VisReason: A Large-Scale Dataset for Visual Chain-of-Thought Reasoning

Lingxiao Li[2], Yifan Wang[1], Xinyan Gao[3], Chen Tang[3], Xiangyu Yue[3], and Chenyu You[1]*

[1]Stony Brook University [2]Boston University [3]MMLab, CUHK

https://TODO-project-page-url

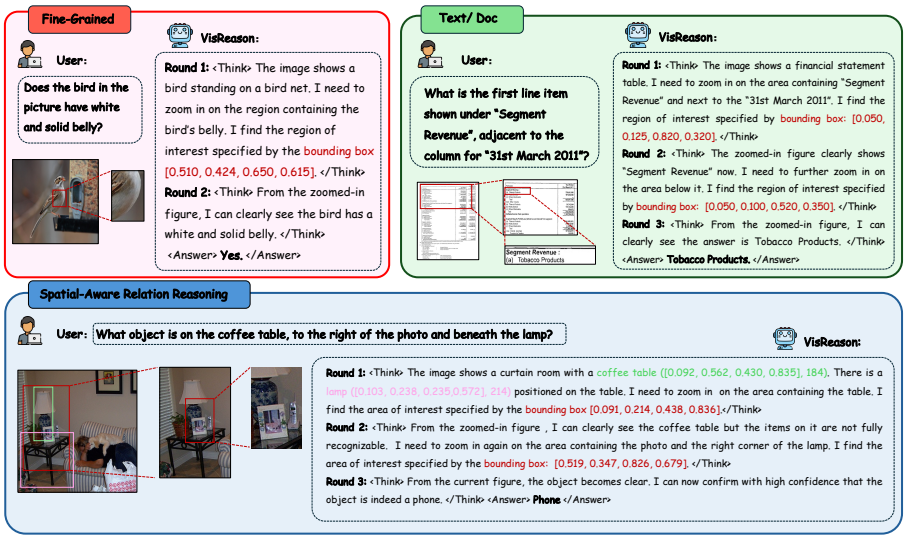

Fig. 1: An MLLM fine-tuned on VisReason emulates a *human-like visual reasoning process* to solve a complex query. Rather than processing the entire image uniformly, the model adopts a **dynamic global-to-local workflow**: it first assesses the overall scene, then progressively focuses on salient regions to collect fine-grained visual evidence. This multi-step, spatially grounded *visual Chain-of-Thought* allows the model to anchor its reasoning in localized visual cues, supporting fine-grained recognition and spatial reasoning tasks that are difficult to solve from a single global view. *(Zoom in for better visibility.)*

Abstract. Chain-of-Thought (CoT) prompting has proven remarkably effective for eliciting complex reasoning in large language models (LLMs). Yet, its potential in multimodal large language models (MLLMs) remains largely untapped, hindered by the absence of large-scale datasets that capture the rich, spatially grounded reasoning intrinsic to visual understanding. Existing visual-CoT resources are typically small, domain-specific, or lack the structured stepwise supervision necessary for compositional visual reasoning. In this paper, we introduce VisReason, a

* Corresponding author.

large-scale dataset designed to advance visual Chain-of-Thought reasoning. VisReason comprises 489K annotated examples spanning four diverse domains, each featuring multi-round, RoI-grounded rationales that guide MLLMs through interpretable visual reasoning steps. Building upon this, we curate VisReason-Pro, a 165K subset produced with a stronger GPT annotator, enriched with detailed reasoning traces and depth-augmented spatial annotations derived from monocular depth and segmentation cues. Fine-tuning strong MLLM backbones on VisReason and VisReason-Pro yields substantial improvements in step-by-step visual reasoning accuracy, RoI localization, interpretability, and fine-grained/spatial reasoning performance. These results demonstrate that VisReason equips MLLMs with more systematic and verifiable visual reasoning capabilities. We envision VisReason as a cornerstone for cultivating human-like visual reasoning, paving the way toward the next generation of multimodal intelligence.

1 Introduction

Recent multimodal large language models (MLLMs) have achieved remarkable progress in practical visual understanding, largely by pairing high-capacity language models with powerful visual encoders through sophisticated alignment pipelines [2, 39, 64, 73]. Foundational models such as LLaVA [31], InternVL [7], Qwen2.5-VL [60], and MiniCPM-V [62], have demonstrated state-of-the-art performance across a wide range of tasks, including visual question answering [24], fine-grained grounding [41], and optical character recognition [69]. These systems have rapidly evolved into versatile visual reasoning assistants for knowledge access, creative work, and embodied interaction.

However, despite architectural advances, the reasoning paradigm of most MLLMs remains rudimentary. In text-only LLMs, Chain-of-Thought (CoT) prompting has revolutionized complex reasoning by training models to articulate explicit intermediate rationales before producing an answer [55]. This stepwise supervision has yielded dramatic improvements in arithmetic, commonsense, and symbolic reasoning. In contrast, the multi-modal domain has yet to experience an equivalent shift [34, 36, 45]. Current MLLMs are predominantly optimized in a direct input-to-answer manner, offering no guidance on intermediate cognitive steps. Although recent work has begun exploring multi-round visual CoTs [43, 48], closely aligned with the broader "thinking-with-images" paradigm [40,49], these attempts still lack the scale, domain breadth, and spatial grounding necessary to model global-to-local visual reasoning. These deficiencies encourage shortcut learning, over-reliance on linguistic priors, and vulnerability to hallucination when confronted with multi-step visual queries [21].

We argue that this limitation arises from a *fundamental misalignment* between prevailing training paradigms and the structure of human visual cognition. Humans solve visually grounded problems through a global-to-local reasoning process: first forming a holistic hypothesis about the scene, then iteratively narrowing focus to inspect salient regions and refine local evidence. This involves

targeted manipulations such as zooming or cropping to probe inter-object relations and depth cues until the task is resolved [43]. While modern VLMs have acquired basic perceptual skills like grounding and OCR during pretraining, they lack supervision that captures this human-like reasoning workflow. The root cause is the inadequacy of current visual reasoning datasets, which suffer from three key deficiencies: ❶ limited scale and domain diversity, constraining the ability to learn generalizable reasoning patterns [36, 44, 68]; ❷ insufficient multi-round supervision, often collapsing reasoning into a single-step QA pair without explicit rationales [28,43,63]; and ❸ predominantly 2D annotations that provide limited supervision for relative-depth and occlusion-aware spatial reasoning [44, 45, 58]. This lack of high-quality, process-level data has become a central obstacle to developing MLLMs capable of explicit, grounded visual reasoning rather than shortcut-based perception.

