# EDITING TO CLARIFY AND HARNESS REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal Large Language Models have advanced visual reasoning, yet a purely textual chain of thought remains a bottleneck for questions that require fine-grained focus or view transformations. The "think with images" paradigm narrows this gap, but existing approaches are either constrained by fixed predefined toolkits or produce noisy intermediate images from unified multimodal methods. We pursue a third option: using a dedicated image editing model and decouple it with an understanding model. However, off-the-shelf image editors fail as reasoning assistants with two complementary gaps: a *language-side* gap, where editors trained as passive instruction-followers cannot map an abstract question to an appropriate visual transformation, and a *generation-side* gap, where edit correctness degrades as reasoning depth grows. Guided by this analysis, we introduce **ETCHR** (**E**diting **T**o **C**larify and **H**arness **R**easoning), a question-conditioned, reasoning-aware image editor decoupled from the downstream understanding model and trained with a two-stage recipe targeted at the two gaps: Reasoning Imitation via supervised fine-tuning on edit trajectories, followed by Reasoning Enhancement with VLM-derived rewards for edit correctness and downstream reasoning accuracy. Since the editor is decoupled, ETCHR plugs into different open- and closed-source MLLMs in a training-free manner. Across five task families (fine-grained perception, chart understanding, logic reasoning, jigsaw restoration, and 3D understanding), ETCHR raises average Pass@1 from 55.95 to 60.77 (+4.82) with Qwen3-VL-8B, from 65.08 to 70.55 (+5.47) with Gemini-3.1-Flash-Lite, and from 76.55 to 81.16 (+4.61) with Kimi K2.5.

## 1 INTRODUCTION

Recent Multimodal Large Language Models (MLLMs) Hurst et al. (2024); Comanici et al. (2025) have substantially improved visual reasoning, but a purely textual chain-of-thought Wei et al. (2022) remains a bottleneck when a question depends on *where* to look or *how* a scene would change under an action. In such cases, the model must verbalize a spatial state it cannot draw, and small descriptive errors compound across steps. A growing line of work therefore enables MLLMs to **"think with images"** OpenAI (2025); Zheng et al. (2025); Hu et al. (2024), generating an intermediate image during inference and feeding it back into the reasoning trajectory, with reported gains on fine-grained visual search Wu & Xie (2024); Zheng et al. (2025); Hu et al. (2024), chart reasoning Zhang et al. (2025); Hong et al. (2025), and spatial navigation Gu et al. (2025); Li et al. (2025a).

The difficulty of the "think with images" paradigm is not simply generating an image, but generating the *right* intermediate image for the question. The system must infer what visual change would advance the answer, render that change faithfully, and produce an intermediate image that the downstream understanding model can exploit. This couples two capabilities that are hard to satisfy simultaneously: **(i)** the *breadth* of visual transformations the system can conceive and execute, ranging from fine-grained highlight of local elements to holistic transitions in spatial perspective; and **(ii)** the *fidelity* with which those transformations are rendered, ensuring the intermediate image can indeed serve as reliable evidence for downstream reasoning. Only when both capabilities are present can a think-with-image framework deliver high-quality assistance across diverse understanding tasks.

Existing approaches address only one side. *Tool-based* methods (Fig. 1 **(a)**) Hu et al. (2024); Wu & Xie (2024); Zheng et al. (2025); Hong et al. (2025); Zhang et al. (2025) fine-tune the understanding model to emit bounding boxes, crop/zoom commands, or executable snippets that a deterministic renderer applies; the actions are controllable but confined to low-level, localized manipulations,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

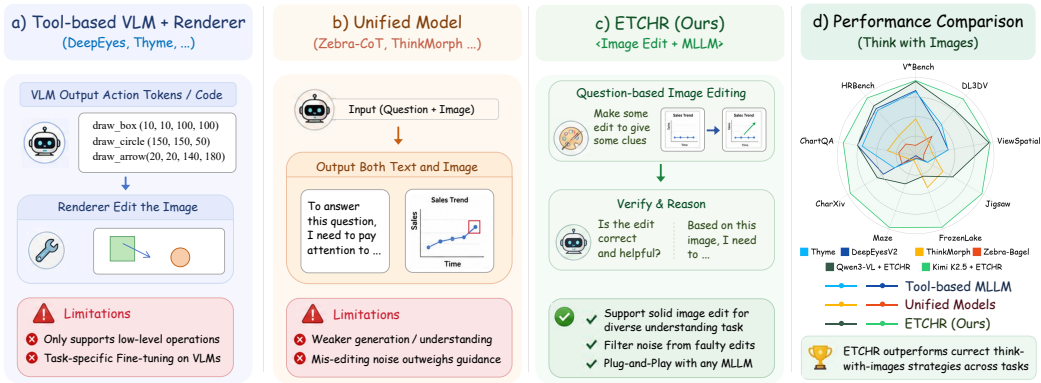



Figure 1: **ETCHR vs. prior "think with images" paradigms.** **(a)** *Tool-based* methods emit action tokens to a renderer, limiting edits to low-level operations and requiring VLM fine-tuning. **(b)** *Unified models* share one backbone for text and images, weakening both and producing noisy intermediates. **(c)** ETCHR decouples a question-conditioned editor from the understanding MLLM and adds a verify-and-reason step, enabling plug-and-play use across tasks. **(d)** Across nine benchmarks, ETCHR (with Qwen3-VL-8B and Kimi K2.5 1T) surpasses tool-based and unified-model baselines.

and task-specific fine-tuning can erode the model's general competence Kirkpatrick et al. (2017). *Unified multimodal* methods (Fig. 1 **(b)**) Team (2024); Chern et al. (2024); Deng et al. (2025); Gu et al. (2025); Li et al. (2025a) instead use a single backbone to interleave text and image tokens, gaining flexibility but inheriting a weaker generative head: recent work Wen et al. (2026) shows that their intermediate images often inject noise rather than guidance, while the unified backbone itself lags specialist understanding and generation models Wu et al. (2025b); Xie et al. (2024). Both families share a further blind spot: **neither verifies whether the intermediate edit is actually correct** before reasoning forward from it Madaan et al. (2023); Shinn et al. (2023); Huang et al. (2023), allowing a noisy edit to propagate directly into the final answer.

These limitations motivate us to pursue a **third option**: using a dedicated image-to-image editor as the intermediate-image generator. A specialist editor expresses a far broader transformation space than the predefined actions of tool-based methods preserves the fidelity of a model dedicated to editing rather than the weaker generative head of a unified backbone, and decouples edit quality from the understanding model so the latter need not be retrained on task-specific editing formats. This option is now plausible architecturally: modern image-to-image editors (e.g., FLUX-class Labs (2025) and Qwen-Image-Edit Wu et al. (2025a)) replace the shallow CLIP-style text encoder Radford et al. (2021) of earlier editors with an MLLM-style encoder, giving the editor enough language-side capacity to parse a complex question and support cross-task visual transformations.

