

Unlocking Complex Image Editing via Natively Interleaved Visual Textual CoT with Deep Confidence Reasoning

Zhentao Zou^{1,2} , Zhengrong Yue^{1,2}, Kunpeng Du², Binglei Bao², Hanting Li²,
Haizhen Xie², Guozheng Xu², Yue Zhou ³, Jie Hu², Xue Jiang ¹, and
Xinghao Chen²

¹ Shanghai Jiaotong University, Shanghai, China

² Huawei Technologies

³ East China Normal University, Shanghai, China
{zhtzou23}@sjtu.edu.cn

Abstract. Language guided image editing has gained significant popularity, yet existing methods struggle with intricate object intersections and fine grained spatial relationships due to the absence of an explicit visual reasoning process. While Chain of Thought (CoT) has been explored to enhance reasoning, purely textual CoT or coordinate based prompts are fundamentally limited in representing intricate visual layouts and lack the pixel level cues necessary for precise editing. To address these challenges, we propose **Unlocking Complex Image Editing via Multimodal Reasoning Edit (MURE)**, a natively multimodal framework that shifts the editing process from purely verbal reasoning to a sequence of **native interleaved textual and visual rationales**. As a unified monolithic model, MURE enables seamless end to end optimization by sharing a latent space and reusing KV cache, which provides a more flexible intermediate visual process and mitigates the error propagation typical of multi tool systems. To further ensure reasoning reliability, we introduce the **Multimodal Deep Confidence (MMDC)** paradigm. This paradigm explores a tree of visual reasoning paths at each step, pruning low quality branches via a deep confidence score from a reward model to ensure a high fidelity trajectory toward the final result. We define the formal representation for native interleaved chains and release **CoT-Edit 14K**, the first dataset comprising 14K high quality multimodal reasoning examples. Extensive experiments demonstrate that MURE establishes a new state of the art in complex editing, achieving an average gain of **7.5%** on the MagicBrush benchmark and showcasing superior spatial reasoning capabilities. The code and CoT-Edit-14K dataset are released at <https://github.com/zhentao-zou/MURE>.

Keywords: Image editing, natively interleaved text-image CoT, unified model.

1 Introduction

¹ Z. Zou and Zh. Yue contributed equally to this work. Corresponding authors: Kunpeng Du and Xue Jiang. Project Leader: Kunpeng Du.

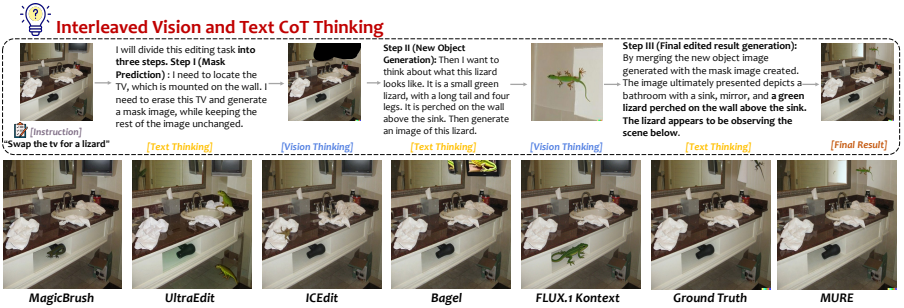


Fig. 1: Visualization of our interleaved text-visual reasoning process and a comparative result. Given the prompt “swap the tv for a lizard”, the MURE model **correctly performs multi-step reasoning to remove the lizard and its reflection in the mirror, generating a final edited image that maintains physical consistency**. In contrast, baseline approaches fail to handle this complex editing tasks, leading to erroneous results.

Image editing with natural language instructions [10, 30, 36, 45] has become increasingly popular, which eliminates the need for manual masks that are typically required by traditional methods [21, 35]. Traditional image editing systems primarily rely on diffusion-based models [3, 15, 29] to directly transfer textual instructions into visual components without an explicit reasoning process. These approaches struggle with complex cases involving intricate object intersections and fine-grained spatial relationships, that humans naturally encounter in daily life. This limitation significantly restricts their practical application potential. Meanwhile, Chain-of-Thought (CoT) reasoning [51] has emerged as a powerful tool for multimodal Large Language Models (MLLMs) [11, 40, 64, 66], enabling them to tackle complex tasks [20, 31, 58, 64] by explicitly generating intermediate thought processes.

Motivated by this, prior work has attempted to leverage CoT for image editing from different perspectives. Some researchers leverage purely textual CoT to analyze instructions, which guides the inference of the final edited image [12, 28]. However, this approach is fundamentally limited in providing precise control over object attributes and spatial relationships. To address this, other works, such as Generation of thought (GoT) [13, 14], augment the textual CoT process with explicit coordinate information, typically in the format of bounding boxes. While these methods can manage the placement of visual elements, their CoT process remains inherently text-based, and is thus inadequate for representing complex visual layouts with intricate object intersections or irregular shapes. This, coupled with their inability to represent fine-grained visual details, can lead to unintended or inaccurate edits. As shown in Figure 1, existing approaches often fail to precisely localize edit regions and, consequently, produce physically inconsistent scenes. For instance, they may erase an object while leaving its reflection intact, thereby violating the laws of physics.

This motivates our insight to alleviate this issue by introducing visual reasoning into the CoT process, as visual cues can provide precise intended edited

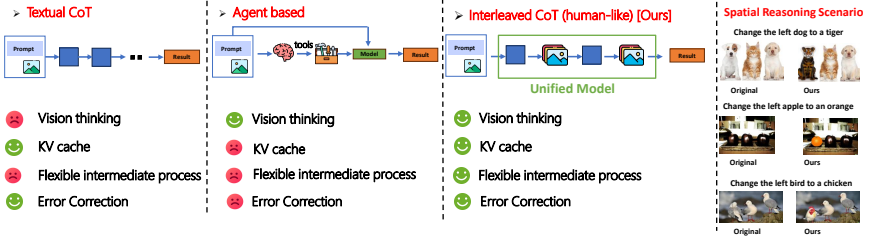


Fig. 2: Comparison of image editing paradigms and visualization results of our proposed MURE in complex spatial reasoning scenarios. (a) Textual CoT methods lack visual rationales during the reasoning process, which limits their ability to represent intricate visual layouts. (b) Agent-based approaches, which rely on external tools (e.g., DeepSeek, SAM) for segmentation, often lack KV cache efficiency and flexible intermediate processes due to their rigid dependency on fixed tool calls. (c) MURE (Ours) generates interleaved textual and visual rationales within a unified model. By iteratively conditioning on previous outputs to generate step-by-step thinking, MURE effectively prevents error accumulation. Leveraging the emergent capabilities of the unified model, it produces a highly flexible visual thinking process tailored to diverse and complex editing tasks.

regions, and also decomposing a complex task into several manageable sub-tasks, since high-quality completion of each step contributes significantly to the final high-quality result. Furthermore, focusing on the quality of intermediate steps is critical, as the inherent randomness of diffusion models [56] and the hallucination phenomenon of large language models (LLMs) [2, 62] inevitably introduce low-quality elements that can degrade the final result.

