

CORE-V: Chain-Of-thought REasoning for Image Editing with Visual Interaction

Tianyang Yan¹, Peisen Zhao^{2†}, Guanghao Zheng¹, Mingxing Xu²,
Zhibo Zhang², Wenrui Dai^{1*}, Junni Zou¹, Hongkai Xiong¹, Xiaopeng Zhang^{2*},
and Qi Tian²

¹ Shanghai Jiao Tong University, Shanghai, China

{yantianyang,zgh990318,daiwenrui,zoujunni,xionghongkai}@sjtu.edu.cn

² Huawei Inc., China

{pszhao93,xmx18829236935,zxphistory}@gmail.com, billz@mail.nwpu.edu.cn

tian.qi1@huawei.com

Abstract. Existing methods for instruction-guided image editing are usually confined to text-based CoT reasoning for semantic understanding and reasoning, and lack of visual awareness necessary for fine-grained synthesis. In this work, we propose CORE-V, a novel Visual CoT framework to explicitly perform visually-interactive reasoning process for controllable image editing. Specifically, CORE-V refines global semantics into detailed entities and generates visual imagination images based on attribute description in the instructions. These visual imagination images, together with spatial conditions of input images (i.e., segmentation masks, depth maps and edge maps) are applied within the reasoning process. To achieve visual reasoning training, we build the CORE-Edit-800K dataset that contains both reference images and spatial controls aligned with reasoning pipelines. Furthermore, we introduce CORE-Bench to assess directional editing similarity. Comprehensive experiments on Emu-Edit, Reason-Edit and CORE-Bench demonstrate superior performance of CORE-V for visual-semantic alignment and structural integrity.

Keywords: Chain-of-Thought reasoning · Controllable image editing · Coarse-to-Fine benchmark

1 Introduction

Recent advancements in text-to-image synthesis have facilitated the research on instruction-guided image editing. To interpret editing instructions with efficient semantic extractors, text encoders are commonly coupled with powerful Diffusion Transformer (DiT) variants such as Stable Diffusion [31, 34] and Flux [16] for high-quality image synthesis. To provide spatial guidance, external conditioning modules like T2I-Adapter [25] or ControlNet [43] are frequently integrated by injecting latent control features into the denoising process.

[†] Project Leader

^{*} Corresponding Authors

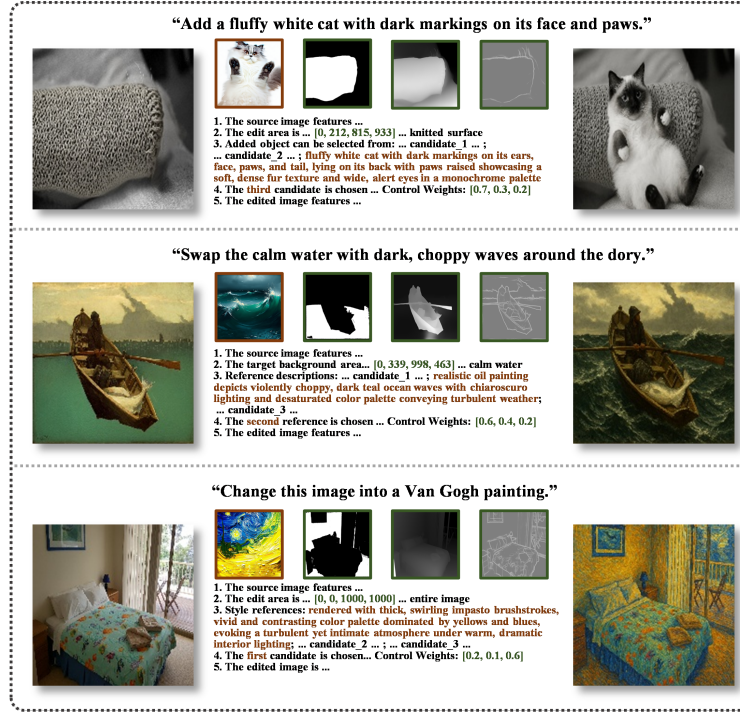


Fig. 1: Illustration of the textual and visual CoT dataset. Given the input image (left), output image (right), and editing instruction, we formulate an explicit CoT reasoning process. The green segments denote spatial control, including bounding boxes and the corresponding weights of three spatial conditions (green-framed). The orange segments represent the selected reference descriptions, where the reference image (orange-framed) exemplifies the visual imagination of specific attributes and context.

As natural language prompts tailored for various editing tasks directly affect the synthesis outcomes, such heavy reliance on textual instructions introduces inherent instability. Subtle changes in texts may lead to divergent semantic interpretations and disparate visual effects. Due to the absence of explicit interaction between semantic logic and visual structures, existing frameworks usually suffer from a loss of structural fidelity and unintended global distortions. This discrepancy poses a significant challenge in bridging the gap between high-level semantic intent and low-level pixel synthesis.

