Thinking in Pictures: A Diagnostic Study of Visual vs. Textual Chain-of-Thought Reasoning in Vision-Language Models

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) reasoning has become a standard technique for eliciting complex reasoning in large language models, and recent work has extended it to vision-language models (VLMs). However, virtually all multimodal CoT methods generate intermediate reasoning steps in natural language, even for inherently visual problems such as spatial reasoning, geometric manipulation, and object tracking. We ask a fundamental question: *when should a VLM reason in words, and when should it reason in pictures?* We present VISCOT-DIAG, a diagnostic benchmark of 1,200 instances across five visual reasoning categories, and compare four CoT paradigms across four VLMs. Our results reveal a striking *modality gap*: textual CoT *degrades* performance by up to 17.5% on spatial transformation and 13.2% on multi-object tracking, while visual CoT yields gains of up to 23.1%. We identify three failure modes (spatial state collapse, transformation hallucination, tracking loss) and show that adaptive modality routing achieves 73.1% accuracy versus 68.9% for V-CoT-everywhere. We recommend practitioners use visual CoT for spatial tasks and textual CoT for compositional counting.

1 Introduction

Chain-of-thought (CoT) prompting has dramatically improved the reasoning capabilities of large language models by encouraging models to decompose problems into intermediate steps before arriving at an answer (Wei et al., 2022; Kojima et al., 2022). This paradigm has been extended to vision-language models (VLMs), where multimodal CoT methods generate textual rationales that incorporate information from both visual and linguistic inputs (Zhang et al., 2024; Lu et al., 2022).

A growing body of work has advanced multimodal CoT through structured reasoning stages (Xu et al., 2025; Thawakar et al., 2025), visual tool use (Hu et al., 2024; Wang et al., 2025), and visualization of reasoning traces (Wu et al., 2024). However, a fundamental question remains underexplored: *when should a VLM reason in words, and when should it reason in pictures?* Our central finding: **textual chain-of-thought actively harms spatial reasoning in VLMs; visual CoT fixes it.**

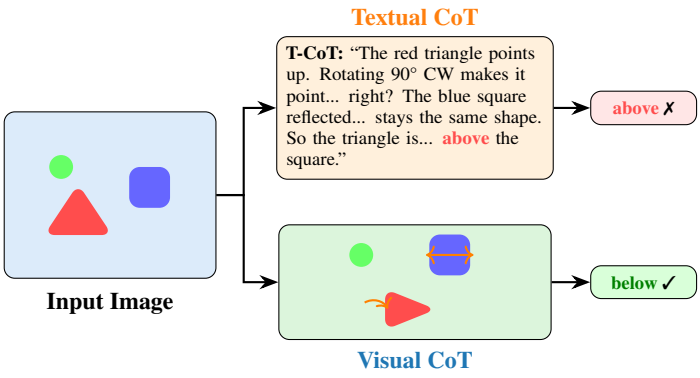

Figure 1: Motivating example. Given a spatial transformation question, textual CoT produces a plausible but incorrect verbal trace (the model hallucinates the post-rotation spatial relationship), while visual CoT draws the actual transformations, grounding reasoning in a veridical visual representation.

Consider the motivating example in Figure 1. A model is shown a scene with colored shapes and asked: "After rotating the red triangle 90° clockwise and reflecting the blue square horizontally, is the triangle above or below the square?" Under textual CoT, the model must verbally describe each transformation and its spatial consequence, a process that is error-prone because language lacks the precision to faithfully encode continuous spatial configurations. Under visual CoT, the model can draw the transformed objects directly, producing a veridical visual representation that grounds subsequent reasoning.

This observation is grounded in cognitive science. When humans determine whether two shapes are identical under rotation, they perform *mental ro-*

tation, a visual process whose response time scales linearly with the angle of rotation (Shepard and Metzler, 1971). They do not describe the shapes verbally. Dual coding theory (Paivio, 1991) posits that humans maintain separate but interconnected verbal and imagistic representational systems, and that certain tasks are better served by one system than the other. The "mind's eye" enables manipulation of mental images for spatial reasoning (Kosslyn, 1995), a capacity that current VLMs largely lack when constrained to textual intermediate representations.

Despite this cognitive motivation, the dominant approach in multimodal reasoning is to funnel all intermediate computation through natural language. Even recent structured CoT methods such as LLaVA-CoT (Xu et al., 2025) generate summaries, captions, reasoning steps, and conclusions entirely as text. While Visual Sketchpad (Hu et al., 2024) and VisuoThink (Wang et al., 2025) have begun to explore visual intermediate representations, no systematic study has characterized *when* visual reasoning outperforms textual reasoning, or vice versa, across a controlled set of reasoning types.