Motivated by these insights, we propose **Multimodal Reasoning Edit** termed as **MURE**, a novel approach that *shifts the visual editing process from purely text-based reasoning to a series of interleaved textual and visual rationales*. Specifically, we formulate MURE as a multimodal reasoning chain that integrates both textual and visual information. Textual descriptions analyze instructions and provide powerful guidance for the subsequent reasonable image generation, while various visual components such as positioned masks for edit regions, representations of new content, or visualizations of intermediate actions provide explicit cues for final edited results. These textual and visual CoT processes are seamlessly combined to form an autoregressive, coherent, and well ordered chain that guides the model toward the final, satisfactory edited result. To enhance robustness within the inference stage phase, existing methods, such as parallel thinking or CoT with self-checking, have attempted to address this. However, parallel thinking can lead to inferior performance when low-quality traces dominate the voting process. As argued by Fu et al. [17], focusing on the local quality of the reasoning chains is far more important than concentrating solely on the final answer. To alleviate this issue, we propose a new paradigm, **Multimodal Deep Confidence (MMDC)** reasoning. This method considers multiple reasoning paths at each visual generation step, forming a comprehensive reasoning

tree. Based on this constructed tree, we prune low-quality branches of visual paths using a deep confidence score from a reward model. By selecting the most promising path, our approach enhances the robustness of the edited results and improves overall reliability.

As shown in Figure 1, in contrast to existing approaches, our proposed **MURE** model tackles such complex tasks by decomposing them into a series of interleaved textual and visual reasoning steps. This allows the model to first generate a precise intermediate state, such as the intended edited region that removes the lizard both in and out of the mirror. This foundational step ensures the generation of subsequent correct and physically plausible editing results. Ultimately, **MURE** step-by-step approach, where generation is sequentially conditioned on all preceding outputs, allows the model to tackle intricate editing challenges and yield high-quality, satisfactory results. As shown in Figure 2, in contrast to modular frameworks that depend on external tools like SAM for region localization, **MURE performs the entire editing process within a unified monolithic model**. This architecture enables efficient KV cache reuse and a more flexible intermediate visual process, while inherently mitigating error propagation through internal self correction. In general, our contributions are three-fold,

- We propose **MURE**, a novel approach that employs a natively multimodal, interleaved text-image CoT to decompose editing tasks into a series of inter-dependent sub-tasks. This approach enables more precise subtask completion at each stage, thereby leading to higher-quality edited results.
- We propose the **MMDC** reasoning paradigm, which operates within a **natively interleaved stream**. By leveraging a reward model’s deep-confidence score to prune low-quality visual paths, MMDC effectively enhances the quality of intermediate reasoning steps, ultimately yielding high-fidelity edited results.
- We define the formulation for interleaved text-image chains and introduce the first **CoT-Edit-14K** dataset, comprising 14K high-quality editing examples to facilitate multimodal editing reasoning. Our experimental results demonstrate significant improvements across three image editing benchmarks.

2 Related Work

Image editing techniques. Since the emergence of diffusion models, various image editing approaches have been proposed that can be broadly categorized into two types. Early works, such as training-free editing methods [6, 67], modify the original images through inversion in latent space and attention manipulation. Another line of work focuses on training-based approaches [8, 16, 47, 50, 52], which involve modifying model architectures or fine-tuning on high-quality datasets. For example, ICEdit [63] employs a LoRA-MoE hybrid tuning strategy to fine-tune the Flux.1 Fill model, while FLUX.1 [29] leverages efficient flow matching technology in a novel latent space to achieve high-quality context-aware image

generation and editing. More recently, models like Step-1X Edit [36] and Seed Edit [49] leverage MLLMs to encode visual features and analyze instructions, which are then injected into a diffusion model. These approaches shows strong instruction-following and image-fidelity capabilities. Zhang et al. [60] and Ji et al. [26] employ large language models for high level planning, which then invoke a sequence of external tools for object localization and segmentation before generating the final image based on multiple conditions.

Reasoning in Multimodal Large Language Models. With the growing prominence of visual reasoning and MLLMs, CoT has been extended to both visual understanding [53,55] and generation [9,22,27,46,61] tasks. Prior work has explored different forms of CoT, DeepSeek R1 [19] uses a Vision-Language Model (VLM) to generate a comprehensive textual CoT for analyzing user instructions before generating a final answer, while MINI-CoT [9] introduces visual tokens via cropping and zooming from original images to solve mathematical problems. MM-R1 [32] introduces a cross-modal CoT comprising visual components that spatially isolate concepts before generating the final image. For text-to-image generation [33], Uni-CoT [43] proposes a hierarchical design with separate CoT levels for task planning and subtask execution. ReasonGen-R1 [61] addresses this by introducing a large-scale CoT image generation dataset and integrating the CoT process into autoregressive image generation. However, the application of multimodal reasoning to image editing remains largely unexplored. Most existing editing approaches rely solely on textual CoT to analyze visual components, lacking the rich intermediate visual processes, such as explicit masks for edit regions or representations of new content, that are vital for synthesizing a high-quality final result.

3 Methods

Most existing editing methods rely on **text-only reasoning**, occasionally supplemented with coordinate information. This approach is fundamentally limited by the inherent difficulty of representing complex spatial relationships and visual details through textual descriptions alone. To address these limitations, we propose **MURE**, a novel approach that *shifts the visual editing process from purely text-based reasoning to a series of interleaved textual and visual rationales*.

