From Illusion to Intention: Visual Rationale Learning for Reliable Evidence Acquisition

Changpeng Wang[1], Junhan Liu[1], Xi Chen[2⋆], Haozhe Wang[3], Donglian Qi[1], and Yunfeng Yan[1]

[1]Zhejiang University [2]The University of Hong Kong
[3]The Hong Kong University of Science and Technology

Abstract. "Thinking with images" has recently emerged as a promising direction for vision-language models, where models actively interact with images via iterative zoom-ins to acquire visual evidence for answering. Yet, prevailing frameworks ignore the quality of intermediate zoom-ins, boosting metrics but leading to an action–outcome disconnect. This failure mode creates the illusion of thinking with images, where models appear visually grounded while relying on context-agnostic or redundant crops. We address this problem by introducing **Visual Rationale Learning (ViRL)**, an end-to-end reinforcement learning paradigm that grounds the evidence acquisition process. ViRL ensures action faithfulness and efficiency through: (1) process supervision with ground-truth visual rationales, (2) rationale fidelity rewards that incentivize evidence-grounded zoom-ins, and (3) fine-grained credit assignment to penalize redundant or erroneous steps. Experimental results show that ViRL achieves state-of-the-art performance across benchmarks spanning perception, hallucination, and perception-grounded reasoning. We also dissect the illusion of thinking with images and position visual rationales as a verifiable foundation for building trustworthy vision-language models.

Keywords: Thinking with Images · Multimodal Large Language Model · Reinforcement Learning

1 Introduction

The advent of Chain-of-Thought (CoT) [53] has marked a pivotal moment for large language models, enabling them to solve complex problems by externalizing their reasoning into verifiable, step-by-step textual traces [29, 48, 50]. However, in vision-language settings, reasoning is not confined to language [1, 2, 4, 14, 24, 47, 58]. To solve real-world multimodal tasks [3, 10, 30, 33], models must not only "think in text" but also "think with images" [36, 49]. This paradigm shift moves beyond purely textual reasoning toward reasoning in the pixel space, where models actively interact with the visual world.

In particular, zoom-in has emerged as a straightforward yet powerful mechanism [16, 18, 49, 56], allowing models to selectively inspect informative image

⋆ Corresponding author.

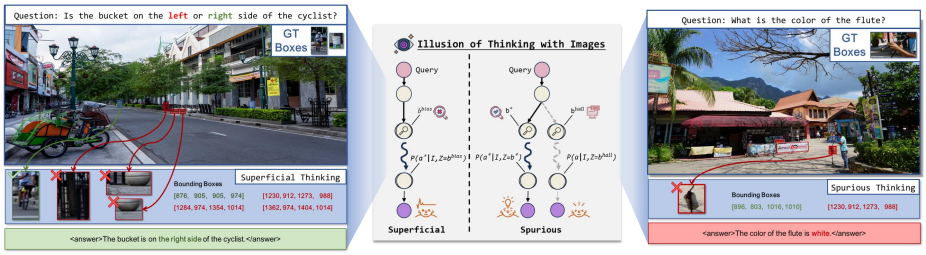

Fig. 1: Examples illustrating the illusion of "thinking with images". Even when models execute visual actions, they often zoom into irrelevant regions (superficial) or select misleading cues (spurious), creating the false impression of grounded reasoning while contributing little to answering the question.

regions and tightly couple perception with reasoning. By performing zoom-in operations, a model can decompose a complex visual scene into a series of focused inquiries, mimicking the human ability to scrutinize details and gather evidence. This active perception promises to yield models that are: (a) more robust, grounding each step of their reasoning in specific visual evidence rather than relying on spurious correlations; (b) more efficient, by concentrating computational resources on relevant details; and (c) more trustworthy, by producing a transparent and verifiable visual audit trail for their decisions.

However, a significant gap exists between this ideal and the current reality. As illustrated in Fig. 1, we identify a fundamental problem we term the illusion of "thinking with images". This illusion arises when models appear to be seeking evidence, yet their visual actions are superficial or even spurious, contributing little to the final decision. This issue stems primarily from how visual actions are optimized in current frameworks. Under outcome-only rewards, supervision is sparse: only final answer correctness is rewarded, providing little signal to assess whether intermediate zoom-in decisions actually acquire useful evidence. Moreover, trajectory-level credit assignment is broadcast uniformly across steps, obscuring the marginal contribution of each action and allowing redundant or superficial zoom-ins to persist as long as the final answer is correct.

The consequence of this view is a stark disconnect between actions and outcomes. As shown in Figure 2, some models achieve high answer accuracy (e.g., Deepeyes [64] at 89.1%) while exhibiting low accuracy in their visual actions (57%). Such high performance on a benchmark belies a fragile process. As shown in our analyses, this illusion results in models that are: (1) brittle, as their shortcut behaviors collapse under distribution shifts; (2) inefficient, inflating inference costs with redundant actions; and (3) untrustworthy, as their spurious reasoning traces undermine interpretability in high-stakes applications.

To break this illusion, we introduce **Visual Rationale** as a verifiable evidence-acquisition trace in pixel space: a structured sequence of zoom-ins that should progressively surface the visual cues needed for answering. Building on this

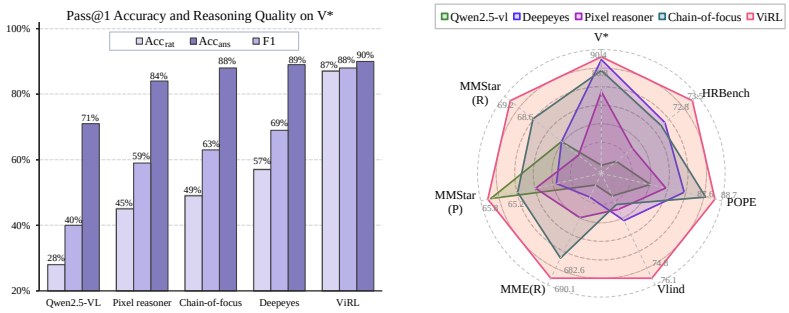

Fig. 2: Overall comparison of existing visual-reasoning methods and ViRL. **Left**: answer accuracy(Acc_{ans}), rationale accuracy (Acc_{rat}), and their joint score ($\mathcal{F}_1$) on V* [55]. **Right**: Radar-chart across multiple benchmarks.

view, we propose **Visual Rationale Learning (ViRL)**, an end-to-end learning paradigm that explicitly optimizes the fidelity and utility of the visual rationale.

