
VLM-R³: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought

Chaoya Jiang^{1,*}, Yongrui Heng^{1,*}, Wei Ye^{1,†}, Han Yang³, Haiyang Xu^{2,†},
Ming Yan², Ji Zhang², Fei Huang², Shikun Zhang¹

¹ National Engineering Research Center for Software Engineering, Peking University

² Alibaba Group

³ ZEEKR Intelligent Technology Holding Limited

{wye}@pku.edu.cn,

{shuofeng.xhy}@alibaba-inc.com

Abstract

Recently, reasoning-based MLLMs have achieved a degree of success in generating long-form textual reasoning chains. However, they still struggle with complex tasks that necessitate dynamic and iterative focusing on and revisiting of visual regions to achieve precise grounding of textual reasoning in visual evidence. We introduce **VLM-R³** (Visual Language Model with **R**egion **R**ecognition and **R**easoning), a framework that equips an MLLM with the ability to (i) decide *when* additional visual evidence is needed, (ii) determine *where* to ground within the image, and (iii) seamlessly weave the relevant sub-image content back into an interleaved chain-of-thought. The core of our method is **Region-Conditioned Reinforcement Policy Optimization (R-GRPO)**, a training paradigm that rewards the model for selecting informative regions, formulating appropriate transformations (e.g. crop, zoom), and integrating the resulting visual context into subsequent reasoning steps. To bootstrap this policy, we compile a modest but carefully curated Visuo-Lingual Interleaved Rationale (VLIR) corpus that provides step-level supervision on region selection and textual justification. Extensive experiments on MathVista, ScienceQA, and other benchmarks show that VLM-R³ sets a new state of the art in zero-shot and few-shot settings, with the largest gains appearing on questions demanding subtle spatial reasoning or fine-grained visual cue extraction.

1 Introduction

Multimodal Large Language Models (MLLMs) have recently emerged as a powerful paradigm, demonstrating remarkable capabilities in understanding and generating content across different modalities, primarily vision and language [38, 27, 23, 65, 5, 7]. Models like O1 [37], QvQ [3], and Gemini 2.5 [1] have showcased impressive performance on a wide array of tasks such as MMMU [66], MathVista [30], and ScienceQA [31]. A key factor contributing to their advanced reasoning abilities is the integration of Chain-of-Thought (CoT) prompting [58], which elicits step-by-step reasoning pathways, often leading to more accurate and interpretable outputs.

Despite these advancements, a critical limitation persists in the way current MLLMs interact with visual information during complex reasoning processes. Most existing approaches [3, 37, 61, 62]

*These authors contributed equally to this work.

†corresponding authors.

employing CoT predominantly confine the reasoning steps to the textual domain, with only an initial static grounding in the visual input. This paradigm falls short in scenarios demanding dynamic, iterative, and fine-grained interaction with specific visual regions throughout the reasoning chain. As shown in Figure 1, examples include sequentially verifying hypotheses against image details, tracking object states across visual cues, or comprehending intricate spatial relationships all of which require a more active and adaptive visual grounding mechanism. Encouragingly, recent models such as O3 [2] which capable of interleaving image analysis with text generation inspire a new frontier where reasoning is not merely conditioned on an image, but is continuously intertwined with ongoing visual perception and localization.

Developing an MLLM that can look again during reasoning faces two notable hurdles: **Region-grounding learning**. The model must learn *where* to focus and *how* to transform the grounded region (crop, zoom) based on partial textual deliberation. **Credit assignment**. Simply supervising final answers does not teach the model whether a chosen region actually contributed to correct reasoning, making it hard to refine the visual-query policy.

To bridge this crucial gap, we make two primary contributions. First, we introduce Visuo-Lingual Interleaved Rationale (VLIR), a pioneering dataset meticulously curated to support the development of MLLMs for interleaved text-image CoT reasoning. VLIR provides explicit annotations for visual region localization, image cropping instructions, and semantic enhancement cues, all embedded within multi-step reasoning narratives. Second, building upon this, we propose **VLM-R³** (Visual Language Model with **R**egion **R**ecognition and **R**easoning), a novel framework designed to master this intricate reasoning style. VLM-R³ is trained using a distinctive strategy that combines cold-start finetuning on our VLIR dataset with a novel Region-Conditioned Reinforcement Policy Optimization (R-GRPO). This empowers VLM-R³ to learn when and where to look within an image, how to process the localized visual evidence (e.g., by cropping or requesting enhancement), and how to integrate this dynamically acquired information into its evolving reasoning chain. Our extensive experiments on diverse multimodal reasoning benchmarks, including MME [14], ScienceQA [31] and MathVista [30], demonstrate that VLM-R³ significantly outperforms existing state-of-the-art models. In summary, our contributions are:

- The introduction of VLIR, the first benchmark dataset tailored for training and evaluating MLLMs on interleaved visual-textual CoT reasoning with explicit region-level interactions.
- The proposal of VLM-R³, a novel MLLM framework, and its associated R-GRPO training strategy, which enables dynamic visual region localization and evidence integration within the reasoning process.
- Comprehensive empirical validation showing that VLM-R³ achieves superior performance on challenging multimodal reasoning tasks, setting a new benchmark for fine-grained, visually-grounded inference.

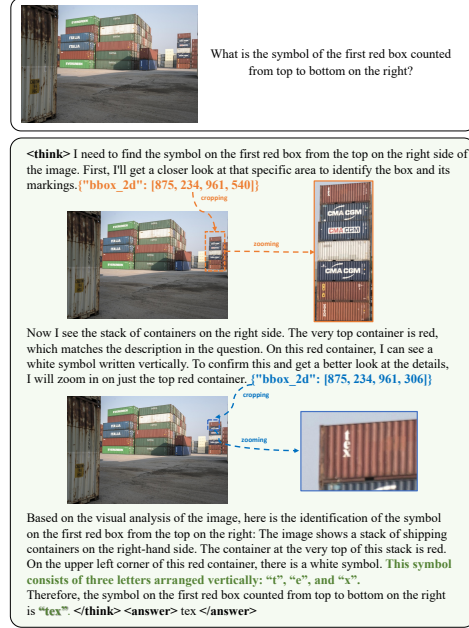


Figure 1: This figure visualizes our proposed VLM-R³ approach, which integrates region grounding and refinement in an interleaved visual-textual reasoning chain. While conventional text-based reasoning fails when analyzing scenes that require dynamic, iterative, and fine-grained interaction with specific visual regions, our approach succeeds by precisely identifying and focusing on critical visual elements, such as 'the first red box counted from top to bottom on the right' in this example, to derive accurate conclusions through targeted visual reasoning.

