Revealing the Seen, Imagining the Beyond: A Survey of Image-Grounded Chain-of-Thought Reasoning in Multimodal LLMs

Qihua Dong, Yitian Zhang, Huimin Zeng, Yizhou Wang, Jianglin Lu, Kuo Yang, Yun Fu

Northeastern University

Abstract

Multimodal large language models (MLLMs) are making rapid strides in complex visual reasoning. This survey synthesizes the emerging paradigm of **Image-Grounded Chain-of-Thought** (IG-CoT), where models ground intermediate inferences by interleaving textual rationales with visual state updates. We formalize IG-CoT, present a method-centric taxonomy covering prompting, supervised fine-tuning, and reinforcement learning, and map these techniques to representative benchmarks. Our analysis identifies two domains where IG-CoT offers significant advantages: detail-oriented reasoning requiring meticulous perception, and imagined-world reasoning for simulating unseen states in games, geometry, and planning. We discuss the practical trade-offs of current methods regarding controllability, data, and compute. We conclude by highlighting key challenges (efficiency, data quality, and generative capabilities) and outlining promising future directions, including lightweight architectures, richer intermediate supervision, and method-aware evaluations that better assess faithfulness and long-horizon reasoning. We maintain a continuously updated paper list at https://github.com/dddraxxx/Awesome-Image-Grounded-CoT.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs), including LLaVA (Liu et al., 2023a), the GPT-4 series (OpenAI, 2024, 2025), and reinforcement-learning paradigms such as GRPO (Shao et al., 2024b), have shifted visual reasoning from tool-calling to end-to-end perception–reasoning. This has given rise to **Image-Grounded Chain-of-Thought (IG-CoT)**, where models iteratively interleave textual rationales with visually updated states so that what they "see" continuously informs what they "think" (Figure 1). We focus on such *MLLM-native* approaches, in contrast to

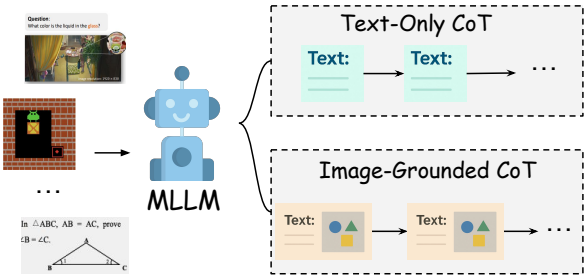

Figure 1: Image-Grounded Chain-of-Thought (IG-CoT) versus Text-Only CoT. IG-CoT enables MLLMs to interleave reasoning with visual information for complex visual tasks like games, scene understanding, abstract problem-solving, and long-video understanding.

text-centric controllers (e.g., ViperGPT (Surís et al., 2023)) that reason over extracted text or external tools.

However, as IG-CoT methods advance, they face significant challenges: (1) high **computational costs** from long, image-heavy reasoning chains; (2) a **scarcity of high-quality data** for supervised training; (3) the **limited ability** of current MLLMs to generate visual outputs during reasoning; and (4) **evaluation mismatch**: choosing appropriate tasks/data to assess IG-CoT; for example, on knowledge-centric VQA suites, image-grounded CoT may yield limited gains. While broad surveys on MLLMs exist (Jiang et al., 2024; Lin et al., 2025b; Zhou et al., 2025; Lu et al., 2025a), a dedicated, method-focused analysis is needed to help researchers navigate these specific problems. This survey fills that gap by providing a structured overview of the field. While our focus is image-grounded reasoning, we also consider temporally extended visual inputs when visual grounding occurs over frames or segments.

To address these challenges, we make three contributions. **First**, we propose a method-centric framework that organizes IG-CoT approaches into prompting, supervised fine-tuning, and reinforcement learning, clarifying trade-offs in controlla-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45055–45070

bility, data needs, and compute with an eye to *efficiency* and *data scarcity*. **Second**, we connect these methods to representative benchmarks across detail-oriented and imagined-world reasoning, offering a task-first map for evaluation with a focus on long-horizon reliability and the axes of detail grounding and imagination (see Figure 2); we also highlight when IG-CoT is likely to help, and when it is not. **Finally**, we synthesize current limitations and outline promising directions toward more efficient architectures, richer intermediate supervision, and method-aware evaluations that better assess faithfulness and long-horizon reasoning.

Roadmap. Following Figure 2, §2 formalizes IG-CoT and its scope; §3 surveys methods by training paradigm (Training-Free, SFT, RL and Hybrids) and states testable hypotheses; §4 maps methods to benchmarks across detail-oriented, imagined-world, video, and general suites; §5 outlines open challenges. The Appendix documents our selection protocol and compiles consolidated tables.

2 IG-CoT Definition and Scope

Definition. We scope Image-Grounded Chain-of-Thought (IG-CoT) to methods where an end-to-end Multimodal LLM both *sees* and *thinks* throughout the reasoning process, allowing visual content to directly inform intermediate thoughts.

Terminology. Throughout this survey, we use *visual state updates* to denote atomic operations that modify the visual context (e.g., crops, zooms, sketches, imagined frames); *visual grounding* to denote keeping each reasoning step explicitly tied to evidence in the current visual state; and *interleaved reasoning* to denote the iterative alternation between textual rationales and such visually grounded updates.

When IG-CoT helps. IG-CoT is most beneficial in two bands: (i) *detail-oriented reasoning* requiring precise perception and localization (e.g., fine-grained VQA, grounding/segmentation, hallucination checks, OCR/numeracy, long video), where zoom/crop and frame selection clarify fine details; and (ii) *imagined-world reasoning* that demands simulating unobserved states or rules (e.g., games, geometry, planning), where imagined or drafted visuals support multi-step prediction. When visual drafts are interleaved with text, models tend to be more faithful and more robust over long horizons.

2.1 Distinguishing from Text-based Visual Reasoning

Our survey contrasts *MLLM-native* IG-CoT with influential *text-centric* pipelines such as *ViperGPT* (Surís et al., 2023), *VisProg* (Gupta and Kembhavi, 2023), *IdealGPT* (You et al., 2023), and chain-of-thought prompting for knowledge-based visual reasoning (Chen et al., 2024b). Those systems follow an *LLM-as-controller* paradigm: a text-only LLM decomposes tasks into symbolic plans (e.g., Python/DSL) for vision tools to execute or reasons purely over text extracted from images. In these pipelines, core reasoning is text-based and detached from direct visual perception; images are handled by external modules, and the reasoner does not perceive during deliberation. By contrast, MLLM-native IG-CoT tightly couples perception and reasoning, interleaving visual state updates (e.g., cropping, sketching, imagining) with textual rationales.

3 Methods

We organize IG-CoT methods by how they externalize and control intermediate visual states and by the supervision required to acquire these capabilities. The decision trade-offs in Table 1 summarize controllability versus supervision/compute footprints and indicate where families tend to excel.

The methods for enabling Image-Grounded Chain-of-Thought (IG-CoT) can be broadly classified into two main categories, as indicated in Table 4 (Appendix A.3): training-free approaches and training-based approaches. This section outlines these methodologies, highlighting their core principles and common techniques.

3.1 Training-Free (Prompting)

Core Idea & Paradigm. As the most direct way to elicit advanced reasoning, prompting-based IG-CoT seeks to *unlock* the latent capabilities of powerful, pre-trained MLLMs without modifying their weights. The core insight is that by structuring the inference process as an iterative loop (plan, act with visual tools, rationalize, and refine), these models can externalize their thinking process. This approach is highly flexible and extends naturally to temporal inputs by adding a preliminary step of curating key frames or shots (e.g., TCoT/CoS (Arnab et al., 2025; Hu et al., 2025)), making it a powerful tool for rapid prototyping and exploring the upper bounds of existing models.