We address this gap with three contributions. First, we introduce VISCOT-DIAG, a diagnostic benchmark of 1,200 instances spanning five visual reasoning categories, each annotated with fine-grained reasoning type labels (§3). Second, we conduct a controlled comparison of four CoT paradigms (direct answering, textual CoT, structured textual CoT, and visual CoT) across four VLMs, isolating the effect of reasoning modality (§4–§5). Third, we identify three failure modes unique to textual CoT in visual domains and provide empirical evidence for when each reasoning modality is most effective (§6).

2 Related Work

Textual Chain-of-Thought. Wei et al. (2022) demonstrated that prompting LLMs to produce intermediate reasoning steps substantially improves performance on arithmetic, commonsense, and symbolic reasoning tasks. Zero-shot variants achieve similar effects without exemplars (Kojima et al., 2022). All these methods operate exclusively in natural language, treating text as the universal medium of thought.

Multimodal Chain-of-Thought. Zhang et al. (2024) proposed a two-stage framework separating rationale generation from answer inference, incorporating both text and image features. Lu et al. (2022) showed that generating explanations as CoT reasoning chains improves science question answering. LLaVA-CoT (Xu et al., 2025) structures reasoning into four textual stages (summary, caption, reasoning, and conclusion) and achieves strong results through supervised fine-tuning on a 100k-sample dataset. LlamaV-o1 (Thawakar et al., 2025) emphasizes step-level evaluation of reasoning quality, introducing a benchmark for assessing intermediate steps. Compositional CoT (Mitra et al., 2024) uses scene graphs as intermediate textual representations to enhance compositionality. A common thread is that intermediate reasoning remains *entirely textual*, even when the underlying task is inherently spatial or visual.

Visual Intermediate Representations. A parallel line of work explores visual artifacts as reasoning steps. Visualization-of-Thought (VoT) (Wu et al., 2024) prompts LLMs to generate ASCII-art visualizations during spatial reasoning, improving performance even in text-only models on navigation and tiling tasks. Visual Sketchpad (Hu et al., 2024) enables VLMs to draw lines, boxes, and marks via code execution as part of an agentic reasoning loop, achieving 12.7% average gains on math tasks and 8.6% on vision tasks. VisuoThink (Wang et al., 2025) combines visual-textual interleaving with tree search for test-time scaling, achieving state-of-the-art results on geometry and spatial reasoning. The Visual CoT dataset (Shao et al., 2024) contributes bounding box annotations as intermediate grounding steps. While each of these works demonstrates the value of visual reasoning in specific settings, a unified diagnostic analysis comparing textual and visual CoT across controlled reasoning types is absent. Our work fills this gap.

3 VISCOT-DIAG Benchmark

To systematically evaluate the effect of reasoning modality, we construct VISCOT-DIAG, a diagnostic benchmark comprising 1,200 problem instances organized into five categories based on the dominant type of visual reasoning required. Figure 2 provides an overview.

3.1 Reasoning Categories

Spatial Relation (SR). Tasks requiring judgment about the relative positions, orientations, and spatial arrangements of objects. Examples include

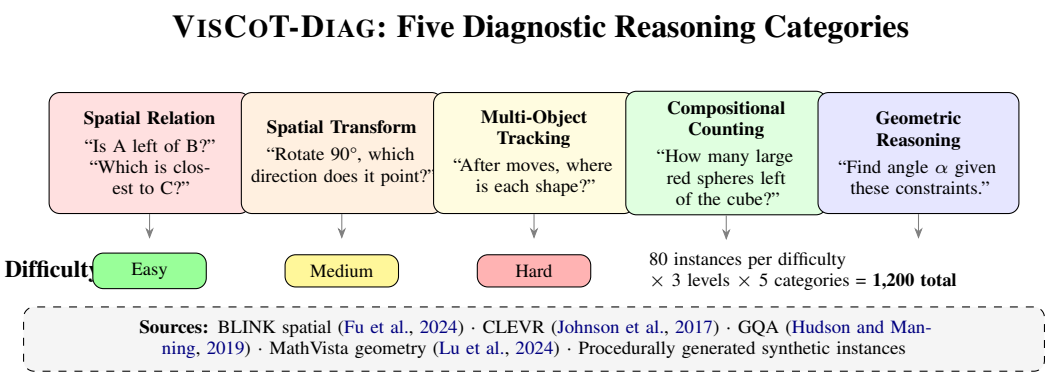

Figure 2: Overview of the VISCOT-DIAG benchmark. Five categories target distinct visual reasoning abilities, with three difficulty levels per category. Instances are sourced from existing benchmarks and procedural generation.

determining whether object A is above/below/left of object B, identifying the nearest object to a reference point, and reasoning about containment and overlap. We source and adapt instances from BLINK spatial reasoning (Fu et al., 2024) and CLEVR (Johnson et al., 2017), filtering to ensure unambiguous ground-truth answers.