2 Related Work

2.1 Large Language Model Reasoning

Reasoning in Large Language Models [50, 70, 67, 18] evolved substantially with Chain of Thought (CoT) prompting [49, 58, 24, 57, 35, 41], which enables models to break down complex problems into intermediate steps, mimicking human reasoning. This foundational approach has expanded to include diverse structures like program-of-thoughts [10], table-of-thoughts [19], and tree-of-thoughts [64], each offering unique advantages for different reasoning scenarios. Recent advances include OpenAI’s O1 [37], which combines reinforcement learning [39, 44, 16] with CoT to optimize decision-making without external guidance, and DeepSeek R1 [12], which employs pure reinforcement learning through Group Relative Policy Optimization (GRPO) [46] to enable autonomous evolution of reasoning capabilities while incorporating rule-based rewards that significantly improve performance across complex reasoning tasks.

2.2 Multi-modal Large Language Model Reasoning

Multi-modal Large Language Model reasoning research [69, 59, 43, 34, 40, 29, 28] has emerged following the success of text-only reasoning models [37, 6, 12, 52], focusing on both effective multi-modal chain-of-thought structures [61, 53, 51, 21] and high-quality training data construction methods [13, 47, 4]. Mainstream approaches have adapted text-based reasoning paradigms to multi-modal contexts, as seen in Virgo [13], which demonstrated that text-only reasoning data can activate certain multi-modal reasoning capabilities, and more structured frameworks like LLaVA-CoT’s [61] four-stage reasoning process and MM-Verify’s [51] verification-enhanced approach. However, these methods largely inherit reasoning paradigms from text-only models without adequately addressing visual information processing, leading to limitations in visually-intensive reasoning tasks.

3 Method

We propose a novel framework, VLM-R³, designed to perform visuo-lingual interleaved reasoning with region grounding. This section details the components of our approach, including the construction of the Visuo-Lingual Interleaved Rationale (VLIR) dataset used for cold-start supervised fine-tuning, the interactive inference pipeline enabling dynamic visual grounding, and the Region-Conditioned Reinforcement Policy Optimization (R-GRPO) strategy employed to enhance reasoning capabilities.

3.1 Visuo-Lingual Interleaved Rationale (VLIR) Dataset

Prior work, such as Visual CoT [45], introduced the concept of incorporating visual grounding (specifically bounding boxes) into reasoning chains. However, these methods typically suffer from several limitations: (1) They often lack explicit linguistic reasoning steps interleaved with visual actions. (2) The visual grounding actions (e.g., cropping based on bounding boxes) are predefined or manually specified, rather than being dynamically generated by the model. (3) They are often restricted to a limited number of visual interactions, typically a single bounding box selection before providing a final answer, lacking the flexibility for multi-step visual querying. To address these limitations and cultivate the ability for models to autonomously and flexibly perform iterative visual retrieval and cropping based on their ongoing reasoning, we introduce the **Visuo-Lingual Interleaved Rationale (VLIR)** dataset. This dataset is specifically curated to provide rich, interleaved sequences of textual reasoning steps interspersed with explicit visual grounding actions and the corresponding cropped visual evidence.

3.1.1 Data Construction

The construction of the VLIR dataset focuses on scenarios that necessitate fine-grained spatial understanding and precise utilization of visual cues. We select data from a diverse set of existing benchmarks to cover a wide range of visual reasoning challenges:

- **Text/Document Understanding:** TextVQA [48], DocVQA [33] for tasks requiring OCR and document structure understanding.

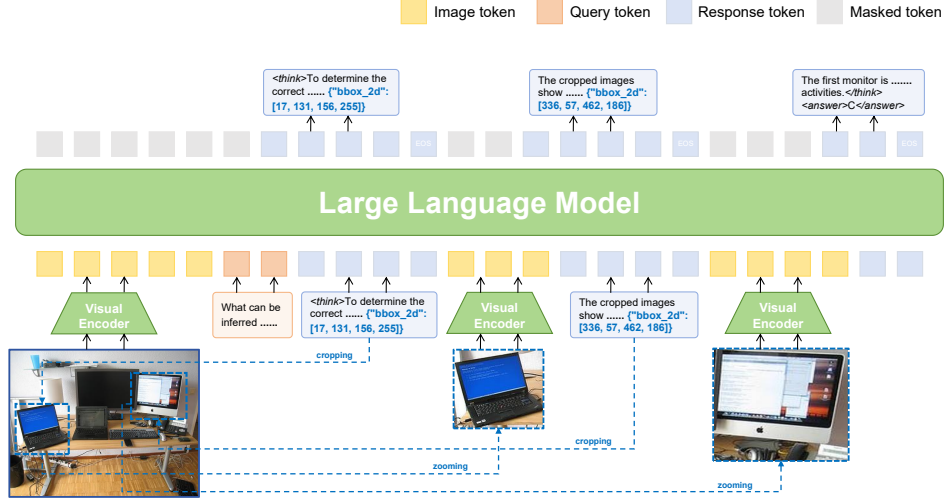


Figure 2: The visualization of the inference pipeline of our proposed methods.

- **General Visual Question Answering:** GQA [17] for complex multistep reasoning over visual scenes.
- **Chart and Infographic Interpretation:** InfographicsVQA [32] for understanding structured visual data.
- **Spatial Relation Reasoning:** VSR [26] for tasks focused on identifying and reasoning about spatial relationships between objects.

We leverage the advanced capabilities of powerful MLLMs, such as Qwen2.5-VL 72B [8], through sophisticated prompt engineering to generate interleaved image-text reasoning chains for data points from benchmarks like GQA and TextVQA, which represent real-world question answering scenarios. We then employ a rejection sampling strategy on the generated samples, filtering for those that align with the ground-truth answers.

For tasks where direct prompt engineering on the original image-question pair is less effective, particularly those involving detailed OCR or tabular data interpretation (e.g., data underlying Visual CoT [45]), we utilize GPT-4o [36] with tailored prompts that incorporate the metadata provided by the source dataset (e.g., the initial bounding boxes from Visual CoT). This allows us to generate detailed, step-by-step interleaved rationales within these challenging domains.

3.1.2 Data Filtering

To ensure the quality and relevance of the interleaved rationales generated, we apply a rigorous filtering process based on the following criteria:

- **Semantic Unit Validity of Regions:** Each proposed bounding box must enclose a complete and semantically meaningful visual unit (e.g., a recognizable object, a block of text, or a distinct part of a chart). To automate this, we utilize a smaller VLM and prompt it with the cropped image corresponding to the proposed bbox, asking it to confirm the presence and identity of a recognizable entity ("Can you identify what is in this image (a specific object or piece of text)? Respond with yes/no."). Samples in which the VLM fails to confirm a meaningful semantic unit are rejected.
- **Logical Coherence and Non-Redundancy of Reasoning:** The generated textual reasoning steps must be logically sound, progressive, and directly contribute to arriving at the final answer, avoiding spurious or redundant text. We employ a powerful text-only LLM, such as DeepSeek V3 [25], through prompt engineering to evaluate the logical flow and relevance of the text rationale preceding each visual interaction and the overall reasoning path. Samples with illogical or padded reasoning are rejected.

3.2 Interactive Inference Pipeline

The VLM-R³ model executes reasoning through an interactive pipeline that enables the model to dynamically select and incorporate visual information during its inference process.