The overall framework of the proposed MURE model, illustrated in Figure 3, can be viewed as a sequential, natively multimodal CoT process. The model *addresses complex editing tasks by breaking them down into a series of interwoven textual and visual reasoning steps*. For a given task, such as “change the table for a dog”, MURE first analyzes the user’s instruction and the input image to plan the necessary subtasks, then executes them to generate the interleaved CoT step-by-step. The MURE model executes the editing task through the following steps: **Step I (Mask Prediction)**: The model first analyzes the input image to identify and localize the object to be replaced. It then generates a textual CoT to describe the object’s characteristics (e.g., “a white footstool with a black frame”). Conditioned on this text, the model generates a corresponding mask, effectively

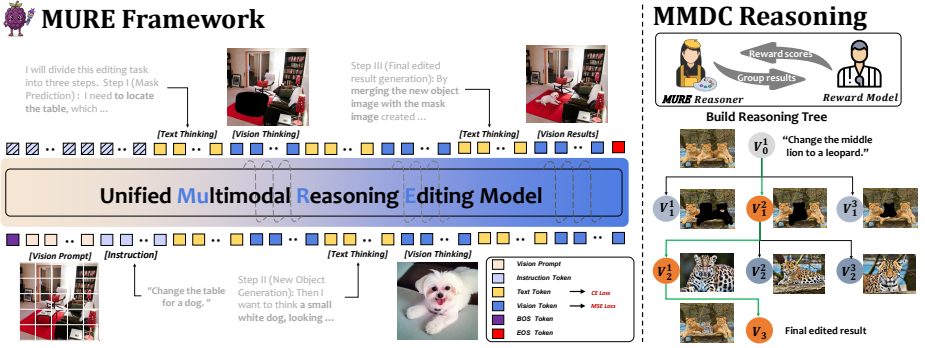


Fig. 3: Overview of the MURE framework. **Left:** Our framework leverages an interleaved text-image CoT to achieve high-fidelity image editing. **Right:** The Multimodal Deep Confidence (MMDC) reasoning explores a tree of visual reasoning paths at each step. It prunes low-quality branches based on a deep confidence score from a reward model, ensuring a superior trajectory toward the final edited image.

segmenting and removing the original object while preserving the rest of the image. **Step II (New Object Generation):** The model then forms a concept for the new object and generates a detailed textual description (e.g., “a small white dog. Its fur is fluffy, and it is wearing a collar”). The model’s generation branch then renders a high-fidelity image of the new object that aligns with this textual description. **Step III (Final edited result generation):** This step involves leveraging the spatial information from the mask and the detailed content of the new object to ensure seamless and contextually appropriate integration. Simultaneously, the model generates a detailed description of the final edited image, using all this crucial information to produce a satisfactory edited result.

From this process, it is evident that the MURE model *decomposes a complex editing task into a sequence of manageable, interdependent subtasks, thereby enabling more precise subtask completion at each stage.* Implementing MURE requires three key components,

A Comprehensive CoT-Edit-14K Dataset. To facilitate interleaved text-image CoT editing, we introduce a comprehensive dataset. This dataset contains a collection of high-quality, carefully designed interleaved text-image CoT examples, each tailored to various editing subtasks. Such a dataset provides the foundational basis for enabling multimodal reasoning within the editing process.

MURE: A Unified Framework for Visual Understanding and Generation. The model is designed to understand and generate an interleaved natively multimodal CoT, enabling efficient inference of paired text and image outputs within a unified model framework. This unified architecture ensures that intermediate CoT representations are efficiently stored and retrieved using a Key-Value (KV) cache, which is crucial for maintaining state and coherence throughout the sequential generation process. Figures in the appendix also demonstrate the **editable capabilities** of our interleaved CoT.

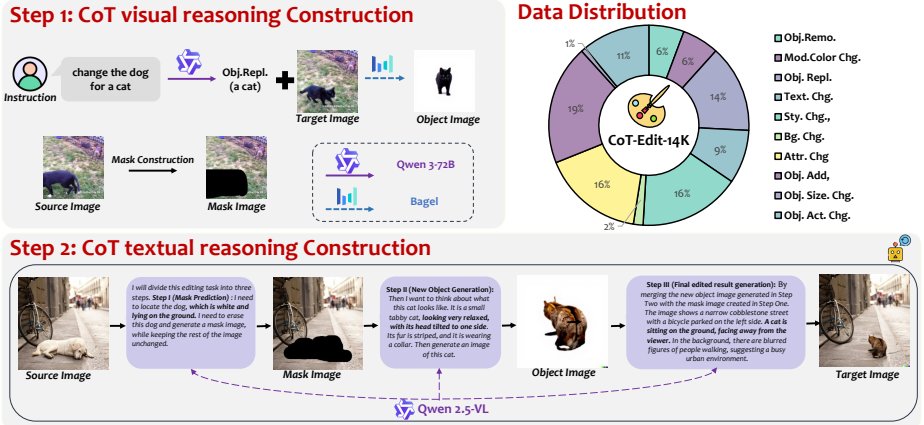


Fig. 4: MURE Dataset Construction Process. **Top:** The visual annotation pipeline constructs explicit visual cues, including positional masks that define intended edited regions and valid representations of new content. **Bottom:** The textual annotation pipeline generates detailed textual descriptions based on the annotated CoT images from the top pipeline. The specific example illustrates the reasoning for an Obj.Repl. task, for detailed overview of the dataset, refer to Figures in the appendix.

MMDC: An Efficient Reasoning Paradigm for Interleaved CoT. The inherent stochasticity of diffusion models often introduces low-quality intermediate visual steps, which can inevitably deteriorate the final result. To mitigate this, we explore multiple visual reasoning paths at each step, pruning low-quality branches with a deep confidence score to enable a more robust framework.

3.1 CoT-Edit-14K Dataset Construction

As illustrated in Figure 4, we propose a multi-stage pipeline to construct a comprehensive dataset of high-quality interleaved text-image CoT. This process is specifically tailored to each of the 10 unique editing sub-tasks and involves two distinct pipelines, one for generating visual reasoning data and another for textual reasoning data. To ensure the quality of our dataset, we manually filtered out approximately 5K erroneous or low-quality examples from the raw data, resulting in 14K samples.