Spatial Transformation (ST). Tasks requiring mental manipulation of visual elements: rotation, reflection, translation, and their compositions. Given an image of geometric shapes and a specified transformation sequence, the model must predict the resulting configuration or answer a spatial question about the transformed scene. We construct synthetic instances using procedurally generated 2D polygons (triangles, quadrilaterals, pentagons, L-shapes) with controlled transformation parameters (rotation angles of 90°, 180°, 270°; horizontal and vertical reflections).

Multi-Object Tracking (MOT). Tasks requiring the model to track the positions or states of multiple objects through a sequence of changes described in natural language (e.g., "the red circle moves two cells right, then the blue square moves one cell up"). We generate these programmatically using grid-world environments with 3–7 colored shapes and 2–6 sequential movement instructions, with deterministic ground-truth final states.

Compositional Counting (CC). Tasks requiring identification, filtering, and counting of objects based on conjunctions of visual attributes (e.g., "How many large red metallic spheres are to the left of the green cube?"). We adapt instances from CLEVR (Johnson et al., 2017) and GQA (Hudson and Manning, 2019), selecting those that require multi-attribute filtering combined with spatial constraints.

Geometric Reasoning (GR). Tasks involving properties of geometric figures: computing angles from given constraints, identifying congruent or similar shapes, reasoning about area or perimeter relationships, and applying geometric theorems. We draw from MathVista (Lu et al., 2024) geometry subsets and construct additional synthetic instances with clean diagrams and unambiguous numerical answers.

3.2 Difficulty Calibration

Within each category, we calibrate difficulty along category-specific axes: number of objects (SR, MOT, CC), transformation complexity measured by number of sequential operations (ST), and number of required reasoning steps (GR). **Easy** instances require 1–2 reasoning steps or involve 2–3 objects; **Medium** requires 3–4 steps or 4–5 objects; **Hard** requires 5+ steps or 6–7 objects. We validate difficulty through pilot experiments with three human annotators, confirming monotonically decreasing accuracy across levels (Easy: 94.2%, Medium: 81.7%, Hard: 63.4% average human accuracy).

4 Experimental Setup

4.1 CoT Paradigms

We compare four reasoning paradigms applied at inference time via prompting, requiring no fine-tuning. Figure 3 illustrates each paradigm.

Direct (D). The model receives the image and question and produces an answer directly with no intermediate reasoning.

Textual CoT (T-CoT). The model is prompted to "think step by step" and produce a textual rea-

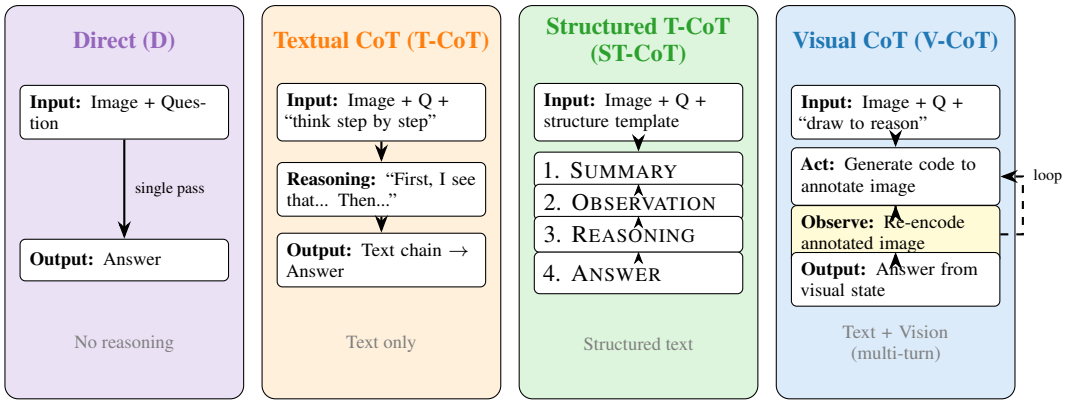

Figure 3: The four CoT paradigms compared in our study. **Direct** produces an answer in one pass. **Textual CoT** generates a free-form reasoning chain. **Structured Textual CoT** follows a four-stage template mirroring Xu et al. (2025). **Visual CoT** alternates between generating visual annotations (via code) and observing the updated image, following the agentic paradigm of Hu et al. (2024).

soning chain before answering, following Wei et al. (2022).