Realizing this option, however, is non-trivial. Image editors are normally trained as *passive instruction-following tools* that expect an explicit edit prompt such as "add a bounding box around the red car" Brooks et al. (2023); Zhang et al. (2023), whereas a model that thinks with images is handed only a *question* and must itself decide what edit would help answer it. Moreover, the generative components of current editing models often lack the capacity for sophisticated reasoning, making it difficult to assist those understanding tasks requiring complex logic or 3D spatial manipulations. The **central challenge** is therefore to turn the editor into an autonomous question-conditioned system with strong reasoning capabilities, whose intermediate images are genuinely reasoning-useful rather than merely visually plausible. Correctness is also critical: an erroneous edit can silently mislead the downstream understanding model without verification Madaan et al. (2023); Shinn et al. (2023); Huang et al. (2023), so the system must also be able to detect and reject its own bad edits before they enter the reasoning trajectory.

We instantiate this design as **ETCHR** (**E**diting **T**o **C**larify and **H**arness **R**easoning, (Fig. 1 **(c)**)): a **question-conditioned, reasoning-aware image editor** coupled with edit-verification at inference. This design uses question-conditioned editing to strengthen *language-side reasoning* and a dedicated specialist editor to preserve *generative fidelity*. Unlike *tool-based* methods, ETCHR is not confined to a predefined action space and infers what to change directly from the question. Unlike *unified multimodal* methods, ETCHR keeps understanding and generation in separate specialist models, so editing fidelity is not compromised by joint optimization. Going beyond both families, ETCHR

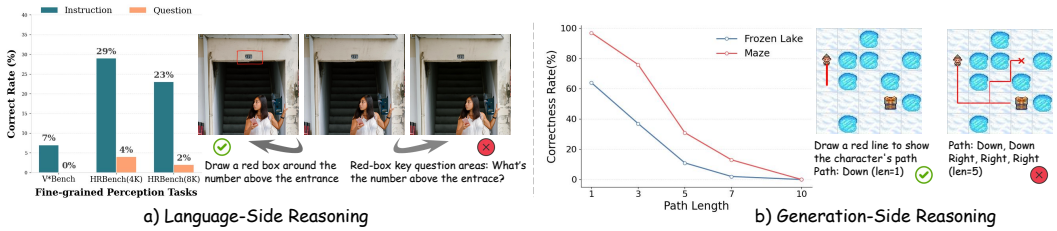



Figure 2: **Diagnosis of two complementary gaps in image editors. (a) Language-side reasoning**: editors trained as instruction-followers perform worse on abstract questions than on concrete instructions, revealing a gap in mapping reasoning questions to visual transformations. **(b) Generation-side reasoning**: even with concrete instructions, edit correctness decreases as path length $L$ grows on Maze Solving, indicating difficulty executing multi-step reasoning during DiT decoding.

incorporates a *reflective inference step* at inference, letting the understanding model reject noisy edits rather than propagate them to the final answer.

ETCHR is trained in two stages and deployed with a reflective inference procedure. (1) **Reasoning Imitation** applies supervised fine-tuning to convert the editor from a passive renderer into a question-conditioned editor that infers the useful transformation from the question alone. (2) **Reasoning Enhancement** then uses reinforcement learning with *reasoning-aware rewards*, rather than only visual plausibility, to push outputs toward edits that are correct in isolation and useful for downstream reasoning. At inference, ETCHR follows an **Edit-Verify-Reason** procedure: the editor proposes an intermediate image, the understanding model verifies whether the proposed edit is reliable, and the answer is produced from the edited image only when verification succeeds; otherwise, the system falls back to the original image.

On a diverse benchmark suite covering fine-grained perception (V*Bench Wu & Xie (2024), HRBench Wang et al. (2025a)), chart understanding (ChartQA Masry et al. (2022), CharXiv Wang et al. (2024)), logic and path reasoning (Maze, Frozen Lake), jigsaw reasoning (built on COCO Lin et al. (2014)), and 3D understanding (ViewSpatial Li et al. (2025b)), ETCHR raises average Pass@1 from 55.95 to 60.77 (+4.82) with Qwen3-VL-8B Bai et al. (2025), from 65.08 to 70.55 (+5.47) with Gemini-3.1-Flash-Lite and from 76.55 to 81.16 (+4.61) with Kimi K2.5, outperforming tool-based Hong et al. (2025); Zhang et al. (2025) and unified-model Li et al. (2025a); Gu et al. (2025) baselines (Fig. 1 **(d)**). Ablations confirm that both training stages, the two reward signals (editing correctness and editing guidance), and the Edit-Verify-Reason reflection at inference, each contribute and are complementary.

Our contributions: **1)** We propose a mechanism for "thinking with images" that uses a dedicated image-to-image editor as the source of intermediate visual evidence, with an inference-time verification step that lets the understanding model reject unreliable edits. **2)** We instantiate this mechanism as **ETCHR**, a reasoning-aware editor built on FLUX.2-klein-base-9B Labs (2025) and trained with a two-stage recipe: *Reasoning Imitation*, supervised fine-tuning on question-conditioned edit trajectories, followed by *Reasoning Enhancement* with two VLM-derived reward signals. **3)** Because ETCHR conditions on the question and is decoupled from the MLLMs, it pairs with different open- and closed-source MLLMs without fine-tuning them, evaluated across reasoning tasks spanning fine-grained perception, spatial and path reasoning, puzzle restoration, and 3D understanding.

## 2 Analysis

To motivate ETCHR, we decompose reasoning-aware editing into two sub-capabilities that current editors lack. *Language-side reasoning* infers, from an abstract question alone, what visual transformation would help answer it. *Generation-side reasoning* faithfully renders that transformation when it requires non-trivial spatial or algorithmic inference (e.g., tracing a maze path). We diagnose both gaps with off-the-shelf editors and understanding models.

**Language-Side Reasoning: From Question to Edit.** Modern image-to-image editors (e.g., FLUX.2-klein-base-9B Labs (2025), InstructPix2Pix Brooks et al. (2023)) are optimized as instruction-following tools: they expect an explicit edit prompt such as "draw a red box around

the trash can" rather than an abstract question such as "Is the trash can on the left or right side of the black chair?" We therefore probe a question-to-edit gap: whether the editor can recover the useful transformation from the raw question alone.

We use Gemini-3.1-Flash-Lite Comanici et al. (2025) as a prompt enhancer that converts each question $q$ into a concrete editing instruction $p_{inst}$, and Qwen3-VL-8B Bai et al. (2025) judges correctness on 100 samples from V*Bench Wu & Xie (2024) and HRBench Wang et al. (2025a). As shown in Fig. 2 (a), the concrete-instruction condition significantly outperforms the abstract-question condition. Shaped by explicit instruction-following, the base editor lacks a reliable question-to-edit mapping; reasoning-aware editing therefore needs more than generic image-editing fidelity.

**Generation-Side Reasoning: Scaling Robustness to Reasoning Depth.** Despite efforts to address the weak language-side reasoning with prompt enhancers, we argue that their reasoning ability is even more lacking during the DiT decoding phase. Even with a concrete instruction, an editor can fail when the transformation itself requires multi-step spatial reasoning during DiT decoding. We measure edit correctness as a function of task complexity on Maze Solving and Frozen Lake Solving.