The interaction is initiated by providing the VLM-R³ with an instruction that defines the reasoning task and the available visual interaction tool:

You need to first think about the reasoning process in your mind, and then provide the answer. When thinking, you should call the "crop" tool (format: {"bbox_2d": [x1, y1, x2, y2]}) to focus on the key areas in the image. The reasoning process and the answer are included in the <think> </think> and <answer> </answer> tags respectively.

When the model generates a string that matches the specified JSON format, the pipeline intercepts the output. The system parses the coordinates $[x1, y1, x2, y2]$ and performs a cropping operation on the original input image. The resulting cropped image is then zoomed in and encoded into visual tokens and appended to the model's input sequence, effectively providing the model with the requested visual detail as a new context. Following the injection of the cropped image, the model resumes generation, which may involve generating further text or issuing additional "Crop" commands. This interactive loop continues until the model generates the final answer, at which point the process terminates. This pipeline structure shows as Figure 2,

3.3 Region-Conditioned Reinforcement Policy Optimization (R-GRPO)

Standard supervised learning on fixed trajectories struggles to optimize the complex state-dependent policy of deciding *when* and *where* to acquire visual information. Our approach, Region-Conditioned Reinforcement Policy Optimization (R-GRPO), adapts a policy optimization framework, building upon Group Relative Policy Optimization (GRPO) [46]. The "Region-Conditioned" aspect implies that π_θ is explicitly conditioned on the visual state, including dynamically incorporated regional evidence.

To estimate the advantage of each reasoning trajectory, we normalize its reward relative to the group as follow:

$$\hat{A}^i = \frac{r^i - \text{mean}(\{r^1, r^2, \dots, r^M\})}{\text{std}(\{r^1, r^2, \dots, r^M\})} \quad (1)$$

Here, r^i is the total reward for the i -th trajectory in a group of M trajectories, and $\hat{A}^i$ serves as a form of advantage function relative to the group performance.

A critical adaptation in R-GRPO concerns the computation of the policy gradient and the actions considered in the objective. In our interleaved image-text sequences, some tokens are generated by the model (textual reasoning, bbox commands), while others (the representations of cropped images) are injected by the environment. The policy gradient should only optimize the likelihood of actions generated by the model. Therefore, when calculating the gradient of $\log \pi_\theta(a_t | s_t)$, we apply a mask: the gradient is computed only for tokens a_t that are text tokens or bounding box command tokens, masking out gradients for tokens corresponding to injected image regions. Conceptually, the sum of actions $\mathcal{A}_s$ in the loss primarily considers the probabilities of generating valid text tokens and bounding box commands, weighted by their advantage. The injected image tokens influence the state s_{t+1} but are not actions a_t for which we compute a policy gradient.

Following this, we optimize the policy model π_θ with the loss function defined as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{Q \in D_S} \left[\sum_{i=1}^M \frac{\pi_\theta(c^i | Q)}{\pi_\theta(c^i | Q)|_{\text{no grad}}} \hat{A}^i - \beta D_{KL}(\pi_\theta || \pi_{\text{ref}}) \right] \quad (2)$$

where D_S is the dataset of question-state pairs, Q represents a specific question and current visual state, c^i is the sequence of generated tokens for the i -th trajectory given Q , and β is a coefficient for the KL divergence term. The first term in the sum uses the normalized reward $\hat{A}^i$ to weight the likelihood of the generated sequence, encouraging sequences with higher relative rewards.

The KL divergence between the policy model and the reference model is estimated as in [46]:

$$D_{KL}(\pi_{\theta}||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(c^i|Q)}{\pi_{\theta}(c^i|Q)} - \log \frac{\pi_{\text{ref}}(c^i|Q)}{\pi_{\theta}(c^i|Q)} - 1 \quad (3)$$

The total reward r^i for a trajectory comprises four components that encourage desired VLM-R³ behaviors:

- **Accuracy Reward (r_{acc}):** A terminal reward: 1 if the final answer is correct, 0 otherwise.
- **Format Reward (r_{format}):** A terminal reward: 1 if the output uses correct <answer> tags, 0 otherwise.
- **Region Validity Reward (r_{region}):** An intermediate reward of 0.5 for each valid, non-redundant bounding box generated (only when the answer is correct), capped at 0.5 per episode.
- **Reasoning Length Reward (r_{length}):** An intermediate reward of 0.001 per character for reasoning steps, capped at 0.25 per episode to prevent verbosity.

The total reward is computed as:

$$r_{overall} = r_{acc} + r_{format} + r_{length} + r_{region} \cdot \mathbb{I}(r_{acc} = 1.0) \quad (4)$$

By optimizing this objective, R-GRPO encourages the VLM to learn a policy that not only leads to correct final answers but also involves generating logical textual reasoning and strategically gathering the necessary visual evidence.

4 Experiments

4.1 Experiment Setting

We covers several public benchmarks. General visionlanguage understanding is measured on MME [14] and MMMU[66]; complex mathematical reasoning on MathVista [30] and MathVision [55]; scientific question answering on ScienceQA [31]; and document understanding on DocQA [33]. We also assess hallucination rates with HallucinationBench[15]. We evaluate our method against three categories of multimodal models. The first category consists of open-source baselines without explicit reasoning capability, including Qwen2.5-VL 7B [8] (also used as our primary baseline), InternVL2.5-8B [11], and LLaVA-Next 8B [22]. The second category comprises closed-source non-reasoning systems, represented by Gemini-2 Flash [1] and GPT-4o [36]. The third category contains models equipped with dedicated reasoning modules, namely LLaVA-CoT 11B [61], Mulberry-Qwen2VL 7B [63], R1-onevision 7B [62]. To probe the upper bound of performance, we also compare our results with two larger closed-source models o1 [37]. Moreover, we also evaluate our model on four vision-centric benchmarks: MMVP [54], V* [60], HR-Bench [56], and MME-Realworld [68] to evaluate the model’s capability in fine-grained visual understanding and spatial reasoning.

4.2 Dataset Details

Our supervised fine-tuning experiments used the VLIR dataset, which comprises 11,810 samples in total. As shown in figure 3, the distribution of crops per image exhibits considerable variation: 11,105 images contain a single crop, 607 images feature two crops, 68 images have three crops, 16 images include four crops, 8 images contain five crops, and 6 images have six or seven crops (3 each). These samples are drawn from five distinct source datasets: GQA (4,057 samples), TextVQA (3,267 samples), DocVQA (1,497 samples), InfographicsVQA (1,497 samples), and VSR (1,492 samples). We categorize the crops based on their relative size, defined as the ratio of the bounding box area to the total image area: "very small" (ratio < 0.05) accounts for 5,280 crops; "small" (0.05 ≤ ratio < 0.25) comprises 4,043 crops; "medium" (0.25 ≤ ratio < 0.5) includes 1,914 crops; and "large" (ratio ≥ 0.5) consists of 573 crops.