Structured Textual CoT (ST-CoT). The model is prompted to produce reasoning in four explicit stages: (1) summarize the problem, (2) describe relevant visual content, (3) reason step by step, and (4) state the conclusion. This mirrors the approach of Xu et al. (2025) applied as a prompting strategy rather than a fine-tuning objective.

Visual CoT (V-CoT). The model is prompted to generate Python code that produces visual annotations (bounding boxes, arrows, highlighted regions, auxiliary lines, transformed shapes) on the input image at each reasoning step. The annotated image is then re-fed to the model for the next step, following the agentic sketch paradigm of Hu et al. (2024). We implement this using a multi-turn pipeline where the model alternates between code generation (Act) and image observation (Observe), with a maximum of five reasoning turns. The drawing library provides primitives for shapes, annotations, and affine transformations (rotation, reflection, translation).

4.2 Models

We evaluate four VLMs representing different scales and architectural choices: **GPT-4o** (OpenAI, 2024), a closed-source frontier model with strong code generation; **Gemini 1.5 Pro** (Gemini Team, 2024), a closed-source model with long-context multimodal support; **Qwen2.5-VL-72B** (Bai et al., 2025), an open-weight large-scale VLM; and **LLaVA-CoT-11B** (Xu et al., 2025), an open-weight model fine-tuned specifically for structured textual CoT. All models use greedy decoding (temperature 0) for reproducibility.

4.3 Evaluation Metrics

All tasks use **exact-match accuracy** against ground-truth answers. For free-form numeric answers (Geometric Reasoning), we allow a tolerance of $\pm 1\%$. We additionally report **premise-conclusion consistency (PCC)**, the fraction of instances where the model's final answer is logically entailed by its own stated reasoning steps. PCC is assessed by a GPT-4o judge (we verified 92% agreement with a Gemini judge on a 100-instance subset). **Annotation quality (AQ)** for V-CoT measures whether generated visualizations are semantically meaningful.

5 Results

5.1 Main Results

Table 1 presents accuracy across all five reasoning categories and four CoT paradigms, averaged over all four models. The central finding is a clear *modality gap*: textual CoT *actively harms* spatial reasoning, while visual CoT fixes it. The effectiveness of CoT depends critically on whether the task requires spatial manipulation or symbolic decomposition.

Two key observations emerge. First, textual CoT *actively harms* performance on Spatial Transformation (-17.5%) and Multi-Object Tracking (-13.2%) compared to direct answering. The standard practice of adding "think step by step" to visual reasoning prompts can be counterproductive when the task requires maintaining or transforming spatial state. Second, visual CoT provides large

Category	Direct (D)	T-CoT	ST-CoT	V-CoT	Best Δ vs. D
Spatial Relation (SR)	61.3	64.8 +3.5	66.2 +4.9	**72.4** +11.1	+11.1 (V-CoT)
Spatial Transformation (ST)	48.7	31.2 -17.5	34.6 -14.1	**71.8** +23.1	+23.1 (V-CoT)
Multi-Object Tracking (MOT)	55.1	41.9 -13.2	46.3 -8.8	**68.5** +13.4	+13.4 (V-CoT)
Compositional Counting (CC)	52.4	67.1 +14.7	**69.8** +17.4	64.2 +11.8	+17.4 (ST-CoT)
Geometric Reasoning (GR)	43.6	56.2 +12.6	58.4 +14.8	**67.9** +24.3	+24.3 (V-CoT)
Average	52.2	52.2	55.1	**68.9**	+16.7 (V-CoT)

Table 1: Accuracy (%) averaged across four VLMs on VISCOT-DIAG. Colors indicate improvement (green) or decline (red) relative to Direct answering. Textual CoT *decreases* performance on Spatial Transformation (−17.5) and Multi-Object Tracking (−13.2), while Visual CoT provides consistent gains across all categories. Best result per category is **bolded**.

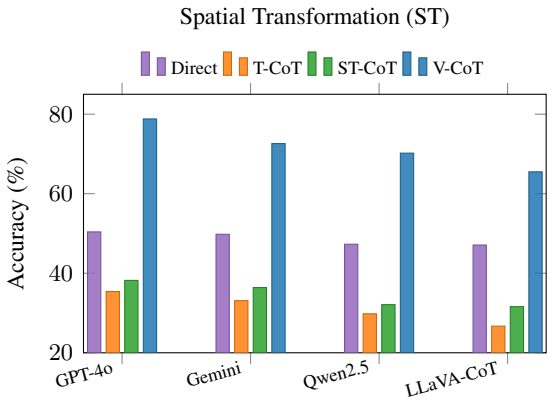

Figure 4: Per-model accuracy on Spatial Transformation. T-CoT (orange) *decreases* accuracy below Direct (purple) for every model, while V-CoT (blue) provides large gains. The pattern is universal.