For these two tasks, we construct a held-out set of 100 samples with shortest-path lengths $L \in \{1, 3, 5, 7, 10\}$ and present the shortest path in the text prompt. A VLM-as-Judge marks an edit correct only if the highlighted path matches the input. Fig. 2 (b) plots edit correctness against $L$: accuracy is near-perfect at $L$=1 and decreases sharply as $L$ grows, approaching zero on the longest paths. This indicates that even when instructions are provided with maximum precision and granularity, current editing models still struggle to execute operations requiring multi-hop reasoning during the DiT decoding phase.

**Summary.** Together, the two gaps motivate ETCHR's two-stage design: Stage I question-conditioned imitation to close the question-to-edit gap and equip with strong reasoning capabilities, and Stage II reasoning-oriented enhancement to further enhance multi-step generation.

# 3 ETCHR: Editing To Clarify and Harness Reasoning

ETCHR trains a question-conditioned image editor in two stages. After preliminaries (Sec. 3.1), **Stage I (Reasoning Imitation, Sec. 3.2)** converts the base editor into a question-conditioned editor via supervised fine-tuning, and **Stage II (Reasoning Enhancement, Sec. 3.3)** aligns it with downstream reasoning utility via RL under VLM-derived rewards. At inference, **Edit-Verify-Reason** (Sec. 3.4) verifies each edit and reverts to the original image on failure.

## 3.1 Preliminaries

We cast reasoning-aware editing as a question-conditioned image-to-image task. Each training instance is a tuple $\{i, i_{gt}, q, a\}$, where $i$ is the input image, $q$ is a natural-language question, $a$ is a ground-truth answer, and $i_{gt}$ is a ground-truth edited image that surfaces the visual evidence needed to derive $a$ from $i$ (e.g., a traced path through a maze). ETCHR learns a question-conditioned editor $\mathcal{E}$ that maps $(i, q)$ to an edited image $i_{edit} = \mathcal{E}(i, q)$ serving the same evidential role as $i_{gt}$. Quality of $i_{edit}$ is judged by an understanding model $\mathcal{M}$: a useful edit satisfies $\mathcal{M}(i, i_{edit}, q) = a$ while the unaided baseline satisfies $\mathcal{M}(i, q) \neq a$. Notations are reused throughout Sec. 3.2 and Sec. 3.3.

## 3.2 Reasoning Imitation Supervised Fine-tuning (SFT)

**Data Preparation.** We build a large-scale SFT corpus of question-conditioned edit trajectories $\{i, i_{gt}, q, a\}$, partitioned into *five reasoning families* spanning the visual transformations required by downstream reasoning tasks: *fine-grained perception* probes localization of small or easily missed referents; *chart understanding* probes grounding in structured plots; *logic reasoning* probes multi-step algorithmic inference; *jigsaw reasoning* probes global geometric reorganization; and *3D understanding* probes viewpoint and camera-pose transformation. Our task coverage, from local annotation to whole-image rearrangement, prevents collapse onto a single edit template and forces the editor to acquire a *meta-capability* of inferring what visual transformation each question demands.

For *fine-grained perception*, we use the $V^*$ training dataset Wu & Xie (2024), covering both $V^*$-GQA and $V^*$-COCO subsets, and synthesize $i_{gt}$ by rendering the annotated bounding boxes onto

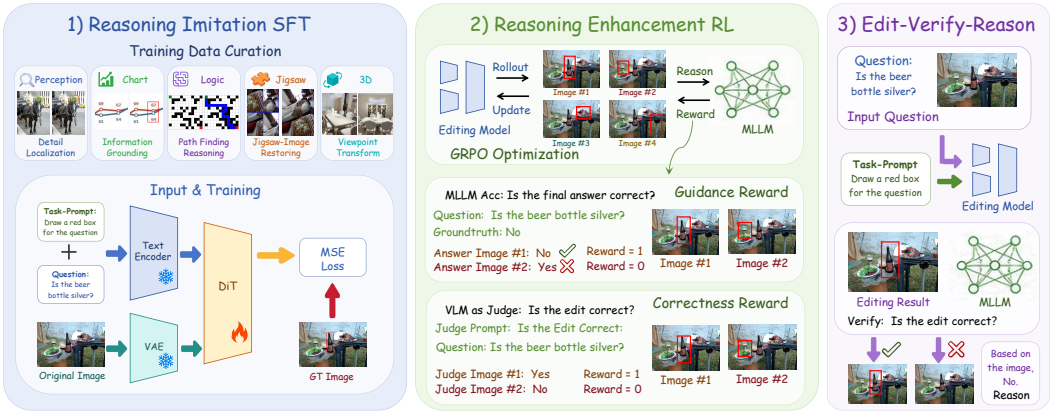



Figure 3: **Overview of ETCHR. (1) Reasoning Imitation SFT** fine-tunes the base editor on curated question-edit trajectories, teaching it to infer the useful visual transformation from the question alone. **(2) Reasoning Enhancement RL** applies GRPO with two VLM-derived rewards, edit correctness and downstream answer accuracy, to align the editor with reasoning utility rather than visual plausibility. **(3) Edit-Verify-Reason** inference lets the understanding model verify each proposed edit and fall back to the original image when verification fails.

$i$. For *chart understanding*, we draw from RefChartQA Vogel et al. (2025) and apply the same bounding-box overlay procedure to obtain $i_{gt}$. For *logic reasoning*, we build an in-house maze corpus where $i$ shows the maze topology and $i_{gt}$ overlays the correct traversal path. For *jigsaw reasoning*, we sample from Spatial-SSRL Liu et al. (2025), taking $i$ as a spatially shuffled image and $i_{gt}$ as its restoration. For *3D understanding*, we use DL3DV-10K Ling et al. (2024), which contains videos of real-world 3D scenes together with per-frame camera poses. We sample $i$ and $i_{gt}$ from the same video and synthesize $q$ and $a$ from the camera extrinsics.

**Task-level Prompt Enhancement.** Using $q$ alone as the editing prompt causes severe cross-task interference: the editor's latent space, shaped by explicit instruction-following, lacks the priors to disambiguate whether a question demands localization, path-tracing, rearrangement, or viewpoint transformation. We therefore prepend a task-level meta-prompt $p_{task}$ to $q$, evoking the editing modality appropriate to each task family. At training, $p_{task}$ acts as a soft task-router that partitions the editor's latent space into task-specific manifolds and suppresses gradient conflicts across families. At inference, it adds no architectural cost: because $p_{task}$ is task-level, the editor needs no access to the understanding model's internal representations or per-instance instructions, letting ETCHR be deployed in a training-free manner atop any open- or closed-source MLLMs.

**Training Strategy.** We use Low-Rank Adaptation (LoRA) Hu et al. (2022) to fine-tune the Diffusion Transformer (DiT) Peebles & Xie (2023) of FLUX.2-klein-base-9B Labs (2025), keeping the VAE and text encoder frozen as inputs are standard frames and prompts are simply $p_{task}$ concatenated with $q$. We apply a large LoRA rank ($r = 768$) to all linear layers in the DiT blocks to provide sufficient capacity for multi-task learning.