To address these challenges, we introduce VisReason, a large-scale dataset explicitly designed to instill human-like, spatially grounded reasoning in MLLMs. VisReason consists of 489K annotated examples across four domains, complemented by VisReason-Pro, a 165K-example expert-curated subset with richer rationales and depth-informed supervision. To build these, our unified construction pipeline begins by enriching each image with pseudo-depth cues derived from monocular depth estimation and semantic segmentation. Using these depth-augmented scene representations, we prompt advanced MLLMs to generate multi-round visual CoT traces that follow global-to-local reasoning workflows. For VisReason-Pro, we employ stronger GPT-4.1-series guidance to produce detailed rationales, ensuring higher semantic fidelity and reasoning quality. Each step provides a concise scene summary, identifies a region of interest, and concludes with an explanatory rationale. This fine-grained supervision discourages shortcut learning, promotes global-to-local "zoom-and-verify" behavior, and improves fine-grained and spatial reasoning. Furthermore, we incorporate ordinal-depth cues – most prominently in VisReason-Pro – to encourage spatial reasoning from single-view depth estimates without claiming metric 3D reconstruction. Building upon this resource, we establish a held-out evaluation suite to assess models' stepwise reasoning, RoI localization, and depth-augmented spatial reasoning abilities, offering a unified platform for advancing visual CoT research.

Our main contributions are as follows:

- We construct and release VisReason, a large-scale dataset of 489K examples across four diverse domains, together with its 165K high-quality subset VisReason-Pro produced under stronger GPT-4.1-series guidance. The dataset provides multi-round, RoI-grounded step-by-step supervision enriched with depth-informed annotations for relative-depth and spatial reasoning.
- We establish a held-out evaluation suite based on VisReason-Pro, designed to evaluate models' fine-grained reasoning, RoI localization, and depth-augmented spatial reasoning through detailed, multi-step visual CoT tasks.

– We design a unified training and inference pipeline that leverages these annotations to instill spatially grounded CoT reasoning in MLLMs. Experiments across strong MLLM backbones demonstrate improved fine-grained recognition, spatial relation reasoning, RoI grounding, interpretability, and transfer beyond a single base model.

2 Related Works

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) have become a central direction in vision-language research [30, 65]. Most modern systems couple a high-capacity visual encoder, such as ViT [11], with a projection module, e.g., an MLP or Q-Former [25], to align visual representations with the language embedding space of an LLM for autoregressive decoding [60]. Recent model families, including LLaVA-OneVision [24], InternVL [7], Qwen-VL [2], LLaVA-UHD [59], InternVL-3 [74], Qwen2.5-VL [3], GPT-4 [39], and Gemini-2.5-Pro [9], have substantially advanced multimodal reasoning and high-resolution perception. Despite this progress, current MLLMs still struggle when answer-critical evidence is small, visually subtle, or embedded in complex spatial relations. A uniform high-resolution encoding pipeline often allocates computation to irrelevant regions and may amplify linguistic priors rather than support deliberate visual inspection. This motivates adaptive reasoning workflows that decide where to look, when to zoom, and how to verify local evidence, which is the focus of this work.

2.2 Multimodal Reasoning

Chain-of-Thought (CoT) prompting has improved the interpretability and reliability of LLM reasoning [15,55,67,71], inspiring analogous efforts in MLLMs. Existing methods can be broadly grouped into model-oriented and dataset-oriented approaches. Model-oriented methods elicit visual reasoning through in-context, tool-based, or compositional prompting [8, 14, 16, 38, 50, 70], or improve reasoning behavior through supervised fine-tuning and reinforcement learning [32, 33, 51, 61, 66, 72]. These approaches encourage intermediate rationales, but typically provide limited explicit supervision over stepwise visual attention.

Dataset-oriented efforts instead supervise the reasoning process more directly. However, many prior datasets reduce visual reasoning to a single localization or attention step, such as $V^\star$, VPD, Visual CoT, DualFocus, and Chain-of-Spot [4, 18, 34, 45, 57]. This simplified supervision can leave multi-step reasoning under-specified and encourages shortcuts when the task requires iterative evidence gathering. More recent work explores multi-round, manipulation-based, or interleaved visual reasoning, including CogCoM and VoCoT [27, 43], visually grounded region replay in VGR [54], interleaved text-image traces in Zebra-CoT [23], and code-driven visual thoughts in CodePlot-CoT [12]. Related efforts

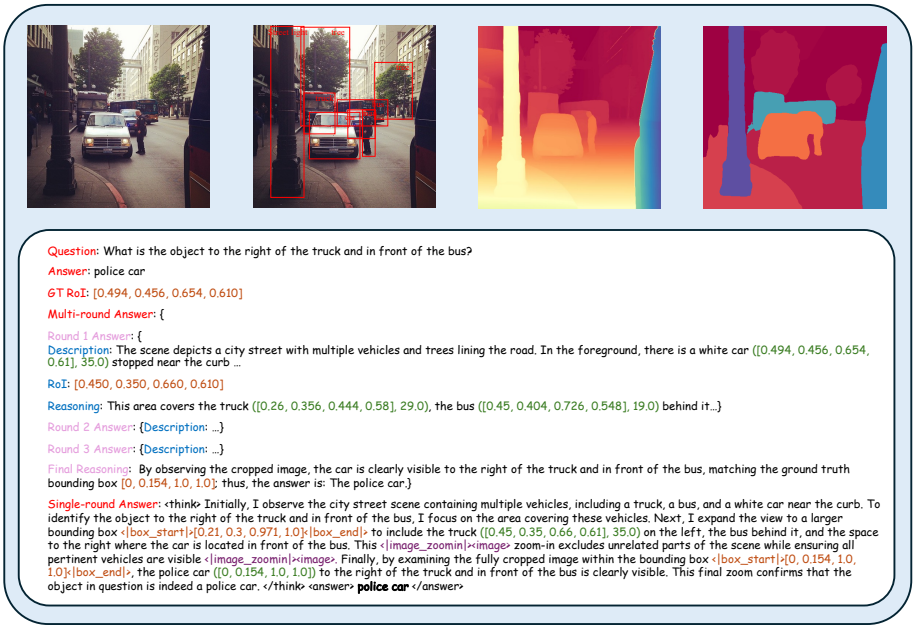

Fig. 2: For each image-question pair, we provide a region of interest (bounding box) and a compact multi-round visual chain-of-thought: each round offers a scene sketch, an optional zoom to a predicted RoI, and a brief rationale. When available, depth cues indicate ordinal ordering. The annotations are concise and process-oriented, enabling spatially grounded reasoning on fine details and complex relations.

further study adaptive visual search, grounded long-chain reasoning, and pixel-space reasoning [10, 36, 48, 68], aligning with the broader "thinking with images" paradigm.