Yet, turning this principle into a learnable objective is non-trivial. In particular, learning meaningful zoom-in policies is hindered by three fundamental obstacles: (i) existing datasets lack answer-supporting evidence anchors for visual rationales, and there is no principled way to curate them, leaving supervision noisy and diluted; (ii) outcome-dominated optimization supplies only an answer correctness reward, so superficial or spurious zoom-ins receive little discriminative feedback when the final answer is correct; and (iii) sequence-level credit assignment entangles the contributions of iterative actions, making it difficult to distinguish correct, redundant, and erroneous zoom-ins during training. To address these obstacles, ViRL makes three key contributions:

- Evidence-grounded dataset: It constructs a dataset with ground-truth visual rationales and reasoning-centric filtering, enabling reliable supervision over intermediate evidence acquisition.
- Rationale fidelity reward: It incentivizes faithful visual rationales by measuring how closely the model's visual foci (zoom-in coordinates) align with the ground-truth evidence over steps.
- Fine-grained credit assignment: It defines step-level utilities for zoom-in actions (e.g., correct, redundant, or erroneous), shaping the policy to make deliberate and efficient evidence-acquisition choices.

This work establishes a new framework for building verifiable and trustworthy vision-language models. As shown in Figure 2, by directly optimizing the alignment between zoom-in actions and task-relevant evidence, ViRL not only sets a new state-of-the-art on key benchmarks but also produces models with demonstrably superior accuracy and efficiency in their visual rationales. More broadly, this work analyzes the source of the illusion of thinking with images, demonstrating that process-level alignment between actions and evidence is essential for reliable "thinking with images".

2 Related Work

Thinking with Images. Recently, many reasoning models have been proposed to handle more complex tasks due to their strong ability to decompose these difficult questions into simple, small queries. Early methods rely only on text to decompose these complex tasks and do not analyze the full inputs [25–27,62]. Like Visual cot [40], VPD [17], V* [55], Inght-V [11], and Llava-cot [59]. These methods decompose the visual questions into sample queries in Chain-of-Thoughts(CoTs). However, they tend to ignore the visual clues when thinking in deep chains. In the latter, many methods introduce visual tools to think with images rather than text only. Visual Sketchpad [16], DetToolChain [56], Cropper [21], toolformer [38], AutoCode [46], and react [60] understand complex tasks and call visual tools like zoom-in, drawing line, and depth perception [8] as additional auxiliary tips through in-context learning without training. However, these methods are mainly made up of visual workflows and cannot truly understand the tools. Later, Models like o3 [36], Pixel Reasoner [49], Deepeyes [64], Chain-of-focus [63], and OpenThinkIMG [44] demonstrate the ability by dynamically applying visual tools to reasoning in multimodal chains of thoughts [13, 45, 57]. While these methods improve visual reasoning, their outcome-supervised rewards make the visual thinking process error-prone and susceptible to hallucination [32].

Chain-of-Thought Reasoning. Reinforcement Learning, as a technique in the post-training of MLLMs [9,31,47,51], has been shown to effectively enhance reasoning ability in visual reasoning tasks. In particular, Proximal Policy Optimization (PPO) [39] has become the most widely adopted algorithm for aligning multimodal reasoning behaviors with reward signals and has been extensively applied in recent vision-language alignment frameworks [35]. More recently, GRPO [9] has been introduced as an efficient alternative, offering faster convergence and improved sample efficiency in complex multimodal reasoning scenarios. Later, to alleviate the vanishing advantages problem [47] during RL training, SSR [47] and DAPO [61] was proposed. Several works have further refined the optimization process by either redesigning the reward function or adjusting the advantage estimation. VLM-R1 [42] incorporates multi-component rewards, such as accuracy, format, and REC, to stabilize training. HICRA [50] discovered emergent reasoning hierarchy in reasoning LLMs and propose to modulate the advantages on critical planning tokens, leading to better strategic exploration and training results. In this paper, we focus on shaping reward and credit assignment to ensure that models not only reach correct answers but also ground their reasoning in coherent visual rationales.

3 Method

3.1 Problem Statement

We first formalize the challenge of "thinking with images" by identifying the core constraint of standard Vision-Language Models (VLMs): inherent partial

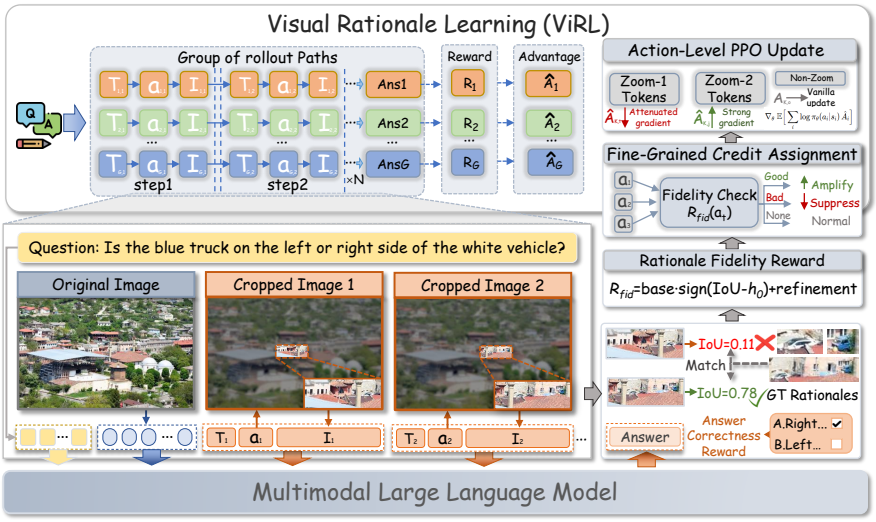

Fig. 3: The end-to-end visual rationale learning framework. The framework combines multi-turn learning with rationale fidelity reward and fine-grained credit assignment to produce interpretable and verifiable visual reasoning chains.

observability. When a VLM first processes an image, its visual encoder creates an information bottleneck. Constrained by finite resolution, this encoding obscures or loses the fine-grained details crucial for complex reasoning. The model, therefore, operates on an incomplete belief state, not the full image.