For the CoT vision reasoning construction pipeline, we first leverage Qwen-3 [54] to categorize editing tasks, as different tasks necessitate distinct CoT structures. We then follow Uniworld [34] to construct position masks using an adaptive editing region weighting strategy. Furthermore, we employ BAGEL with carefully designed in-context prompts to synthesize new objects or styles. For tasks involving motion, we also capture the intermediate process of motion changes from video data.

For the CoT text reasoning construction pipeline, we utilize Qwen2.5-VL [1] to generate the textual component of the interleaved CoT. This generation is conditioned on the editing instruction, the original image, and the CoT vision

image from the previous pipeline. The completed text and image pairs are then seamlessly combined to form the complete CoT process.

We showcase our pipeline for constructing interleaved text-image CoT using the Object Replacement task as a representative example in Figure 4. Additional examples demonstrating the versatility of the MURE model across various tasks are provided in the appendix.

3.2 MURE Framework

The implementation of an **interleaved text-image CoT** necessitates a model with unified multi-modal understanding and generation capabilities. We thus adopt a unified model, such as BAGEL [12], as our backbone and finetune it for our specific task. The output of the designed **MURE** model, denoted $\mathbf{f}_\theta$, is a sequence of interleaved textual reasoning steps $s^{(i)}$ and visual tokens $v^{(i)}$, followed by the final edited image O . Specifically, $v^{(i)}$ comprises either a mask or new content depending on the editing step. The textual step $s^{(i)}$ and visual token $v^{(i)}$ are clearly delimited by the special tokens $\langle \text{visual start} \rangle$ and $\langle \text{visual end} \rangle$, respectively.

$$\left\{ s^{(1)}, v^{(1)}, s^{(2)}, v^{(2)}, \dots, v^{(k-1)}, s^{(k)} \right\}, O \sim \mathbf{f}_\theta(\cdot \mid I, P) \quad (1)$$

where I, P represent the input images and edit instruction, respectively.

Training Objectives. We train our model to generate text tokens autoregressively, optimizing the process with a Cross-Entropy (CE) Loss. This loss is computed at positions $\mathbf{T} \subset \{1, 2, \dots, k\}$ that correspond to all textual segments $s^{(i)}$. Let $Y = \{y_1, y_2, \dots, y_T\}$ denote the complete interleaved text-image CoT sequence. The CE loss for next-token prediction is formally defined as,

$$\mathcal{L}_{\text{CE}}^{\text{text}} = - \sum_{t \in \mathbf{T}} \log P_\theta(s_t \mid y_{<t}, I, P) \quad (2)$$

For visual generation, we utilized Rectified Flow paradigm [37], Given a clean latent representation $z_0^{(i)}$ of the i -th visual image v^i , and a Gaussian noise sample $z_1^{(i)}$, the noisy latent variable $z_t^{(i)}$ is constructed by linear interpolation, defining a straight-line path between the two,

$$z_t^{(i)} = t \cdot z_0^{(i)} + (1 - t) \cdot z_1^{(i)}, \quad t \in [0, 1] \quad (3)$$

The model f_θ is trained to predict the velocity field on this straight-line path by minimizing the Mean Squared Error (MSE) loss. This loss is applied to the latent variables of all predicted visual segments within the sequence, including the final edited image, and is formally defined as,

$$\mathcal{L}_{\text{MSE}}^{\text{image}} = \mathbb{E} \left[\left\| \mathbf{f}_\theta \left(\mathbf{z}_t^{(i)} \mid y_{<t}, I, P \right) - \left(\mathbf{z}_0^{(i)} - \mathbf{z}_1^{(i)} \right) \right\|^2 \right] \quad (4)$$

The context for our MURE model is the full preceding sequence, which comprises the original image, editing instructions, and a complete interleaved text-image

CoT sequence, $Y = \{y_1, y_2, \dots, y_T\}$, this sequence contains visual components, such as intended edit regions or the representations of new content, paired with their corresponding textual guidance. The overall objective for our training is a weighted combination of the CE and MSE loss functions.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}}^{\text{text}} + \mathcal{L}_{\text{MSE}}^{\text{image}} \quad (5)$$

where the hyperparameter λ_{CE} serves as a scaling coefficient to balance the relative contribution of the textual and visual objectives during training.

Inference. At the inference stage, given an initial image I and an editing instruction P , the MURE model autoregressively generates an interleaved sequence of text and images $Y = \{s^{(1)}, v^{(1)}, \dots, s^{(k)}, v^{(k)}, O\}$, by leveraging special tokens to flexibly switch between text and image generation. The process concludes automatically upon generating the final edited image O , which is signaled by a dedicated end-of-sequence token.

Compared to models that use purely CoT, such as the unified BAGEL model [12], our MURE framework takes a more sophisticated approach. It conditions the final edited result on a rich, interleaved text-image context, which is efficiently managed by a KV cache to ensure coherence throughout the sequential generation process.

3.3 MMDC Reasoning Paradigm

Low-quality intermediate visual steps may deteriorate the final result. Therefore, we introduce MMDC reasoning paradigm, which evaluates multiple reasoning paths and filter low-quality sub-path at each visual image generation step. Specifically, we leverage an independent reward model, R_θ , which acts as an evaluator to score each candidate branch. This model is employed to assess the quality of each candidate visual image $v^{(k,i)}$ and its alignment with the paired textual instruction. We assign a deep confidence score, $S_{k,i}$ to each candidate image, which reflects its potential to contribute to a high-quality final output. This score is formally defined as:

$$S_{k,i} = R_\theta \left(v^{(k,i)} \mid s^{(k)}, y_{<k}, I, P \right) \quad (6)$$

In our interleaved text-image generation process, this score represents the model’s confidence in a given reasoning path. We employ a greedy pruning strategy: at each visual generation step, we retain only the branch with the highest score and use it to guide subsequent text reasoning and visual generation. Ultimately, the optimal candidate image for the current step is selected by maximizing this deep confidence score. This selection process is represented by the following formula,

$$i^* = \arg \max_{i \in \{1, \dots, N\}} S_{k,i} \quad (7)$$

Through this phased tree search, our model can effectively self-evaluate and correct errors during the generation process, selecting the path most likely to

lead to a high-quality final output. This significantly enhances the robustness of the generated results. Note that different candidate branches in the reasoning process are executed in parallel. The reward model’s prompt guidance is detailed in appendix, which includes **out-of-distribution (OOD)** generalization validation.