Nevertheless, existing dataset-oriented resources still face a trade-off among scale, domain coverage, explicit multi-round RoI supervision, and spatial grounding. In particular, most visual-CoT datasets remain primarily 2D and provide limited supervision for ordinal-depth or occlusion-aware spatial relations [44, 45, 58]. In contrast, VisReason provides a large-scale, multi-domain benchmark with compact global-to-local reasoning traces, RoI-grounded supervision, and depth-informed annotations, enabling MLLMs to learn more faithful fine-grained and spatial reasoning workflows.

3 VisReason

As detailed in Sec. 1, existing visual reasoning datasets suffer from three persistent limitations – insufficient scale and domain coverage, lack of multi-round stepwise supervision, and limited supervision for depth-augmented spatial reasoning – necessitating a resource that trains MLLMs to follow global-to-local

visual reasoning. We address these gaps by curating **VisReason** – a large, spatially aware visual chain-of-thought (CoT) corpus that explicitly supervises the *process* of visual reasoning rather than only final answers. As illustrated in Fig. 2, each sample consists of a question, an answer, and a *multi-round* CoT that mirrors global-to-local problem solving: every round provides (i) a brief *scene description*, (ii) a predicted *region of interest* (RoI, via bounding box) when zoom is warranted, and (iii) a detailed *rationale* explaining why that RoI suffices. Importantly, zoom-in operations are triggered only when the target object or text is small or visually subtle, allowing the model to retain its efficiency and reliability on simpler cases while engaging in fine-grained reasoning when necessary. Beyond 2D cues, we attach pseudo-depth and segmentation signals – monocular depth and semantic segmentation – so that the chain can reference ordinal depth and part/region evidence when needed. This unified annotation format encourages models to localize, zoom, and verify iteratively, reducing shortcut learning and promoting depth-augmented spatial reasoning.

To ensure broad coverage while keeping the focus on process supervision, VisReason spans *four domains* – text/doc understanding, fine-grained recognition, general VQA, and spatial-aware relational reasoning – continuing and extending prior category choices (Tab. 1). In total, the primary set contains **489k** examples, and we further release a **165k** high-fidelity subset, VisReason-Pro, with richer rationales and stronger depth-informed grounding. Together, these resources offer detailed, stepwise supervision (see Fig. 2) designed to cultivate global-to-local "zoom–and–verify" behaviors and robust reasoning over small objects, 2D spatial relations, and ordinal-depth relations.

3.1 Dataset Generation

VisReason. Building on the Visual-CoT seed [45], we expand each image–question–answer triple with *process-level* supervision. For every example, the model (GPT-4.1-Nano [1]) is prompted to produce a concise scene description, a normalized region of interest (RoI; $[x_1, y_1, x_2, y_2] \in [0, 1]^4$), and a brief rationale. We enforce coverage by adjusting the RoI to tightly contain the ground-truth box and iteratively refining via global-to-local zoom. The refinement terminates when the RoI area is no more than twice the GT area or when a small round budget (≤ 3) is reached. For easier cases where the target object occupies a sufficiently large portion of the image (*e.g.*, $> 30\%$ of the area), we skip iterative cropping and instead provide a single detailed reasoning step followed directly by the final answer. This yields multi-round chains that are compact yet faithful, providing stepwise evidence aligned with the final answer while discouraging shortcut learning.

VisReason-Pro. VisReason-Pro. As illustrated in Fig. 3, VisReason-Pro augments the above pipeline with explicit spatial priors to elicit depth-augmented reasoning. This subset is constructed primarily from the GQA portion of Visual-CoT, which provides richer annotations (e.g., bounding boxes, relations). We first derive pseudo-depth and segmentation cues per image – monocular depth and semantic segmentation (object IDs, categories, pixel boxes, and

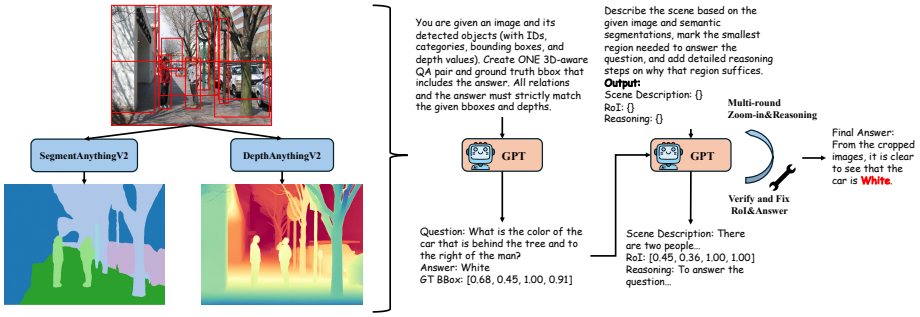

Fig. 3: Pipeline for VisReason and VisReason-Pro data generation and supervision. Given an input image, we derive semantic segments and monocular depth to form an object list with categories, bounding boxes, and ordinal depth; a generator then produces a depth-augmented spatial QA pair and target box. A second stage emits a compact, multi-round visual CoT – scene sketch, predicted RoI, and rationale – while iteratively zooming and verifying (with RoI/answer fix) until the final answer and finalized annotations are obtained.

ordinal depths) – and feed these structured signals, together with the image, to a stronger generator (GPT-4.1-Mini [1]). The model is instructed to create depth-augmented questions whose relations jointly involve 2D layout (*left of*, *above*) and ordinal depth (*in front of*, *behind*), outputting a consistent GT box for the target. We then apply the same verify-and-fix routine with multi-round zoom (round ≤ 4) to obtain concise descriptions, RoIs, and rationales at each step. In addition to these *multi-round* traces, we provide a *single-round* distilled variant that compacts the multi-step chain into one rationale and a final RoI. As shown in *Single-round Answer* in Fig. 2, our VisReason-Pro also enables single-pass answering while preserving explicit process supervision. The result is a depth-informed visual CoT corpus tailored for small-object queries, 2D spatial relations, and ordinal-depth reasoning. More details of the prompt design and algorithms are provided in the Supplementary Material (SM).