The rapid evolution of Multimodal Large Language Models (MLLMs) like open-source Qwen-VL [1, 2] and closed-source GPT-4o [12] benefits current image editing models from capturing complex semantics. Data-driven paradigm is prominent to utilize MLLMs to curate large-scale datasets with refined captions and editing instructions [3, 36, 37, 42, 47], whereas architectural integration emphasizes to adopt MLLMs as unified backbones to obtain latent representations [23, 24, 39]. Recent generation and editing models incorporate visual CoT

architectures to construct reasoning pipelines for refining visual-semantic relationship, including reinforcement learning-based strategies for concurrent generation and assessment [5, 44] and construction of token-level correspondences by interleaving specific visual tokens within semantic embeddings [14].

Nevertheless, existing methods for image editing are still restricted by the imperfect alignment between visual and semantic modalities. From the architectural perspective, there often lack of adaptive scheduling and interaction across modalities. The common design of passively concatenating implicit semantics with visual features fails to produce the spatially conditioned structures in complex scenarios. Meanwhile, the underlying knowledge, from data curation to the inference process, remains fundamentally rooted in textual logic. This dependence precludes models from deriving effective representations of abstract visual attributes through linguistic descriptions alone. Such text reliance may trigger vision-agnostic hallucinations and affect fine-grained structural synthesis.

To bridge this gap, we introduce CORE-V, a visual CoT framework designed to synchronize semantic reasoning with explicit visual interaction for controllable image editing. Specifically, CORE-V reformulates complex editing instructions into a five-stage reasoning trajectory: Description, Grounding, Imagination, Control, and Generation. As illustrated in Fig. 1, our pipeline extracts structured intermediate representations by integrating global captions with detailed visual conditions. Leveraging fine-tuned MLLMs, CORE-V deduces object and style attributes such as color, shape, and texture to generate a consistent visual imagination. Simultaneously, expert tools are incorporated to derive precise spatial constraints, including position, lighting, and reflections. These components collectively constitute a unified CoT framework with visual-semantic interaction for high-fidelity image manipulation.

We curated a visual CoT dataset comprising 800K samples across six primary editing categories with each annotated with both textual and visual intermediate reasoning steps. To train the unified architecture, we propose a dual-stage optimization strategy employing a composite loss function to ensure the seamless integration of visual constraints and semantic reasoning within the final editing results. Furthermore, to quantify the efficacy of visual reasoning in achieving fine-grained editing, we introduce the first one-to-many image editing benchmark CORE-Bench. This benchmark evaluates whether the editing directions conform to intricate attributes, by quantifying the embedded feature difference of editing outputs. Compared editing results are conducted from a single source image under hierarchical coarse-to-fine instructions. Extensive experiments demonstrate that our method achieves highly competitive performance on diverse evaluations.

The main contributions of this work are summarized as below.

- We propose CORE-V, a unified framework which schedules semantic reasoning with interleaved visual control and attribute imagination, thereby achieving superior visual-semantic alignment for fine-grained image editing.
- We construct a comprehensive dataset CORE-Edit-800K that manipulates both textual and visual structured CoT reasoning processes, developed through a rigorous curated and filtered pipeline.

- We establish CORE-Bench, a pioneering coarse-to-fine benchmark specifically designed to assess fine-grained editing capabilities. Both quantitative and qualitative evaluations across diverse benchmarks demonstrate that CORE-V consistently surpasses existing image editing baselines.

2 Related Work

2.1 Image Editing Models

For precise visual manipulation, instruction-guided image editing methods utilize fine-tuned diffusion models [8, 33] to decode fused latent embeddings from both textual instructions and source images. Early works such as Instruct-Pix2Pix [3], MagicBrush [42], and UltraEdit [47], typically leverage high-quality editing instruction-image pairs synthesized by MLLMs via specialized generation and filtering pipelines. These methods strive to bridge the gap between semantic instructions and pixel-level synthesis, which requires extensive training on massive and curated datasets.

Recent research has prioritized architectural innovation by optimizing both diffusion modules and semantic encoders. On the one hand, diffusion backbones have transitioned from U-Net based diffusion models [31], which are often constrained by local inductive biases, toward scalable DiT modules [27] with superior text-to-image synthesis capabilities. As a leading foundation for image editing, FLUX-based models [15, 16] utilize rectified flow transformers [6] to streamline the inference process, which mitigates semantic drift through continuous flow matching trajectories [21]. On the other hand, MLLMs endowed with fine-grained visual perception and semantic understanding abilities are increasingly superseding traditional text encoders such as CLIP [28] and T5 [29]. Methods like UniWorld-V1 [19] and Step1X-Edit [23] incorporate pretrained MLLMs including LLaVA [22] and Qwen-VL [2] to extract multimodal latent conditions that synchronize linguistic intent with visual context. While architectural advancements have improved editing performance to some extent, achieving controllable synthesis still necessitates training on high quality datasets.

Current methods rely on limited semantic alignment, condensing complex edit instructions into static latent representations. Such implicit mapping fails to disentangle specific visual cues, particularly structural preservation and attribute manipulation. Significant challenges remain in modulating semantic conditions and deriving detailed object or stylistic attributes through explicit pipelines, ensuring controllable generation aligned with fine-grained edit instructions.