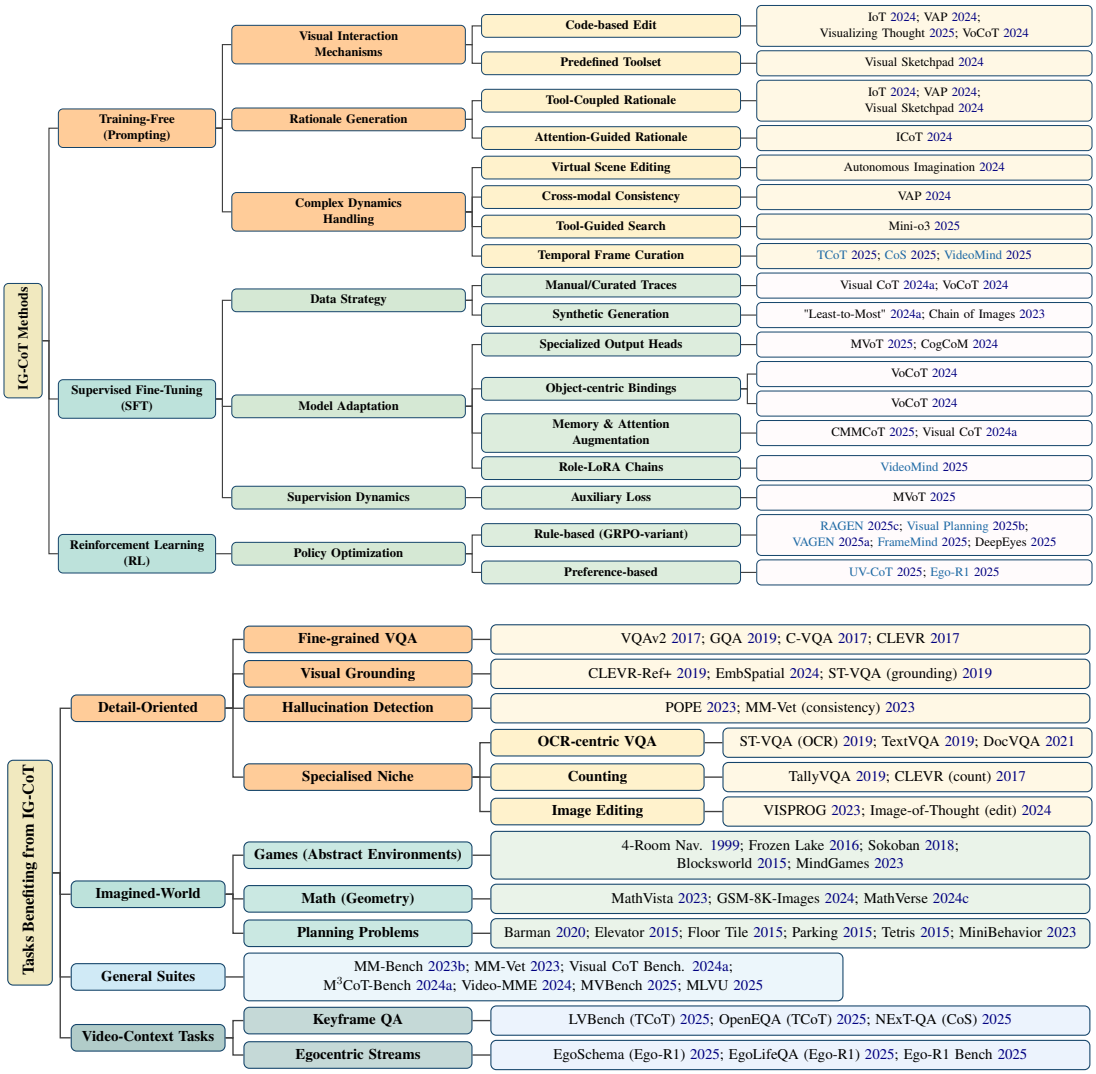

Figure 2: Taxonomies of Image-Grounded Chain-of-Thought (IG-CoT). **Top**: A method-centric view organizing approaches by training paradigm. Video-compatible methods appear in blue. **Bottom**: A task-centric view categorizing problems by their core reasoning demands.

Key Designs & Mechanisms. The design of prompting methods centers on three core components that directly mirror our taxonomy: (1) **Visual Interaction Mechanisms**, which define how the model "sees" and manipulates visual information. A key trend is the move toward richer "visual scratchpads," from code-based generation (VAP (Xiao et al., 2024)) and free-form drawing (VISUAL SKETCHPAD (Hu et al., 2024)) to lightweight attention-based cropping (ICoT (Gao et al., 2024)) and hierarchical agentic context management (HVA (Dong et al., 2026a)); (2) **Rationale Generation**, where the MLLM justifies its actions, either coupled with tool use or guided by internal attention; and (3) **Complex Dynamics Handling**, where techniques like virtual scene editing (AUTONOMOUS IMAGINATION (Liu et al., 2024)) or cross-modal consistency checks (Xiao et al., 2024) manage state changes over time.

Performance & Promise. Empirical results indicate that giving models a way to "see" their own intermediate reasoning can yield substantial gains: IMAGE-OF-THOUGHT (Zhou et al., 2024) improves GPT-4o scores on MME by over 100 points, and VISUAL SKETCHPAD (Hu et al., 2024) reports state-of-the-art results on challenging geometry and visual reasoning benchmarks. This underscores a critical insight: externalizing the reasoning process into a sequence of visual states is a powerful paradigm for improving both performance and faithfulness, as it allows the model to iteratively ground, correct, and build upon its own

Method	Type	Reasoning Length	Control Modality	Controllability	Train Data Size (k)	Compute (GPU-h)	Video Ready	Detail Bench	Imagined Bench
2024 VAP	Prompting	M	Code-generated	H	–	–	✗	–	✓
2024 Image-of-Thought	Prompting	M	Image transforms	M	–	–	✗	✓	–
2024 Visual Sketchpad	Prompting	M	Model-generated	H	–	–	✗	✓	✓
2024b Mind's Eye	Prompting	H	ASCII grids	L	–	–	✗	–	✓
2024 ICoT	Prompting	L	Attention crops	L	–	–	✗	✓	–
2024 Auto. Imagination	Prompting	H	Model-generated	H	–	–	✗	✓	✓
2025 Visualizing Thought	Prompting	H	Code-generated	M	–	–	✗	–	✓
2025a ItS-CoMT	Prompting	L	Model-generated	H	–	–	✗	✓	✓
2025b VisuoThink	Prompting	H	Model-generated	H	–	–	✗	✓	✓
2025 Temporal CoT	Prompting	H	Frame crops	M	–	–	✓	–	✓
2023 Chain of Images	SFT	M	SVG code	H	-	-	✗	–	✓
2024a Visual CoT	SFT	M	Zoom crops	M	438	≥400	✗	✓	–
2024 VoCoT	SFT	M	Token chunks	M	80	720	✗	✓	–
2024 CogCoM	SFT	M	Zoom crops	M	40,000	102,144	✗	✓	–
2024a Least-to-Most	SFT	M	Zoom/OCR	M	50	9	✗	✓	–
2024 V* Guided Search	SFT	M	Zoom/Detector	M	387	-	✗	✓	–
2025 MVoT	SFT	H	Fixed Tokens	H	17	-	✗	–	✓
2025 CMMCoT	SFT	M	Token chunks	M	260	-	✗	✓	–
2025 VideoMind	SFT	H	Frame/Zoom crops	M	-	-	✓	✓	✓
2025a VAGEN	RL	M	Sim. grids	L	-	-	✗	–	✓
2025b Visual Planning	SFT + RL	L	Model-generated	H	-	-	✗	–	✓
2025c RAGEN	RL	M	ASCII grids	L	-	-	✗	–	✓
2025 UV-CoT	RL	L	Zoom crops	M	249(DPO)	≥480	✗	✓	–
2025 FrameMind	RL	H	Retrieved clips	M	-	-	✓	–	✓
2025 Ego-R1	SFT+RL	H	Frame/Zoom crops	H	25(SFT)+4.4(RL)	-	✓	✓	✓
2025 DeepEyes	RL	L	Zoom crops	M	47(RL)	-	✗	✓	–
2025 Mini-o3	SFT+RL	H	Zoom crops	M	6(SFT)+12(RL)	-	✓	✓	–

Table 1: Trade-off comparison of representative Image-Grounded Chain-of-Thought methods. H/M/Ldenote high/medium/low reasoning length (L is 1-2 steps, M is 3-4 steps, H is 5+ steps) or controllability. Benchmark columns tick domains where the method reports gains; "–" indicates training-free methods without reported data or compute, or stats unreported. Train data size (in thousands of examples) is shown where applicable.

thoughts.