3.2 Dataset Analysis

We visualize corpus statistics in Fig. 4 and summarize coverage in Tab. 1. RoIs are strongly biased toward small regions – especially in text/doc tasks – showing that answer-critical evidence often occupies only a minor part of the image ($\sim 13.2\%$ on average), reinforcing the need for models to *localize, zoom, and verify*. Most samples resolve in 2–3 rounds, with harder spatial or depth-augmented cases extending to 4, while easier ones naturally default to a single-step rationale. Notably, the *per-round response length* remains consistently substantial, indicating detailed supervision rather than brief hints and surpassing prior process-level datasets [43, 45, 68]. Overall, VisReason provides large-scale, explicit multi-round reasoning across diverse domains, complemented by a depth-aware subset

Table 1: Overview of the VisReason dataset. It spans four distinct domains and aggregates diverse source datasets, providing broad coverage of visual data styles while supplying spatial-aware, process-level supervision for robust visual chain-of-thought reasoning.

Domain	Source Dataset	Train/Val Size		GPT Model	Dataset Description
Text/Doc	TextVQA [47]	16k	526	4.1-nano	Images with text
	TextCaps [46]	32k	846	4.1-nano	Images with text
	DocVQA [37]	50k	846	4.1-nano	Doc Images
	DUDE [52]	11k	559	4.1-nano	Doc Images
	SROIE [19]	2k	685	4.1-nano	Invoice Images
Fine-Grained Understanding	Birds-200-2011 [53]	10k	491	4.1-nano	Images of birds
General VQA	Flickr30k [42]	126k	1455	4.1-nano	Images
	Visual7W [75]	30k	994	4.1-nano	Images
Spatial Relation Reasoning	VSR [29]	3k	404	4.1-nano	Images
	GQA [20] **(Pro)**	165k	978	4.1-mini	Images **(with spatial-aware detailed reasoning steps)**
	Open Images [22]	43k	944	4.1-nano	Images

(VisReason-Pro) that strengthens grounded 2D and ordinal-depth spatial reasoning.

4 Enhancing MLLMs with CoT Reasoning Capabilities

Formulation and Training. Given an image I and textual query Q, our model generates a multi-step reasoning process $Y = (a_0, a_1, \ldots, a_T)$ to derive the final answer (Fig. 5). At step t, the action $a_t = (r_t, b_{t+1})$ consists of a textual reasoning snippet r_t and a bounding box b_{t+1} for the next region of interest. Generation is conditioned on prior actions and their visual inputs: the visual context at step t is obtained by cropping I with b_t from the previous step, and we denote features by $\mathcal{V}(\text{crop}(I, b_t))$. The process is initialized with b_0 as the full image. The model auto-regressively outputs the tokens of each a_t – both the rationale and the serialized box coordinates – based on the initial query and the full history of preceding visual and textual data:

$$a_t \sim P_\theta(\cdot|Q, a_0, \ldots, a_{t-1}, \mathcal{V}(\text{crop}(I, b_0)), \ldots, \mathcal{V}(\text{crop}(I, b_t))). \tag{1}$$

We fine-tune this model on VisReason via Supervised Fine-Tuning (SFT) using Qwen2.5-VL-7B [3] as the base. During fine-tuning, we apply LoRA [17] for efficient adaptation. The objective maximizes the likelihood of the ground-truth sequence Y given (I, Q) in a standard autoregressive manner, predicting the next token at each step. Concretely, we minimize the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{(I,Q)\in\mathcal{D}} \sum_{t=1}^{|Y|} \log P_\theta(Y_t|I, Q, Y_{<t}), \tag{2}$$

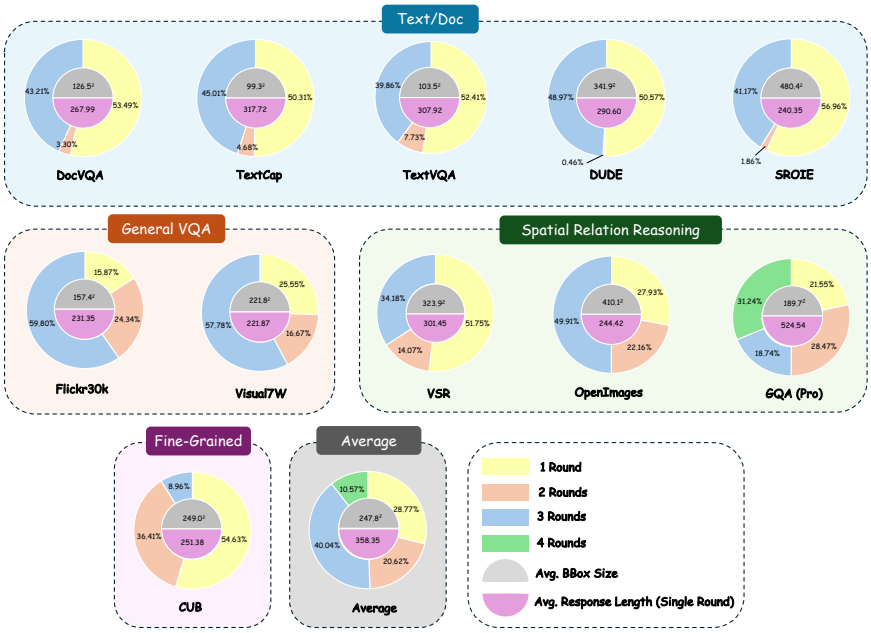

Fig. 4: Statistics of the proposed VisReason dataset. We report the distribution of CoT rounds (1–4), the average bounding-box size, and the average response length per round for each source dataset, showing that VisReason offers rich multi-round supervision and consistently long, detailed reasoning steps across diverse domains.