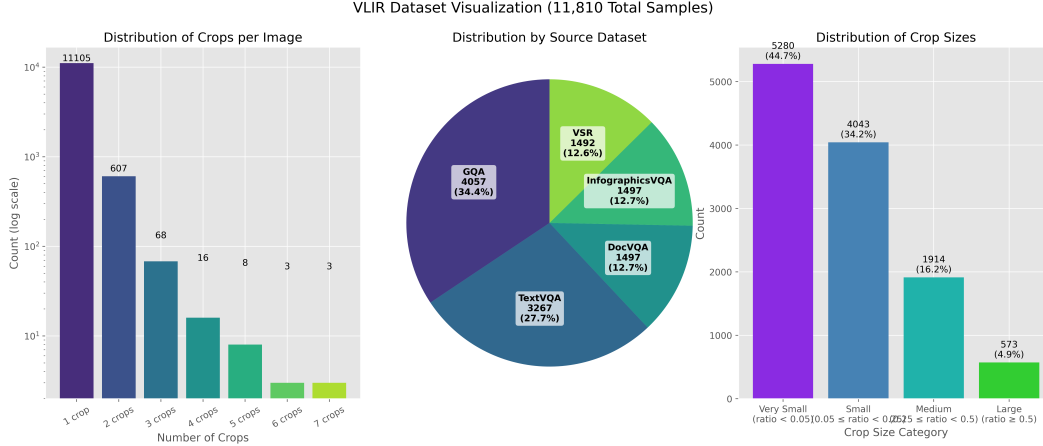


Figure 3: Distribution of the VLIR dataset: (a) number of crops per image, (b) samples across different source datasets, and (c) categorization of crops based on relative size.

Table 1: Performance comparison of various multimodal models across different benchmarks. Our model (in **bold**) is compared against non-reasoning models (both open-source and closed-source) and reasoning-based MLLMs. The highest values in each column are highlighted with green background.

Model	Params	Benchmarks						
		MathVista	MathVision	MMMU	MME	ScienceQA	DocVQA	HallusionBench
Closed-Source Non-Reasoning MLLMs								
Gemini-2 Flash	-	73.1	41.3	71.7	-	-	92.1	-
GPT-4o	-	63.8	30.4	70.3	2328	66.2	91.1	56.2
Larger Closed-Source Models								
o1	-	71.8	63.2	77.6	-	-	81.6	-
Open-Source Non-Reasoning MLLMs								
InternVL2.5	8B	58.3	17.1	51.8	2210	-	-	-
LLaVA-Next	8B	37.5	-	41.7	1957	72.8	-	-
Open-Source Reasoning MLLMs								
LLaVA-CoT	11B	54.8	-	-	-	-	-	47.8
R1-onevision	7B	64.1	29.9	-	-	-	-	-
Vision-R1	7B	73.5	-	-	2190	-	-	49.5
Mulberry	7B	63.1	-	55.0	-	-	-	54.1
Qwen2.5-VL	7B	68.2	25.1	58.6	2347	73.6	95.7	61.3
Ours	7B	70.4	30.2	62.2	2432	87.9	96.8	62.0

4.3 Main Result

Our VLM-R³ model, built upon the Qwen2.5-VL 7B architecture, consistently outperforms its base model across all benchmarks, with particularly significant gains in domains requiring precise visual reasoning and fine-grained understanding. Specifically, we observe a 2.2% improvement on MathVista (70.4% vs. 68.2%) and a remarkable 5.1% improvement on MathVision (30.2% vs. 25.1%), highlighting our method’s effectiveness in mathematical reasoning tasks that demand careful attention to visual details. The substantial performance gain of 14.33% on ScienceQA (87.90% vs. 73.57%) further demonstrates VLM-R³’s superior capability in scientific reasoning, where dynamic grounding of visual evidence is critical. When compared to other open-source reasoning-focused models like Vision-R1 and Mulberry, VLM-R³ exhibits competitive performance on MathVista and surpasses Mulberry on HallusionBench (62.0% vs. 54.1%), indicating enhanced reliability in avoiding visual hallucinations. Our approach also narrows the gap with closed-source models like Gemini-2 Flash and o1, despite having significantly fewer parameters and being fully transparent in its architecture.

Table 2: Performance comparison and ablation study on vision-centric benchmarks. The highest values in each column are highlighted with green background.

Model	MMVP	V*	HR-4k	HR-8k	MME-Real	Avg.
<i>Open-source General Models</i>						
LLaVA-OneVision-72B	–	73.8	66.3	60.9	48.7	–
InternVL3-8B	–	72.3	70.8	62.0	47.9	–
InternVL3-38B	–	77.5	76.3	67.0	51.0	–
Qwen2.5VL-7B (Base)	66.7	74.3	69.8	64.6	42.3	63.5
Base + Vanilla Text RL	72.0	78.5	72.9	64.1	46.2	66.7
Base + Text-only Bbox	72.4	79.5	73.2	65.2	47.2	67.5
Base + Re-insert Full Image	73.9	77.8	72.5	64.8	45.8	67.0
Ours	75.0	83.8	73.4	66.8	51.6	70.1
Δ vs. Vanilla Text RL	+3.0	+5.3	+0.5	+2.7	+5.4	+3.4
Δ vs. Text-only Bbox	+2.6	+4.3	+0.2	+1.6	+4.4	+2.6
Δ vs. Re-insert Full Image	+1.1	+6.0	+0.9	+2.0	+5.8	+3.1

Table 3: Ablation study on MathVista, MMMU, ScienceQA, and DocVQA benchmarks. We evaluate the contribution of each key component: Interleaved Chain-of-Thought, VLIR fine-tuning, and R-GRPO.

Model Variant	MathVista	MMMU	ScienceQA	DocVQA	Avg.
Base Model (Qwen2.5-VL)	68.2	58.6	73.6	95.7	74.0
w/o Interleaved Chain-of-Thought	67.1 ($\downarrow$ 3.3)	59.4 ($\downarrow$ 2.8)	75.4 ($\downarrow$ 12.5)	95.9 ($\downarrow$ 0.9)	74.4 ($\downarrow$ 4.9)
w/o VLIR Fine-tuning	65.8 ($\downarrow$ 4.6)	57.0 ($\downarrow$ 5.2)	72.2 ($\downarrow$ 15.7)	93.3 ($\downarrow$ 3.5)	72.1 ($\downarrow$ 7.2)
w/o R-GRPO	69.7 ($\downarrow$ 0.7)	60.8 ($\downarrow$ 1.4)	84.6 ($\downarrow$ 3.3)	96.1 ($\downarrow$ 0.7)	77.8 ($\downarrow$ 1.5)
Full VLM-R ³ (Ours)	70.4	62.2	87.9	96.8	79.3

Table 2 presents the performance of our VLM-R³ model on vision-centric benchmarks, including MMVP [54], V* [60], HRBench [56], and MME-Realworld [68]. Our model shows superior capability in handling complex visual reasoning tasks. Notably, VLM-R³ attains an average score of 70.1%, significantly outperforming the base Qwen2.5-VL model (63.5%) and larger open-source models like LLaVA-OneVision-72B [20] and InternVL3-38B [71].