Challenges & Trade-offs. The flexibility of prompting comes at a cost, trading training-time supervision for significant inference-time compute and engineering overhead. The core tension is between *controllability* and *cost*. High-fidelity generated drafts offer maximum control but are slow and token-intensive, directly impacting *efficiency*. This "pay-at-inference" model also makes prompting susceptible to common failure modes like tool hallucination and semantic drift in long chains. The complexity is thus shifted from model training to sophisticated prompt and tool engineering, a direct response to the *data-scarcity* problem that nonetheless introduces its own challenges in scalability and robustness.

3.2 Training-Based Methods

Training-based methods involve fine-tuning or further training LMMs on specialized datasets to explicitly instill IG-CoT abilities. These methods typically fall under Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL).

3.2.1 Supervised Fine-Tuning (SFT)

Core Idea & Paradigm. Whereas prompting elicits latent skills, SFT aims to internalize them: moving IG-CoT from an emergent, prompt-dependent capability to a reliable, intrinsic behavior. By training models on high-quality demonstrations, SFT teaches them to autonomously generate visually grounded reasoning steps. This paradigm is particularly effective for creating specialized models that excel at specific, well-defined reasoning patterns, including those that unfold over time, where supervision can be augmented with frame selections or specialized modules to manage temporal context (e.g., VIDEOMIND (Liu et al., 2025)).

Key Designs & Mechanisms. Success in SFT is driven by three key factors, as outlined in our taxonomy: (1) **Data Strategy**, which is paramount. The field is exploring a spectrum from expensive, high-quality manual traces (Visual CoT (Shao et al., 2024a)) to scalable synthetic data ("Least-to-Most" (Cheng et al., 2024a)); (2) **Model Adaptation**, where architectures are augmented with specialized heads for drawing (MVoT (Li et al., 2025)), object-centric bindings (VoCoT (Li et al., 2024)), or explicit memory modules (CMMCoT (Zhang et al.,

2025)); and (3) **Supervision Dynamics**, which typically rely on a primary cross-entropy loss, sometimes augmented with auxiliary losses to guide intermediate steps.

Performance & Promise. SFT has proven highly effective, consistently delivering substantial, often double-digit, accuracy gains and pushing the state of the art. CogCoM (17B) (Qi et al., 2024) reports strong GQA (71.7%) and ST-VQA (71.1%) results as a generalist model, and VoCoT (7B) (Li et al., 2024) surpasses GPT-4V on complex visual reasoning. The key insight from these successes is that forcing models to "show their work" by supervising intermediate visual steps leads to significant improvements in both final accuracy and the faithfulness of the reasoning process. This highlights the immense value of grounded intermediate supervision.

Challenges & Trade-offs. The predominant challenge for SFT is the *data-quality bottleneck*, making it a direct but costly answer to the *data-scarcity* problem. The high cost of manual annotation and the large *compute* budget required for training create a core trade-off between annotation expense and model capability. SFT also trades the fine-grained *controllability* of prompting for more autonomous, baked-in behaviors, which can be less flexible if the training data does not cover a sufficiently diverse set of reasoning patterns. While synthetic data improves scalability, it risks teaching models superficial logic, failing to solve the core need for high-quality reasoning traces.

3.2.2 Reinforcement Learning and Hybrids

Core Idea & Paradigm. RL represents a paradigm shift from imitation to *discovery*. Instead of being taught *how* to reason, the MLLM agent learns to generate effective IG-CoT strategies by optimizing a policy to achieve a goal in a rule-based environment. This makes RL ideal for tasks where the optimal reasoning path is unknown, but success is clearly definable, such as puzzles, games, or mathematical proofs. The recent advent of **GRPO** (Shao et al., 2024b), a simple and stable gradient-free algorithm, has significantly lowered the barrier to entry and is fueling a surge of interest in RL for IG-CoT. Recent methods span pure-RL pipelines and *hybrids* (**SFT+RL**), where the model is first trained with supervised fine-tuning to bootstrap a usable policy and then further optimized with RL. We treat both together in this subsection.

Key Designs & Mechanisms. In RL, the central design challenges revolve around two components: (1) **Reward Design**, which is critical for guiding the model's exploration and can be sparse (final outcome only) or dense (rewarding intermediate progress); and (2) **Policy Optimization**, where the stability and simplicity of GRPO and its variants have made them the *de facto* standard in recent work, often augmented with retrieval-based loops for temporal tasks.

Performance & Promise. When a good reward signal is available, RL can deliver remarkable performance gains, enabling models to discover non-obvious strategies. For example, Visual Planning (Xu et al., 2025b) uses visual-only steps to reach 74.5% exact-match on small Mazes (3×3–6×6) and 91.6% on FrozenLake, substantially above text-only baselines. Furthermore, RL can be highly data-efficient in terms of human labels; UV-CoT (Zhao et al., 2025) uses fully unsupervised score-DPO and improves over a strong supervised baseline on six visual-reasoning benchmarks without any human-labeled bounding-box traces, offering a compelling answer to the *data-scarcity* challenge. This showcases RL's promise for teaching models to *plan* and *strategize* visually.

Challenges & Trade-offs. The power of RL is balanced by its significant practical challenges. The core trade-off is between *flexibility and stability*. While RL offers the highest potential for autonomous discovery, this comes at the cost of low user *controllability* and immense *compute* demands for training, which impacts its overall *efficiency*. RL methods are notoriously unstable and sample-inefficient, which has largely confined their application to "toy-scale" digital worlds with short horizons. The key open research question is how to bridge the gap from these simple, rule-based environments to complex, open-ended tasks. Until then, the promise of RL for general-purpose IG-CoT remains compelling but largely unrealized.

A detailed side-by-side comparison of eight RL and SFT+RL hybrid methods along reward design, sample efficiency and scale, training stability, and recurring failure modes is deferred to Appendix A.2 (Table 3); the high-level takeaway is captured by Hypothesis H3 in §3.4.

3.3 Cross-cutting Design Patterns

Three design patterns recur across families. First, light-weight *verification* (self-alignment checks on caption–plan conflict, visual–text consistency audits, evaluator-in-the-loop preferences) consistently

improves robustness. Second, the choice of *visual draft* (code/ASCII diagrams for higher control and lower cost versus free-drawn or intrinsic manipulations for higher fidelity but higher token/latency cost) should match task demands and budget. Third, the same patterns lift to video via frame/shot selection (TCoT/CoS (Arnab et al., 2025; Hu et al., 2025)), role-structured adapters (VIDEOMIND (Liu et al., 2025)), or agentic retrieval (FRAMEMIND/EGO-R1 (Ge et al., 2025; Tian et al., 2025)), where controlling turn budgets and retrieval granularity is key to stability and cost.

3.4 Emerging Patterns and Testable Hypotheses

Across the three families surveyed above, several cross-paper regularities recur about *when* IG-CoT helps, *how*, and *at what cost*. We extract these regularities as three falsifiable hypotheses, each tied directly to evidence already compiled in our method and benchmark tables.