To overcome this, the model transforms from a passive observer into an active reasoner that sequentially gathers new evidence to resolve uncertainty. We define an action space that allows the model to interleave internal logic with active perception. At any step, the policy π_θ chooses between generating a **textual rationale** to hypothesize, strategize, and integrate multimodal evidence and executing a **visual rationale**, where it performs a zoom-in operation on region b_k to acquire high-fidelity visual evidence.

Inspired by LLMs such as DeepSeek-R1, it is tempting to apply Group Relative Policy Optimization (GRPO) [41] with outcome-only rewards to elicit reasoning, since it can be trained directly from readily available benchmark answers without costly process-level annotations. However, while outcome-only RL successfully elicits self-contained reasoning in pure text, it struggles in "thinking with images". Because MLLMs are prone to shortcut learning, they often guess answers using textual priors, entirely bypassing actual visual evidence acquisition. A straightforward fix of adding a coarse tool-use bonus frequently causes reward hacking: the model performs superficial zoom-ins to collect bonuses while relying on hallucinated correlations. Since outcome signals cannot verify if visual actions actually acquired supporting evidence, outcome-only optimization perversely discourages the grounded policy we aim to elicit.

3.2 Visual Rationale Learning

To break this illusion, we shift the training objective from outcome accuracy to process-level fidelity and utility, ensuring that each zoom-in action contributes meaningfully to the reasoning trajectory. We introduce Visual Rationale Learning (ViRL), a principled framework that learns a policy by explicitly rewarding the reasoning process. ViRL is built on two insightful components designed to solve the challenges of hallucinated rationales and coarse credit assignment.

Rationale fidelity reward. Leveraging the process-grounded data with visual rationales, we provide step-wise supervision by rewarding each zoom-in according to its alignment with the intended visual evidence. Our objective is to learn a rational policy that maximizes task utility while maintaining faithful and concise visual reasoning. Given a query $\mathcal{Q}$ and trajectory τ, we optimize:

$$\max_{\phi} \ \mathbb{E}_{\tau \sim \pi_\phi(\cdot|\mathcal{Q})}\big[\mathcal{R}_{\text{total}}(\tau)\big] \tag{1}$$

where the total reward jointly evaluates answer correctness, format compliance, and process fidelity:

$$\mathcal{R}_{\text{total}} = \mathcal{R}_{\text{acc}} + \mathcal{R}_{\text{fmt}} + \bar{\mathcal{R}}_{\text{fid}} \tag{2}$$

This formulation emphasizes process-level supervision, where each visual rationale directly supports the model's perceptual and inferential steps. Each zoom-in action a_k receives a fidelity reward based on its spatial alignment $u_k = \text{IoU}(b_k, b_k^*)$ with the best matching reference rationale b_k^*.

Instead of using raw IoU as a continuous reward, we adopt a base-plus-refinement design that first learns the basic zoom-in operation and distinguishes roughly correct evidence from wrong regions and then progressively rewards finer alignment. Specifically, it comprises a signed correctness term that directly reinforces plausible reasoning while penalizing misaligned ones, and a discrete refinement bonus activated beyond a soft threshold h_0, which further encourages precise spatial alignment as IoU increases. Formally,

$$\mathcal{R}_{\text{fid}}(a_k) = \mathcal{R}_{\text{base}} \cdot \text{sign}(u_k - h_0) + \eta \Big\lfloor \frac{\max(0, u_k - h_0)}{\Delta h} \Big\rfloor \tag{3}$$

where $\text{sign}(u_k - h_0)$ provides the signed correctness signal, η controls the magnitude of the refinement bonus, and $\lfloor \cdot \rfloor$ denotes the discretization operator that introduces stepwise gains. Specifically, $\lfloor (u_k - h_0)/\Delta h \rfloor$ increases by one for every Δh improvement in IoU, thereby modeling progressive alignment refinements beyond the soft threshold h_0. If no visual action is taken, $\mathcal{R}_{\text{fid}}(a_k) = 0$.

Aggregating the step-level rewards across the reasoning trajectory yields the overall fidelity score:

$$\bar{\mathcal{R}}_{\text{fid}} = \frac{1}{N} \sum_{k=1}^{N} \big[\mathcal{R}_{\text{fid}}(a_k) - \rho(C_k)\big] \tag{4}$$

where averaging ensures fair credit assignment across variable-length trajectories. A smooth redundancy penalty $\rho(C_k)$ grows beyond a soft budget, discouraging repetitive or overlapping rationales and promoting concise reasoning.

Fine-grained credit assignment. With the trajectory reward $R(\tau)$, we confront the credit assignment problem: the trajectory-level advantage is too coarse, indiscriminately crediting or blaming all steps and encouraging erroneous visual rationales within successful trajectories. As shown in Figure 3, we solve this with a bi-level credit assignment strategy that modulates the advantage estimation based on rationale-specific quality.