4.4 Ablation Study

To assess the contribution of each component in our VLM-R³ framework, we conduct comprehensive ablation experiments. Table 2 and Table 3 summarize our findings.

4.4.1 Effectiveness of Interleaved Chain-of-Thought

To isolate the impact of our interleaved reasoning approach, we conduct an experiment where we maintain the region localization capabilities (bounding boxes) but remove the associated region images from the reasoning chain. This variant relies solely on textual descriptions of identified regions without visually grounding each reasoning step. As shown in Table 3, removing the interleaved visual evidence leads to a consistent performance drop across all benchmarks, with particularly notable decreases on ScienceQA (-12.5%) and MMMU (-2.8%). This degradation is most pronounced in tasks requiring fine-grained visual understanding, such as scientific diagrams in ScienceQA, where purely textual descriptions of regions fail to capture crucial visual patterns and spatial relationships.

In Table 2, we further compare this variant against other ablations on vision-centric benchmarks. The results indicate that the interleaved reasoning chain achieves an average improvement of 3.4% over a vanilla text-based reinforcement learning approach. Furthermore, it outperforms variants that rely solely on text-based bounding box descriptions (+2.6%) or re-inserting the full image at each reasoning step (+3.1%). These findings underscore the critical role of dynamically incorporating visual regions into the reasoning process for effective multimodal understanding.

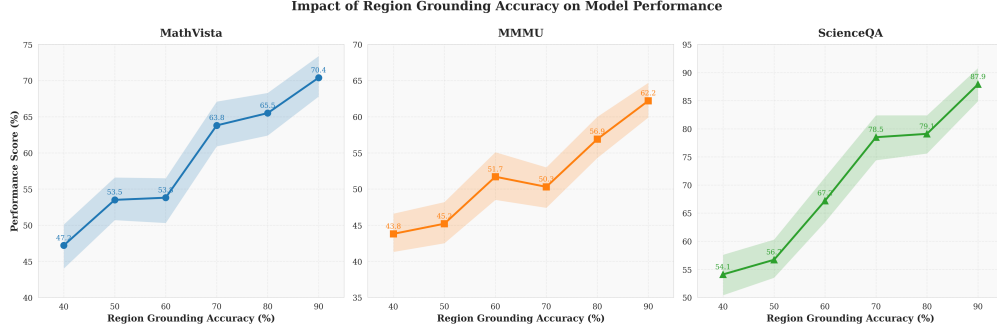


Figure 4: Impact of region grounding accuracy on model performance across three benchmarks. Each subplot shows the performance trajectory from 40% to 90% grounding accuracy with confidence intervals (shaded regions).

4.4.2 Effectiveness of Finetuning on VLIR

Our approach leverages the VLIR corpus to bootstrap the model’s ability to identify informative regions and incorporate them into coherent reasoning chains. To evaluate the specific contribution of VLIR fine-tuning, we experiment with a variant that skips this initialization phase and proceeds directly to R-GRPO training. The results in Table 3 demonstrate that omitting VLIR fine-tuning leads to performance degradation across all benchmarks, with particularly significant decreases observed in ScienceQA (-15.7%) and MMMU (-5.2%). More critically, we observed that ablating VLIR fine-tuning impairs the model’s instruction-following capabilities, leading to substantial deficiencies such as failures to adhere to required formatting conventions for bounding box specifications. This accounts for the substantial performance deterioration observed in our experimental results.

4.4.3 Effectiveness of R-GRPO

To assess the impact of our Region-Conditioned Reinforcement Policy Optimization (R-GRPO), we evaluate a variant that relies solely on supervised fine-tuning using the VLIR corpus without the subsequent reinforcement learning stage. This allows us to isolate the specific benefits of our reinforcement learning approach over purely supervised learning. The experimental results show that removing R-GRPO reduces performance in all benchmarks, with the highest decreases observed in ScienceQA (-3.28%) and MathVista (-0.7%). This suggests that while VLIR fine-tuning provides a strong foundation, the reinforcement learning stage is essential for optimizing the model’s region selection and reasoning policies beyond what can be achieved through imitation learning alone.

4.5 Discussion

4.5.1 Impact of Region Grounding Accuracy on the Reasoning Chain

The quality of region grounding, represented by the accuracy of bounding boxes (bbox), plays a critical role in multimodal reasoning capabilities. Our analysis investigates how varying levels of grounding accuracy impact the performance of the VLM-R³ model across multiple benchmarks. We systematically evaluated model performance by controlling grounding accuracy from 40% to 90% and measuring outcomes on three key benchmarks: ScienceQA, MathVista, and MMMU. Grounding accuracy was manipulated by randomly replacing or perturbing a controlled percentage of bounding boxes in the input. As shown in Figure 4, there is a clear positive correlation between region grounding accuracy and model performance across all three benchmarks. ScienceQA demonstrates the most substantial improvement, with performance increasing from 54.1% at 40% grounding accuracy to 87.9% at 90% grounding accuracy. MathVista shows a similar upward trend, rising from 47.9% to 70.4%, while MMMU exhibits more modest but consistent gains from 43.8% to 62.2%. These results underscore the fundamental importance of precise region grounding for effective multimodal reasoning, with higher-level reasoning tasks showing greater sensitivity to grounding quality.



Figure 5: Comparison of attention distribution patterns between the interleaved reasoning chain with visual region localization (top) and general textual reasoning chain (bottom).

4.5.2 Why is the Interleaved Reasoning Chain with Visual Region Localization Effective?

To understand the efficacy of our VLM-R³ approach, we conducted a comparative analysis between the interleaved reasoning chain with visual region localization and traditional textual reasoning chains. Figure 5 visualizes the attention distribution patterns for both approaches when answering the same visual query. Our analysis reveals a critical insight: in traditional approaches where the image is positioned at the beginning of the sequence, attention to visual information diminishes significantly as the reasoning chain progresses. As shown in the lower portion of Figure 4, general MLLMs tend to make incorrect inferences (identifying a "menu board" instead of an umbrella) as they lose visual context during extended reasoning. In contrast, VLM-R³ maintains persistent visual attention throughout the reasoning process by dynamically localizing and incorporating relevant visual regions. The attention heatmap demonstrates that tokens generated later in the reasoning process maintain strong attention connections to the cropped visual regions. This region-specific attention enables the model to correctly identify the blue object as an umbrella by explicitly focusing on the area behind the person, cropping it for detailed examination, and making accurate observations about its features.