4 Experiments

4.1 Experimental Setup

We use BAGEL [12], a leading open-source unified model, to initialize our unified visual understanding and generation model. We trained this model on our CoT-Edit-14K dataset, which comprises approximately 14K interleaved text-image CoT pairs. To ensure the model retains its ability to perform single-step editing, we also incorporated data with text-only CoT, using system prompts to differentiate between the two modes. We utilize the Qwen2.5-VL-7B model as our reward model, eliciting its zero-shot capabilities through our specific prompt design to implement the guidance mechanism detailed in the Appendix.

4.2 Evaluation settings

To validate the effectiveness of the proposed approach, we evaluate its performance on three benchmark datasets for image editing, MagicBrush test set [57], Emu [44], and SmartEdit benchmark [25]. To evaluate performance on the MagicBrush benchmark, we report a standard set of metrics utilized by prior works [63, 65]. We report CLIP [23, 38], DINO [7, 41], and L1 distance to measure the similarity between the edited images and their manually generated ground-truth (GT) images. Adhering to the evaluation protocols established by [44, 63, 65], for the Emu edit benchmark, we employ CLIP-I and DINO to assess the similarity between the source and edited images, and CLIP-Out to quantify the distance between the generated output caption and the edited image. While for the SmartEdit benchmark, to assess both visual fidelity and semantic alignment, we utilize PSNR, SSIM [24], and LPIPS [59] to measure perceptual consistency with the original image. We also calculate the CLIP Score [23] to evaluate the semantic alignment between the edited image’s foreground and the GT text label.

4.3 Comparisons with State-of-the-Art

Quantitative Results on Three Public Benchmarks. We evaluate our MURE model against various approaches, including UNet-based [5, 18, 57, 65], DiT-based [4, 39, 42, 48, 63], autoregressive-based [25], and unified model approaches [12]. Our performance is detailed in Table 1, and 2. The proposed MURE model achieves performance on par with state-of-the-art (SOTA) across all three datasets. On the MagicBrush dataset, our output

Table 1: Quantitative results on two benchmarks. We compare our method against baselines on the **MagicBrush** and **Emu** test sets. Best results for each metric are highlighted in bold.

Methods	CoT	MagicBrush Test Set			Emu Test Set		
		L1 ↓	CLIP-I ↑	DINO ↑	CLIP-I ↑	CLIP-Out ↑	DINO ↑
InstructP2P (CVPR23)	✗	0.114	0.851	0.744	0.856	0.292	0.773
MagicBrush (NeurIPS23)	✗	0.074	0.908	0.847	0.877	0.298	0.807
UltraEdit (NeurIPS24)	✗	0.066	0.904	0.852	0.880	0.304	0.847
FluxEdit (HuggingFace)	✗	0.114	0.779	0.663	0.852	0.282	0.760
FLUX.1 Fill (HuggingFace)	✗	0.192	0.795	0.669	0.794	0.273	0.659
RF-Solver Edit (arXiv25)	✗	0.112	0.766	0.675	0.797	0.309	0.683
ACE++ (arXiv25)	✗	0.195	0.741	0.591	0.791	0.280	0.687
ICEdit (arXiv25)	✗	0.060	0.928	0.853	0.907	0.305	0.866
Bagel (arXiv25)	Text	0.067	0.923	0.856	0.869	0.308	0.824
MURE	Text & Images	0.049	0.943	0.877	0.920	0.301	0.897

demonstrates strong alignment with GT images. On the Emu test set, our model achieves SOTA-level text alignment while exhibiting superior image fidelity with original images. This strong performance is attributed to our approach’s ability to predict intended edit regions within the interleaved text-image CoT. This explicit conditioning serves a dual purpose, it allows for precise regional modifications to mitigate editing errors and enhances background consistency by preserving unedited areas. Furthermore, on the SmartEdit benchmark, the proposed approach achieves superior performance compared to competing methods on both the understanding and reasoning scenarios, demonstrating robust performance on these complex tasks. This consistent strength is attributed to the application of the interleaved text-image CoT, which effectively breaks down complex problems into smaller sub-tasks. This multi-step process enables more precise task completion at each stage, allowing our approach to successfully accomplish highly intricate editing tasks. We also report L1 scores on the MagicBrush test relative to the average number of output tokens. Figure 5 illustrates a positive correlation between the model’s performance and the combined textual and visual output token count, suggesting that a marginal increase in tokens yields a substantial improvement.

Visual Comparison. As Figure 6 demonstrates, the proposed MURE model achieves significant superiority over other open-source SOTA approaches in both editing accuracy and visual quality. This robust performance is attributable not only to our interleaved text-image CoT but also to the novel MMDC reasoning paradigm. The MMDC framework ensures the model efficiently follows a high-confidence inference path, effectively mitigating incoherent or sub-optimal outputs.

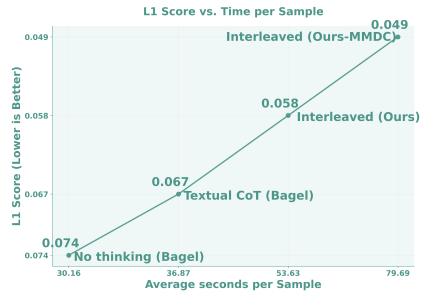


Fig. 5: L1 Score vs. Output Token Cost on MagicBrush test set.