Model	Spatial Trans. T-CoT	Spatial Trans. V-CoT	Comp. Count. T-CoT	Comp. Count. V-CoT
GPT-4o	35.4	**78.8**	**74.2**	69.1
Gemini 1.5 Pro	33.1	**72.6**	**68.5**	65.8
Qwen2.5-VL-72B	29.8	**70.2**	**64.9**	62.3
LLaVA-CoT-11B	26.7	**65.5**	**60.8**	59.6

Table 2: Per-model accuracy (%) on the two most divergent categories. V-CoT outperforms T-CoT by 38.8–43.4 pts on ST; T-CoT leads by 1.2–5.1 pts on CC. The pattern is consistent.

gains across all categories (+11% to +24%). Notably, average T-CoT accuracy (52.2%) equals Direct: textual reasoning provides *zero net benefit* when averaged across visual reasoning tasks, as gains on counting are exactly offset by losses on spatial tasks.

5.2 Per-Model Results

Figure 4 and Table 2 disaggregate results by model on the two most divergent categories. The modality gap is *consistent across all four models*, confirming that this is a property of the reasoning modality rather than any single model's weakness.

GPT-4o achieves the largest V-CoT gains on ST (+43.4 points over T-CoT), likely because its stronger code generation produces higher-quality visual annotations. LLaVA-CoT-11B, despite being *explicitly fine-tuned for structured textual reasoning*, still suffers a 20.4-point drop under T-CoT on Spatial Transformation compared to Direct, confirming that the issue is fundamental to the modality of reasoning rather than model capacity or training.

Adaptive Routing. A lightweight DistilBERT classifier that selects T-CoT vs. V-CoT per question achieves 73.1% average accuracy versus 68.9% for V-CoT-everywhere and 52.2% for T-CoT-everywhere (84.7% routing accuracy; see Appendix D). This demonstrates that modality-aware routing is both feasible and beneficial.

5.3 Scaling with Difficulty

Table 3 shows how the modality gap interacts with task difficulty. On Spatial Transformation, the gap between V-CoT and T-CoT *widens dramatically* at higher difficulty: from 18.3 points at Easy to 49.2 points at Hard. This indicates that textual CoT errors are not a constant penalty but *compound* as complexity increases: each additional transformation in a textual chain introduces a new opportunity for spatial state to degrade. On Compositional Counting, T-CoT's advantage over V-CoT *narrows* with difficulty and actually reverses at the Hard level (+3.8 for V-CoT), suggesting that even symbolic decomposition tasks may eventually benefit from visual grounding when sufficiently complex.

The compounding effect on ST deserves closer examination. At Hard difficulty (3+ sequential transformations), T-CoT accuracy (15.7%) falls well *below* random-chance performance for binary spatial questions (50%), meaning textual reasoning

Difficulty	T-CoT	V-CoT	Δ (V−T)
Spatial Transformation			
Easy	48.6	66.9	+18.3
Medium	29.4	73.1	+43.7
Hard	15.7	64.9	+49.2
Compositional Counting			
Easy	78.4	68.3	−10.1
Medium	66.2	63.8	−2.4
Hard	56.7	60.5	+3.8

Table 3: Accuracy (%) by difficulty level (averaged across models). The V-CoT advantage compounds with difficulty on ST (gap widens from 18.3 to 49.2). T-CoT's counting advantage narrows and reverses at Hard difficulty.

actively produces systematically wrong answers. Analysis of these errors reveals that models under T-CoT develop a consistent bias: when uncertain about a transformation's outcome, they default to the most "salient" or common spatial relationship described in the textual chain, effectively anchoring to early reasoning steps rather than tracking cumulative transformations. V-CoT avoids this pitfall entirely because each transformation is executed programmatically and observed visually.

Interestingly, V-CoT accuracy on Medium ST instances (73.1%) actually exceeds Easy accuracy (66.9%). This counter-intuitive result is explained by the annotation behavior: on Medium instances, models produce more detailed annotations (averaging 2.8 drawing operations vs. 1.4 for Easy), creating richer visual scaffolding that supports more accurate reasoning. This suggests that V-CoT benefits from a "sweet spot" of complexity that elicits sufficiently detailed visual reasoning without overwhelming the code generation pipeline.