### 3.3 Reasoning Enhancement Reinforcement Learning (RL)

**Data Preparation.** We curate the RL set by sampling 2,000 instances from each of the five families (fine-grained perception, chart understanding, logic reasoning, jigsaw, and 3D understanding) in the SFT corpus, yielding 10,000 pairs. An instance $(i, i_{gt}, q, a)$ is retained only if it satisfies:

$$\mathcal{C}(i, i_{gt}, q, a) = \mathbf{1}[\mathcal{M}(i, q) \neq a \wedge \mathcal{M}(i, i_{gt}, q) = a], \tag{1}$$

where the notation follows Sec. 3.1. We keep only samples where (i) the understanding model fails on the raw image and (ii) succeeds when conditioned on the ground-truth edit. The first clause discards instances already solvable without visual assistance, avoiding wasted gradients on redundant edits; the second guarantees a verifiable upper-bound signal per sample, yielding a denser reward landscape and lower-variance policy gradients.

**Reward Design.** Edit quality admits no direct scalar metric, so we infer it through two rewards that probe distinct facets of edit quality and compensate for each other's blind spots.

*Editing Guidance Reward ($r_{guide}$).* The first reward measures the downstream reasoning utility of $i_{edit}$: it is one iff $\mathcal{M}$ answers correctly given $(i, i_{edit})$,

$$r_{guide}(i, i_{edit}, q, a) = \mathbf{1}[\mathcal{M}(i, i_{edit}, q) = a]. \tag{2}$$

$r_{guide}$ optimizes the end-to-end objective of producing edits that help $\mathcal{M}$ answer correctly. Its fidelity, however, is bounded by $\mathcal{M}$'s capability ceiling: on easy questions, $\mathcal{M}$ may succeed with a partially erroneous edit, and on hard ones it may fail even with a perfect edit. The filter $\mathcal{C}$ in Eq. (1) reduces but cannot remove this coupling, motivating a second, decoupled signal.

*Editing Correctness Reward ($r_{correct}$).* The second reward evaluates the edit *in isolation*, without solving the underlying task. Following the VLM-as-Judge paradigm Zheng et al. (2023), a judge VLM $\mathcal{J}$ assesses only whether $i_{edit}$ contains the visual information needed to answer $q$:

$$r_{correct}(i, i_{edit}, q) = \mathbf{1}[\mathcal{J}(i, i_{edit}, q) = 1]. \tag{3}$$

$r_{correct}$ lifts the capability ceiling: $\mathcal{J}$ can recognize that a bounding box highlights the correct referent or an annotation traces the valid path, even without the reasoning depth to derive the answer, giving a broader notion of edit adequacy. The cost is judge noise: $\mathcal{J}$ may accept a plausible-but-uninformative edit, or reject a correct one due to superficial visual differences.

The two rewards are complementary: $r_{guide}$ is faithful but ceiling-bound, while $r_{correct}$ is ceiling-lifting but judge-noisy. We combine them as a convex sum:

$$\mathcal{R}(i, i_{edit}, q, a) = \alpha\, r_{guide} + \beta\, r_{correct}, \tag{4}$$

with $\alpha = \beta = 0.5$ by default, so that each term partially compensates for the other's blind spot. We soften both indicators into probabilities: $\mathcal{M}$ and $\mathcal{J}$ are each queried with $K$ stochastic decodings, and $r_{guide}$, $r_{correct}$ are the fractions of correct answers and positive verdicts, respectively.

**Optimization.** We use Pref-GRPO Wang et al. (2025b), a pairwise-preference extension of GRPO Shao et al. (2024). Given a rollout group of $G$ edited images $\{i^j_{edit}\}_{j=1}^G$ from the same $(i, q)$, we compute a win rate for each image under the reward $\mathcal{R}$. The win rate of image $m$ is:

$$w_m = \frac{1}{G-1} \sum_{n \neq m} \mathbf{1}[\mathcal{R}(i, i^m_{edit}, q, a) > \mathcal{R}(i, i^n_{edit}, q, a)]. \tag{5}$$

Win rates are normalized within the group to give the advantage:

$$\hat{A}_m = \frac{w_m - \mathrm{mean}(\{w_j\}_{j=1}^G)}{\mathrm{std}(\{w_j\}_{j=1}^G)}. \tag{6}$$

By replacing absolute reward with pairwise preference, Pref-GRPO amplifies reward variance and absorbs reward noise, mitigating the illusory advantage problem and stabilizing training.

### 3.4 Edit-Verify-Reason Inference

Even a well-trained editor can produce misleading intermediates. To prevent such errors from contaminating reasoning, ETCHR adds a reflective inference procedure that verifies edit reliability before use. The pipeline has three steps: **(1) Edit.** The editor takes the concatenated prompt $[p_{task}, q]$ with image $i$ and produces a candidate edit $i_{edit}$. **(2) Verify.** $\mathcal{M}$ performs a binary check on whether $i_{edit}$ contains the visual information needed to answer $q$, without producing the answer itself. **(3) Reason.** If verification passes, $\mathcal{M}$ conditions on both $i$ and $i_{edit}$ to produce the answer: $a = \mathcal{M}(i, i_{edit}, q)$. Otherwise, it falls back to the original image alone: $a = \mathcal{M}(i, q)$. Verification is vital as edit errors are asymmetric in cost: a correct edit gives decisive visual guidance, while an incorrect one introduces structured confounders that MLLMs struggle to override.

## 4 Experiments

**Models.** We build our pipeline on top of multiple understanding models: the open-source Qwen3-VL-8B-Instruct Bai et al. (2025), Kimi K2.5 Team et al. (2026), and the closed-source Gemini-3.1-Flash-Lite-Preview Comanici et al. (2025).

Table 1: **Main results on five task families.** ETCHR is compared against tool-based and programmatic baselines and unified multimodal models. † marks tasks outside a baseline's supported image-editing scope; all numbers are Pass@1 at temperature 0.