2.2 Visual CoT Reasoning

CoT reasoning process [13, 38, 46] guides MLLMs to decompose complex instructions into explicit intermediate reasoning steps, which significantly bolsters the cross-modal alignment quality and procedural interpretability. Current visual models primarily leverage CoT prompting to enhance text-centric knowledge reasoning capabilities. Image-CoT [44] develops potential assessment reward model

to evaluate each step in autoregressive image generation. T2I-R1 [14] constructs a semantic-level CoT grounded in textual reasoning, together with a token-level CoT that generates pixel details in a patch-by-patch manner. Recently, Bagel [4] incorporates the thinking mode of MLLMs to enable knowledge-based reasoning for image generation and editing. However, existing visual CoT pipelines still struggle to jointly ensure logical consistency of both semantic reasoning and visual information. In most cases, the apparent reasoning capability is largely inherited from training on textual CoT datasets, which primarily capture text-centric knowledge. When applied to complex editing tasks, the semantic reasoning required by the generation modules should instead focus on visual information representing the intermediate states during the editing operations, rather than purely textual reasoning. GoT [5, 7] explicitly incorporates editing bounding boxes into textual CoT and extracts the corresponding local visual features for downstream generation and editing. We find this idea compelling, and suggest that it can be extended by more fine-grained intermediate visual conditions, tailored to the needs of different editing tasks.

Notably, current visual CoT methods still rely on training with textual reasoning datasets, and lack supervision from offline visual data to guide the intermediate process. Moreover, there is no dedicated one-to-many image editing benchmark to evaluate editing fidelity with coarse-to-fine attributes. Such evaluation is essential for assessing whether improved reasoning capabilities truly translate into better image editing quality.

3 Methods

In this section, we describe the detailed visual CoT training paradigm for image editing. Sec 3.1 demonstrates the curated process of the CORE-Edit-800K dataset. Based on the dataset, we propose CORE-V framework in Sec 3.2 to achieve visual imagination reasoning capacity. Furthermore, we introduce CORE-Bench in Sec 3.3, which assesses the execution of fine-grained editing and quantifies structural consistency across diverse instructions.

3.1 CORE-Edit-800K Dataset

To achieve visual imagination process in image editing, we establish four phases for the construction of CORE-Edit-800K Dataset: data collection, CoT annotation workflow, visual condition generation and filtering criteria. The overall annotation workflow is illustrated in Fig. 2 (left).

Data Collection. We collect the instruction-image pairs from the public dataset GPT-Image-Edit-1.5M [36] which contains three sub-datasets OmniEdit [37], HQ-Edit [11] and UltraEdit [47]. To ensure a structured reasoning process, we categorize editing instructions into six primary editing operations as shown in Fig. 2 (right): object addition, object removal, object replacement, background replacement, attribute modification (*e.g.*, texture, color, or facial expression), and global transfer (*e.g.*, style, weather, or environment changes).

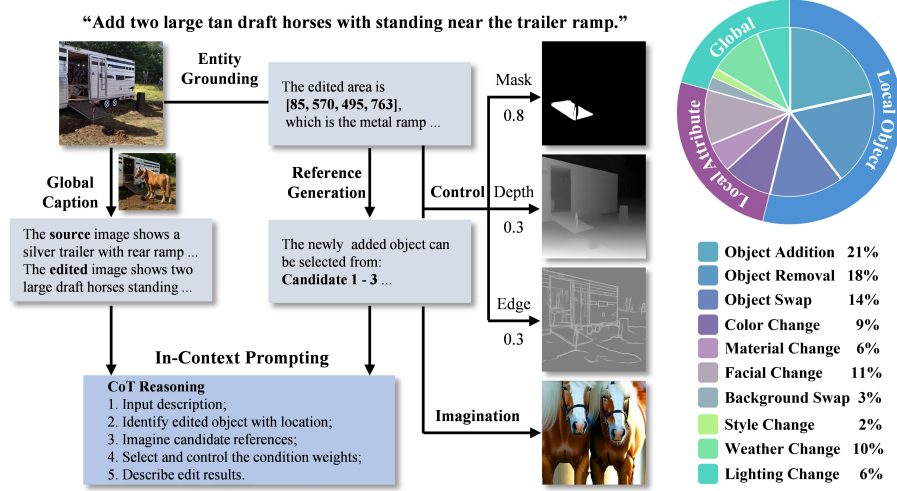


Fig. 2: Overview of the CORE-Edit-800K annotation workflow and dataset statistics. (Left) Construction of five-step CoT reasoning using Qwen3-VL-8B with visual CoT conditions (mask, depth, edge, and reference images) from expert tools. (Right) The dataset distributions categorized by local object, local attribute, and global levels.

CoT Annotation Workflow. We first localize the source and target entities via bounding box scaled to with $[0, 1000]$, guided by the editing instruction and source image. Subsequently, we generate detailed source and target descriptions with entity attributes to ensure semantic consistency. A reference description is then distilled from ground truth attributes isolating core intrinsic features like color, shape or texture out of other unrelated context. Rewriting on this reference, we generate alternative candidate descriptions to form a selection process with ground truth option. Concurrently, we leverage MLLMs to determine the optimal weights within $(0, 1]$ for mask, depth, and edge conditions to be integrated into the downstream connector for precise control. The final textual CoT is assembled as sequential steps with in-context logic. We constrain that the object removal and attribute modification tasks consist of only four steps, as they do not involve new objects or styles.