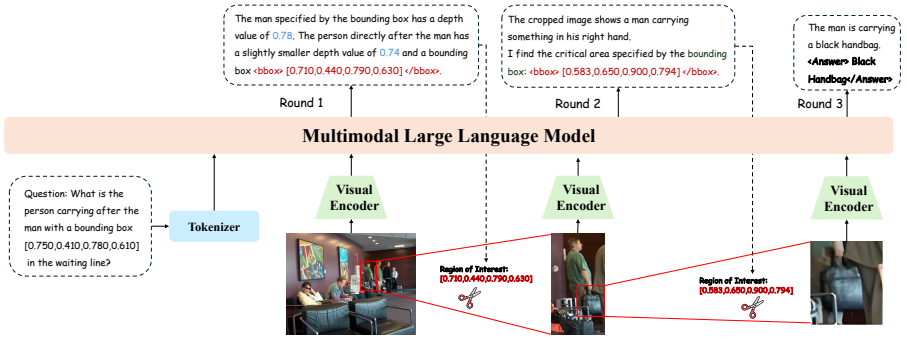

Fig. 5: Overview of VisReason paradigm. The model iteratively processes the query by first generating a textual rationale and a bounding box for the next region of interest. It then crops the original image to this region, extracts new visual features, and appends them to the context to inform the next reasoning step, creating a zoom-and-verify sequence.

where $\mathcal{D}$ is the training set, θ the trainable parameters, and Y_t the t-th token of Y. The sequence Y is formed by serializing the multi-step CoT output, convert-

ing each b_i into discrete tokens, so the model is trained end-to-end to produce both textual reasoning steps and serialized RoI coordinates for focusing visual attention.

5 Experiment

Training Details. We fine-tune Qwen2.5-VL-7B [3] on VisReason and VisReason-Pro. The baseline model is first trained for 2 epochs on VisReason excluding VisReason-Pro, followed by an additional epoch on the full combined datasets. We adopt a learning rate of 2×10^{-5} for both the LLM backbone and the projector, and freeze the ViT encoder. To test backbone transferability, we additionally fine-tune InternVL-2.5-8B [74] on VisReason for one epoch. Additional training configurations and ablations are provided in the Supplementary Materials.

Benchmarks. We follow the Visual-CoT evaluation protocol [45] and benchmark our models on its 11 source test sets (Fig. 4). As motivated in Sec. 1, we group tasks into four domains: text/document understanding, fine-grained recognition, general VQA, and spatial relational reasoning – to capture both perceptual competence and multi-step spatial inference. To assess behavior beyond this evaluation suite, we additionally report results on MME [13] and V*Bench [56], two widely used LVLM benchmarks with short-answer or multiple-choice evaluation protocols. For automatic evaluation, we follow prior MLLM work [26, 35, 45] and use GPT-4o-mini as an LLM-based judge to assign a scalar score in $[0, 1]$ for each example. We run the judge five times and report averaged scores. The judge prompt, decoding setting, parser, scoring script, and human-calibration results are provided in the Supplementary Materials.

5.1 Comparison with State-of-the-art MLLMs

Comparison on Visual-CoT-Benchmark. As shown in Tab. 2, our models achieve the strongest overall performance among the compared open-source MLLMs. VisReason-Pro–7B obtains the highest average score (0.807), improving over its Qwen2.5-VL-7B backbone (0.770) and outperforming InternVL-2.5-8B, LLaVA-NeXT-8B, and VisCoT-7B. The gains are most pronounced on fine-grained recognition (Birds-200-2011: 0.681 $\rightarrow$ 0.831) and spatial relation reasoning (OpenImages: 0.498 $\rightarrow$ 0.805), which are the settings most aligned with our RoI-grounded, global-to-local supervision.

Error Analysis on Doc/Text Tasks. We observe a drop on Doc/Text benchmarks under the default multi-round CoT inference compared to the Qwen2.5-VL-7B base model. As shown in Tab. 3, this drop is largely format-dependent: when the same VisReason-Pro checkpoint is evaluated with a matched direct-QA prompt, Text/Doc performance nearly recovers to the Qwen baseline. This suggests that the degradation is not an intrinsic loss of text/document understanding ability, but mainly arises from evaluating verbose RoI-grounded reasoning traces under short-answer QA protocols. At the same time, multi-round

Table 2: Comparison with state-of-the-art MLLMs on Visual CoT benchmark. **Bold** indicates the best results and underline indicates the second best results.

MLLM	Doc/Text					Fine-grained
	DocVQA	TextCaps	TextVQA	DUDE	SROIE	Birds-200-2011
MiniGPTv2 [5]	0.118	0.378	0.360	0.134	0.01	0.678
VisCoT-7B [45]	0.476	0.675	0.775	0.386	0.470	0.559
LLaVA-NeXT-8B [6]	0.728	0.775	0.850	0.581	0.666	0.715
InternVL-2.5-8B [74]	0.846	0.829	0.907	0.716	0.907	0.747
CoF-SFT-7B [68]	0.955	0.867	0.934	0.813	0.979	0.641
Qwen2.5-VL-7B [3]	**0.964**	**0.871**	**0.952**	**0.817**	**0.987**	0.681
VisReason-7B	0.926	0.835	0.905	0.778	0.949	0.792
VisReason-Pro-7B	0.928	0.847	0.922	0.786	0.966	**0.831**

MLLM	General VQA		Spatial Relation Reasoning			Average
	Flickr30k	Visual7W	GQA	Open Images	VSR	
MiniGPTv2 [5]	0.563	0.635	0.656	0.615	0.626	0.466
VisCoT-7B [45]	0.668	0.558	0.631	**0.822**	0.614	0.580
LLaVA-NeXT-8B [6]	0.755	0.703	**0.736**	0.559	0.647	0.705
InternVL-2.5-8B [74]	0.713	0.681	0.689	0.502	**0.737**	0.738
CoF-SFT-7B [68]	0.606	0.686	0.674	0.503	0.657	0.748
Qwen2.5-VL-7B [3]	0.772	0.690	0.651	0.498	0.705	0.770
VisReason-7B	0.766	0.695	0.619	0.798	0.654	0.791
VisReason-Pro-7B	**0.777**	**0.698**	0.670	0.805	0.654	**0.807**

CoT remains preferable for fine-grained and spatial reasoning, where localized visual inspection is central to the task.

Evaluation on the VisReason-Pro Benchmark. The VisReason-Pro benchmark evaluates two key abilities: (i) accurate RoI localization and (ii) explicit 3D grounding (grounded ratio, average IoU, and depth error). To ensure fair comparison and mitigate evaluation bias, all baselines are explicitly prompted to output bounding boxes using their native grounding templates. As shown in Tab. 4 and Tab. 5, our models achieve the best RoI localization and stronger alignment with ordinal-depth cues. These results support the effectiveness of depth-informed process supervision, while we emphasize that the depth signal is pseudo-depth supervision rather than metric 3D ground truth.