| Models | Fine-grained Perception | | Chart Understanding | | Logic | | Jigsaw | 3D Understanding | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| | V*Bench | HRBench (4K / 8K) | ChartQA | CharXiv (DQ / RQ) | Maze | Frozen Lake | COCO Jigsaw | ViewSpatial Person-RelDir | DL3DV-2k | |
| *Tool-based & Programmatic Models* | | | | | | | | | | |
| DeepEyesV2 | 81.8 | 77.9 / 73.8 | 88.4 | 78.6 / 48.9 | 0.5† | 0.0† | 6.5† | 39.3† | 69.0† | **51.34** |
| Thyme | 81.2 | 77.1 / 71.9 | 86.5 | 66.1 / 45.3 | 0.0† | 2.0† | 6.5† | 39.2† | 70.8† | **49.69** |
| *Unified-Models* | | | | | | | | | | |
| Bagel-Zebra-CoT | 50.8 | 51.9 / 45.8 | 71.8 | 62.8 / 37.4 | 0.0 | 1.5 | 5.5 | 28.6 | 64.9 | **38.27** |
| ThinkMorph-7B | 65.8 | 59.6 / 50.9 | 76.4 | 61.2 / 28.1 | 6.5 | 31.0 | 10.5 | 32.2 | 62.4 | **44.05** |
| *Baselines & Our Image Editing Model* | | | | | | | | | | |
| Qwen3-VL-8B | 84.8 | 78.1 / 74.5 | 86.7 | 85.0 / 47.9 | 27.5 | 9.5 | 9.5 | 41.1 | 70.8 | **55.95** |
| +ETCHR | 86.9 | 80.6 / 75.3 | 88.4 | 86.8 / 49.7 | 38.5 | 13.5 | 13.0 | 57.2 | 78.6 | **60.77** |
| *Improvement* | +2.1 | +2.5 / +0.8 | +1.7 | +1.8 / +1.8 | +11.0 | +4.0 | +3.5 | +16.1 | +8.2 | **+4.82** |
| Gemini-3.1-Flash-Lite | 69.6 | 81.5 / 74.5 | 81.2 | 93.8 / 68.6 | 40.0 | 62.5 | 27.5 | 46.6 | 70.1 | **65.08** |
| +ETCHR | 72.3 | 82.9 / 75.1 | 83.3 | 94.6 / 70.8 | 51.5 | 69.0 | 36.0 | 57.8 | 82.7 | **70.55** |
| *Improvement* | +2.7 | +1.4 / 0.6 | +2.1 | +0.8 / 2.2 | +11.5 | +6.5 | +8.5 | +11.2 | +12.6 | **+5.47** |
| Kimi K2.5 | 85.9 | 85.6 / 80.6 | 93.7 | 98.3 / 77.1 | 95.5 | 52.5 | 44.5 | 54.9 | 73.4 | **76.55** |
| +ETCHR | 87.4 | 86.8 / 81.1 | 94.1 | 98.4 / 77.9 | 98.0 | 57.0 | 70.5 | 57.7 | 83.9 | **81.16** |
| *Improvement* | +1.5 | +1.2 / 0.5 | +0.4 | +0.1 / 0.8 | +2.5 | +4.5 | +26.0 | +2.8 | +10.5 | **+4.61** |

Table 2: **Comparison with a closed-source image editor.** Pass@1 of Qwen3-VL-8B and Gemini-3.1-Flash-Lite with Nano Banana 2 (NB2) vs. ETCHR editor on a 100-sample-per-benchmark subset.

| Model | Per. | Cha. | Log. | Jig. | 3D | **Avg.** |
|---|---|---|---|---|---|---|
| Qwen3-VL | 80.3 | 75.7 | 17.5 | 11.0 | 56.0 | 48.10 |
| + NB2 | 82.0 | 77.0 | 20.5 | 13.0 | 55.5 | 49.60 |
| **+ Ours** | 82.0 | 77.3 | 24.0 | 15.0 | 66.0 | **52.86** |
| Gemini-3.1 | 75.3 | 83.0 | 50.5 | 28.0 | 60.0 | 59.36 |
| + NB2 | 77.7 | 84.0 | 53.5 | 32.0 | 59.0 | 61.24 |
| **+ Ours** | 77.3 | 84.3 | 61.0 | 35.0 | 68.5 | **65.22** |

Table 3: **Ablation of the reflection mechanism.** Pass@1 with and without the verification step in the Edit-Verify-Reason pipeline across Perception (Per.), Chart (Cha.), Logic (Log.), Jigsaw (Jig.), and 3D.

| Model | Per. | Cha. | Log. | Jig. | 3D | **Avg.** |
|---|---|---|---|---|---|---|
| Qwen3-VL | 79.1 | 73.2 | 18.5 | 9.5 | 56.0 | 47.26 |
| w.o. reflection | 79.3 | 73.2 | 25.5 | 13.0 | 68.2 | 51.84 |
| w. reflection | 80.9 | 75.0 | 26.0 | 13.0 | 67.9 | **52.56** |
| Gemini-3.1 | 75.2 | 81.2 | 51.3 | 27.5 | 58.4 | 58.72 |
| w.o. reflection | 76.0 | 81.8 | 59.0 | 34.5 | 70.7 | 64.40 |
| w. reflection | 76.8 | 82.9 | 60.3 | 36.0 | 70.3 | **65.26** |

**Benchmarks.** We evaluate our model on five categories of understanding tasks. **(1)** *Fine-grained perception:* V*Bench Wu & Xie (2024), which probes recognition of small details in high-resolution natural images, and HRBench (4K/8K) Wang et al. (2025a), which extends this setting to 4K and 8K inputs. **(2)** *Chart understanding:* ChartQA Masry et al. (2022) for question answering over bar, line, and pie charts, and CharXiv Wang et al. (2024) (Descriptive/Reasoning), which targets charts from arXiv papers. **(3)** *Logical reasoning:* 200 Maze and 200 Frozen Lake tasks that we construct in-house, requiring the model to navigate from a start cell to a goal while respecting walls or hazards. **(4)** *Jigsaw:* 200 puzzles built from MS COCO Lin et al. (2014) images by applying tile-shuffling strategies, which probe whether the model can recover image structure from permuted patches. **(5)** *3D perception and reasoning:* the Person-Perspective Relative-Direction subtask of ViewSpatial-Bench Li et al. (2025b), together with DL3DV-2k Ling et al. (2024), a benchmark we constructed using scenes and images from DL3DV, comprising 2000 questions targeting the reasoning capability over 3D perspective-transformation.

### 4.1 Main Results

**Overall Performance.** Tab. 1 compares ETCHR against the two prevailing paradigms. Tool-based and programmatic models are competitive on chart understanding and within a few points on fine-grained perception, but their predefined action spaces do not cover tasks requiring non-local edits (Logic, Jigsaw, 3D), which we mark with † to indicate unsupported task scope rather than evaluated failure. Unified models avoid this coverage limit by natively interleaving generation, yet on average they trail specialized understanding backbones by more than ten points (e.g., ThinkMorph-7B at 44.05 and Bagel-Zebra-CoT at 38.27 vs. Qwen3-VL-8B at 55.95), reflecting the well-known tension

Table 4: **Ablation of the two-stage training recipe.** Pass@1 of the base FLUX.2-klein-base-9B editor, after Stage I SFT (+*sft*), and after Stage II RL (+*sft&rl*).

| Model | Per. | Cha. | Log. | Jig. | 3D | **Avg.** |
|---|---|---|---|---|---|---|
| Qwen3-VL | 79.1 | 73.2 | 18.5 | 9.5 | 56.0 | 47.26 |
| Flux | 74.3 | 68.9 | 18.5 | 9.5 | 55.1 | 45.26 |
| + sft | 79.7 | 74.1 | 25.8 | 13.0 | 67.3 | 51.98 |
| + sft&rl | 80.9 | 75.0 | 26.0 | 13.0 | 67.9 | **52.56** |
| Gemini-3.1 | 75.2 | 81.2 | 51.3 | 27.5 | 58.4 | 58.72 |
| Flu | 70.3 | 74.5 | 51.3 | 27.0 | 59.5 | 56.52 |
| + sft | 76.0 | 82.4 | 59.5 | 36.0 | 70.0 | 64.78 |
| + sft&rl | 76.8 | 82.9 | 60.3 | 36.0 | 70.3 | **65.26** |

Table 5: **Ablation of the two RL reward signals in Stage II RL.** Pass@1 with Editing Correctness only (Cor.), Editing Guidance only (Gui.), or both, , paired with Qwen3-VL-8B.