H1: Visual intermediates improve accuracy and faithfulness. For detail-oriented and imagined-world tasks with strong visual signal, methods that externalize intermediate visual states (crops, sketches, imagined frames) yield higher accuracy and faithfulness than text-only chain-of-thought at a fixed backbone and data budget. Image-of-Thought (Zhou et al., 2024) reports a $>$100-point MME boost for GPT-4o by introducing explicit image-of-thought intermediates; Visual Sketchpad (Hu et al., 2024) attains 80.3% on V*Bench by iteratively sketching visual reasoning steps; Visual CoT (Shao et al., 2024a) delivers consistent gains over its SFT baseline on its dedicated benchmark by supervising bounding-box "thoughts"; and CogCoM (17B) (Qi et al., 2024) reaches 71.7% on GQA and 71.1% on ST-VQA through explicit, step-wise visual grounding.

H2: Step-wise visual verification reduces hallucination and shortcutting. Incorporating step-wise verification mechanisms (caption–plan consistency checks, step–image alignment probes, trace perturbation, evaluator-in-the-loop preferences) systematically reduces hallucination and shortcutting relative to unverified IG-CoT and text-only CoT. Methods that redraw or revise when visual and textual plans conflict (e.g., VAP (Xiao et al., 2024)) report improved robustness on POPE (Li et al., 2023) and MM-Vet (Yu et al., 2023) consistency subsets, and our evaluation recommendations in Section 5 emphasize step–image alignment scores, counterfactual sensitivity, and trace-perturbation tests as diagnostic tools.

H3: RL-based IG-CoT is powerful but currently confined to small, rule-based environments. Reinforcement-learning-based IG-CoT excels in small, rule-based environments with short horizons, but has not yet robustly transferred to long-horizon, open-world tasks due to stability and sample-efficiency limitations. Visual Planning (Xu et al., 2025b) reaches 74.5% exact-match on small 3×3–6×6 Mazes using visual-only steps, and VAGEN (Wang et al., 2025a) and RAGEN (Wang et al., 2025c) show clear gains on Sokoban and FrozenLake in small rule-based environments. All of these gains are obtained in constrained, simulator-backed settings with well-defined rewards and short episodes; in contrast, performance on broader suites such as VSP (Wu et al., 2024a) and M^{3}CoT-Bench (Chen et al., 2024a) remains low. The RL/hybrid comparison in Table 3 further shows that these methods rely on composite rewards and substantial SFT initialization, and share recurring failure modes (reward hacking, entropy collapse, tool misuse).

4 Evaluation and Benchmarks

As IG-CoT methods mature, evaluation must evolve from task-specific leaderboards to a more nuanced, method-aware analysis that diagnoses *how* models reason. This section connects our method taxonomy back to the core reasoning domains identified in the introduction: detail-oriented and imagined-world problems. We argue that robust evaluation requires not just measuring final accuracy, but also assessing the faithfulness and efficiency of the reasoning process itself, directly addressing the **evaluation mismatch** and **computational cost** challenges. Detailed benchmark summaries are in Table 2.

4.1 Evaluating Image-Grounded, Detail–Oriented Reasoning

This domain directly tests the "meticulous perception" central to IG-CoT, a core theme of our survey. Success here demonstrates that a model's reasoning is genuinely grounded in visual evidence, rather than relying on textual shortcuts.

Fine-grained Visual Question Answering (VQA) benchmarks like VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and the synthetic CLEVR (Johnson et al., 2017) are

Benchmark	Type	Key Metrics	Depth (steps)	# Samples	Reported Accuracy	
					Gemini-1.5	GPT-4o
SpatialEval (2024)	Detail-oriented	Facet-wise acc.	< 3	13.9k items	40%	64.5%
VisuLogic (2025a)	Detail-oriented	Overall & per-category acc.	< 3	1k Qs	—	—
MathVista (2023)	Detail-oriented	Exact-match acc. (leaderboard)	< 3	6.1k problems	63.9%	63.8%
V*Bench (2024)	Detail-oriented	Attr. acc.; spatial acc.; overall acc.	< 3	191 QA	—	80.3%
MageBench (2024b)	Imagined-world	Scenario success; best-of-N	> 5	483 scenarios	48.3%	48.6%
VSP (2024a)	Imagined-world	Plan success; percept./reason. acc.	> 5	10 subtasks	63%	72%
ING-VP (2024a)	Imagined-world	Level-clear acc.; win-rate	> 20	6 games, 300 levels	1.1%	2.8%
Sokoban (2018)	Imagined-world	Plan success	> 5	50 maps	—	—
Blocksworld (IPC-7) (2015)	Imagined-world	Plan success	> 5	30 tasks	—	—
MiniBEHAVIOR (2023)	Imagined-world	Goal success	> 5	100 scenarios	—	—
EgoSchema (2025)	Egocentric Streams	MC QA accuracy	shot loop ($\approx$10)	5k clips	-	
LVBench (2025)	Keyframe QA	MC QA accuracy (long videos)	keyframe loop ($\approx$12)	240 clips	54.1%	—
NExT-QA (2025)	Keyframe QA	MC QA accuracy	shot loop ($\approx$8)	5.4k clips	-	66.1%
OpenEQA (2025)	Keyframe QA	0–100 rating	keyframe loop ($\approx$14)	1k trajectories	71.6 (TCoT)	—
Video-MME (Long) (2024)	General suite	Holistic score (long-form)	keyframe loop ($\approx$15)	36 long videos	-	67.4% (o3)
MLVU (2025)	General suite	Composite score (M-Avg)	multi-turn ($>$12)	200 tasks	-	—
MVBench (2025)	General suite	Overall score	shot loop ($\approx$9)	200 videos	-	45.5%
CoMT (2024b)	General suite	Answer acc.; rationale quality	≈ 8	3.9k Qs, 14.8k imgs	—	—
Visual CoT (2024a)	General suite	Answer acc.; box-IoU (visual)	> 3	438k QA pairs	—	—
M^{3}CoT (2024a)	General suite	Answer acc.; step-counts	6–15	11.5k Qs, 11.3k imgs	—	—
MMBench (2023b)	General suite	Multiple-choice acc.	< 3	2 974 Q	77.1%	81.8%
MM-Vet v2 (2023)	General suite	Overall acc.	< 3	218 QA	—	71.0%

Table 2: Vision-language reasoning benchmarks benefited from IG-CoT methods, with their category, reasoning-depth, dataset size, and the latest publicly reported baselines from Gemini-1.5 and GPT-4o ("—" = not reported as of 20 May 2025).

foundational for assessing compositional reasoning (understanding objects, attributes, and relationships). CLEVR tests pure logic on simple shapes, while VQAv2/GQA use real-world images for more complex, common-sense relational questions. SFT approaches like VoCoT (Li et al., 2024) (object-centric embeddings) and Visual CoT (Shao et al., 2024a) (bounding-box "thoughts") excel by explicitly grounding steps; CogCoM (Qi et al., 2024) (17B) reports 71.7% on GQA as a generalist. Prompting (e.g., IoT (Zhou et al., 2024)) and data synthesis (Least-to-Most (Cheng et al., 2024a), small but consistent GQA gains over SFT baselines) also help, with UV-CoT (Zhao et al., 2025) and tool-supervised RL (Dong et al., 2026c) showing RL's utility. However, complex compositional questions in real-world VQA needing multi-step deep reasoning or robust common sense remain challenging.

Visual Grounding and Segmentation tasks, on benchmarks like CLEVR-Ref+ (Liu et al., 2019) and EmbSpatial (Du et al., 2024), demand precise localization of referred entities; recent work applies IG-CoT directly to referring expressions (Dong et al., 2025, 2026b). Success on these benchmarks is a direct measure of a model's ability to create a tight link between language and specific image regions, a foundational skill for faithful IG-CoT.