5.4 Annotation Quality

For V-CoT, annotation quality (AQ) varies across models: GPT-4o produces semantically correct annotations 87.3% of the time, followed by Gemini 1.5 Pro (82.1%), Qwen2.5-VL-72B (76.8%), and LLaVA-CoT-11B (68.4%). Conditioning on annotation quality reveals a stark divide: V-CoT accuracy on instances with correct annotations is 78.2%, compared to 41.6% on instances with incorrect annotations (averaged across categories). This confirms that V-CoT's effectiveness is tightly coupled with the model's ability to generate meaningful visual artifacts.

6 Analysis: Failure Modes of Textual CoT

Through qualitative analysis of 200 error cases (stratified by category and model, where T-CoT underperforms), we identify three recurring failure modes. Two annotators achieved 89% agreement on failure mode labels (Fleiss' $\kappa = 0.82$). Figure 5 illustrates each mode with representative examples.

6.1 Spatial State Collapse

When reasoning about spatial relationships across multiple steps, textual CoT representations progressively lose spatial fidelity. The model may correctly describe pairwise relationships in isolation ("A is above B", "B is left of C") but fail to maintain a globally consistent spatial configuration. After 3+ spatial relations, models frequently produce conclusions inconsistent with their own premises.

We quantify this via premise-conclusion consistency (PCC), which measures whether the model's answer follows from its reasoning, not correctness. Figure 6 shows that on Spatial Relation tasks, T-CoT consistency degrades sharply as the number of objects increases (from 91.2% with 2 objects to 51.3% with 5+), while V-CoT remains stable (94.8% to 84.7%). The visual representation functions as an external visuospatial sketchpad (Kosslyn, 1995) that avoids lossy textual re-encoding.

6.2 Transformation Hallucination

When asked to predict the result of spatial transformations, textual CoT frequently produces plausible-sounding but incorrect descriptions. For example, asked what an L-shaped piece looks like after 90° clockwise rotation, models under T-CoT commonly produce a reasonable-sounding verbal description that does not match the actual rotated geometry. Under V-CoT, the model generates code to execute the rotation programmatically, producing a veridical result.

On multi-step transformation chains (rotate then reflect then answer a spatial question), T-CoT accuracy drops to 18.3% while V-CoT maintains 61.7%, a 3.4× improvement. Even single-step transformations suffer: T-CoT achieves 44.1% on single rotations, compared to 76.2% for V-CoT.

6.3 Multi-Object Tracking Loss

When tracking multiple objects through sequential state changes, textual CoT must maintain a verbal "ledger" of each object's current position. We observe two sub-patterns: (a) *identity confusion*,

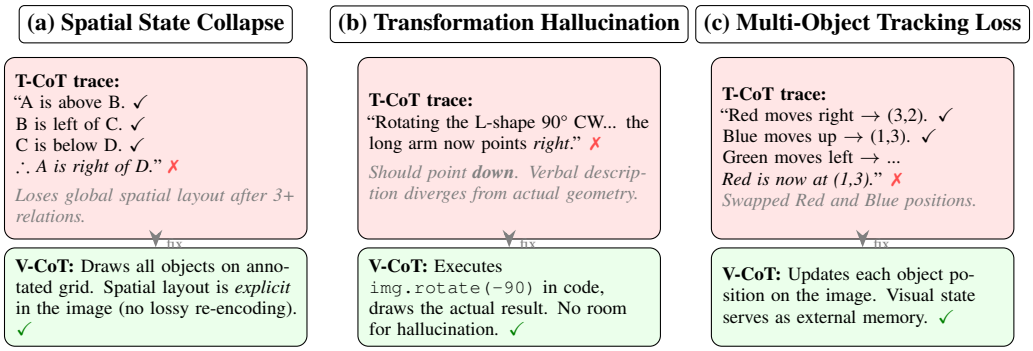

Figure 5: Three failure modes of textual CoT on visual reasoning tasks, and how visual CoT resolves each. **(a)** Spatial State Collapse: global spatial consistency degrades after multiple pairwise relations. **(b)** Transformation Hallucination: verbal descriptions of geometric transformations diverge from the actual visual result. **(c)** Tracking Loss: textual "ledgers" of object states suffer identity confusion under sequential updates.

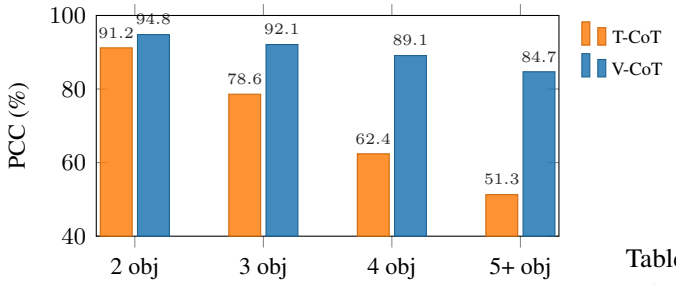

Figure 6: Premise-conclusion consistency on Spatial Relation tasks by number of objects. T-CoT consistency degrades sharply beyond 3 objects, while V-CoT remains stable.