| Cor. | Gui. | Per. | Cha. | Log. | Jig. | 3D | **Avg.** |
|---|---|---|---|---|---|---|---|
| *Qwen3-VL-8B-Instruct* | | | | | | | |
| ✓ | ✗ | 80.7 | 74.7 | 25.8 | 13.0 | 66.8 | 52.20 |
| ✗ | ✓ | 80.2 | 74.2 | 26.0 | 13.0 | 67.6 | 52.20 |
| ✓ | ✓ | 80.9 | 75.0 | 26.0 | 13.0 | 67.9 | **52.56** |
| *Gemini-3.1-flash-lite-preview* | | | | | | | |
| ✓ | ✗ | 76.6 | 82.5 | 60.0 | 35.5 | 69.7 | 64.86 |
| ✗ | ✓ | 76.1 | 82.5 | 60.0 | 36.0 | 69.9 | 64.90 |
| ✓ | ✓ | 76.8 | 82.9 | 60.3 | 36.0 | 70.3 | **65.26** |

between a jointly optimized generative head and pure understanding quality. ETCHR sidesteps both pitfalls by treating a specialist editing model as a question-conditioned reasoning module, and improves every evaluated backbone (Qwen3-VL-8B, Gemini-3.1-Flash-Lite, Kimi K2.5 across the task families where comparisons are available.

**Comparison with a Closed-Source Editor.** Tab. 2 compares our editor against the closed-source Nano Banana 2 Google (2026) as the visual-thought provider for Qwen3-VL-8B and Gemini-3.1-Flash-Lite. To contain API cost, both editors are evaluated on the same 100-sample-per-benchmark subset, so absolute numbers should be read as trends rather than significance-tested estimates. Both editors improve over the no-edit baseline on most tasks, confirming that question-conditioned edits are useful even when produced by a generic frontier editor. Our editor is comparable to Nano Banana 2 on Perception and Chart and shows larger margins on Logic, Jigsaw, and 3D, where edits must encode task structure rather than re-render local regions. Results are consistent with our hypothesis that reasoning-aware training, not editor scale, drives guidance quality on structural tasks.

### 4.2 Ablation Analysis

To isolate each design choice in ETCHR, we ablate (i) the two-stage training recipe (Tab. 4), (ii) the two RL reward signals, Editing Correctness and Editing Guidance (Tab. 5), and (iii) the inference-time reflection mechanism (Tab. 3). All settings are evaluated with both understanding backbones, Qwen3-VL-8B and Gemini-3.1-Flash-Lite.

**Training Recipe:** Tab. 4 compares the base FLUX.2-klein-base-9B Labs (2025) editor with its Stage I SFT and Stage II RL checkpoints. The base editor (where evaluable) tracks the no-edit backbone. Stage I SFT improves all five task families on both backbones, suggesting the LLM encoder retains the reasoning capacity needed for question-conditioned editing and that supervised imitation suffices to reactivate it. Stage II RL adds under one point on Perception and Chart and is essentially flat on Logic, Jigsaw, and 3D. We attribute this to the *sampling granularity* of GRPO Shao et al. (2024): Perception and Chart involve localized edits whose group-sampled variants are spatially distinguishable by the reward, whereas Jigsaw and 3D require structural edits, which group sampling appears less able to produce, motivating future work on semantic-aware sampling.

**Reward Mechanism:** Tab. 5 isolates the two RL reward signals introduced in Sec. 3.3: *Editing Correctness* (isolated edit quality judged by a VLM-as-Judge Zheng et al. (2023)) and *Editing Guidance* (downstream answer accuracy). Each reward has complementary blind spots, and neither alone matches the full model. Guidance-only is on par with Correctness-only on Logic, where the downstream model benefits directly from edits that expose task-relevant structure, whereas Correctness-only is slightly stronger on Perception and Chart, where the judge can reliably assess localized visual edits such as bounding boxes and chart annotations. Combining the two matches or exceeds either reward alone across all task families: Correctness provides a fidelity floor by filtering implausible edits, while Guidance raises the task ceiling by ensuring each edit advances the reasoning chain.

**Reflection Mechanism.** Tab. 3 ablates the Edit-Verify-Reason pipeline by removing the verification step, so every edited image is passed to the understanding model. Reflection yields consistent gains on Perception and Chart, where the judge can reliably flag mis-localized edits and where backbone accuracy is already high (around 80%); in this regime, suppressing harmful edits matters more than

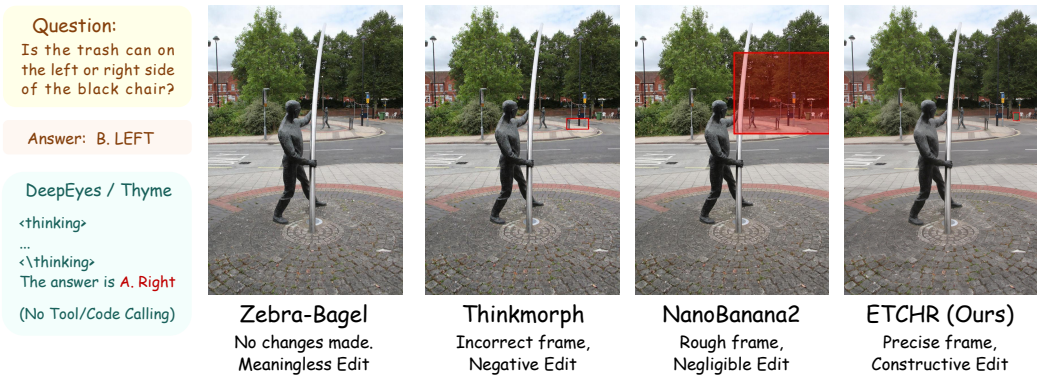



Figure 4: **Qualitative comparison** of ETCHR with existing methods. ETCHR provides the most precise guidance for this problem, assisting the reasoning model in generating the correct answer.

supplying additional hints. Gains are smaller on Logic and absent on Jigsaw for Qwen3-VL-8B, with a small regression on 3D. Baseline accuracy is low on these tasks. The model rarely solves these tasks from the original image alone, and we hypothesize that even imperfect edits are more useful as hints than discarded. This suggests reflection should be applied selectively, gated by backbone confidence or task family rather than uniformly.

**Case Analysis.** Fig. 4 demonstrates how various 'think with images' strategies assist comprehension through image editing. DeepEyesV2 and Thyme failed to invoke the editing tools, leading to incorrect final answers. Bagel and ThinkMorph performed incorrect edits, also resulting in wrong answers. While Nano Banana 2 executed the edit correctly, the highlighted regions lacked precision, providing minimal guidance for the reasoning model. In contrast, our editing model provides extremely precise edits, enabling the reasoning model to answer the question correctly. We defer extended cases across the five main tasks to Appendix D.