Visual Condition Generation We further generate visual conditions of source images to enforce strict visual-semantic alignment. Specifically, we leverage expert models (i.e. SAM2 [30], Depth Pro [?], and PiDiNet [35]) to produce segmentation masks, depth maps, and edge maps that match the original image resolution. It is worth noting that the segmentation masks are derived from the contents contained in the extracted bounding boxes. To facilitate visual imagination, we utilize SD3.5 [34] to generate 512×512 reference images according to the reference description, which serve as fine-grained attribute conceptualization of the object or style, ensuring visual similarity and semantic consistency with the edited output.

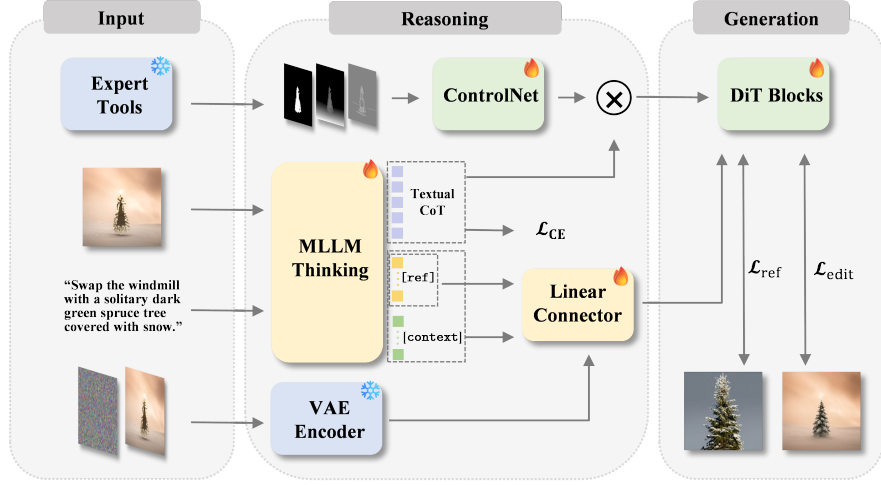


Fig. 3: Framework architecture and optimization objectives of CORE-V. The model is trained using a combined objective: $\mathcal{L}_{CE}$ for textual CoT reasoning, $\mathcal{L}_{ref}$ for reference image generation via [ref] tokens, and $\mathcal{L}_{edit}$ for image editing guided by both [ref] and [context] tokens. The color scheme indicates training stages: frozen (blue), pre-trained in the first stage (yellow), and jointly fine-tuned in the second stage (green).

Filtering Criteria. To ensure the precision and logical coherence of our visual CoT steps, we implement a rigorous filtering process. First, we verify that all source and target entities from reasoning are present in the input and output images. We verify if edit entities are correctly grounded in the grounded area by re-captioning the cropped image. To ensure compatibility, we restrict the description lengths, while enforcing significant attribute-level divergence among candidates. Reference descriptions with irrelevant objects or mismatched attributes are excluded to ensure structural consistency. Furthermore, to mitigate the semantic ambiguity arising from the conflated instructions in [36], we revert to the original instructions from sub-datasets with standardized formats.

3.2 CORE-V Framework

As shown in Fig. 3, the CORE-V consists of three modules: MLLM reasoning, Condition generation and Diffusion generation. The MLLM extracts latent semantic tokens from the input image and instruction, which are then integrated into the diffusion module alongside spatial condition features and VAE-encoded visual latents for final synthesis.

Model Structure. We employ Qwen-VL for text reasoning and visual latents generation. Following the text-based CoT phase, we incorporate 64 [ref] tokens and 64 [context] tokens to provide semantic guidance for downstream generation. Adopting the sequential reasoning paradigm, the [ref] tokens are utilized to derive a visual imagination of the target attributes. Subsequently, the

remaining [context] tokens are incorporated as supplementary contextual features, forming a complete 128-token sequence to control the final edit synthesis. This causal architecture effectively anchors the reference features, ensuring that the synthesized attributes remain faithful to the detailed reference description in semantic reasoning.

Specifically, within the linear connector, we condense the MLLM embeddings via average pooling and a double-layer MLP to derive a semantic guidance. To ensure semantic alignment, an affine transformation is performed on the visual tokens, effectively modulating their feature distribution toward the semantic feature. Finally, the visual latents are integrated with the semantic guidance vector via channel-wise concatenation for downstream DiT processing.

In parallel, we employ off-the-shelf specialist models to generate auxiliary segmentation masks, depth maps, and edge maps. These spatial priors provide the structural information that MLLMs typically struggle to perceive or synthesize directly. To integrate multiple conditions, we utilize the open-source ControlNet backbones [17, 18, 41] with an explicit weighting mechanism. Specifically, each spatial condition is independently encoded and aggregated via a weighted summation governed by weight scalars $[w_{\text{mask}}, w_{\text{depth}}, w_{\text{edge}}]$ from CoT reasoning. By adaptively modulating the output features of spatial control signals prior to their injection into the DiT blocks, our framework allows the model to selectively prioritize spatial control cues. This synergy achieves a balance between instruction alignment and geometric integrity.