- **Trajectory-level advantage:** First, we compute a coarse advantage signal for the entire trajectory τ_i by comparing its reward to the average of a group of G trajectories [41]:

$$A_i = R(\tau_i) - \frac{1}{G}\sum_{j=1}^{G} R(\tau_j) \tag{5}$$

  This indicates whether the trajectory is globally advantageous ($A_i > 0$) or disadvantageous ($A_i < 0$) in the group.
- **Rationale-level adjustment:** To obtain fine-grained credit assignment, we differentiate the coarse advantage signal. Textual rationales preserve the original advantage A_i, whereas visual rationales are adaptively modulated by their fidelity (Eq. 3). The resulting adjusted advantage for each action a_t is denoted as $\hat{A}_{i,t}$:

$$\hat{A}_{i,t} = A_i \cdot h(a_t) \tag{6}$$

where $h(a_t)$ is a simple modulator based on the rationale:

$$h(a_t) = \begin{cases} h_{\text{good}}, & \text{if } a_t \text{ is good visual rationale } (\mathcal{R}_{\text{fid}} > 0) \\ h_{\text{bad}}, & \text{if } a_t \text{ is bad visual rationale } (\mathcal{R}_{\text{fid}} \leq 0) \\ 1, & \text{if } a_t \text{ is textual rationale} \end{cases} \tag{7}$$

This logic provides a highly principled teaching signal:

- **In an advantageous trajectory ($A_i > 0$):** the model amplifies credit for high-fidelity visual rationales ($h_{\text{good}} > 1$) while attenuating it for low-fidelity ones ($h_{\text{bad}} < 1$). This guides the policy to attribute success *to* its precise and faithful visual reasoning.
- **In a disadvantageous trajectory ($A_i < 0$):** the model amplifies blame for bad visual rationales ($h_{\text{bad}} > 1$) and mitigate blame for any good ones ($h_{\text{good}} < 1$). This guides the policy to recognize that failure stems from inaccurate or misleading visual reasoning.

Policy optimization. This fine-grained advantage $\hat{A}_{i,t}$ is then used to update the policy π_θ via a standard PPO-style objective. This bi-level ViRL framework compels the policy to learn the true contribution of each visual rationale, moving the model from illusion to intention by ensuring it generates reasoning that is not just faithful, but also utility-enhancing.

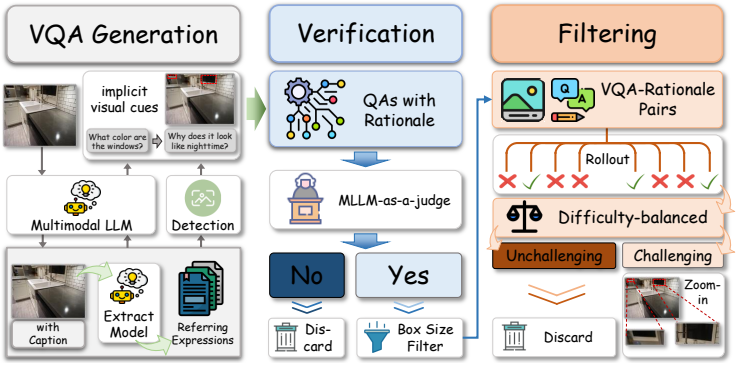

Fig. 4: A three-stage pipeline for evidence-grounded data curation, involving generation, verification, and reasoning-centric filtering.

3.3 Evidence-grounded Dataset Construction

Constructing such rationales is non-trivial, as it requires anchoring visual evidence to question-answer semantics and isolating zoom-dependent tasks—those that cannot be resolved without localized visual cues. As illustrated in Figure 4, the dataset is built via a three-stage pipeline comprising generation, verification, and filtering. The pipeline ensures high-quality data and explicit process-level supervision, forming the foundation of the visual rationale learning paradigm.

Evidence-grounded qa construction. To foster robust perception-grounded reasoning, we develop an automated pipeline to synthesize reasoning-centric VQA triplets. Specifically, we first generate fine-grained region captions with large multimodal models based on GRIT [54] images, which provide rich attributes within localized areas. For each region, we extract referring expressions with SpaCy [15] and prompt large multimodal models to generate questions requiring implicit visual reasoning rather than explicit object naming, encouraging the model to rely on localized visual cues rather than category memorization. The corresponding visual rationale is obtained by detecting the referring expression and padding the box for more context, serving as an evidence anchor aligned with the question-answer semantics. Finally, we apply NMS to filter overlapping candidates. This process yields question, answer, rationale region triplets that serve as supervision for learning visual evidence acquisition behaviors.

Verification. To ensure training-ready supervision, we perform MLLM-based consistency verification using the generated visual rationale as an explicit visual hint. Concretely, we overlay the predicted visual rationale on the original image (highlighting the hinted evidence) and provide the judge with the QA. The judge then evaluates data quality under two criteria: (i) answer correctness–whether the answer is correct with respect to the question; and (ii) rationale consistency–whether the highlighted region aligns with the visual evidence implied by the question and the answer. Only samples that satisfy both criteria are retained; otherwise, they are discarded as mismatched or spurious.

Reasoning-centric filtering. Finally, we apply reasoning-centric filtering to ensure the remaining samples genuinely require perception-grounded reasoning. First, we remove samples with overly large rationale regions, since they can often be solved without localized evidence. Second, we perform offline difficulty balancing via n rollouts **without** grounding cues; samples answered correctly in all rollouts are removed as unchallenging. For the remaining hard cases, we run additional rollouts conditioned on the visual rationales as an explicit visual hint. We retain samples if the predicted answer is correct with this hint, indicating that the question is zoom-in-dependent. Samples that remain incorrect even with the rationales are discarded as overly noisy or inconsistent. This process yields a dataset that encourages models to think with images, providing reliable step-level visual rationales for evidence-grounded learning.

Following this procedure, we construct a 200k-scale dataset by collecting samples from Visual-Cot [40], GQA [19], TextVQA [43], Visual7W [65], InfographicsVQA [34], and GRIT [54], covering general understanding, visual grounding, perception-grounded VQA, and OCR tasks. After processing within the data generation pipeline, 8k high-quality samples are retained. Since all data have open-vocabulary answers that are difficult to evaluate, we repackage them as multiple-choice questions by synthesizing 3–7 plausible distractors from the VQA and global descriptions to preserve task difficulty.

4 Experiments

4.1 Implementation Details.

Training paradigm. We fine-tune Qwen2.5-VL-7B [6] on 8 A100 GPUs for 80 iterations. Following DAPO [61], we adopt clip-higher, dynamic sampling, and token-level policy loss to enhance data utilization. Training uses a batch size of 64 with 16 rollouts per prompt, allowing up to 6 rounds of cropping. The learning rate is 1×10^{-6}, the maximum response length is 20,480 tokens, and neither KL nor entropy regularization is applied.