Generalization and the Evaluation Format Mismatch. External benchmarks such as MME and V* rely on short-answer or multiple-choice parsing, whereas our default inference uses verbose multi-round CoT with explicit RoI coordinates. To separate output-format mismatch from capability regression, we evaluate the same checkpoints under a matched direct-QA protocol without retraining. As shown in Tab. 7, direct-QA inference largely recovers MME and V* performance, while multi-round CoT remains stronger on the Visual-CoT benchmark. This indicates that the external drop mainly comes from protocol

Table 3: Effect of inference format on the Visual-CoT evaluation suite. We evaluate the same VisReason-Pro-7B checkpoint with either the default multi-round visual-CoT format or a matched direct-QA prompt following the vanilla Qwen answer format. Direct-QA largely recovers Text/Doc performance, while multi-round CoT remains stronger on fine-grained and spatial reasoning. Full per-dataset results are provided in the Supplementary Materials.

Model / Prompt	Text/Doc	Fine-grained	General VQA	Spatial Rel.	Avg.
Qwen2.5-VL-7B, direct QA	0.918	0.681	0.731	0.618	0.770
VisReason-Pro-7B, multi-round CoT	0.890	**0.831**	**0.738**	**0.710**	**0.807**
VisReason-Pro-7B, direct QA	0.914	0.709	0.729	0.621	0.782

Table 4: Detection performance on the VisReason-Pro held-out suite. The ground-truth bounding boxes used for computing the metric are the final-round CoT region-of-interest bounding boxes annotated in VisReason-Pro.

Accuracy	MiniGPTv2	LLaVA-Next	InternVL-2.5	VisReason	VisReason-Pro
IoU@0.5 ↑	0.14	0.29	0.08	0.27	**0.34**
IoU@0.75 ↑	0.06	0.19	0.03	0.13	**0.23**

mismatch, and that VisReason should be viewed as a specialization for RoI-grounded multi-step visual reasoning rather than a replacement for direct-QA inference in all settings.

Second-backbone Transfer. As shown in Tab. 6, fine-tuning InternVL-2.5-8B on VisReason for one epoch produces a similar specialization pattern: fine-grained and spatial reasoning improve, while Doc/Text performance is less favorable under the multi-round format. This suggests that the effect of VisReason is not specific to the Qwen2.5-VL backbone.

5.2 Ablation Study

Tab. 9 isolates the contributions of our two dataset variants and the "zoom-in when needed" strategy. Training on VisReason improves over the Qwen baseline, particularly on relation reasoning and fine-grained tasks, indicating that multi-round spatial supervision provides substantial benefit. Incorporating VisReason-Pro further strengthens general VQA, relation reasoning, and fine-grained recognition, reflecting the value of higher-fidelity rationales and depth-aware cues. Adding the adaptive zoom-in mechanism yields the best overall performance, boosting fine-grained and relation-heavy benchmarks while largely preserving Doc/Text accuracy. These results confirm that both richer supervision and selective zooming contribute meaningfully to stronger and more balanced visual reasoning.

Table 5: Ordinal-depth and RoI grounding performance on the VisReason-Pro held-out suite. Depth error is computed from monocular-depth-derived ordinal cues and does not indicate metric 3D reconstruction accuracy.

	LLaVA-Next	InternVL-2.5	Qwen-VL-2.5	**VisReason-Pro**
Grounded Ratio ↑	0.039	0.011	0.035	**0.276**
BBox (IoU) ↑	0.207	0.214	0.115	**0.278**
Depth (Abs Diff) ↓	0.394	0.290	0.294	**0.266**

Table 6: Second-backbone transfer.

Model	**Text/Doc**	**Fine-grained**	**General VQA**	**Spatial Rel.**	**Overall**
InternVL-2.5-8B	**0.842**	0.747	**0.700**	0.621	0.738
InternVL-2.5-8B + VisReason	0.815	**0.823**	0.684	**0.669**	**0.740**

5.3 User Study

We conducted a blinded study with 30 raters on 20 sampled items per method, evaluating Answer Accuracy (AA), Grounded Faithfulness (GF), and Stepwise Clarity & Sufficiency (SCS). As shown in Tab. 8, models trained on VisReason achieve clear gains in GF and SCS – reflecting tighter RoIs and more coherent global-to-local chains – which also boosts AA. VisReason-Pro further strengthens all three criteria, with raters noting more reliable grounding and more complete reasoning steps. These results confirm that multi-round, depth-aware supervision substantially improves the quality and faithfulness of visual reasoning.

5.4 Visualization

Fig. 6 illustrates how our model performs visual CoT reasoning by progressively localizing critical regions and integrating information from both the original and zoomed-in views. We further compare against VisCoT with and without CoT supervision, showing that accurate RoI prediction and coherent stepwise refinement lead to more reliable grounded reasoning and answer quality.

5.5 Discussion and Limitations

Data Quality, Ordinal-depth Noise, & Human Validation. We conduct a stratified blind human audit on 2,200 examples, sampling 200 examples from each source dataset and aggregating the results into four domains. The audit shows high answer and RoI quality overall: 99.1% answer consistency, 98.5% target containment, and 95.0% RoI tightness, with 86.5% rationale necessity and 86.5% rationale faithfulness. The main weakness appears in Text/Doc rationales, where local crops can miss global layout cues required for document reasoning. Regarding depth supervision, monocular depth provides useful ordinal cues but remains noisy and should not be interpreted as metric 3D ground

Table 7: Effect of inference protocol on external benchmarks. DQA = direct QA; MR = multi-round CoT; VR-Pro = VisReason-Pro.

Model	Prot.	V-CoT	MME	V*
Qwen2.5	DQA	0.770	**0.861**	**0.791**
VR-Pro	MR	**0.807**	0.777	0.603
VR-Pro	DQA	0.769	0.856	**0.791**
InternVL	DQA	0.738	0.848	0.597
InternVL+VR	MR	0.774	0.765	0.539
InternVL+VR	DQA	–	0.843	0.598

Table 8: Human evaluation (1–5). AA = Answer Accuracy; GF = Grounded Faithfulness; SCS = Stepwise Clarity & Sufficiency.