Training Paradigm. To facilitate the integration of textual and visual CoT while maintaining high-fidelity editing results, we define a joint training objective of three components as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{edit}}\mathcal{L}_{\text{edit}} \quad (1)$$

Here, $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss for supervising the textual CoT sequences generated by the MLLM. $\mathcal{L}_{\text{ref}}$ represents the reconstruction loss for intermediate imagination, which is calculated by decoding only the [ref] tokens while omitting ControlNet conditions to preserve the fine-grained reasoning attributes from the interference of spatial conditions. Finally, $\mathcal{L}_{\text{edit}}$ characterizes the final editing loss, which supervises the final output by integrating multimodal guidance including visual, semantic, and spatial control. The loss triplet corresponds to the textual, reference, and edited ground truths in our CoT dataset.

To ensure training stability, we adopt a progressive two-stage optimization strategy. In the first stage, we focus on reasoning pre-training, where only the MLLM decoder is updated by optimizing $\mathcal{L}_{\text{CE}}$. The second stage jointly upgrades the entire framework using the total loss, shown in Eq. 1. Regarding parameter efficiency, we apply low-rank adaption (LoRA) [9] to both the MLLM decoder and DiT blocks, while the connector and ControlNet network are fully fine-tuned. All other components and auxiliary tools remain frozen throughout the whole training process.

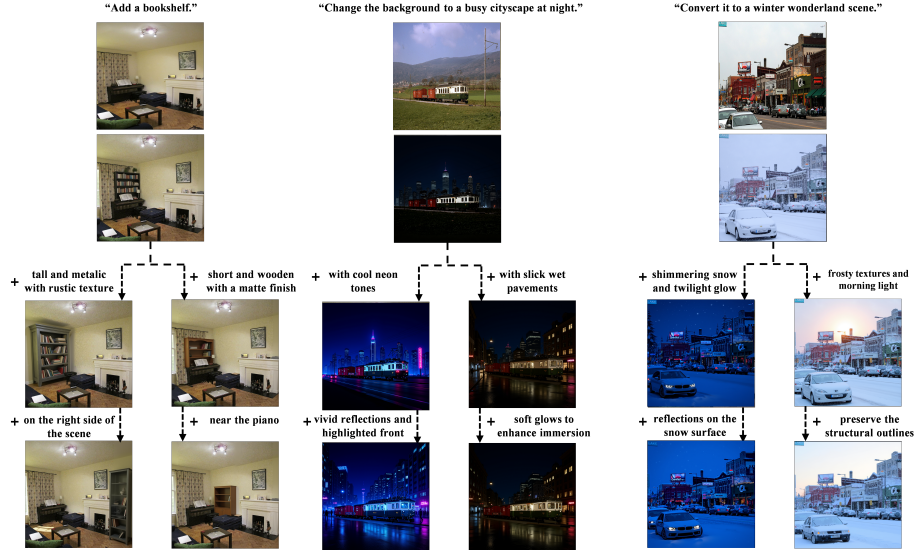


Fig. 4: Qualitative examples from the proposed CORE-Bench. Starting from the same input image (first row), we apply editing instructions with increasing granularity to generate multi-level results across two distinct branches (rows 2–4). These results provide a visual evaluation of our methods on CORE-Bench for coarse-to-fine editing.

3.3 Coarse-to-Fine Benchmark

As current image editing benchmarks are confined to one-to-one image-instruction pairings, they lack direct quantification of fine-grained attribute manipulation. To address these limitations, we introduce CORE-Bench, a hierarchical coarse-to-fine benchmark designed for multi-level assessment of editing performance.

Data annotation. The images in CORE-Bench are sourced from the MSCOCO dataset [20] with object annotations. To rewrite finer instructions, we initialize our benchmark with five randomly selected base categories: object addition, object removal, object swap, background swap, and global change. As depicted in Fig. 4, our evaluation begins by formulating an atomic instruction, which is restricted to only object or stylistic classes with no detailed descriptions. The base instruction is then augmented with instance-specific details, such as color and shape, to demand intrinsic visual precision. The final level continues to introduce advanced priors, including position, lighting, and environmental interactions, requiring the model to maintain high-level global consistency and realism.

Through this process, we leverage GPT-4o to generate instructions across three granularities: base-level, attribute-level, and context-level. For each image, we rewrite two progressive paths derived from a single base instruction. Since removal tasks preclude the need for additive descriptions, we prioritize multi-instance scenes and utilize attribute descriptions solely for target disambiguation, while bypassing the redundant context stage. Following rigorous manual

verification and refinement to eliminate physical inconsistencies and ensure distinct boundaries between levels, we finalized a high-quality dataset of 1,090 coarse-to-fine editing instructions.

Evaluation Metric. Conventionally, CLIP-dir [28] score is employed to measure text-image direction similarity from the CLIP feature agreement between caption difference and image difference. However, CLIP-dir score suffers from the imprecision of fine-grained ground-truth captions (e.g. identical captions ‘a train station in city’ in Emu-Edit test [32]), a limitation that has discouraged its adoption in recent research [45]. Accordingly, we propose CLIP_{Delta} to accurately quantify coarse-to-fine directions by measuring the agreement between CLIP feature increments in editing instructions and edit results:

$$\text{CLIP}_{\text{Delta}} = \cos(\mathcal{E}_t(T_{\text{fine}}) - \mathcal{E}_t(T_{\text{coarse}}), \mathcal{E}_v(I_{\text{fine}}) - \mathcal{E}_v(I_{\text{coarse}})), \quad (2)$$

where $\mathcal{E}_t$ and $\mathcal{E}_v$ denote the CLIP text and visual encoders, respectively. T and I represent the editing instructions at different granularities and their corresponding edited results derived from the same source image.