## 5 Related Work

**Thinking with Images.** A growing line of work lets MLLMs generate intermediate images during reasoning, splitting into two families. *Tool-based* pipelines Hu et al. (2024); Wu & Xie (2024); Zheng et al. (2025); Hong et al. (2025); Zhang et al. (2025) fine-tune the understanding model to emit structured commands (bounding boxes, crops, code snippets) that a deterministic renderer applies, confined to predefined, localized actions. *Unified multi-modal* models Team (2024); Xie et al. (2024); Wu et al. (2025b); Gu et al. (2025); Li et al. (2025a) interleave text and image tokens in a single backbone, gaining flexibility but inheriting a weaker generative head whose intermediate images often inject noise rather than guidance Wen et al. (2026). ETCHR departs from both: it uses a dedicated image-to-image editor as the intermediate-image generator, decoupled from the understanding model. We defer extended discussion to Appendix A.

## 6 Conclusion

We present ETCHR, a question-conditioned image editor for think with images reasoning. Our analysis identifies two gaps in previous works, a *language-side* gap in mapping abstract questions to edits and a *generation-side* gap under increasing reasoning depth, and motivates a two-stage recipe (Reasoning Imitation SFT and Reasoning Enhancement RL) paired with an Edit-Verify-Reason inference procedure. Decoupling the editor from the understanding model lets ETCHR plug into open- and closed-source MLLMs in a training-free manner.

**Limitations:** *GRPO sampling granularity:* GRPO's sampling exhibits limited semantic diversity for structural editing (e.g., jigsaw restoration), which constrains policy exploration and prevents uniform RL gains across all reasoning families. *Downstream capability bound:* ETCHR end-to-end performance remains capped by the understanding model; even optimally edited intermediates cannot overcome the reasoning ceiling of the downstream MLLM.

### AI USE STATEMENT

In this work, we used generative AI tools solely for minor language polishing. They were not used for research ideation, experimental design, or any task with recommended disclosure; the remaining disclosure items are not applicable to this work. We reviewed all AI-assisted edits, which were limited to minor language polishing, and we take full responsibility for the final content, including any text, claims, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses only publicly available datasets. To the best of our knowledge, the use of these datasets does not raise concerns regarding discrimination, bias, fairness, privacy, or security. No new data involving human participants were collected for this study.

### REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we will publicly release our code and model checkpoints upon publication.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed, Feb 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *ArXiv*, abs/2505.21500, 2025b. URL https://api.semanticscholar.org/CorpusID:278911718.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepseekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. Refchartqa: Grounding visual answer on chart images through instruction tuning. In *International Conference on Document Analysis and Recognition*, pp. 523–537. Springer, 2025.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024.

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

Zimo Wen, Boxiu Li, Wanbo Zhang, Junxiang Lei, Xiaoyu Chen, Yijia Fan, Qi Zhang, Yujiang Wang, Lili Qiu, Bo Li, et al. Unig2u-bench: Do unified models advance multimodal understanding? *arXiv preprint arXiv:2603.03241*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

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

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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

# A Extended Related Work

**Tool-based Pipelines.** Tool-based "think with images" methods couple the understanding MLLM with a deterministic external renderer: the MLLM emits a structured action and the renderer executes it on the input image, producing the intermediate visual aid that feeds the next reasoning step. Visual Sketchpad Hu et al. (2024) is an early instance, prompting the MLLM to draw spatial annotations and auxiliary lines for geometric and mathematical reasoning. V* Wu & Xie (2024) and DeepEyes Zheng et al. (2025) specialize this idea to fine-grained visual search by emitting bounding-box and crop-zoom commands that progressively localize the queried region, while Thyme Zhang et al. (2025) and DeepEyes-V2 Hong et al. (2025) extend the action space to executable code snippets for chart parsing and multi-step chart understanding. Although the resulting actions are interpretable and reproducible, two limitations persist. First, the action space is fixed at design time, so global, free-form, or task-novel transformations (e.g., redrawing a maze path, reordering jigsaw tiles, re-rendering a 3D viewpoint) lie outside what the renderer can produce. Second, each new family of actions typically requires task-specific fine-tuning of the understanding model, which can erode its general competence Kirkpatrick et al. (2017) and forces a separate adaptation per deployment. ETCHR avoids both costs by replacing the deterministic renderer with a question-conditioned image editor whose transformation space is continuous, and by keeping the understanding model frozen so a single trained editor plugs into different open- and closed-source MLLMs without further fine-tuning.

**Unified Multi-modal Models.** A second family pursues "think with images" with a single backbone that natively interleaves text and image tokens, removing the external renderer altogether. Chameleon Team (2024) and Show-o Xie et al. (2024) demonstrate joint autoregressive and diffusion modeling over a shared token space, while Janus Wu et al. (2025b) decouples the visual encoder for understanding from the one used for generation to relieve the optimization tension between the two objectives. More recently, ThinkMorph Gu et al. (2025) and Zebra-CoT Li et al. (2025a) fine-tune unified models on interleaved reasoning traces and report emergent adaptive modality switching. Despite this flexibility, the jointly optimized generative head trails specialist image editors in fidelity, so its intermediate images frequently inject noise rather than evidence, as quantified by UniG2U-Bench Wen et al. (2026); the same backbone also lags specialist understanding models on pure perception tasks Wu et al. (2025b); Xie et al. (2024). ETCHR sidesteps this failure by keeping understanding and generation in separate specialist models: the editor is trained on reasoning-aware objectives without touching the understanding MLLM, and edit fidelity is preserved because the editor is not asked to also serve as the answer producer.

**More Image Editing Models.** Beyond the MLLM-encoder editors discussed in the main paper, instruction-tuned diffusion editors form the closest prior art: InstructPix2Pix Brooks et al. (2023) fine-tunes a latent diffusion model on synthetic before/after pairs, and MagicBrush Zhang et al. (2023) improves edit faithfulness with human-annotated real-image edits. Both assume the user supplies an explicit edit prompt, so the editor is optimized as a faithful instruction renderer rather than a question-conditioned reasoner, the assumption ETCHR removes.

# B Prompts

**Task-level Prompt:**

*Fine-grained Perception:* Draw a red box to mark the important regions for this question in the figure.

*Chart Understanding:* Draw a red box to mark the important regions for this question in the figure.

*Logic:* Draw the shortest path of the maze in blue.

*Jigsaw:* Draw the original image after restoring this jigsaw puzzle task.

*3D Understanding:* Imagine a new perspective of the original image that helps answer the question.

**Verify Prompt:**

*Fine-grained Perception:* Please judge whether the information relevant to solving this problem is marked with a red box in the picture. If the red box contains valid information, reply with 1; if there

is no red box in the picture, or the area enclosed by the red box does not contain valid information, reply with 0.

*Chart Understanding:* Please judge whether the information relevant to solving this problem is marked with a red box in the picture. If the red box contains valid information, reply with 1; if there is no red box in the picture, or the area enclosed by the red box does not contain valid information, reply with 0.

*Maze:* Please determine whether the following image contains a valid maze path: the middle path (blue) connects the starting point (green) to the destination (red), is continuous in four directions (up, down, left, right) with no diagonal movement allowed, and no black obstacles or white path tiles have been altered compared to Figure 1 (the original maze) in Figure 2 (the path diagram). If the maze if valid, reply with 1. Otherwise, return with 0.