Evaluation. For evaluation, we adopt a comprehensive benchmark suite spanning three dimensions of vision–language understanding: (i) perception-oriented benchmarks, including fine-grained perception (V*) [55] and high-resolution perception (HRBench) [52]; (ii) reliability-oriented benchmarks, covering hallucination tendency (POPE) [28] and language-prior reliance (VLind) [20]; and (iii) perception-grounded reasoning benchmarks, including MME(R) [12] and MM-Star [7]. We evaluate their perception-grounded reasoning subsets, which require integrating localized visual evidence with question understanding while remaining primarily perception-driven rather than symbolic or mathematical reasoning. This comprehensive design ensures a fair and multi-dimensional comparison.

In addition to external benchmarks, we further report an internal diagnostic that quantifies how models engage in perception-grounded reasoning: Rationale Accuracy, which measures whether visual zoom-ins successfully hit the ground-truth evidence. We also provide a joint $\mathcal{F}_1$ score that summarizes answer cor-

Table 1: Our main results on the six evaluated benchmarks. MMStar† evaluates four core categories (coarse perception, fine-grained perception, instance reasoning, and logical reasoning), focusing on general perception-oriented reasoning.

Model	Size	Perception		Reliability		Reasoning	
		V*	HRBench	POPE	VLind	MME(R)	MMStar†
Models w/o "Thinking with images"							
GPT-4o [37]	–	45.0	46.8	84.6	**89.8**	674.6	**73.0**
GPT4o-mini [37]	–	50.8	48.0	83.3	64.4	564.3	62.6
LLaVA-OV [22]	7B	72.8	64.7	88.3	54.8	415.4	62.1
Qwen2.5-VL [6]	7B	71.4	69.2	86.6	67.4	610.1	66.6
Qwen2.5-VL [6]	32B	81.2	73.4	85.7	81.2	645.6	68.8
Models with "Thinking with images"							
Sketchpad [16]	–	80.4	–	–	–	–	–
SEAL [55]	7B	75.4	–	–	–	–	–
DyFo [23]	7B	81.2	–	–	–	–	–
CoF [63]	7B	88.0	70.5	88.4	68.3	673.5	66.6
Qwen3-VL [5]	4B	80.1	78.3	89.4	78.1	640.0	67.7
Deepeyes* [64]	7B	88.9	73.1	87.7	70.0	620.7	65.4
Pixel reasoner [49]	7B	84.3	72.8	87.1	68.8	638.5	64.8
Ours: Backbone generalization							
base + ViRL	7B	90.1	75.3	88.7	76.1	691.0	67.5
$+\Delta$ vs. Qwen2.5-VL		+18.7	+6.1	+2.1	+8.7	+80.9	+0.9
base + ViRL	4B	**90.8**	**79.9**	**89.6**	79.8	**704.8**	68.1
$+\Delta$ vs. Qwen3-VL		+10.7	+1.6	+0.2	+1.7	+64.8	+0.4

rectness and rationale fidelity. Complete definitions of these metrics are included in Appendix A.

4.2 Main Results

Table 1 shows that ViRL consistently improves fine-grained perception, reliability, and perception-grounded reasoning. On perception-oriented benchmarks (V*, HRBench), model scaling alone brings limited gains (e.g., Qwen2.5-VL-7B vs. 32B), whereas explicit visual reasoning yields clear improvements; ViRL achieves the best scores (90.1 on V*, 75.3 on HRBench) by decomposing complex scenes via targeted zoom-ins, which expose high-resolution cues that are often lost in the initial encoder bottleneck. On reliability-oriented benchmarks, ViRL attains the highest POPE and VLind scores, indicating both lower hallucination and weaker reliance on language priors. Importantly, VLind score explicitly probes whether a model can override misleading textual regularities by grounding its prediction in visual evidence. ViRL's strong VLind performance (+8.7 over Qwen2.5-VL-7B) therefore suggests that its zoom-in behavior is not cosmetic.

Table 2: Efficiency trade-off on V*. K is the maximum zoom rounds at inference. We report accuracy, total tool calls, mean latency, and prompt/completion tokens. ViRL achieves higher accuracy with lower latency.

Method	K	Acc.	Tool↓	Lat.(s)↓	P/C-tok↓
PixelReasoner	–	85.7	246	1.9	4996 / 76
CoF	–	89.8	403	6.8	9311 / 194
ViRL	1	88.7	**221**	**1.5**	4781 / 101
ViRL	3	89.8	279	2.1	6153 / 124
ViRL	6	**90.1**	381	2.6	8875 / 123

The model actively uses localized evidence to guide decisions when language priors are insufficient or incorrect. Finally, ViRL also leads on perception-grounded reasoning (691.0 on MME(R), 67.5 on MMStar†). Unlike outcome-centric tool training that treats zoom-in as an auxiliary action, ViRL provides step-level fidelity signals and decomposes trajectory-level advantages into action-wise utilities, encouraging coherent multi-step evidence accumulation and yielding more stable gains on reasoning categories that require integrating localized cues with question understanding.

Backbone generalization. Table 1 further reports ViRL on multiple backbones. For each backbone, we report both the base model and its ViRL-trained counterpart using the same training recipe. In all cases, ViRL yields consistent gains across perception, reliability, and perception-grounded reasoning benchmarks, indicating that the improvements are not specific to a single architecture.

Efficiency and budgeted inference. Multi-round zoom-ins may increase inference cost, thus we analyze the trade-off between accuracy and cost on V*. By explicitly supervising action quality, ViRL favors evidence-improving zoom-ins and suppresses redundant ones, translating step-level grounding into lower tool usage and latency. As a result, successful trajectories reinforce only the zoom-ins that truly acquire task-relevant cues, while redundant or misaligned actions are explicitly down-weighted. As shown in Tab. 2, ViRL achieves comparable or higher accuracy with markedly fewer tool calls and lower latency, which indicates that ViRL trains a more precise evidence-acquisition policy.