Consequently, we can quantify the execution fidelity across hierarchical levels and analyze the semantic shifts between distinct instructions at the same stage. To ensure reliable assessment, we also employ DINO similarity [26] to measure geometric consistency and GPT-4o [12] for overall quality control.

4 Experiments

4.1 Experimental Settings

Implementation Details. We utilize Qwen3-VL-4B [1] as our MLLM module with coefficients $\lambda_{\text{ref}} = 1.0$ and $\lambda_{\text{edit}} = 2.5$. Both training stages use the Adam optimizer with a total batch size of 128. In the first stage, we optimize the Qwen3-VL decoder and linear connector for 50,000 steps with a learning rate of 1×10^{-4} . The second stage incorporates the ControlNet and DiT blocks to train the whole model for 20,000 steps with a maximum learning rate of 2×10^{-5} . Both MLLM and DiT blocks are fine-tuned using LoRA with rank $r = 64$ and alpha $\alpha = 128$.

Benchmarks and Metrics. To evaluate general editing performance on public benchmarks, we also utilize the Emu-Edit [32] and Reason-Edit [10] benchmarks to measure fundamental and reasoning capabilities. Unlike these public benchmarks tend to compute CLIP and DINO similarities between the edited output and the original input or instruction, CORE-Bench evaluates coarse-to-fine performance through cross-level comparisons. Our benchmark calculates CLIP_{Delta} and DINO by comparing results at the semantic and context levels against the original base level instructions and corresponding outputs. In CORE-Bench, we employ GPT-4o [12] to assess the overall editing quality focusing on attribute or contextual constraints given the source image and instruction.

Table 1: Quantitative results on the Emu-Edit benchmark [32]. Existing editing frameworks typically leverage various text encoders, coupled with base models for generation.

Methods	Text Encoder	Architecture	CLIP _{in} ↑	CLIP _{out} ↑	DINO↑
Instruct-Pix2Pix [3]	CLIP	U-Net	0.856	0.292	0.773
MagicBrush [42]	CLIP	U-Net	0.877	0.298	0.807
UltraEdit [47]	CLIP+T5	U-Net	0.880	0.304	0.847
OmniGen [40]	—	DiT	0.831	0.286	0.765
GoT [7]	MLLM	U-Net	0.864	0.287	0.784
IC-Edit [45]	CLIP+T5	DiT	0.907	0.305	0.866
Step1X-Edit [23]	MLLM	DiT	0.860	0.304	0.782
BAGEL [4]	LLM	DiT	0.839	0.307	0.753
Flux.1 Kontext [16]	CLIP+T5	DiT	0.875	0.296	0.792
Ours	MLLM	DiT	0.902	0.313	0.885

**Fig. 5:** Qualitative results of image editing models on Emu-Edit benchmark. By synthesizing visual imagination for fine-grained attributes consistent with final outputs, our model facilitates superior semantic alignment and editing quality.

4.2 Public Benchmarks

We benchmark our method against several representative open-source image editing models to evaluate its general editing performance and reasoning ability.

Results on Emu-Edit. Table 1 summarizes the quantitative results on the Emu-Edit benchmark. Overall, our framework CORE-V achieves state-of-

Table 2: Quantitative results on the Reason-Edit [10]. CORE-V consistently outperforms other reasoning-enhanced models particularly in reasoning scenarios.

Methods	Understanding Scenarios			Reasoning Scenarios			GPT-Eval↑
	PSNR↑	SSIM↑	CLIP↑	PSNR↑	SSIM↑	CLIP↑	
Instruct-Pix2Pix [3]	21.576	0.721	22.762	24.234	0.707	19.413	0.286
MagicBrush [42]	18.120	0.680	22.620	22.101	0.694	19.755	0.334
SmartEdit [10]	22.049	0.731	23.611	25.258	0.742	20.950	0.572
GoT [7]	22.255	0.713	22.943	26.067	0.804	20.699	0.561
BAGEL [4]	22.685	0.826	23.593	27.331	0.787	20.958	0.553
Ours	22.918	0.893	23.575	28.356	0.849	21.316	0.586

the-art (SOTA) performance, particularly in CLIP_{out} and DINO scores which indicate instruction alignment and structural consistency. Furthermore, in the qualitative comparisons Fig. 5, we visualize the reference imagination generated by CORE-V during the CoT reasoning. As CORE-V surpasses baselines in both structural fidelity and fine-grained detail, the attribute consistency between the intermediate imaginations and the final results validates the efficacy of our visual CoT paradigm.

Results on Reason-Edit. We further evaluate CORE-V on the Reason-Edit benchmark, as detailed in Table 2. CORE-V consistently outperforms baseline methods in both understanding and reasoning scenarios. While models like BAGEL utilize the thinking mode from pretrained MLLMs, the absence of a structured reasoning process hinders the effective translation of understanding into reasoning. Our methods address the limitation through well-defined CoT steps. With explicit textual reasoning for solving complex problems, CORE-V achieves superior synthesis with visual-semantic alignment.

