Task-aware gating for unified generative mixture-of-experts. *arXiv preprint arXiv:2601.08881*, 2026. 3

[60] Siwei Yang, Mude Hui, Bingchen Zhao, Yuyin Zhou, Nataniel Ruiz, and Cihang Xie. Complex-edit: Cot-like instruction generation for complexity-controllable image editing benchmark. *arXiv preprint arXiv:2504.13143*, 2025. 3, 5

[61] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 2, 5, 6

[62] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025. 6, 7

[63] 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. 6, 7

[64] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025. 1

[65] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 3, 6, 7

[66] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. 6, 7

[67] 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. 5

[68] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 2, 3

[69] Zhentao Zou, Zhengrong Yue, Kunpeng Du, Binlei Bao, Hanting Li, Haizhen Xie, Guozheng Xu, Yue Zhou, Yali Wang, Jie Hu, et al. Beyond textual cot: Interleaved text-image chains with deep confidence reasoning for image editing. *arXiv preprint arXiv:2510.08157*, 2025. 3