Hallucination Detection and Consistency Checking, using benchmarks like POPE (Li et al., 2023) and MM-Vet (Yu et al., 2023)'s consistency subset, focus on faithfulness. These are critical for verifying that IG-CoT chains are not just plausible but correct, ensuring that the move to MLLM-native reasoning yields more trustworthy and verifiable outputs than older, text-centric pipelines.

Specialised Niche Reasoning includes: *OCR-centric VQA* (e.g., ST-VQA (Biten et al., 2019), TextVQA (Singh et al., 2019)), where CogCoM (Qi et al., 2024) reports 71.1% on ST-VQA and 84.0% on TallyQA-simple; *Counting and Numeracy* (e.g., TallyVQA (Acharya et al., 2019)); and *Image Editing/Program Synthesis* (e.g., VISPROG (Gupta and Kembhavi, 2023)'s Twitter-Memes). These specialized areas often require dedicated architectural components or training strategies, and robust performance across all niches is still developing.

4.2 Evaluating Imagined-World Reasoning

This domain tests the ability to "simulate unseen states," a key advantage of IG-CoT for tasks requiring forecasting or abstract thought.

Games (Abstract Environments) like Mazes (Xu et al., 2025b), Sokoban (Schrader, 2018), Frozen Lake (Brockman et al., 2016), and Blocksworld (López et al., 2015) test planning, prediction, and understanding of dynamics, ranging from simple grid navigation to complex stateful planning often requiring long-horizon reasoning. RL methods show significant promise: VAGEN (Wang et al., 2025a) improves success rates on Sokoban and FrozenLake; RAGEN (Wang et al., 2025c) stabilizes multi-turn RL; Visual Planning (Xu et al., 2025b) reaches 74.5% exact-

match on small 3×3–6×6 Mazes with visual-only steps. SFT methods like MVoT (Li et al., 2025) (intermediate "imagination" images, e.g., +7.7 pp over Direct on FrozenLake) and prompting with visual intermediate states (Mind's Eye (Wu et al., 2024b), Visualizing Thought (Borazjanizadeh et al., 2025) boosting GPT-4o Blocksworld accuracy from 35.5% to 90.2%) also yield gains. However, many complex PDDL domains or games with intricate rules are far from solved. RL methods often struggle with stability/sample efficiency, and the VSP benchmark (Wu et al., 2024a) (most VLMs $\leq$32% success) highlights a persistent perception-reasoning gap. Reported settings commonly use small grids (3×3–5×5) or moderate mazes (up to 32×32) and short episodes, which should be considered when interpreting gains.

Math (Geometry) problems, on benchmarks like MathVista (Lu et al., 2023), require integrating visual perception with mathematical reasoning from diagrams, often involving multi-step calculations. Prompting methods allowing externalized thought via sketches (Visual Sketchpad (Hu et al., 2024), SOTA on V*Bench 80.3%; VAP (Xiao et al., 2024); VisuoThink (Wang et al., 2025b), multimodal tree search with visual tool use) are particularly effective. SFT methods like Chain of Images (Meng et al., 2023) (generating intermediate image sequences) also help. Robustly translating complex visual information from noisy/ambiguous diagrams into solvable mathematical formalisms remains challenging, as seen on MathVista.

Planning Problems, including PDDL-based planning sets and embodied navigation (e.g., MiniBehavior (Jin et al., 2023)), assess long-horizon reasoning in structured environments demanding abstract state representation and multi-step action sequencing. Visualizing Thought (Borazjanizadeh et al., 2025) shows strong PDDL performance, while MVoT (Li et al., 2025) and Visual Planning (Xu et al., 2025b) address MiniBehavior. These domains are generally very challenging, especially for complex PDDL problems requiring deep search or embodied tasks with rich sensory input and long completion times.

4.3 Evaluating Video-Context Reasoning

Video extends IG-CoT's challenges over time, making the **computational cost** of processing long sequences a primary concern. Methods are often distinguished by their strategy for managing this complexity. Training-free approaches like **TCoT** and **CoS** (Arnab et al., 2025; Hu et al., 2025) manage cost by first selecting salient frames, evaluated on QA over keyframes (LVBench, NExT-QA). Broader suites like **Video-MME** (Gemini Team, 2024) and **MVBench** (Hu et al., 2025) test a wider range of temporal skills. Egocentric video (e.g., EgoSchema, evaluated by **Ego-R1** (Tian et al., 2025)) pushes these limits further. Effective evaluation in this space must report not just accuracy but also efficiency metrics like token budgets and retrieval costs, which are the primary trade-offs.

4.4 Evaluating General Multimodal Competence Suites

Comprehensive benchmarks like MM-Bench (Liu et al., 2023b), MM-Vet (Yu et al., 2023), the Visual CoT Benchmark (Shao et al., 2024a), and M^3CoT-Bench (Chen et al., 2024a) provide a broad assessment of multimodal reasoning from basic perception to complex reasoning. IG-CoT methods generally improve performance: IoT (Zhou et al., 2024) delivers a >100-point MME boost on GPT-4o; the specialized Visual CoT model (Shao et al., 2024a) delivers consistent gains over its SFT baseline; CMMCoT (Zhang et al., 2025) lifts multi-image QA scores over strong Qwen2-VL/Qwen2.5-VL baselines by a few points on average. Many benchmarks (e.g., M^3CoT) reveal that reasoning with depths >10 steps or across multiple images is still a frontier, and VSP (Wu et al., 2024a) also indicates that complex spatial planning with perception is far from solved even with SFT.

In summary, while IG-CoT methods excel on many benchmarks, they face persistent challenges in long-horizon planning, robust reasoning in complex or noisy environments, achieving true faithfulness, and managing the computational cost of multi-step visual grounding. Evolving benchmarks are increasingly highlighting these limitations, guiding research towards more robust and reliable systems.

5 Future Directions

Looking ahead, we highlight three thrusts likely to advance Image-Grounded Chain-of-Thought (IG-CoT) reasoning: (i) core architectural and representational advances; (ii) data- and environment-centric methodology; and (iii) evaluation and trustworthiness.

5.1 Core Architectural and Representational Advances

Multimodal Generative Reasoning: A central direction is to move from *interpreting* to deliberately *generating* visual state during reasoning. Two complementary paths are promising: (1) tighter, programmable control over external visual tools for drawing, diagramming, and transformation (e.g., Visual Sketchpad (Hu et al., 2024)); and (2) intrinsic visual generative capabilities that let models synthesize, edit, and carry visual state internally across steps. Key questions include how to represent intermediate visuals (e.g., vector primitives, sketches, or latent feature maps (Lu et al., 2025b)), how to control them (programmatic APIs versus natural-language interfaces), and how to bound cost. Hybrid designs that combine lightweight, compositional internal representations with on-demand tool calls for high-fidelity renders may strike a favorable balance among fidelity, controllability, and efficiency.

Efficient and Scalable IG-CoT Architectures: Iterative, image-heavy chains stress both memory and latency. Promising directions include cross-step attention/key-value reuse and token caching; selective re-encoding with state compression or sketch-like summaries; hierarchical visual memories that store and retrieve intermediate states; dynamic-resolution processing (image pyramids, learned cropping/zoom) and visual token routing; and modular designs that disentangle perception from symbolic reasoning while enabling shared state. Reporting and optimizing for end-to-end efficiency metrics (tokens, tool calls, wall-clock latency) should be first-class concerns alongside accuracy.

5.2 Data-centric and Environment-centric Innovations

Advanced Data Creation and Augmentation Strategies: Datasets should *require* stepwise visual grounding and resist shortcutting. Useful tactics include counterfactuals and perturbations that break text-only heuristics, partial-evidence settings that compel visual inspection, and chain-level supervision with stepwise critiques or preferences. Programmatic synthesis can provide breadth while preserving structure, and solver- or simulator-backed pipelines can bootstrap high-quality traces.