4.3 Ablation Study

We also conduct our ablation studies of model structure and training objective coefficients on the public benchmarks with results in Table 3.

Ablation on Model Structure. As our core architectural innovations center on the CoT reasoning process, we conduct an incremental ablation study on the Reason-Edit benchmark to dissect the individual contributions of these key designs and evaluate their collective impact on reasoning efficacy. For fair comparison, we only update the MLLM decoder and linear connector while freezing other components. Replacing the MLLM encoder with CLIP impairs the overall editing quality. While applying pretrained MLLMs with textual CoT reasoning improves semantic alignment, the model lacks the visual information necessary for structural preservation. The best performance is achieved by adding spatial conditions and visual imagination, proving the effectiveness of both textual and visual CoT designs.

Ablation on Loss Coefficients. Each term within the total objective 1 corresponds to the specific component of our proposed CoT framework. While

Table 3: Ablation studies of CORE-V framework. (Left) Impact of loss coefficients on Emu-Edit. (Right) Effectiveness of key structural components on Reason-Edit.

λ_{ref}	λ_{edit}	CLIP _{im} ↑	CLIP _{out} ↑	DINO↑	MLLM Enc.	Textual CoT	Visual Imag.	Spatial Ctrl.	PSNR↑	SSIM↑	CLIP↑
1.0	2.5	0.902	0.313	0.885	✗	✗	✗	✗	21.307	0.754	21.78
1.0	3.0	0.908	0.309	0.869	✓	✗	✗	✗	23.257	0.795	22.22
1.0	4.0	0.886	0.297	0.879	✓	✓	✗	✗	23.544	0.803	22.54
0.8	2.5	0.887	0.298	0.880	✓	✓	✓	✗	25.012	0.858	22.46
0.6	2.5	0.869	0.291	0.865	✓	✓	✓	✓	24.567	0.880	22.89

Table 4: Quantitative results on CORE-Bench. The CLIP_{Delta} score is multiplied by 10^2 , and the total GPT-Eval scores are calculated on a 5-point scale. Following [42,47], we utilize ViT-B/32 and dino-vits16 for feature embeddings.

Methods	Semantic Level		Context Level		GPT-Eval↑
	CLIP _{Delta} ↑	DINO↑	CLIP _{Delta} ↑	DINO↑	
Instruct-Pix2Pix [3]	5.402	0.658	5.223	0.643	1.86
MagicBrush [42]	5.816	0.695	5.333	0.612	1.93
OmniGen [40]	6.080	0.672	5.516	0.596	1.97
Step1X-Edit [23]	7.312	0.821	7.179	0.891	2.85
Qwen-Image-Edit [39]	7.772	0.814	6.591	0.824	2.72
BAGEL [4]	8.131	0.756	7.565	0.833	3.00
Flux.1 Kontext [16]	6.557	0.733	6.533	0.805	2.65
Ours	7.998	0.876	7.901	0.881	3.24

fixing the text weight to 1.0, we evaluate the impact of coefficients λ_{ref} and λ_{edit} on the Emu-Edit. Ablation results show that over-weighting the final editing loss tends to compromise semantic alignment and reduced λ_{ref} leads to structural degradation. Consequently, we fix $\lambda_{\text{ref}} = 1.0$ and $\lambda_{\text{edit}} = 2.5$ to achieve an optimal balance between structural integrity and semantic logic.

4.4 CORE-Bench

We conduct a comprehensive quantitative comparison between our CORE-V framework and SOTA baselines including Qwen-Image-Edit [39] on the CORE-Bench. Table 4 summarizes the evaluation results across semantic and context levels, which are derived from the expansion of base instructions. While the semantic level focuses on apparent attributes of entities, the context level introduces the subtle environmental information. Abstract information like reflection, lighting and atmosphere, challenge models to derive fine-grained details from limited texts.

Compared with recent MLLM-based image editing methods such as Qwen-Image-Edit, CORE-V consistently achieves better semantic alignment and structural fidelity. Although our model slightly trails BAGEL in CLIP score at the

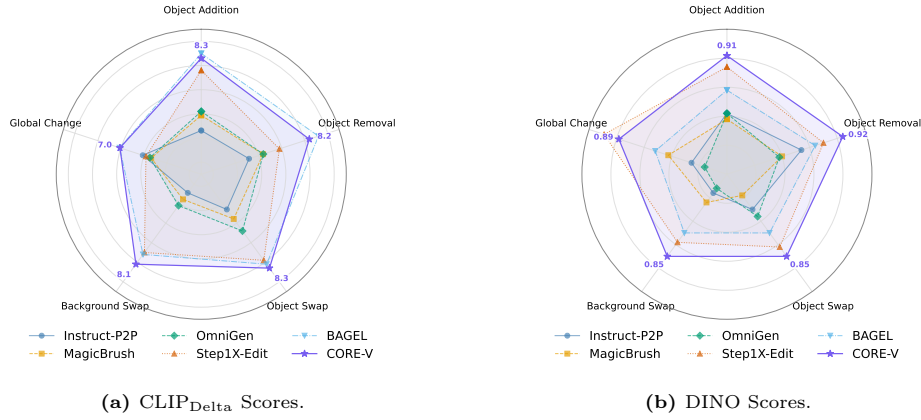


Fig. 6: Performance breakdown on the CORE-Bench. Both charts illustrate the robust performance of our framework in maintaining both semantic alignment and structural fidelity across five primary tasks.

semantic level, CORE-V demonstrates superior robustness in complex context-level tasks. Benefiting from our visual imagination design, our editing direction alignment remains stable and significantly outperforms other baselines. Regarding the structural fidelity, as measured by DINO similarity, our methods maintain consistent structural integrity with the base-level outputs. This is attributed to the introduction of visual conditions adaptively scheduled by CoT reasoning. Ultimately, these strengths are reflected in our top-tier GPT score, confirming the overall coarse-to-fine editing ability.