4.3 Ablation Study

Core components analysis. As shown in Table 3, we ablate ViRL's learning mechanisms while keeping the training data fixed. Replacing the rationale fidelity reward with the answer-only reward collapses the model's ability to think with images. Although answer accuracy remains superficially strong (87.6%), the absence of explicit visual rationale signals leads to degenerate shortcut behaviors, where the model "answers in context" without reliably invoking zoom-in evidence. A seemingly stronger baseline that adds a tool-call bonus further exposes reward hacking: it overuses zoom-ins (Zoom = 2.6) yet achieves lower

accuracy (85.3%) and poor rationale alignment ($\mathcal{F}_1 = 0.55$), indicating that incentivizing tool usage without measuring evidential utility makes the verification process noisy and unstable. In contrast, fine-grained credit assignment (FGCA) improves action-level discrimination within successful trajectories; removing FGCA reduces rationale quality ($\text{Acc}_{rat} = 78.2$; $\mathcal{F}_1 = 0.83$), showing that ViRL benefits from explicitly separating helpful vs. redundant zoom-ins to better align visual rationales with QA semantics.

Fidelity reward design. Table 4 analyzes how fidelity reward shaping and the soft threshold h_0 influence learning visual evidence acquisition. A raw IoU reward is suboptimal, as it is overly sensitive to small localization errors and provides unstable supervision for early-stage action learning. In contrast, our signed-plus-refinement formulation yields a more robust coarse-to-fine signal, improving both answer and rationale performance with $Acc_{ans} = 89.8$ and $Acc_{rat} = 83.7$. Sweeping h_0 reveals a clear trade-off: a low threshold $h_0 = 0.2$ tolerates weak alignments and blurs quality distinctions with $\text{Acc}_{rat} = 75.5$ and $\mathcal{F}_1 = 0.8$, while a moderate threshold ($h_0 = 0.3$) offers the best balance. When h_0 is too high (≥ 0.4), positive feedback becomes too sparse, and the policy collapses to an answer-only shortcut (Zoom = 0). Importantly, this design reduces sensitivity to extensive reward tuning: rather than requiring carefully calibrated continuous IoU weights, our shaping uses a single interpretable threshold to provide stable learning dynamics—first teaching the model when a zoom-in is warranted, and then progressively rewarding how well it aligns, effectively inducing a curriculum-like transition from tool acquisition to precise evidence grounding.

Data construction factors. Table 5 compares object-grounded supervision (Visual-CoT) with our visual rationale data pipeline. Visual-CoT boxes are typically tied to explicitly named objects, which can bias learning toward detection-style consistency and lexical shortcuts. In contrast, our visual rationale avoids direct object-name leakage and treats boxes as answer-supporting cues, boosting VLind by 3.5% and MMStar by 0.7% via enhanced visual reliance. Building on these QAs, consistency verification removes QA–rationales mismatches and substantially improves rationale fidelity, elevating Acc_{rat} from 55.3 to **87.6** and $\mathcal{F}_1$ to **0.89**. Finally, zoom-in dependency filtering suppresses trivial instances to incentivize active exploration of localized cues, keeping V* nearly unchanged while further improving perception-grounded reasoning.

From latent ability to structured reasoning. As shown in Fig. 5a, we observed a characteristic: an initial spike in zoom-in invocations followed by a steep decay to near-zero invocation frequency, even as answer accuracy remains non-trivial. We term this behavior "**visual thinking collapse**". Our empirical study points to three mechanistic contributors:

- **Prompt effects.** As shown by the "Rationale Count" in Fig. 5a, Clear prompts more strongly tie zoom-ins to reasoning, fostering stronger early exploration, whereas prior designs treat visual thinking as a vague action and yield weak early engagement.

Table 3: Core mechanism ablations. We evaluate answer and rationale accuracy to verify faithful evidence acquisition for question answering.

Method	Acc_{ans}	Acc_{rat}	$\mathcal{F}_1$	Zoom ↓
Base model	71.4	–	–	0
ViRL w/o rationale fidelity reward				
Answer-only reward	87.6	–	–	0
Answer + tool-call bonus	85.3	40.8	0.55	2.6
ViRL w/o FGCA	88.9	78.2	0.83	1.12
ViRL (Ours)	**90.1**	**87.4**	**0.88**	1.05

Table 4: Ablation on fidelity reward shaping. We compare different reward designs to evaluate their effectiveness in learning visual evidence alignment.

Reward design	Acc_{ans}	Acc_{rat}	$\mathcal{F}_1$
Raw IoU reward	84.3	50.8	0.63
Ours (signed + refinement)	**89.8**	**83.7**	**0.86**
$h_0 = 0.2$ (low)	87.2	75.5	0.80
$h_0 = 0.3$	88.4	77.3	0.82
$h_0 = 0.4$	86.6	–	–
$h_0 = 0.5$ (high)	85.4	–	–

Table 5: Data ablations across benchmarks. We compare object-grounded QA supervision with our evidence-seeking data pipeline. Our dataset pipeline yields more reliable gains, especially on visual evidence acquisition.

Training data variant	V*			VLind	MMStar†
	Acc_{ans}	Acc_{rat}	$\mathcal{F}_1$		
Visual-CoT (MCQ-adapted)	79.9	47.3	0.59	65.7	64.2
Evidence-seeking QA	82.6	55.3	0.66	69.2	64.9
+ Consistency verification	**90.8**	**87.6**	**0.89**	75.8	66.8
+ Zoom-in dependency filtering	90.1	87.3	0.88	**77.2**	**67.1**

– **Unfamiliarity with visual thinking.** Early zoom-ins often inject misleading regional evidence, yielding low-utility explorations and increasing "Answer Wrong w/ Rationale" errors (Fig. 5b).
– **Outcome-only supervision.** Outcome-only rewards encourage avoiding rationales: accuracy improves without zoom-ins but degrades with them, reinforcing risk-averse, answer-first behavior (Fig. 5b).