Method	AA	GF	SCS	Mean
MiniGPTv2 [5]	2.37	1.94	1.88	2.06
VisCoT-7B [45]	2.83	2.58	2.34	2.58
LLaVA-NeXT-8B [6]	3.52	2.93	3.08	3.18
InternVL-2.5-8B [74]	3.87	3.32	3.24	3.48
VisReason-7B	4.07	4.18	4.12	4.12
VisReason-Pro-7B	**4.19**	**4.46**	**4.37**	**4.34**

Table 9: Ablation study on dataset selection and Zoom-in strategy. (**Baseline** refers to VisReason, **Pro** refers to VisReason-Pro, and **AZ** refers to **A**daptive **Z**oom-In Strategy)

Baseline	Pro	AZ	Doc/ Text	General VQA	Relation Reasoning	Fine-grained	Average
			0.920	0.739	0.598	0.681	0.770
✔			0.864	0.744	0.678	0.798	0.779
✔		✔	0.882	0.738	<u>0.705</u>	0.792	<u>0.791</u>
✔	✔		0.856	**0.750**	0.693	<u>0.809</u>	0.781
✔	✔	✔	<u>0.908</u>	<u>0.745</u>	**0.729**	**0.831**	**0.807**

truth. Full per-domain audit results with Wilson 95% confidence intervals are provided in the Supplementary Materials. We also validate the GPT-4o-mini judge against human ratings and provide the prompt, parser, and scoring script for reproducibility.

Inference Cost & SFT vs. RL. Iterative zoom-and-verify increases inference latency, but this adaptive trade-off allocates more computation to visually complex queries. While Reinforcement Learning (RL) could in principle learn cropping policies, learning them from sparse final-answer rewards is difficult and may encourage reward hacking. Large-scale SFT datasets like VisReason therefore provide useful RoI-grounded trajectory supervision for future multimodal RL or agentic training.

Cascading Localization Failures. Our dynamic *zoom-and-verify* paradigm introduces a mechanistic vulnerability: *cascading localization errors*. Since subsequent steps are conditioned on previous cropped features, an initially misaligned RoI can trap the model's perception in an irrelevant region. Unlike single-pass MLLMs that retain global context, our iterative mechanism lacks a "zoom-out" or backtracking policy to recover from early failures. Developing self-correcting rollback mechanisms remains an important direction for future work.

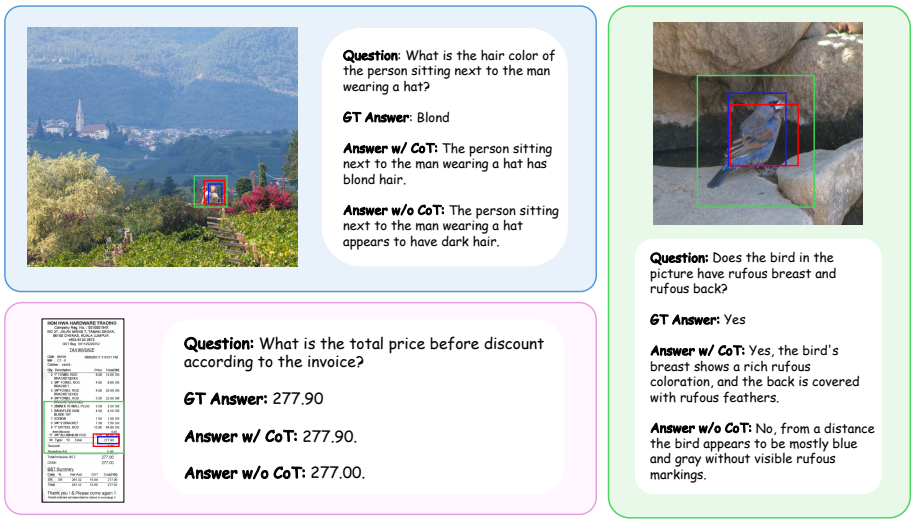

Fig. 6: Visualization results of VisReason to illustrate the difference between various inference modes. Model-generated bounding boxes are shown in green (first-round) and red (second-round), while ground truth (GT) bounding boxes are in blue. Best viewed in color and zoomed in.

6 Conclusion

In summary, we address key limitations of existing visual CoT – insufficient scale, lack of multi-round supervision, and weak depth-aware spatial reasoning – by introducing **VisReason** and **VisReason-Pro** with compact global-to-local rationales, RoI boxes, and, in VisReason-Pro, pseudo-depth and segmentation cues derived from monocular depth estimation and semantic segmentation. Spanning text/doc, fine-grained, general VQA, and spatial relational tasks, these resources provide rich process-level supervision for faithful and spatially grounded inference. Building upon them, we establish a held-out evaluation suite for assessing stepwise reasoning, RoI localization, and ordinal-depth-augmented spatial reasoning. Experiments across different MLLM backbones show consistent gains in fine-grained recognition, spatial reasoning, RoI grounding, and human-rated faithfulness. We hope this dataset family and evaluation suite offer a useful foundation for advancing spatially aware multimodal reasoning.

Acknowledgments

This work was supported in part by [Funding Agency] under Grant/Award No. [XXXX], and in part by [Funding Agency] under Grant/Award No. [XXXX]. The authors thank Mr. Yuan Qing at Boston University for his careful reading of the manuscript and constructive suggestions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Yang, A., Team, Q.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Cao, Y., Zhang, P., Dong, X., Lin, D., Wang, J.: Dualfocus: Integrating macro and micro perspectives in multi-modal large language models. arXiv preprint arXiv:2402.14767 (2024)
5. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
6. Chen, L., Xing, L.: Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. `https://github.com/xiaoachen98/Open-LLaVA-NeXT` (2024). `https://doi.org/10.5281/zenodo.13935471`
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
8. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1254–1262 (2024)
9. Comanici, G., Team, G.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. 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. In: CVPR (2025)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
12. Duan, C., Sun, K., Fang, R., Zhang, M., Feng, Y., Luo, Y., Liu, Y., Wang, K., Pei, P., Cai, X., Li, H., Ma, Y., Liu, X.: Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. arXiv preprint arXiv:2510.11718 (2025)
13. Fu, C., Yang, X., Chen, X., Sun, M., Qiu, L., Huang, Y., Li, Y., Cheng, T., Shi, J., Xiao, Z., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
14. Gao, Z., Du, Y., Zhang, X., Ma, X., Han, W., Zhu, S.C., Li, Q.: Clova: A closed-loop visual assistant with tool usage and update. In: CVPR. pp. 13258–13268 (2024)
15. Gao, Z., Zhang, B., Li, P., Ma, X., Yuan, T., Fan, Y., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In: ICLR (2025)
16. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: CVPR. pp. 14953–14962 (2023)