Embodied AI and Interactive Long-Horizon Reasoning: Extending IG-CoT to embodied agents introduces partial observability, action-conditioned perception, and long-horizon memory. Agents must perceive, act, and update visual state over time, learning when to retrieve, compress, or refresh context. Progress on settings like MiniBehavior (Jin et al., 2023), complex games, and robotics will hinge on budgeted retrieval policies, segment- or episode-level memory, and safety-aware evaluation.

5.3 Evaluation, Trustworthiness, and Foundational Understanding

Comprehensive Benchmarking and Evaluation: Evaluation should move beyond final accuracy to assess faithfulness, robustness, and cost. Standardized reporting should include chain length, images processed, tool calls, tokens, and wall-clock time. Faithfulness can be probed via step–image alignment checks, counterfactual sensitivity, and trace-perturbation tests that detect shortcutting. Robustness analyses should vary retrieval granularity and budget constraints (turns/tokens) for both images and video. Releasing prompts, seeds, and evaluator configurations will improve reproducibility and comparability across methods.

6 Conclusion

This survey provided a structured overview of Image-Grounded Chain-of-Thought (IG-CoT), a rapidly evolving paradigm for Multimodal LLMs. We introduced a dual taxonomy, categorizing methods by their training paradigm (prompting, SFT, and RL) and mapping them to two core reasoning domains: detail-oriented and imagined-world tasks. Our analysis illuminated the fundamental trade-offs between these approaches, weighing the flexibility of prompting against its inference costs, the reliability of SFT against its data bottleneck, and the discovery potential of RL against its computational demands. We underscored the critical need for evaluation to evolve beyond final accuracy, advocating for faithfulness and efficiency. By synthesizing current methods and identifying key challenges, we outlined a forward-looking research agenda aimed at developing better IG-CoT models.

7 Role of LLMs in Aiding Writing

LLMs assisted with drafting and polishing prose; the authors directed content, designed the framework, and validated all technical claims.

8 Ethical Statement

This survey of Image-Grounded Chain-of-Thought (IG-CoT) reasoning in Multimodal Large Language Models (MLLMs) raises important ethical considerations. As MLLMs improve at visual reasoning, risks include perpetuating biases, generating misleading content, and potential misuse in sensitive applications. We urge the community to prioritize fairness, transparency, and accountability, including bias auditing, explainability, and responsible deployment. The authors are committed to promoting ethical progress in this field.

Limitations

This survey offers an overview of Image-Grounded Chain-of-Thought (IG-CoT) reasoning within Multimodal Large Language Models. However, readers should consider several inherent limitations. First, the rapid pace of advancements in this domain means that, despite diligent efforts to include current research, some very recent breakthroughs might not be covered. Second, space considerations necessitate a concise presentation of the surveyed methodologies, precluding exhaustive technical discussions for every approach, which may affect the depth of detail for some readers. Our review predominantly covers literature from the past three years, drawing mainly from prominent AI and NLP conference proceedings (such as ACL, EMNLP, CVPR, NeurIPS, ICLR) and arXiv pre-prints; ongoing work will aim to track these venues for emerging contributions. Furthermore, the analyses and forward-looking statements herein are primarily informed by the empirical results and discussions within the reviewed papers. While this approach offers a solid foundation, it might not encompass every subtlety of the challenges discussed. Finally, as with any comprehensive review, the perspectives and conclusions presented reflect our interpretation of the field, and alternative viewpoints may exist. Notwithstanding these constraints, we are confident that this work furnishes a useful and broad perspective on the current landscape of IG-CoT reasoning in MLLMs. In particular for RL-based IG-CoT, most evidence comes from small, rule-based environments with short horizons; training remains compute-intensive and unstable, and generalization to larger, out-of-distribution tasks is not yet well established.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. *AAAI*.

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *arXiv preprint arXiv:1704.08243*.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2025. Temporal chain of thought: Long-video understanding by thinking in frames.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *ICCV*.

Nasim Borazjanizadeh, Roei Herzig, Eduard Oks, Trevor Darrell, Rogerio Feris, and Leonid Karlinsky. 2025. Visualizing thought: Conceptual diagrams enable robust planning in lmms. *arXiv preprint arXiv:2503.11790*.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M^3cot: A novel benchmark for multi–domain multi–step multi–modal chain–of–thought. *arXiv preprint arXiv:2405.16473*.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *AAAI*.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. 2024a. From the least to the most: Building a plug-and-play visual reasoner via data synthesis. *arXiv preprint arXiv:2406.19934*.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2024b. Comt: A novel benchmark for chain of multi–modal thought on large vision–language models. *arXiv preprint arXiv:2412.12932*.

Qihua Dong, Luis Figueroa, Handong Zhao, Kushal Kafle, Jason Kuen, Zhihong Ding, Scott Cohen, and Yun Fu. 2025. CoT referring: Chain-of-thought grounding improves localization in referring expression tasks. *arXiv preprint arXiv:2510.06243*.

Qihua Dong, Ruozhen He, Junwen Chen, Yizhou Wang, Xu Ma, Songyao Jiang, and Yun Fu. 2026a. Hierarchical visual agent: Managing contexts in joint image-text space for advanced chart reasoning. In *Findings of the Association for Computational Linguistics (ACL)*.

Qihua Dong, Yang Kuo, Ju Lin, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. 2026b. Ref-adv: Exploring MLLM visual reasoning in referring expression tasks. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Qihua Dong, Gozde Sahin, Pei Wang, Zhaowei Cai, Robik Shrestha, Hao Yang, and Davide Modolo. 2026c. Visual reasoning through tool-supervised reinforcement learning. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*.

Jun Gao, Ming-Hao Li, Wenzhang Luan, Geng Tu, Yijie Wang, and Jitao Sang. 2024. Interleaved-modal chain-of-thought. *arXiv preprint arXiv:2411.19488*.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. 2025. Framemind: Frame-interleaved video reasoning via reinforcement learning.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google DeepMind. Technical Report.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962.

Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. 2025. Cos: Chain-of-shot prompting for long video understanding.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*. NeurIPS 2024.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo J Taylor, and Tanwi Mallick. 2024. Towards rationality in language and multimodal agents. *arXiv preprint arXiv:2406.00252*.

Emily Jin, Jiaheng Hu, Zhuoyi Huang, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and Roberto Martín-Martín. 2023. Mini-behavior: A procedurally generated benchmark for long-horizon decision-making in embodied ai. *arXiv preprint arXiv:2310.01824*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 2025. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *EMNLP*.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuan-Jing Huang, and Zhongyu Wei. 2024. VoCoT: Unleashing visually grounded multi-step reasoning in large multimodal models. *arXiv preprint arXiv:2405.16919*.

Yujie Lin, Ante Wang, Moye Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. 2025a. Investigating inference-time scaling for chain of multimodal thought: A preliminary study. *arXiv preprint arXiv:2502.11514*.

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. 2025b. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv:2304.08485*.

Jingming Liu, Yumeng Li, Boyuan Xiao, Yichang Jian, Ziang Qin, Tianjia Shao, Yao-Xiang Ding, and Kun Zhou. 2024. Autonomous imagination: Closed-loop decomposition of visual-to-textual conversion in visual reasoning for multimodal large language models. *arXiv preprint arXiv:2411.18142*.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025. Videomind: A chain-of-lora agent for long video reasoning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Carlos Linares López, Sergio Jiménez Celorrio, and Ángel García Olaya. 2015. The deterministic part of the seventh international planning competition. *Artificial Intelligence*, 223:82–119.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. 2025a. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16669–16684. Association for Computational Linguistics.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. 2025b. The indra representation hypothesis for multimodal alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai–Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. 2023. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*.

OpenAI. 2024. Openai o1 system card. https://arxiv.org/abs/2412.16720. Safety and capability report for the OpenAI o1 model family.

OpenAI. 2025. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Model-card report for the OpenAI o3 and o4-mini models.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. 2024. CogCoM: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*.