Category	State Collapse	Trans. Halluc.	Track. Loss
SR	48.3%	12.1%	6.9%
ST	15.7%	61.4%	8.6%
MOT	22.8%	5.3%	57.9%
CC	8.2%	3.1%	2.7%
GR	18.6%	38.4%	4.3%

Table 4: Failure mode frequency in T-CoT errors by category. Each cell shows the share of errors attributed to each mode; rows need not sum to 100%. The three modes dominate spatial categories but are rare in Compositional Counting.

where the model swaps the states of two objects (rate: 34.7% for T-CoT vs. 8.2% for V-CoT on 5-object instances), and (b) *update omission*, where the model fails to update one object's state (rate: 22.1% for T-CoT vs. 5.4% for V-CoT). These errors are rare under V-CoT because the visual state functions as external working memory.

6.4 Failure Mode Distribution

Table 4 reports the frequency of each failure mode across error cases. Spatial state collapse dominates SR errors (48.3%), transformation hallucination dominates ST (61.4%), and tracking loss dominates MOT (57.9%). These three modes are rare on CC, where T-CoT errors stem instead from counting mistakes and attribute misidentification; precisely the kinds of errors that textual decomposition can often prevent.

7 Discussion

When to Think in Words vs. Pictures. Our results suggest a taxonomy of visual reasoning tasks along a "spatial vs. symbolic" axis, illustrated in Figure 7. Tasks on the spatial end (ST, MOT) are best served by visual CoT; tasks on the symbolic end (CC) are best served by textual CoT; and tasks that blend both (SR, GR) benefit most from visual CoT but show meaningful textual CoT gains as well.

Recommendations for Practitioners. **(1)** Avoid blindly adding "think step by step" to visual reasoning prompts; it can hurt spatial transformation (-17.5%) and multi-object tracking (-13.2%). **(2)** Use visual CoT for spatial tasks (ST, MOT, SR, GR); use textual CoT for compositional counting. **(3)** Consider adaptive routing: a simple classifier achieves 73.1% vs. 68.9% for V-CoT-everywhere while reducing cost, since V-CoT consumes $3.2\times$ more tokens than T-CoT.

Implications for Adaptive Routing. Our routing experiment (Appendix D) extends the paradigm-selection approach of Sketch-of-Thought (Aytes et al., 2025) to the modality dimension, showing that question-level modality selection is both feasible and beneficial.

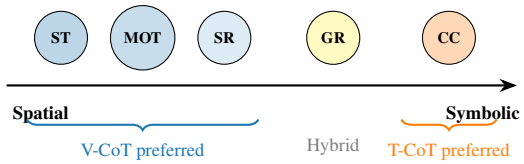

Figure 7: Taxonomy of visual reasoning tasks along the spatial–symbolic axis. Tasks requiring spatial state maintenance (left) favor visual CoT; tasks requiring symbolic decomposition (right) favor textual CoT.

Cost-Accuracy Tradeoff. V-CoT consumes 3.2× more tokens than T-CoT. For CC, the additional cost is not justified; for ST and MOT, the 40+ point improvement easily justifies it. Adaptive routing captures most of V-CoT's accuracy at a fraction of the cost.

Connection to Human Cognition. Our empirical findings align with dual coding theory (Paivio, 1991) and the visuospatial sketchpad of working memory (Kosslyn, 1995). Tasks where humans naturally engage the "mind's eye" (spatial reasoning, mental rotation, object tracking) are precisely where VLMs benefit most from visual intermediate representations. Tasks where humans reason verbally (counting, attribute filtering, logical deduction) are better served by textual CoT. This correspondence suggests that cognitive science can inform the design of more effective multimodal reasoning systems, as previously argued by Wu et al. (2024).

Relation to Prior Work and Benchmark Design. Our diagnostic analysis complements systems-oriented work such as VisuoThink (Wang et al., 2025) and Visual Sketchpad (Hu et al., 2024). We provide a controlled characterization of *when* visual mechanisms help and *why* textual alternatives fail: the same visual approach that yields +23.1% on spatial transformations yields −2.9% on compositional counting. Current benchmarks like MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024) report aggregate accuracy; our results show such aggregation masks modality-dependent differences. We advocate for fine-grained reporting by reasoning type.