5 Conclusion

This paper introduced VLM-R³, a novel framework enabling MLLMs to perform dynamic visual reasoning through region recognition, reasoning, and refinement. By integrating our custom VLIR dataset and Region-Conditioned Reinforcement Policy Optimization (R-GRPO), we demonstrated that interleaved visual-textual chains-of-thought significantly outperform traditional approaches. VLM-R³ achieves state-of-the-art results across multiple benchmarks, particularly excelling in tasks requiring fine-grained spatial reasoning and visual evidence integration. Our work opens promising directions for developing more sophisticated visually-grounded reasoning systems that can adaptively focus on relevant regions during multi-step inference processes.

6 Acknowledgement

This work is supported by the National Natural Science Foundation of China (NO.623B2007), Chinese Institute of Electronics (CIE) - Tencent Doctoral Research Incentive Program (Hunyuan Large Language Model Special Track) (NO.841010445) and CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202415).

References

- [1] Gemini 2.5: Our most intelligent ai model, 2024. <https://deepmind.google/technologies/gemini/>.
- [2] Introducing openai o3 and o4-mini, 2024. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [3] Qvq: To see the world with wisdom, 2024. <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- [4] Seekworld: Geolocation is a natural rl task for o3-like visual clue-tracking reasoning, 2025. <https://huggingface.co/datasets/TheEighthDay/SeekWorld>.
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, 2022.
- [6] Gemini Team Google Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, and Johan Schalkwyk... Gemini: A family of highly capable multi-modal models. *ArXiv*, abs/2312.11805, 2023.
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023.
- [8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.
- [9] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M³cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought, 2024.
- [10] Wenhui Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, 2023, 2022.
- [11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.
- [12] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jia Shi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jiong Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

- Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, Ruiqi Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948, 2025.
- [13] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm, 2025.
- [14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xianwu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [15] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-bench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024.
- [16] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.
- [17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [18] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.
- [19] Ziqi Jin and Wei Lu. Tab-cot: Zero-shot tabular chain of thought. In *Annual Meeting of the Association for Computational Linguistics*, 2023.
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [21] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulic, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought, 2025.
- [22] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024.
- [23] Junnan Li, Dongxu Li, Silvio Savarese, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [24] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step, 2023.

- [25] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [26] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, et al. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [28] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement, 2025.
- [29] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025.
- [30] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.
- [31] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022.
- [32] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021.
- [33] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.
- [34] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*, 2024.
- [35] Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. Skeleton-of-thought: Prompting llms for efficient parallel generation. *arXiv preprint arXiv:2307.15337*, 2023.
- [36] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisopoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gierler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Han-nah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner,

Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kevin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patherdhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyei Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yuri Malkov. Gpt-4o system card, 2024.

- [37] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo

Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitvyr Pong, Vlad Fomenko, Weiye Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024.

- [38] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.
- [39] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
- [40] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl, 2025.
- [41] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*, 2024.
- [42] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.
- [43] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.
- [44] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- [45] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024.
- [46] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- [47] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025.
- [48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [49] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv*, abs/2408.03314, 2024.
- [50] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.
- [51] Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification, 2025.
- [52] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [53] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025.
- [54] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.
- [55] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024.
- [56] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint*, 2024.
- [57] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [59] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Ruihan Jin, Feihu Che, Zengqi Wen, and Jianhua Tao. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*, 2025.
- [60] 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*, pages 13084–13094, 2024.

- [61] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2025.
- [62] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025.
- [63] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search, 2024.
- [64] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601, 2023.
- [65] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [66] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.
- [67] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.
- [68] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.
- [69] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [70] Fan Zhou, Haoyu Dong, Qian Liu, Zhoujun Cheng, Shi Han, and Dongmei Zhang. Reflection of thought: Inversely eliciting numerical reasoning in language models via solving linear systems. *arXiv preprint arXiv:2210.05075*, 2022.
- [71] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (12 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Experiment Settings

A.1 Pipeline Settings

A.1.1 Model Hyperparameter Settings

Our base model is Qwen2.5VL-7B[8], which supports dynamic resolution for input images. In all experiments, we constrained the pixel dimensions of each image to a minimum of 3136 pixels and a maximum of 1605632 pixels. Because the value of the bounding box is related to the number of pixels in the input image, the setting of the range of pixels needs to be unified.

A.1.2 Zoom Scaling Rule

In our pipeline, when a region is selected for closer inspection (e.g., via a "Crop" operation), a zoom operation is applied. The scaling factor for this zoom, denoted as *scale*, is determined dynamically based on the relative area of the selected bounding box (A_{bbox}) compared to the area of the original image (A_{orig}). Let $r = \frac{A_{\text{bbox}}}{A_{\text{orig}}}$ be this area ratio. The scale is calculated using the following piecewise function:

$$\text{scale} = \begin{cases} 2.0, & \text{if } r < 0.125 \\ 1.0, & \text{if } r \geq 0.5 \\ 2.0 - \frac{r - 0.125}{0.375}, & \text{otherwise} \end{cases} \quad (5)$$

This rule implies that smaller selected regions (smaller r) are scaled up more significantly (up to a factor of 2.0), while larger regions (larger r) are scaled up less, or not at all if they already occupy a substantial portion of the original image. The intermediate case provides a linear interpolation of the scaling factor.

A.2 Training Setting for Supervised Fine-tuning Stage

In the supervised fine-tuning stage, we used the complete VLIR dataset. Our experiments were conducted on 4 NVIDIA A100 GPUs, each equipped with 80GB of memory, leveraging DeepSpeed[42] for efficient training. We used a batch size of 2 with a gradient accumulation of 8, a learning rate of 2×10^{-7} , and trained for 3 epochs. During this phase, the vision encoder and MLP projector were frozen, and only the Large Language Model (LLM) component was trained.

A.3 Training Setting for R-GRPO Stage

For the R-GRPO stage, we sampled approximately 5,000 data points from TextVQA [48], GQA [17], VSR [26], DocVQA [33] and M³CoT [9] datasets. Regarding the hyperparameters for the GRPO formulation(2), we set $M = 5$. Following the experience of related studies, we set $\beta = 0.0$, i.e., we eliminate the KL divergence constraint.

Our experiments for R-GRPO were performed on 6 NVIDIA A100 GPUs, each with 80GB of memory, also utilizing DeepSpeed[42]. The batch size per device was set to 1, with a gradient accumulation of 16. The learning rate was 1×10^{-6} , and training continued for 300 steps. We employ a rule-based reinforcement learning approach, where the correctness of the final answer was judged using an exact match criterion. Similar to the supervised fine-tuning stage, the vision encoder and MLP projector were frozen, and only the LLM component was trained.

B Prompt Templates for VLIR Dataset Construction and Filtering

B.1 Data Construction Prompts

Given an {image, question, answer} triplet, the following prompt was used to construct the interleaved visual-linguistic chain of thought:

You are performing "Multimodal Interleaved Reasoning". During the thinking

```

process, you need to keep an eye on the visual cues in the original image,
find regions of the image that help answer the question, and use the "Crop"
tool to crop and zoom in for detailed analysis.
When using the tool, you must output a JSON object in the following format:
{"bbox_2d": [x1, y1, x2, y2]}
Ensure that you "Crop" at least once.
Continue thinking after each operation until you reach the final answer.
Output the thinking process within a pair of <think> </think> tags and then
output the final answer within a pair of <answer> </answer> tags.
{question}

```

Listing 1: Prompt for dataset construction.