Max-Philipp B. Schrader. 2018. gym-sokoban. https://github.com/mpSchrader/gym-sokoban.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi–modal language models with a comprehensive dataset and benchmark for chain–of–thought reasoning. *arXiv preprint arXiv:2403.16999*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. *CVPR*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16317–16327.

Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *arXiv preprint arXiv:2406.14852*.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, and Manling Li. 2025a. VAGEN: Reinforcing world model reasoning for multi-turn vlm agents. *arXiv preprint arXiv:2510.16907*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025b. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. 2025c. RAGEN: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.

Max Waters, Lin Padgham, and Sebastian Sardina. 2020. The barman-HTN domain for IPC 2020. In *The 10th International Planning Competition - Planner and Domains Abstracts*.

Penghao Wu and Saining Xie. 2024. 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)*.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. 2024a. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024b. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Wing Yin Yu, Tao Zhong, Sai Wu, Yuan Jessica Wang, Jianwei Yin, and Gang Chen. 2024. Enhancing llm reasoning via vision-augmented prompting. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2025a. Visulogic: A benchmark for evaluating visual reasoning in multi–modal large language models. *arXiv preprint arXiv:2504.15279*.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. 2025b. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hritik Ayyubi, Kai-Wei Chang, Sung-Ju Lee, and Yang Song. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, and Hao Jiang. 2025. CMM-CoT: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *arXiv preprint arXiv:2503.05255*.

Haoran Zhang, Hangyu Guo, Shuyue Guo, Meng Cao, Wenhao Huang, Jiaheng Liu, and Ge Zhang. 2024a. Ing–vp: Mllms cannot play easy vision–based games yet. *arXiv preprint arXiv:2410.06555*.

Miaosen Zhang, Qi Dai, Yifan Yang, Jianmin Bao, Dongdong Chen, Kai Qiu, Chong Luo, Xin Geng, and Baining Guo. 2024b. Magebench: Bridging large multimodal models to agents. *arXiv preprint arXiv:2412.04531*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024c. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. 2025. Unsupervised visual chain-of-thought reasoning via preference optimization. *arXiv preprint arXiv:2504.18397*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "Thinking with Images" via reinforcement learning.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. 2025. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

A Appendix

A.1 Selection Protocol

We document the protocol used to compile the papers surveyed in this work so that readers and future updates can replicate or extend our coverage. *Time window.* Because Image-Grounded Chain-of-Thought is a recent development, we restrict the survey to papers published in roughly the last two years (after the original Chain-of-Thought paper), supplemented by a handful of foundational earlier works cited for context. *Sources.* We identify candidate papers through keyword search on arXiv and in major AI/NLP/vision venues (ACL, EMNLP, NAACL, CVPR, ICCV, ECCV, NeurIPS, ICLR), and we expand the pool through backward and forward citation tracking from a seed set of prompting, SFT, and RL-based IG-CoT papers. *Inclusion criteria.* We include a paper if it follows our IG-CoT definition (Section 2): an end-to-end Multimodal LLM that iteratively interleaves textual rationales with visually updated states so that what the model "sees" continuously informs what it "thinks." *Exclusion criteria.* We exclude LLM-as-controller and purely text-centric pipelines in which visual content is reduced to captions or tool outputs before reasoning (e.g., ViperGPT (Surís et al., 2023), VisProg (Gupta and Kembhavi, 2023)), and we exclude methods that rely only on a single forward pass over the image without any intermediate visual state update. The originating paper and base model for each surveyed method are annotated in Table 1 in the main text and Table 4 in this appendix.

A.2 RL and Hybrid IG-CoT Comparison

Table 3 gives the full side-by-side comparison of RL and hybrid IG-CoT methods referenced in §3.2.2. For each method we report the training paradigm (pure-RL variants of PPO / GRPO / StarPO, or SFT+RL hybrids), the reward design decomposed into base task reward, auxiliary shaping signals, and regularization terms, the approximate scale and sample efficiency of reported training runs, a qualitative stability label, and recurring failure modes or limits reported in the original papers.

Two patterns stand out. First, almost all entries rely on *composite* rewards: a task-outcome base reward is augmented with substantial shaping (progress maps, format / tool-usage / exploration bonuses, step-wise preferences) and regularization (KL, clipping, invalid-action or over-turn masking); pure task reward is rarely sufficient for stable IG-CoT training. Second, the strongest and most stable gains appear in small, simulator-backed, short-horizon environments (grid mazes, Sokoban, FrozenLake), while performance in long-video reasoning, DocVQA-style layouts, and long-horizon visual search depends heavily on SFT initialization and careful masking. Recurring failure modes (reward hacking, entropy collapse or "echo trap," tool misuse, and degradation without exploration or turn-limit shaping) echo across otherwise quite different pipelines. These patterns directly support Hypothesis H3 in §3.4: RL-based IG-CoT is currently most effective in constrained settings and has not yet robustly transferred to open-world, long-horizon tasks.

A.3 Surveyed IG-CoT Methods

Table 4 enumerates the representative IG-CoT papers we review, capturing their training paradigms, base models, data sources, and reported benchmarks.

Method	Paradigm	Reward Design	Scale / Sample Eff.	Stability	Key Failures / Limits
VAGEN (Wang et al., 2025a)	PPO (Bi-Level GAE)	Base: env task reward. Shaping: dense world-modeling (LLM judge) + format. Reg: KL + Bi-Level GAE credit.	Low eff.: ∼30–40 H100 h/env on small grids (Sokoban, FrozenLake, ManiSkill); ∼10 h for SVG; 3B VLM.	Moderate: VAGEN-Full stable, but Bi-Level GAE alone brittle under sparse rewards.	Response convergence (low-entropy templates) and reward hacking when rewards are sparse or misaligned.
Visual Planning (Xu et al., 2025b)	SFT+GRPO	Base: env task reward. Shaping: parser-based progress + progress map. Reg: invalid-action penalties + KL.	Med eff.: 10 SFT + 10 RL epochs on small grid nav (FrozenLake, Maze, MiniBehavior) with 8×A100 (LoRA).	Stable: smooth GRPO curves with KL; consistent gains over VPFT / text-SFT.	Residual invalid actions; heavy reliance on hand-crafted parser / progress map limits generalization.
RAGEN (Wang et al., 2025c)	StarPO	Base: env reward. Shaping: reasoning-aware trajectory weighting. Reg: format penalty + PPO/GRPO clipping.	Low eff.: 100–200 rollout–update iters with 8×16 rollouts, ≤5 turns / 10 actions; 0.5B–3B Qwen.	Moderate: vanilla PPO/GRPO can collapse; StarPO-S filtering / critic / clipping improves stability.	"Echo Trap": entropy collapse, reward-variance cliffs, gradient spikes, repetitive templates under poorly shaped rewards.
UV-CoT (Zhao et al., 2025)	sDPO (iterative)	Base: preference over answer quality. Shaping: step-wise box scores from evaluator MLLM. Reg: sDPO log-prob ratio margin.	High label eff.: auto-generated preferences at each CoT step; no human boxes; iterative retraining.	Stable: reported more stable and better aligned than vanilla DPO for visual CoT.	On DocVQA / InfographicsVQA, using predicted boxes underperforms GT boxes due to imperfect region localization on complex layouts.
FrameMind (Ge et al., 2025)	GRPO (DRFS)	Base: QA accuracy. Shaping: format, tool-usage, exploration, and turn-efficiency bonuses. Reg: GRPO KL / clipping.	Med eff.: Qwen2.5-VL-7B GRPO; DRFS ladder reuses 32/48/64-frame data to raise accuracy at fixed budgets.	Moderate: DRFS-GRPO robust vs. fixed-res GRPO but fails to learn without exploration bonus.	Without exploration bonus, tool policy fails to learn; fixed 32-frame GRPO underperforms on long-video or long-context segments.
Ego-R1 (Tian et al., 2025)	SFT+GRPO	Base: QA outcome reward. Shaping: multi-step CoTT trajectory structure. Reg: standard GRPO regularization.	Med eff.: 25k SFT traces + RL on 4.4k Ego-QA instances (∼7.4 steps each).	Stable: standard RL training; no dedicated instability issues reported beyond hyperparameter tuning.	Limits center on narrow egocentric domain and dependence on existing RAG + video tools; no systematic RL failure taxonomy.
DeepEyes (Zheng et al., 2025)	GRPO	Base: task success. Shaping: tool-use-oriented bonuses. Reg: standard policy-gradient regularization.	Qual.: direct RL on curated fine-grained perception / reasoning suites; no published GPU-hour table.	Stable: monotonic gains in grounding IoU / tool efficiency; only qualitative failure analysis.	Qualitative failures from tool over- / under-use; limited deeper analysis of RL-specific failure modes.
Mini-o3 (Lai et al., 2025)	SFT+GRPO	Base: visual-search success. Shaping: curriculum + over-turn masking. Reg: standard PG regularization.	Med eff.: SFT + RL on VisualProbe-style search; RL capped at 6 turns but generalizes via self-play.	Moderate: curriculum + over-turn masking stable; long-horizon search degrades without masking.	Without over-turn masking, long-horizon strategies discouraged and hardest cases revert to shallow, monotone search.