Invocation frequency ≠ invocation quality. A naive fix is to reward zoom-ins, but Fig. 5c shows that this boosts activity without improving, and often harming task accuracy. The model learns to invoke for the sake of invoking: rationale counts rise while per-invocation fidelity stagnates. This reveals the illusion of "thinking with images": the model appears to be actively engaging tools, yet the underlying reasoning quality stagnates or deteriorates. Thus, genuine "thinking with images" emerges not from more invocations, but from learning to strategically ground reasoning in the right visual evidence. Efficient visual reasoning requires **sparse yet precise** grounding. As shown in "Visual Rationale Accuracy" and "Answer Wrong w/ Rationale" (Fig. 5c,d), high-performing agents exhibit an early increase in invocation frequency during exploration, followed by convergence to a stable regime with markedly higher per-invocation accuracy and fewer illusion-induced errors. This transition reflects a shift from noisy trial-and-error to targeted, high-value visual reasoning. In contrast, naive

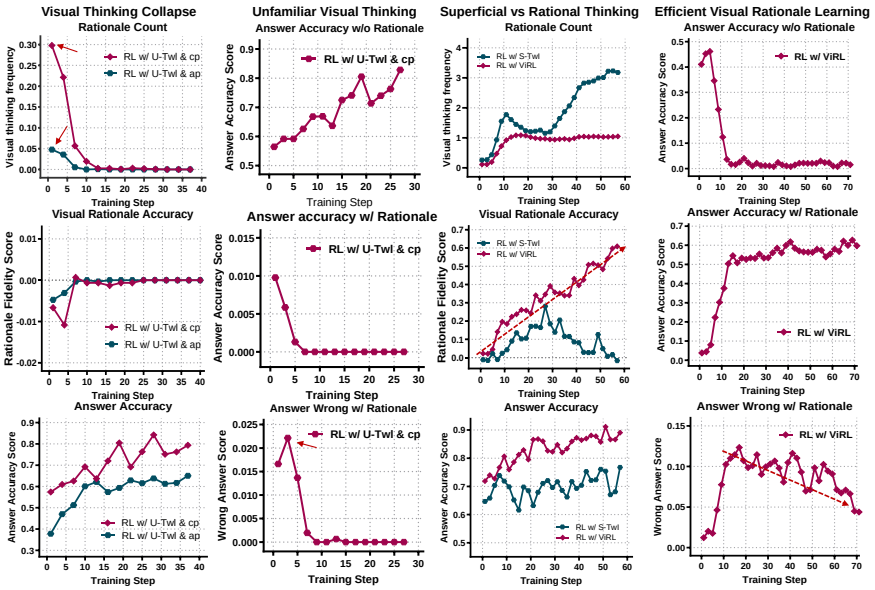

Fig. 5: Training dynamics of visual rationale learning. From left to right: (a) collapse of visual thinking, (b) the underlying cause: unfamiliarity with visual thinking, (c) superficial vs. ViRL-grounded reasoning, and (d) the emergence of faithful reasoning under our ViRL framework. U-TwI and S-TwI denote outcome-only and naive step-wise reward baselines. cp/ap represent clear/ambiguous prompts.

invocation-level incentives lead to persistent over-activation, inflating tool use without improving reasoning quality.

5 Conclusion

In this work, we identify and mitigate the illusion of "thinking with images" arising from outcome-dominated supervision and coarse credit assignment. We formalize visual rationale as a verifiable evidence-acquisition trace in pixel space, and propose Visual Rationale Learning (ViRL), a process-aware RL framework that aligns visual actions with evidence. Concretely, ViRL couples step-level rationale fidelity rewards with fine-grained credit assignment that decomposes trajectory-level advantage into action-wise utility, enabling the optimizer to reward helpful zoom-ins while down-weighting redundant or erroneous ones. As a result, ViRL achieves state-of-the-art results across perception, reliability, and perception-grounded reasoning benchmarks while establishing a principled approach to verifiable visual reasoning. More broadly, our analysis shows that process-level alignment between actions and evidence is essential for reliable "thinking with images," and future work will explore extending this paradigm to richer visual action spaces and interactive settings.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62127803, 624B2124).

References

1. An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., et al.: Mc-llava: Multi-concept personalized vision-language model. arXiv preprint arXiv:2411.11706 (2024)
2. An, R., Yang, S., Zhang, R., Shen, Z., Lu, M., Dai, G., Liang, H., Guo, Z., Yan, S., Luo, Y., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. arXiv preprint arXiv:2505.14671 (2025)
3. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. arXiv preprint arXiv:2304.05884 (2023)
4. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. In: arXiv (2025)
5. 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)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
7. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
8. Chen, Z., Zhao, R., Luo, C., Sun, M., Yu, X., Kang, Y., Huang, R.: Sifthinker: Spatially-aware image focus for visual reasoning (2025), https://arxiv.org/abs/2508.06259
9. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen,

X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948

10. Ding, T., He, L., Ma, W., Zhou, X.: Musellm: Sdf generation and understanding via multi-scale tokenization with position-aware guidance. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. p. 192–201. ICMR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3731715.3733405, https://doi.org/10.1145/3731715.3733405
11. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models (2025), https://arxiv.org/abs/2411.14432
12. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), https://arxiv.org/abs/2306.13394
13. Geng, X., Xia, P., Zhang, Z., Wang, X., Wang, Q., Ding, R., Wang, C., Wu, J., Zhao, Y., Li, K., et al.: Webwatcher: Breaking new frontier of vision-language deep research agent. arXiv preprint arXiv:2508.05748 (2025)
14. Gu, T., Yang, K., Zhang, K., An, X., Feng, Z., Zhang, Y., Cai, W., Deng, J., Bing, L.: Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning (2025)
15. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
16. Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models (2024), https://arxiv.org/abs/2406.09403
17. Hu, Y., Stretcu, O., Lu, C.T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., Fuxman, A.: Visual program distillation: Distilling tools and programmatic reasoning into vision-language models (2023)
18. Huang, Z., Ji, Y., Rajan, A.S., Cai, Z., Xiao, W., Wang, H., Hu, J., Lee, Y.J.: Visualtoolagent (vista): A reinforcement learning framework for visual tool selection (2025), https://arxiv.org/abs/2505.20289
19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019), https://arxiv.org/abs/1902.09506
20. Lee, K.i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., Jung, K.: Vlindbench: Measuring language priors in large vision-language models. arXiv preprint arXiv:2406.08702 (2024)
21. Lee, S.H., Jiang, J., Xu, Y., Li, Z., Ke, J., Li, Y., He, J., Hickson, S., Datsenko, K., Kim, S., Yang, M.H., Essa, I., Yang, F.: Cropper: Vision-language model for image cropping through in-context learning (2025), https://arxiv.org/abs/2408.07790
22. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), https://arxiv.org/abs/2408.03326

23. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding (2025), https://arxiv.org/abs/2504.14920
24. Li, J., Zhang, D., Wang, X., Hao, Z., Lei, J., Tan, Q., Zhou, C., Liu, W., Yang, Y., Xiong, X., et al.: Chemvlm: Exploring the power of multimodal large language models in chemistry area. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 415–423 (2025)
25. Li, Y., Cao, Y., He, H., Cheng, Q., Fu, X., Xiao, X., Wang, T., Tang, R.: M^2IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=9ffYcEiNw9
26. Li, Y., Yang, J., Yang, Z., Li, B., He, H., Yao, Z., Han, L., Chen, Y.V., Fei, S., Liu, D., Tang, R.: Cama: Enhancing multimodal in-context learning with context-aware modulated attention (2025), https://arxiv.org/abs/2505.17097
27. Li, Y., Yun, T., Yang, J., Feng, P., Huang, J., Tang, R.: Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. arXiv preprint arXiv:2505.17098 (2025)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023), https://arxiv.org/abs/2305.10355
29. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The Twelfth International Conference on Learning Representations (2023)
30. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos (2025), https://arxiv.org/abs/2506.05302
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024), https://arxiv.org/abs/2310.03744
32. Liu, Z., Pan, J., She, Q., Gao, Y., Xia, G.: On the faithfulness of visual thinking: Measurement and enhancement (2025), https://arxiv.org/abs/2510.23482
33. Ma, W., Chen, R., Zhang, K., Wu, S., Ding, S.: Instruct where the model fails: Generative data augmentation via guided self-contrastive fine-tuning. Proceedings of the AAAI Conference on Artificial Intelligence **39**(6), 5991–5999 (Apr 2025). https://doi.org/10.1609/aaai.v39i6.32640, https://ojs.aaai.org/index.php/AAAI/article/view/32640
34. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa (2021), https://arxiv.org/abs/2104.12756
35. OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A.T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C.M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F.P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G.,

Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H.W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I.C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J.Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M.Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R.G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S.R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., Li, Z.: Openai o1 system card (2024), https://arxiv.org/abs/2412.16720

36. OpenAI: Thinking with images. https://openai.com/index/thinking-with-images/ (2025), accessed: 2025-09-24

37. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B.,

McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774

38. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools (2023), https://arxiv.org/abs/2302.04761
39. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), https://arxiv.org/abs/1707.06347
40. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning (2024), https://arxiv.org/abs/2403.16999
41. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
42. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., Xu, R., Zhao, T.: Vlm-r1: A stable and generalizable r1-style large vision-language model (2025), https://arxiv.org/abs/2504.07615
43. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read (2019), https://arxiv.org/abs/1904.08920
44. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., Cheng, Y.: Openthinkimg: Learning to think with images via visual tool reinforcement learning (2025), https://arxiv.org/abs/2505.08617
45. Team, T.D.: Tongyi deepresearch: A new era of open-source ai researchers. https://github.com/Alibaba-NLP/DeepResearch (2025)
46. Wang, H., Li, L., Qu, C., Zhu, F., Xu, W., Chu, W., Lin, F.: To code or not to code? adaptive tool integration for math language models via expectation-maximization. arXiv preprint arXiv:2502.00691 (2025)
47. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)

48. Wang, H., Que, H., Xu, Q., Liu, M., Zhou, W., Feng, J., Zhong, W., Ye, W., Yang, T., Huang, W., et al.: Reverse-engineered reasoning for open-ended generation. arXiv preprint arXiv:2509.06160 (2025)
49. Wang, H., Su, A., , Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025)
50. Wang, H., Xu, Q., Liu, C., Wu, J., Lin, F., Chen, W.: Emergent hierarchical reasoning in llms through reinforcement learning. arXiv preprint arXiv:2509.03646 (2025)
51. Wang, H., Zhou, J., He, X.: Learning context-aware task reasoning for efficient meta-reinforcement learning. arXiv preprint arXiv:2003.01373 (2020)
52. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models (2024), https://arxiv.org/abs/2408.15556
53. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
54. Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding (2022), https://arxiv.org/abs/2212.00280
55. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms (2023), https://arxiv.org/abs/2312.14135
56. Wu, Y., Wang, Y., Tang, S., Wu, W., He, T., Ouyang, W., Torr, P., Wu, J.: Dettoolchain: A new prompting paradigm to unleash detection ability of mllm (2024), https://arxiv.org/abs/2403.12488
57. Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., Yao, H.: Mmed-rag: Versatile multimodal rag system for medical vision language models. arXiv preprint arXiv:2410.13085 (2024)
58. Xie, Y., Yang, K., Yang, N., Deng, W., Dai, X., Gu, T., Wang, Y., An, X., Zhao, Y., Feng, Z., et al.: Croc: Pretraining large multimodal models with cross-modal comprehension. arXiv preprint arXiv:2410.14332 (2024)
59. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step (2025), https://arxiv.org/abs/2411.10440
60. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models (2023), https://arxiv.org/abs/2210.03629
61. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Qiao, M., Wu, Y., Wang, M.: Dapo: An open-source llm reinforcement learning system at scale (2025), https://arxiv.org/abs/2503.14476
62. Zhang, D., Lei, J., Li, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., et al.: Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9050–9061 (2025)
63. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl (2025), https://arxiv.org/abs/2505.15436

64. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning (2025), https://arxiv.org/abs/2505.14362
65. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images (2016), https://arxiv.org/abs/1511.03416