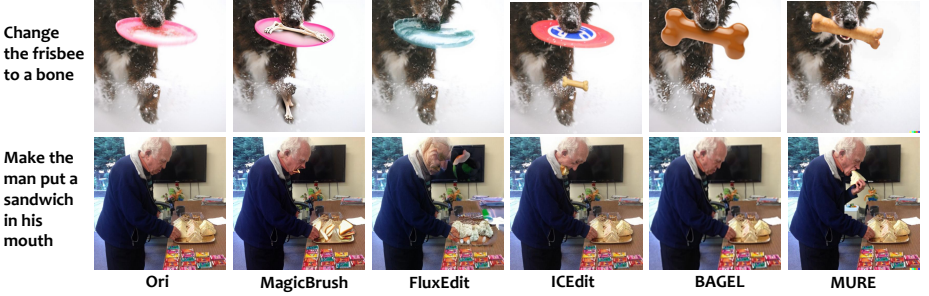


Fig. 6: Visual comparison results. Additional examples are provided appendix.

Table 2: Quantitative comparison of different methods on Understanding and Reasoning scenarios across **SmartEdit** benchmark. The best results for each metric are highlighted in bold.

Methods	Spatial Reasoning Scenarios					World Model Scenarios				
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP Score $\uparrow$		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP Score $\uparrow$	
InstructP2P (CVPR23)	21.576	0.721	0.089	22.762		24.234	0.707	0.083	19.413	
MagicBrush (NeurIPS23)	18.120	0.680	0.143	22.620		22.101	0.694	0.113	19.755	
InstructDiffusion (CVPR24)	23.258	0.743	0.067	23.080		21.453	0.666	0.117	19.523	
SmartEdit-7B (CVPR24)	22.049	0.731	0.087	23.611		25.258	0.742	0.055	20.950	
SmartEdit-13B (CVPR24)	23.596	0.751	0.068	23.536		25.757	0.747	0.051	20.777	
Bagel (arXiv25)	23.823	0.892	0.083	23.842		28.076	0.839	0.060	20.767	
MURE	25.611	0.897	0.065	23.947		28.694	0.883	0.062	21.298	

5 Ablation Study

Framework Design. As shown in Table 3, we conducted an ablation study to analyze the impact of our framework’s key components on the MagicBrush and Emu test sets. We started with a BAGEL model trained only with textual CoT. By incorporating interleaved text-image CoT, we significantly improved image quality, as evidenced by a 13.4% decrease in the L1 metric and a 1.3% increase in the CLIP-I score. This demonstrates that the MURE model’s ability to decompose complex tasks into smaller sub-tasks leads to more precise task completion and a substantial performance gain. The introduction of the MMDC framework further enhanced our model’s performance, particularly on the Emu test set. By leveraging a confidence score from a reward model to guide visual image generation, this framework ensures the model follows a high-confidence path within the inference process, resulting in superior edited images.

Gains via interleaved CoT in challenging cases. To demonstrate the advantages of MURE in complex editing, we evaluated specialized subsets from MagicBrush and Emu Edit, curated via GPT-4o to identify particularly demanding cases. As shown in Table 5, MURE (w/o MMDC) yields substantial improvements in the SmartEdit "Spatial Reasoning" task and these challenging subsets, validating its efficacy in resolving intricate spatial relationships. While our proposed interleaved CoT achieves modest gains of 2%–8% across entire

Table 3: Ablation study of our proposed MURE framework on MagicBrush and Emu test. The acronyms ICT denote Interleaved text-image CoT.

ICT	MMDC	MagicBrush test			Emu test		
		L1 ↓	CLIP-I ↑	DINO ↑	CLIP-I ↑	CLIP-Out ↑	DINO ↑
✗	✗	0.067	0.923	0.856	0.869	0.308	0.824
✗	✓	0.058	0.936	0.856	0.880	0.303	0.836
✓	✓	0.049	0.943	0.877	0.920	0.301	0.897

Table 4: Ablation study of our proposed MURE framework on MagicBrush and Emu test related to deep confidence reasoning search width.

Search Width	MagicBrush test			Emu test		
	L1 ↓	CLIP-I ↑	DINO ↑	CLIP-I ↑	CLIP-Out ↑	DINO ↑
1	0.058	0.936	0.856	0.880	0.303	0.836
3	0.052	0.941	0.872	0.913	0.302	0.887
5	0.049	0.943	0.877	0.920	0.301	0.897

image editing benchmarks (Table 3), it demonstrates significantly higher potential in complex scenarios. Specifically, Table 5 shows that our approach achieves remarkable gains of 5%–20% in metrics such as L1 and CLIP-I compared to textual CoT, underscoring the value of interleaved reasoning in sophisticated image editing scenes.

Table 5: Ablation study of the MURE framework on challenging subsets requiring complex spatial reasoning.

Method	CoT	SmartEdit (Spa Reason.)		MagicBrush (Sub.)		Emu Edit (Sub.)		
		PSNR ↑	LPIPS ↓	L1 ↓	CLIP-I ↑	CLIP-I ↑	CLIP-O ↑	DINO ↑
Bagel w/ CoT	Text	23.823	0.083	0.081	0.912	0.831	0.288	0.789
MURE w/o MMDC	T&I	25.230	0.068	0.063	0.932	0.872	0.302	0.828
MURE w. MMDC	T&I	25.610	0.065	0.052	0.941	0.906	0.301	0.851

Note: Challenging cases are selected from MagicBrush (128/535) and Emu Edit (1231/3589) via GPT-4o with the prompt: “Does this editing task require strong spatial reasoning ability for successful completion?”

Search Width for Deep Confidence Reasoning. Our deep confidence reasoning strategy uses a reward model to assess and prune intermediate image paths during interleaved text-image CoT inference. The results, as detailed in Table 4, demonstrate that increasing the search width (N) monotonically improves model performance. We hypothesize that employing a more capable reward model could further improve the overall effectiveness of our strategy.

Qwen VL Bias and Reliability. To validate the effectiveness of our reward model, we assessed its alignment with human judgment across various editing categories using 121 randomly sampled MagicBrush cases. As indicated in Table 6, Qwen VL exhibits a high degree of congruency with human scores, with its evaluations closely matching human trends across all tasks. Notably, the performance margins captured by Qwen VL in tasks like *Bg.* and *Add.* are nearly identical to human perceived gains. This consistency effectively captures the advancements brought by our MMDC module, confirming Qwen VL’s reliability as an automated evaluation proxy. However, our analysis also reveals that while Qwen VL is generally robust, it possesses a subtle **semantic over spatial bias**, occasionally prioritizing **object presence** over **positional precision** in tasks such as *Obj.Remove/Repl/Add*. Despite this characteristic, the consistent performance improvements, which faithfully mirror human judgment, validate

Qwen VL as a high fidelity proxy for verifying the scaling up potential of our interleaved reasoning framework.