Table 3: RL and hybrid IG-CoT methods compared along reward design, sample efficiency / scale, training stability, and recurring failure modes. Entries span pure-RL pipelines (PPO, GRPO, StarPO, iterative sDPO) and SFT+RL hybrids (SFT+GRPO); the **Paradigm** column reports the training setup disclosed by each paper. In **Reward Design**, *Base* is the primary task reward, *Shaping* lists auxiliary signals added to improve learning, and *Reg* summarizes regularization terms (e.g., KL, clipping, invalid-action penalties). In **Scale / Sample Eff.**, *Low* / *Med* eff. and *High label* eff. give coarse labels for sample efficiency and training cost; *Stable* vs. *Moderate* in **Stability** reflects qualitative stability reports from the original papers.

Paper	TF	RL	SFT	Base model	Visual ops	Training data	Benchmarks
ItS-CoMT (2025a)	✓	✗	✗	Qwen2-VL-72B / GPT-4o	draw; mask; depth	Not applicable	Geometry3K, Maxflow, BLINK (+7)
Visualizing Thought (2025)	✓	✗	✗	GPT-4o	code-gen; diagram	Not applicable	Blocksworld, Parking, Sokoban
VisuoThink (2025b)	✓	✗	✗	GPT-4o / Qwen2-VL-72B / Claude-3.5	draw; highlight; zoom	Not applicable	Geomverse-109, Geometry3K, Visual Nav.
Interleaved-Modal CoT (CVPR (2024))	✓	✗	✗	BLIP-2, LLaVA-1.5 (plug-and-play)	crop; attention	Not applicable	VQA-v2, GQA, ScienceQA
Autonomous Imagination (2024)	✓	✗	✗	LLaVA-1.5, GPT-4V	edit; focus	Not applicable	dense counting, jigsaw, placement
Vision-Augmented Prompting (2024)	✓	✗	✗	GPT-4-Turbo / Llama-2-70B	code-gen; diagram	Not applicable	geometry, Sudoku, TSP
Visualization-of-Thought, (2024b)	✓	✗	✗	GPT-4-Turbo, Llama-2	ascii; grid	Not applicable	NL navigation, TileWorld, NL-VR2
Visual Sketchpad (2024)	✓	✗	✗	GPT-4o, LLaVA-1.5	draw; parse	Not applicable	V*Bench, BLINK, geometry & chess
Temporal CoT (TCoT) (2025)	✓	✗	✗	Gemini 1.5 Pro / GPT-4o	retrieve; keyframe	Not applicable	LVBench (+11.4), EgoSchema, OpenEQA
Chain-of-Shot (CoS) (2025)	✓	✗	✗	GPT-4o, Claude-3.5, InternVL	retrieve; shots	Not applicable	VideoMME, MLVU, MVBench, NExT-QA, LongVideoBench
Mini-o3 (2025)	✗	✓	✓	Qwen2.5-VL-7B	probe; zoom	6k SFT cold-start + 12k RL (8k DeepEyes + 4k VisualProbe)	VisualProbe, V*Bench, HR-Bench, MME-Realworld
Image-of-Thought Prompting (2024)	✓	✗	✗	LLaVA-1.5, BLIP-2, GPT-4V	crop; zoom	Not applicable	VQAv2, TDIUC, CLEVR-HYP
UV-CoT (2025)	✗	✓	✗	LLaVA-1.5-7B	crop; zoom	249k iterative sDPO preference pairs (no human labels)	DocVQA, TextVQA, InfographicsVQA, Flickr30k, GQA, VSR
Visual Planning (2025b)	✗	✓	✗	ViT-based LVLM (post-trained)	env; grid	RL environment roll-outs	FrozenLake, Maze, MiniBehavior
RAGEN (2025c)	✗	✓	✗	Qwen2.5 0.5B–3B	env; grid	RL-generated trajectories (grid worlds)	Sokoban, FrozenLake
VAGEN (2025a)	✗	✓	✗	Qwen2.5-VL-3B	env; grid	Self-generated trajectories	Sokoban, FrozenLake, ManiSkill
FrameMind (2025)	✗	✓	✗	Qwen2.5-VL-7B	retrieve; clips	GRPO with DRFS 32/48/64-frame ladder	Video-MME, MLVU, MVBench
Ego-R1 (2025)	✗	✓	✓	Qwen2.5-VL + tools	retrieve; clips	25k SFT + 4.4k RL on egocentric traces	Video-MME (long), EgoSchema, EgoLifeQA, Ego-R1 Bench
CMMCoT (2025)	✗	✗	✓	Qwen-VL / LLaVA-1.5	crop; memory	Curated multi-image dataset	Multi-image comprehension tasks
CogCoM (2024)	✗	✗	✓	CogCoM-17B	ocr; line; box	70k auto + 6k manual CoM samples (on top of 40M Stage-1 grounded pretrain)	9 VQA/REC tasks
MVoT (2025)	✗	✗	✓	Chameleon-7B (+ auxiliary token discrepancy loss)	gen; diagram	∼18k spatial trace samples (Maze + MiniBehavior + FrozenLake)	mazes, MiniBehavior, FrozenLake
VoCoT (2024)	✗	✗	✓	LLaVA-1.5	crop; zoom	80k VoCoT pairs	CLEVR, EmbSpatial, V*Bench
From the Least to the Most (EMNLP (2024a))	✗	✗	✓	BLIP-2 / LLaVA-1.5 / InstructBLIP / Otter	box; ocr	50k synthetic least-to-most trajectories	VQA-v2, OK-VQA, TextVQA, GQA
Chain of Images (2023)	✗	✗	✓	SyMLLM	gen; diagram	CoI dataset (15 domains)	Geometry, Chess, Common-sense
Visual CoT (2024a)	✗	✗	✓	CLIP ViT-L/14 + Vicuna-7/13B	crop; zoom	438k Visual CoT samples	Visual CoT benchmark (5 domains)
V* Guided Visual Search (2024)	✗	✗	✓	BLIP-2/CLIP-ViT + Llama-2-7B	crop; zoom	387k curated instruction pairs	V*Bench
VideoMind (2025)	✗	✗	✓	Qwen2.5-7B + role LoRAs	retrieve; verify	260k long-video traces	CG-Bench, ReXTime, NExT-GQA, Charades-STA, ActivityNet-Captions, Video-MME, MLVU, LVBench

Table 4: Surveyed Image-Grounded Chain-of-Thought (IG-CoT) papers. **TF** = training-free (✓ yes, ✗ no). "Visual ops" lists the primary visual interactions ($\leq$3 tokens).