17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
18. 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. In: CVPR. pp. 9590–9601 (2024)
19. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019)
20. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR. pp. 6700–6709 (2019)
21. Ke, F., Hsu, J., Cai, Z., Ma, Z., Zheng, X., Wu, X., Huang, S., Wang, W., Delir Haghighi, P., Haffari, G., Krishna, R., Wu, J., Rezatofighi, H.: Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298 (2025)
22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
23. Li, A., Wang, C.L., Goldblum, M., et al.: Zebra-cot: A dataset for interleaved vision language reasoning. arXiv preprint arXiv:2507.16746 (2025)
24. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
25. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
26. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
27. Li, Z., Luo, R., Zhang, J., Qiu, M., Huang, X., Wei, Z.: Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3769–3798. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://doi.org/10.18653/v1/2025.naacl-long.192, https://aclanthology.org/2025.naacl-long.192/
28. Li, Z., Luo, R., Zhang, J., Qiu, M., Wei, Z.: Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. arXiv preprint arXiv:2405.16919 (2024)
29. Liu, F., Emerson, G.E.T., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. vol. 36 (2023), https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html
31. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al.: Llava-plus: Learning to use tools for creating multimodal agents. arXiv preprint arXiv:2311.05437 (2023)
32. Liu, Z., Zhang, Y., Liu, F., Zhang, C., Sun, Y., Wang, J.: Othink-mr1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. arXiv preprint arXiv:2503.16081 (2025)

33. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
34. Liu, Z., Dong, Y., Rao, Y., Zhou, J., Lu, J.: Chain-of-spot: Interactive reasoning improves large vision-language models. arXiv preprint arXiv:2403.12966 (2024)
35. Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207 (2023)
36. Man, Y., Huang, D.A., Liu, G., Sheng, S., Liu, S., Gui, L.Y., Kautz, J., Wang, Y.X., Yu, Z.: Argus: Vision-centric reasoning with grounded chain-of-thought. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14268–14280 (2025)
37. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV. pp. 2200–2209 (2021)
38. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: CVPR. pp. 14420–14431 (2024)
39. OpenAI: Gpt-4v(ision) system card (2023), `https://api.semanticscholar.org/CorpusID:263218031`
40. OpenAI: Thinking with images (2025)
41. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding multimodal large language models to the world. In: ICLR (2024)
42. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
43. Qi, J., Ding, M., Wang, W., Bai, Y., Lv, Q., Hong, W., Xu, B., Hou, L., Li, J., Dong, Y., et al.: Cogcom: Train large vision-language models diving into details through chain of manipulations. arXiv preprint arXiv:2402.04236 (2024)
44. Sarch, G., Saha, S., Khandelwal, N., Jain, A., Tarr, M.J., Kumar, A., Fragkiadaki, K.: Grounded reinforcement learning for visual reasoning. arXiv preprint arXiv:2505.23678 (2025)
45. 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. arXiv preprint arXiv:2403.16999 (2024)
46. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV. pp. 742–758. Springer (2020)
47. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
48. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025)
49. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., et al.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918 (2025)
50. Sun, S., De Araujo, G., Xu, J., Zhou, S., Zhang, H., Huang, Z., You, C., Xie, X.: Coma: Compositional human motion generation with multi-modal agents. arXiv preprint arXiv:2412.07320 (2024)
51. Sun, S., Wang, Y., Zhang, H., Xiong, Y., Ren, Q., Fang, R., Xie, X., You, C.: Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse

rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10386–10397 (2025)
52. Van Landeghem, J., Tito, R., Borchmann, Ł., Pietruszka, M., Joziak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Anckaert, B., Valveny, E., et al.: Document understanding dataset and evaluation (dude). In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19528–19540 (2023)
53. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
54. Wang, J., Kang, Z., Wang, H., Wu, B., et al.: Vgr: Visual grounded reasoning. arXiv preprint arXiv:2506.11991 (2025)
55. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 24824–24837 (2022)
56. Wu, P., Li, Z., Wang, W., Li, Y., Lin, F., Zhang, T., Zeng, Z., Zhang, K., Lu, L., Qiao, Y., Dai, J.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
57. Wu, P., Xie, S.: $v^\star$: Guided visual search as a core mechanism in multimodal llms. In: CVPR. pp. 13084–13094 (2024)
58. Wu, Q., Yang, X., Zhou, Y., Fang, C., Song, B., Sun, X., Ji, R.: Grounded chain-of-thought for multimodal large language models. arXiv preprint arXiv:2503.12799 (2025)
59. Xu, R., Yao, Y., Guo, Z., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Sun, M., Huang, G.: Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. arXiv preprint arXiv:2403.11703 (2024)
60. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
61. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
62. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
63. Ye, X., Gan, Y., Huang, X., Ge, Y., Shan, Y., Tang, Y.: Voco-llama: Towards vision compression with large language models. arXiv preprint arXiv:2406.12275 (2024)
64. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)
65. You, C., Dai, H., Min, Y., Sekhon, J.S., Joshi, S., Duncan, J.S.: Uncovering memorization effect in the presence of spurious correlations. Nature Communications **16**(1), 5424 (2025)
66. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025)
67. Zhang, X., Cao, J., Wei, J., Xu, Y., You, C.: Tokenization constraints in llms: A study of symbolic and arithmetic reasoning limits. arXiv preprint arXiv:2505.14178 (2025)
68. 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. arXiv preprint arXiv:2505.15436 (2025)

69. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint (2023)
70. Zhang, Y., Qian, S., Peng, B., Liu, S., Jia, J.: Prompt highlighter: Interactive control for multi-modal llms. In: CVPR. pp. 13215–13224 (2024)
71. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493 (2022)
72. Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., Hsieh, C.J.: R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132 (2025)
73. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
74. Zhu, J., Chen, Z., Wang, W., et al.: Internvl3: Exploring advanced training and test-time strategies for open multimodal models. arXiv preprint arXiv:2504.10479 (2025)
75. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: CVPR. pp. 4995–5004 (2016)