Table 6: Comparison of model performance across different editing categories using Qwen-VL and human evaluation (LPIPS/L1 metric).

Method	Remo.	Color.	Repl.	Act.	Bg.	Add.	Size.
Bagel (w/ CoT)	0.0505	0.0723	0.0716	0.0461	0.1162	0.0699	0.0549
MURE (w/o MMDC)	0.0485	0.0632	0.0653	0.0402	0.0732	0.0595	0.0509
MURE + MMDC (Qwen VL)	0.0401	0.0581	0.0552	0.0382	0.0585	0.0427	0.0499
MURE + MMDC (Humans)	0.0371	0.0579	0.0501	0.0375	0.0581	0.0395	0.0485

- The unified interleaved CoT training acts as a **Logic Regularizer**, fundamentally enhancing the model’s general visual physical understanding, which improves performance across **both** textual CoT **and** interleaved CoT-driven tasks.

Table 7: Comparison of L1 between MURE and Bagel across different editing categories on MagicBrush test set. The MURE approach achieves different editing tasks by adaptively switching between interleaved CoT (†) and textual CoT (◦). **L1 is a metric where lower values indicate better performance.** MURE[†] represents results when the MURE approach is explicitly forced to perform inference through interleaved text-image chains using a specific system prompt (e.g., “You should achieve this editing task through a chain of thoughts interwoven with text and images.”).

Category	Bagel (w/CoT)	MURE (vs. Bagel)	MURE [†] (vs. Bagel)
Obj.Remo.	0.0505 [◦]	0.0401 [†] (+20.59%)	0.0400 [†] (+20.79%)
Mod.Color Chg.	0.0723 [◦]	0.0581 [†] (+19.64%)	0.0576 [†] (+20.33%)
Obj.Repl.	0.0716 [◦]	0.0552 [†] (+22.86%)	0.0551 [†] (+23.04%)
Obj.Act.Chg.	0.0461 [◦]	0.0382 [◦] (+17.14%)	0.0367 [†] (+20.39%)
Bg.Chg.	0.1162 [◦]	0.0585 [†] (+49.66%)	0.0583 [†] (+49.83%)
Obj.Add.	0.0699 [◦]	0.0427 [†] (+38.91%)	0.0427 [†] (+38.91%)
Obj.Size Chg.	0.0549 [◦]	0.0499 [◦] (+9.11%)	0.0479 [†] (+12.75%)

The substantial efficacy of natively interleaved CoT across a diverse editing tasks. MURE tackles complex editing tasks by adaptively switching between interleaved CoT and textual CoT reasoning. The unified training functions as a Logic Regularizer, forcing the model to acquire a deep visual-physical understanding from the complex constraints inherent in the interleaved CoT data. This advanced knowledge is then efficiently transferred across the model. As detailed in Table 7, **tasks utilizing interleaved CoT exhibit a superior degree of improvement compared to those solved via textual CoT, demonstrating the advantage of the interleaved approach.** Crucially,

the +17.13% gain achieved by the text-reliant Obj.Act.Chg. task over the textual CoT baseline (Bagel) demonstrates the universal benefit of the interleaved CoT paradigm. This confirms that the intensive visual logic training fundamentally enhances reasoning robustness, extending its positive impact to purely text-guided operations. Our approach facilitates **interleaved CoT inference across the majority of editing tasks**. The sole exception is style transfer, as this category primarily relies on global feature manipulation and is independent of the spatial awareness priors required for sequential visual reasoning. More visualization results are provided in the Appendix.

6 Conclusion and limitation

In this paper, we propose MURE, a novel framework that shifts the visual editing paradigm from purely textual reasoning to a series of interleaved textual and visual rationales. By decomposing complex editing tasks into manageable, interdependent subtasks, our approach enables more precise task completion. We demonstrate superior performance on three challenging benchmarks, MagicBrush, Emu, and SmartEdit, and attribute this success to two key innovations: an interleaved text-image CoT and our MMDC reasoning paradigm. The interleaved CoT enables fine-grained modifications and consistent background generation, while MMDC mitigates incoherence by ensuring a high-confidence inference path. Ultimately, MURE marks a significant step toward more effective, human-like reasoning for complex visual editing tasks.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), <https://arxiv.org/abs/2502.13923>
2. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey (2025), <https://arxiv.org/abs/2404.18930>
3. Bazyleva, V., Bonettini, N., Bharaj, G.: X-edit: Detecting and localizing edits in images altered by text-guided diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2729–2739 (2025)
4. Black-Forest-Labs: Flux. <https://github.com/black-forest-labs/flux>. Accessed: 2025-05-15. (2024)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)

7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling (2025), <https://arxiv.org/abs/2501.17811>
9. Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., Li, H.: Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *Advances in neural information processing systems* **38**, 69110–69139 (2026)
10. Comanici, G., Bieber, E., Schaekermann, M., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), <https://arxiv.org/abs/2507.06261>
11. DeepMind, G.: Nano banana pro (2025), <https://deepmind.google/models/gemini-image/pro/>
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining (2025), <https://arxiv.org/abs/2505.14683>
13. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: GoT-R1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning (2026), <https://arxiv.org/abs/2505.17022>
14. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., Liu, X., Li, H.: GoT: Unleashing reasoning capability of multimodal large language model for visual generation and editing (2025), <https://arxiv.org/abs/2503.10639>
15. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025)
16. Fu, T.J., Hu, W., Du, X., Wang, W., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. In: International Conference on Learning Representations. vol. 2024, pp. 54820–54833 (2024)
17. Fu, Y., Wang, X., Tian, Y., Zhao, J.: Deep think with confidence (2025), <https://arxiv.org/abs/2508.15260>
18. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Li, H., Hu, H., et al.: Instructdiffusion: A generalist modeling interface for vision tasks. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 12709–12720 (2024)
19. Guo, D., Yang, D., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* **645**(8081), 633–638 (2025). <https://doi.org/10.1038/s41586-025-09422-z>, <http://dx.doi.org/10.1038/s41586-025-09422-z>
20. Guo, L., Xu, X., Wang, L., Lin, J., Zhou, J., Zhang, Z., Su, B., Chen, Y.: ComfyMind: Toward general-purpose generation via tree-based planning and reactive feedback. *Advances in Neural Information Processing Systems* **38**, 45128–45164 (2026)
21. Guo, Q., Lin, T.: Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6986–6996 (2024)
22. Guo, Z., Zhang, R., Tong, C., Zhao, Z., Huang, R., Zhang, H., Zhang, M., Liu, J., Zhang, S., Gao, P., Li, H., Heng, P.A.: Can we generate images with cot? let’s verify and reinforce image generation step by step (2025), <https://arxiv.org/abs/2501.13926>

23. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
24. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
25. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., 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. pp. 8362–8371 (2024)
26. Ji, L., Qi, C., Chen, Q.: Instruction-based image editing with planning, reasoning, and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17506–17515 (2025)
27. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. Advances in Neural Information Processing Systems **38**, 39856–39890 (2026)
28. Kang, M., Zhang, X., Wei, F., Xu, S., Liu, Y.: Enhancing image editing with chain-of-thought reasoning and multimodal large language models. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
29. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742>
30. Li, M., Gu, X., Chen, F., Xing, X., Wen, L., Chen, C., Zhu, S.: Superedit: Rectifying and facilitating supervision for instruction-based image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19206–19215 (2025)
31. Li, P., Yan, S., Cai, J., Zhang, R., An, R., Guo, Z., Gao, X.: Adaptive classifier-free guidance via dynamic low-confidence masking. Advances in Neural Information Processing Systems **38**, 47362–47381 (2026)
32. Liang, Q., Wu, Y., Li, K., Wei, J., He, S., Guo, J., Xie, N.: MM-R1: Unleashing the power of unified multimodal large language models for personalized image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6835–6843 (2026)
33. Liao, J., Yang, Z., Li, L., Li, D., Lin, K., Cheng, Y., Wang, L.: Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17214–17223 (2025)
34. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., Pang, Y., Yuan, L.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation (2025), <https://arxiv.org/abs/2506.03147>
35. Lin, Y., Chen, Y.W., Tsai, Y.H., Jiang, L., Yang, M.H.: Text-driven image editing via learnable regions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7059–7068 (2024)
36. 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., Jiang, D.: Step1x-edit: A practical framework for general image editing (2025), <https://arxiv.org/abs/2504.17761>

37. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022), <https://arxiv.org/abs/2209.03003>
38. (Mahi)Shafiullah, N., Paxton, C., Pinto, L., Chintala, S., Szlam, A.: Clip-fields: Weakly supervised semantic fields for robotic memory. In: Robotics: Science and Systems XIX. Robotics: Science and Systems Foundation (2023). <https://doi.org/10.15607/rss.2023.xix.074>, <http://dx.doi.org/10.15607/RSS.2023.XIX.074>
39. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1958–1966 (2025)
40. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision (2024), <https://arxiv.org/abs/2304.07193>
42. Paul, S.: Flux.1-dev-edit-v0. Accessed: 2025-02- 21. (2025)
43. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision (2026), <https://arxiv.org/abs/2508.05606>
44. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
45. Shi, Y., Wang, P., Huang, W.: Seedit: Align image re-generation to image editing (2024), <https://arxiv.org/abs/2411.06686>
46. Tong, C., Guo, Z., Zhang, R., Shan, W., Wei, X., Xing, Z., Li, H., Heng, P.A.: Delving into rl for image generation with cot: A study on dpo vs. grpo. *Advances in Neural Information Processing Systems* **38**, 115632–115655 (2026)
47. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
48. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing (2025), <https://arxiv.org/abs/2411.04746>
49. Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., Yang, J.: Seedit 3.0: Fast and high-quality generative image editing (2025), <https://arxiv.org/abs/2506.05083>
50. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., Zhao, Y., Ao, Y., Min, X., Li, T., Wu, B., Zhao, B., Zhang, B., Wang, L., Liu, G., He, Z., Yang, X., Liu, J., Lin, Y., Huang, T., Wang, Z.: Emu3: Next-token prediction is all you need (2024), <https://arxiv.org/abs/2409.18869>
51. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)

52. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems* **38**, 47490–47518 (2026)
53. Xu, Y., Li, C., Zhou, H., Wan, X., Zhang, C., Korhonen, A., Vulić, I.: Visual planning: Let’s think only with images (2026), <https://arxiv.org/abs/2505.11409>
54. Yang, A., Li, A., Yang, B., et al. Qwen3 technical report (2025), <https://arxiv.org/abs/2505.09388>
55. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems* **38**, 29918–29952 (2026)
56. Zhang, B., Luo, W., Zhang, Z.: Purify++: Improving diffusion-purification with advanced diffusion models and control of randomness (2023), <https://arxiv.org/abs/2310.18762>
57. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* **36**, 31428–31449 (2023)
58. Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Qiao, Y., Li, H., Gao, P.: Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In: *The Twelfth International Conference on Learning Representations* (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
60. Zhang, X., Kang, M., Wei, F., Xu, S., Liu, Y., Ma, L.: Tie: Revolutionizing text-based image editing for complex-prompt following and high-fidelity editing (2024), <https://arxiv.org/abs/2405.16803>
61. Zhang, Y., Li, Y., Yang, Y., Wang, R., Yang, Y., Qi, D., Bao, J., Chen, D., Luo, C., Qiu, L.: Reasongen-rl: Cot for autoregressive image generation models through sft and rl (2025), <https://arxiv.org/abs/2505.24875>
62. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al.: Siren’s song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics* pp. 1–46 (2025)
63. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *Advances in Neural Information Processing Systems* **38**, 139195–139227 (2026)
64. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models (2024), <https://arxiv.org/abs/2302.00923>
65. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems* **37**, 3058–3093 (2024)
66. Zheng, G., Yang, B., Tang, J., Zhou, H.Y., Yang, S.: Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems* **36**, 5168–5191 (2023)
67. Zhu, T., Zhang, S., Shao, J., Tang, Y.: Kv-edit: Training-free image editing for precise background preservation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16607–16617 (2025)