To provide a deeper analysis for the editing characteristics of our benchmark, we calculate the evaluation metrics average in the five main classes as Fig. 6. The results reveal that while local object operations are generally more tractable, tasks such as background swapping and global change exhibit higher complexity due to the requirement for environmental interaction and consistency. CORE-V demonstrates robust adaptability across these diverse scenarios. Fig. 4 visualizes results of the CORE-V model on the CORE-Bench. Further cross-level visualizations and corresponding evaluation prompts are deferred to the Appendix.

Finally, we conduct a user study involving 24 participants to evaluate 50 randomly sampled editing results from CORE-Bench. Following the scoring protocol detailed in the Appendix, participants rated each editing pair on a 5-point scale in terms of fine-grained compliance, visual naturalness, and structural consistency. As shown in Table 5, CORE-V outperforms the Qwen-Image-Edit, GoT, and Step1X-Edit across all evaluation criteria in human evaluation.

5 Conclusion

In this paper, we propose CORE-V, a CoT framework for controllable image editing. Our methods carefully design structured reasoning steps to induce the

Table 5: Quantitative results of user study on the CORE-Bench.

Method	GoT	Step1X-Edit	Qwen-Image-Edit	CORE-V
Semantic	2.32	3.00	3.31	3.87
Context	2.22	2.88	2.62	3.19
Overall	2.26	2.93	2.90	3.46

visual-semantic interaction. With a large-scale curated CoT dataset of both textual and visual reasoning, we build a unified framework to efficiently compute semantic guidance and visual conditions in reasoning. Specifically, with ground truth reference images consistent with the edit outputs, we optimize the model to concurrently generate visual imagination and final edit. In this case, our framework provides fine-grained attribute selections for controllable editing. Unlike current benchmarks limited to one-to-one editing, a coarse-to-fine benchmark, CORE-Bench, is built to measure semantic alignment and structural integrity across progressive instruction granularities. Together, the CORE-V framework and CORE-Bench establish a robust foundation for investigating visual reasoning with MLLMs and its manifestation in image editing capabilities.

Acknowledgments

This work was in part supported by the National Natural Science Foundation of China, under Grants 62125109, 62431017, 62320106003, 62371288, 62401357, in part by the National Key Research and Development Program of China under Grant 2025YFF0515600, and in part by the Fundamental Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM611).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. 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)
4. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
5. 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. In: International Conference on Learning Representations, ICLR (2026)

6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
7. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. In: Advances in Neural Information Processing Systems (2025)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations, ICLR (2022)
10. 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)
11. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Zhou, Y., Xie, C.: Hq-edit: A high-quality dataset for instruction-based image editing. In: International Conference on Learning Representations, ICLR (2025)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
13. 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)
14. 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. In: Advances in Neural Information Processing Systems. vol. 38, pp. 39856–39890 (2026)
15. Labs, B.F.: Flux: Official inference repository for flux.1 models. <https://github.com/black-forest-labs/flux> (2024), accessed: March 2026
16. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)
17. Labs, S.: Flux.1-dev-controlnet-union-pro-2.0. <https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro-2.0> (2025), accessed: March 2026
18. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai.github.io/controlnet_plus_plus. In: European Conference on Computer Vision. pp. 129–147. Springer (2024)
19. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147* (2025)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations, ICLR (2023)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
23. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761* (2025)
24. Lu, P.: Qwen2VL-Flux: Unifying image and text guidance for controllable image generation (2024), <https://github.com/erwold/qwen2vl-flux/blob/main/technical-report.pdf>
25. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
30. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations, ICLR (2025)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. 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)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations, ICLR (2021)
34. Stability AI: Stable diffusion 3.5. <https://huggingface.co/stabilityai/stable-diffusion-3.5-large> (2024), accessed: March 2026
35. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5117–5127 (2021)
36. Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., Xie, C.: Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033* (2025)
37. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: International Conference on Learning Representations, ICLR (2024)

38. 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)
39. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
40. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13294–13304 (2025)
41. xinsir: controlnet-union-sdxl-1.0. <https://huggingface.co/xinsir/controlnet-union-sdxl-1.0> (2024), accessed: March 2026
42. 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)
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 3836–3847 (2023)
44. Zhang, R., Tong, C., Zhao, Z., Guo, Z., Zhang, H., Zhang, M., Liu, J., Gao, P., Li, H.: Let’s verify and reinforce image generation step by step. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 28662–28672 (2025)
45. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025)
46. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research* (2024)
47. 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)