Given an {image, question, answer, bounding box annotation} quadruplet, the following prompt was used:

```

I will now provide you with an image, a question, and a "Crop" operation
string. Your task is to write the reasoning process used to answer the
question as instructed. During the reasoning process, the respondent
utilizes a "Crop" operation to assist with reasoning. The format of
the operation is as follows:
{"bbox_2d": [x1, y1, x2, y2]}
This bounding box indicates the key region that needs to be focused
on to correctly answer the question.
You must think step by step from the perspective of the respondent,
using the "Crop" operation at appropriate moments in your reasoning
process to eventually reach the correct answer. Important notes:
1. You must not modify the content or format of the "Crop" operation
in any way.
2. In a real setting, the respondent only has access to the image and
the question. This bounding box indicates the area where the correct
answer information is located. In this task, they are provided to ensure
the correctness of your reasoning process. When writing the reasoning,
pretend you are the respondent who independently identifies when to use
the "Crop" operation and how to reach the answer step by step.
3. Make sure the reasoning is fluent, logical, and concise.
4. Format of the reasoning process: <think>...</think><answer>...</answer>

Here is an example:
Question: Are there any black numbers or letters?
"Crop" operation: {"bbox_2d": [247, 384, 307, 444]}
Reasoning: <think>
Step 1: To determine if there are black numbers or letters, I need to
focus on the text visible in the image. The dog is wearing a heart-shaped
tag that has some text on it. I will crop and zoom in on the tag for a
closer look at the text details. {"bbox_2d": [247, 384, 307, 444]}
Step 2: After cropping, I can see that the letters "G PLUS" are in red,
and the numbers "6 223 13" are also in red. There are no black numbers
or letters on the tag. Review the rest of the image, there are no black
numbers or letters either.</think>
<answer>no</answer>

Question:{question}
"Crop" operation:{crop}
Now Output the reasoning process:

```

Listing 2: Prompt for dataset construction.

B.2 Data Filtering Prompts

The prompt for assessing the recognizability of the cropped images is as follows:

You need to determine whether the content in a picture is a complete and semantically meaningful visual unit. Please look carefully at this cropped image and determine whether it contains a recognizable object, block of text, or specific part of a diagram. If it is recognizable, answer 'yes'; if not, answer 'no'.
Now output 'yes' or 'no' directly.

Listing 3: Prompt for assessing cropped image recognizability.

The prompt for assessing the quality of the reasoning process is as follows:

You need to make an in-depth assessment of this reasoning process. First, determine whether its logic is rigorous and whether each step of reasoning leads naturally and smoothly to the next; second, check whether the reasoning process progresses gradually towards arriving at the final answer; and lastly, check whether there is any false information or repetitive redundancy in the text that is not relevant to the reasoning. If this textual reasoning meets the requirements in terms of logic, advancement and content streamlining, output 'yes'; whenever one of these is not met, output 'no'.
{question}
{ground-truth answer}
{reasoning process}
Now output 'yes' or 'no' directly.

Listing 4: Prompt for assessing reasoning process quality.

C More Demonstrations

This section provides qualitative examples illustrating the capabilities of VLM-R³.



What is the total volume of the measuring cup?

To determine the total volume of the measuring cup, I need to focus on the markings on the side of the cup. The markings indicate the volume in milliliters (ml) and grams (g). The maximum volume is clearly marked at the top of the cup.
Step 1: Use the Crop tool to zoom in on the markings on the measuring cup to ensure accuracy.
{ "bbox_2d": [616, 212, 1021, 599] }



Step 2: Analyze the markings. The maximum volume is indicated by the highest number on the scale, which is 1000 ml.
</think>
<answer>1000 ml</answer>



What is the total volume of the measuring cup?

<think>First, I'll apply the crop tool to isolate the key areas of the image:
{ "bbox_2d": [208, 641, 482, 983] }



The label clearly states "Chablis Premier Cru" and "William Fevre," which are French wine appellations. Additionally, the phrase "Produit de France" is prominently displayed on the label. This information strongly indicates that the wine is from France. The vintage year, 2010, does not affect the country of origin.</think>
<answer>France</answer>

Figure 6: This figure presents how VLM-R³ captures details in a natural image and reasons about them.

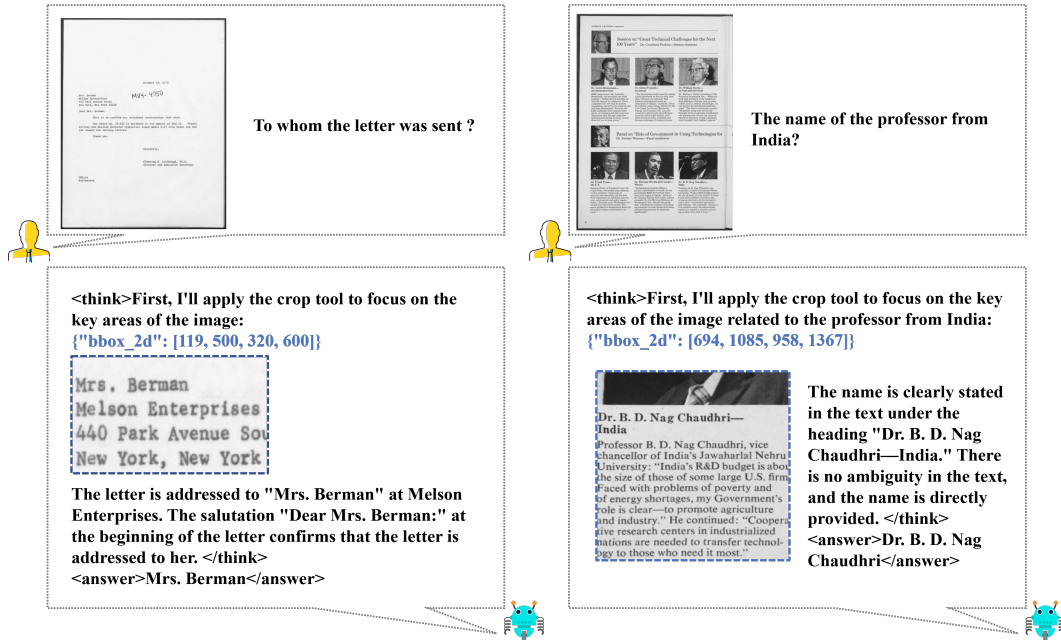


Figure 7: This figure presents how VLM-R³ gets the text details from the document image.