*Frozen Lake:* Please determine whether the path drawn with red lines in Figure 2 connects the treasure to the character without passing through any holes. If the frozen lake if valid, reply with 1. Otherwise, reply with 0.

*Jigsaw:* Please determine whether Figure 2 is a correct restoration of the jigsaw puzzle task in Figure 1. If Fig.2 is correct, reply with 1. Otherwise, reply with 0.

*3D Understanding:* judge: Please judge whether you can see the objects mentioned in the question from the perspective of the second image. If the objects are presented in the image, reply with 1. Otherwise, reply with 0.

**Reason Prompt:**

*Fine-grained Perception:* Answer the question based on the two images above. The first image is the original one corresponding to the question, while the second one uses a red box to mark the area containing the key information for the question in the image. Please first focus on the boxed part in the second image, and answer the question based on the corresponding information in the two images.

*Chart Understanding:* Answer the question based on the two images above. The first image is the original one corresponding to the question, while the second one uses a red box to mark the area containing the key information for the question in the image. Please first focus on the boxed part in the second image, and answer the question based on the corresponding information in the two images.

*Maze:* Solve the maze problem in the first image. The blue path in the second image is very likely the correct solution to the maze.

*Frozen Lake:* Solve the frozen lake problem and output the path from the character to the treasure. The blue path in the second image is very likely the correct solution to the frozen lake problem.

*Jigsaw:* Restore the jigsaw image based on the two images. The second image is likely the correct restoration of the jigsaw puzzle task from the first image.

*3D Understanding:* Answer the question based on the two images above. The first image is the original one corresponding to the question, while the second one provides a novel perspective to help you solve the problem. Please focus on the second image to answer the question.

## C Experiment Details

**Evaluation Details.** We report accuracy (Pass@1, temperature 0) across all experiments. For baselines, we set the maximum number of tool invocations to 10 for structured-action models, and the maximum number of image generations to 10 for unified models.

**Training Details.** For *Reasoning Imitation Supervised Fine-tuning*, we use Low-Rank Adaptation (LoRA) to fine-tune the Diffusion Transformer of FLUX.2-klein-base-9B for 1 epoch with lora rank $r = 768$, learning rate $lr = 1e^{-4}$, Classifier-free guidance $scale = 1$, inference steps $step = 30$. The supervised fine-tuning framework is built on DiffSynth-Studio. For *Reasoning Enhancement Reinforcement Learning*, we use Pref-GRPO to further fine-tune the DiT. The RL framework is built on Pref-GRPO Wang et al. (2025b). We also adapt LoRA fine-tuning strategy, with lora rank

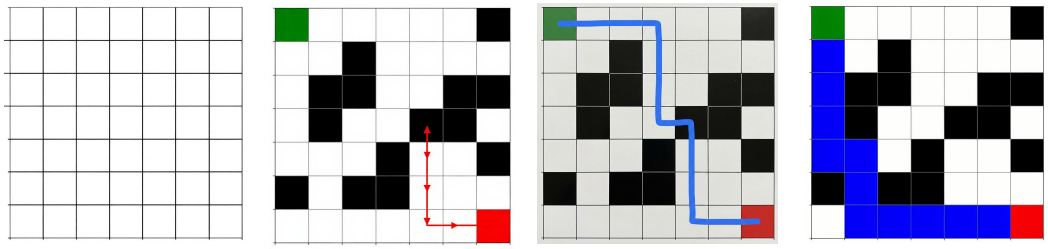



Figure 5: ETCHR enables precise localization of granular information in charts, facilitating a better comprehension of the visual data for understanding models.

$r = 128$, lora alpha $\alpha = 128$, learning rate $lr = 1e^{-4}$, Group Number for each rollout $G = 8$. We use Qwen3-VL-8B-Instruct as both the understanding model $\mathcal{M}$ and judge model $\mathcal{J}$ to provide editing guidance reward and editing correctness reward.

## D More Cases

We further provide examples of ETCHR assisting models in performing reasoning across tasks such as chart understanding (Fig. 5), logical reasoning (Fig. 6), jigsaw puzzles(Fig. 7), and 3D perception (Fig. 8).



**Zebra-Bagel**
Corrupted Image

**ThinkMorph**
Wrong Maze Solving

**Nano-Banana2**
Wrong Maze Solving

**ETCHR(Ours)**
Correct Maze Solving

ETCHR Response:

To find the shortest path from the green square (top-left) to the red square (bottom-right), we follow the path indicated in blue in the second image.
Let's map the coordinates (row, column) starting from (0,0) at the top-left:
Start: (0,0) - Green  Move Down: (1,0)
.....
(5,3) to (5,4): **R**  (5,4) to (5,5): **R**  (5,5) to (5,6): **R**
Combining these, the sequence of moves is D,D,D,D,R,D,D,R,R,R,R,R.

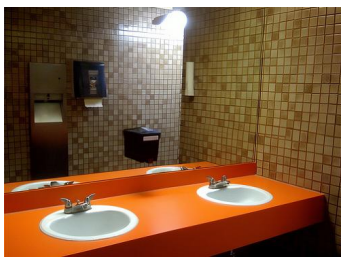
Figure 6: ETCHR can draw the correct path on the maze map, helping understanding model accurately identify the route to exit the maze.



Raw Image

Raw Answer:

Looking at the four patches, I need to determine the correct 2 × 2 arrangement based on mirror reflections and spatial relationships.
**Key observations:** 1. **Patch 3** shows a sink with a mirror that clearly reflects another sink - this reflected sink is **Patch 1** 2. **Patch 1** shows a sink with a mirror that reflects a paper towel ...
.....
This arrangement satisfies all mirror reflections: - Patch 3's mirror reflects the back wall, showing Patch 1's sink - Patch 1's mirror reflects the front wall, showing Patch 0's paper towel dispenser 0-3-2-1



ETCHR Edit Image

ETCHR Answer:

Looking at the two images, I need to identify which of the four patches (0, 1, 2, 3) in the first image is flipped.
Let me analyze each patch by comparing it to the second image
**Patch 0**: Shows paper towel dispensers ...
.....
The arrangement matches the complete image where: - Upper left has the paper dispensers - Upper right has the white wall dispenser - Lower left has the left sink - Lower right has the right sink
0-2-3-1

Figure 7: ETCHR is capable of restoring scrambled jigsaw images, thereby guiding the understanding model to correctly solve jigsaw tasks.



Raw Image



ETCHR Edit Image

Question:

From the perspective of this woman dressed in green, where is the woman dressed in red located?
A. left B. front
C. front-right D. back

Raw Answer:

From the perspective of the woman dressed in green (who is on the left side of the image), the woman dressed in red is to her right. Since the woman in red is also in the foreground, closer to the camera than the background elements, she is in the front-right relative to the woman in green.
Therefore, the correct answer is: **C. front-right**

ETCHR Answer:

To answer this question, we need to analyze the second image, which shows the two women from behind as they walk down a city street. The woman dressed in green is on the right side of the image. The woman dressed in red (or maroon) is on the left side of the image. From the perspective of the woman in green, the woman in red is positioned to her left.
Therefore, the correct answer is A. left.

Figure 8: ETCHR enables perspective shifting based on the specific question, helping to ensure the model correctly understands spatial positioning.