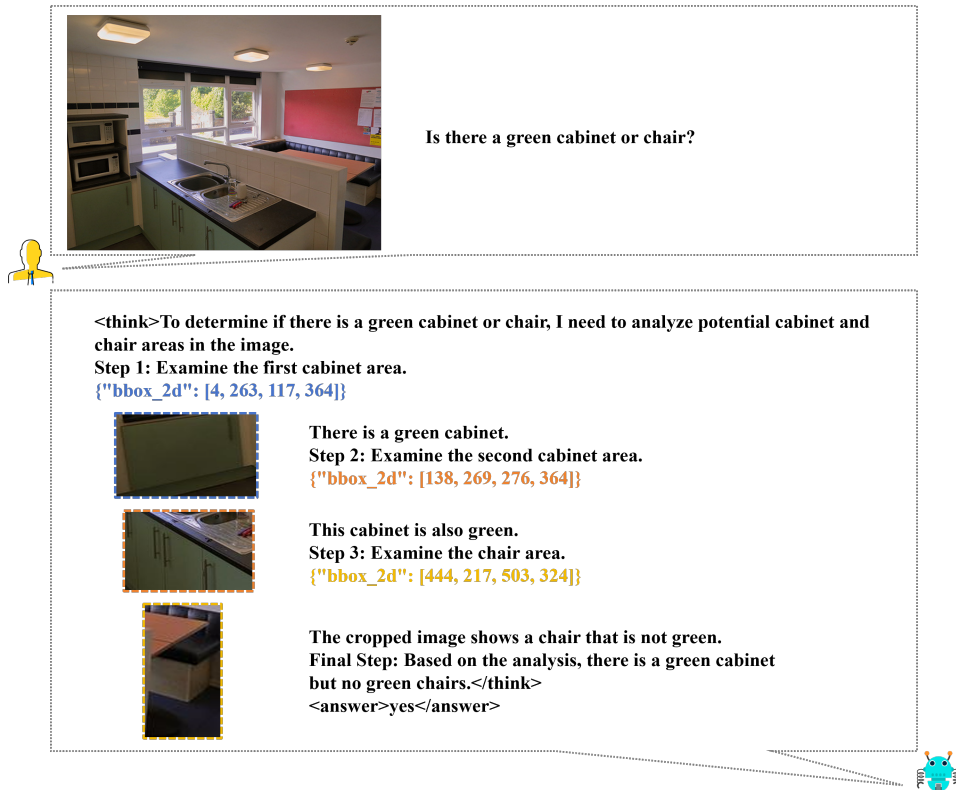


Figure 8: This figure presents how VLM-R³ iteratively tracks multiple visual cues in an image.

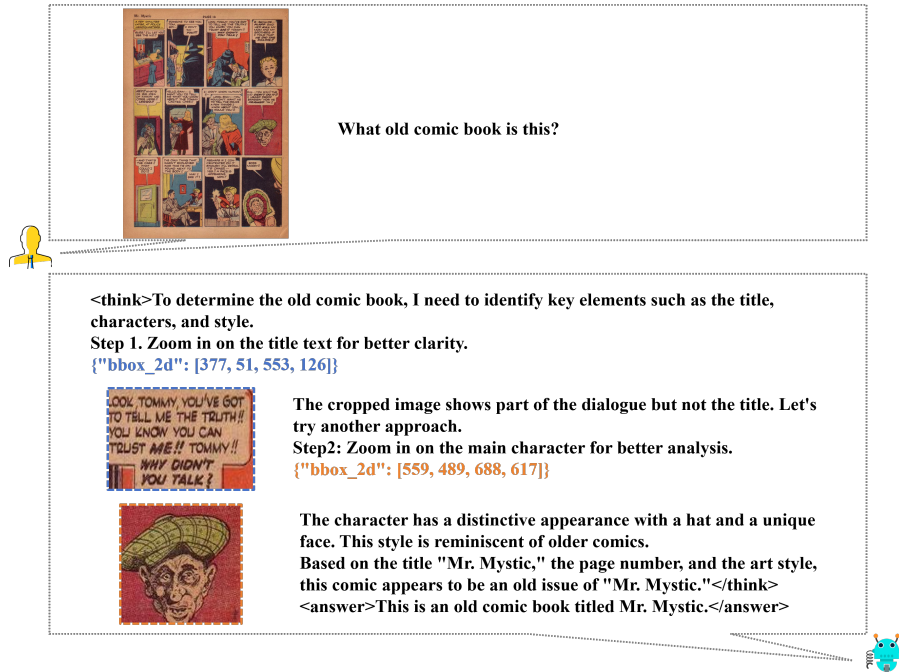


Figure 9: This figure presents how VLM-R³ performs complex interleaved text-image CoT reasoning.

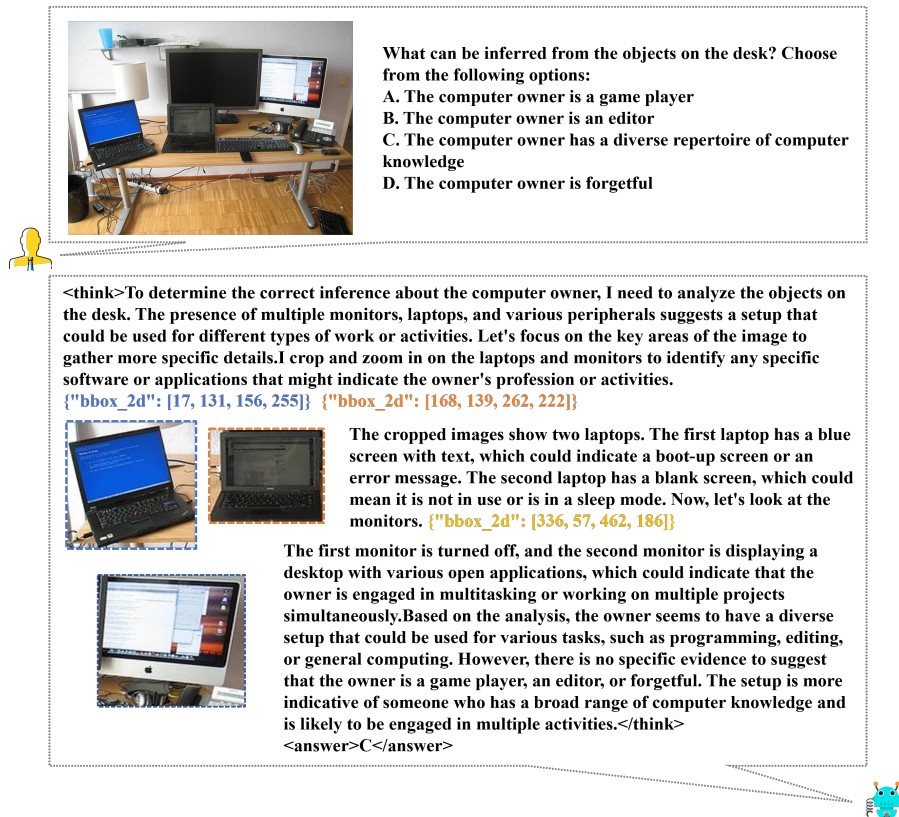


Figure 10: This figure presents how VLM-R³ performs complex interleaved text-image CoT reasoning.