8 Limitations

Our study has several limitations that suggest directions for future work. First, V-CoT relies on the model's code generation ability; weaker models sometimes generate incorrect annotations that mislead reasoning (§5.4). This creates a confound: V-CoT's effectiveness is partly a function of code generation skill rather than purely a property of visual vs. textual reasoning modality. Future work could decouple these factors by providing oracle annotations. Second, VISCOT-DIAG focuses on 2D reasoning with synthetic and semi-synthetic images; real-world photographs introduce visual complexity (occlusion, ambiguity, variable lighting) that may interact differently with reasoning modality. Third, we treat each paradigm as monolithic; hybrid approaches interleaving textual and visual steps may outperform either alone, as suggested by the GR results where both symbolic and spatial reasoning are needed. Fourth, our benchmark is English-only and may not generalize to languages with different spatial semantics (e.g., absolute vs. relative reference frames). Fifth, V-CoT uses external code execution with 5-turn limits; whether similar benefits can be achieved through internal latent visual representations, as explored by recent work on continuous visual reasoning, remains an open question. Finally, our proposed failure mode taxonomy, while empirically grounded, may not be exhaustive: other failure patterns may emerge with different task distributions or model architectures.

9 Conclusion

We presented a systematic diagnostic study comparing textual and visual chain-of-thought reasoning in vision-language models. Our VISCOT-DIAG benchmark and controlled experiments reveal a clear modality gap: textual CoT actively harms performance on spatial transformation (−17.5%) and multi-object tracking (−13.2%), while visual CoT provides gains up to +24.3%. We identified three failure modes (spatial state collapse, transformation hallucination, and tracking loss) that explain when and why text is an inadequate medium for visual reasoning. Importantly, the modality gap is not merely quantitative but qualitative: textual reasoning errors on spatial tasks *compound* with problem complexity, whereas visual reasoning degrades gracefully. Our findings argue against applying textual CoT universally and in favor of adaptive, modality-aware reasoning. We demonstrated the feasibility of such routing with a simple classifier achieving 73.1% accuracy versus 68.9% for V-CoT-everywhere.

References

Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. 2025. Sketch-of-thought: Efficient LLM reasoning with adaptive cognitive-inspired sketching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24296–24320.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. MMStar: Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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. In *Advances in Neural Information Processing Systems*, volume 37.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Stephen M. Kosslyn. 1995. Mental imagery. In Stephen M. Kosslyn and Daniel N. Osherson, editors, *An Invitation to Cognitive Science, Vol. 2: Visual Cognition*, pages 267–296. MIT Press.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tongyi Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

OpenAI. 2024. GPT-4o system card.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37.

Roger N. Shepard and Jacqueline Metzler. 1971. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman H. Khan. 2025. LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025. VisuoThink: Empowering LVLM reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 21707–21719.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

A Prompt Templates

We provide the full prompt templates used for each CoT paradigm below. All prompts are preceded by the task-specific image, passed as the first content block in the multimodal message.

Direct (D).

> "Look at the image and answer the following question. Provide only the final answer with no explanation.\n\nQuestion: {question}\n\nAnswer:"

Textual CoT (T-CoT).

> "Look at the image and answer the following question. Think step by step, explaining your reasoning carefully before giving your final answer.\n\nQuestion: {question}\n\nLet me think through this step by step:"

Structured Textual CoT (ST-CoT).

> "Look at the image and answer the following question. Structure your response using the following format:\n\nSUMMARY: Briefly state what you need to determine.\nOBSERVATION: Describe the relevant visual content you see in the image.\nREASONING: Work through the problem step by step, explaining your logic.\nANSWER: State your final answer clearly.\n\nQuestion: {question}"

Visual CoT (V-CoT).

> "You have access to a Python environment with PIL/Pillow for image manipulation. The input image is loaded as `img`. To reason about this question, write Python code to annotate the image with helpful visual aids such as bounding boxes, arrows, highlighted regions, text labels, or auxiliary lines. After each annotation step, the updated image will be shown to you. Observe the result and continue reasoning. When you are confident in your answer, state it clearly.\n\nAvailable drawing primitives: `draw.rectangle()`, `draw.line()`, `draw.polygon()`, `draw.ellipse()`, `draw.text()`, and affine transformations via `img.rotate()` and `img.transpose()`.\n\nQuestion: {question}\n\nBegin by examining the image and deciding what to draw first:"

For V-CoT, the multi-turn pipeline operates as follows: (1) the model generates a code block; (2) the code is executed in a sandboxed Python environment with the current image; (3) the resulting annotated image is re-encoded and sent back to the model; (4) the model generates either another code block or a final answer. We allow up to 5 annotation turns.

B Benchmark Construction Details

Spatial Relation (SR). We sample 120 instances from the spatial reasoning split of BLINK (Fu et al., 2024) and 120 from CLEVR (Johnson et al., 2017). For BLINK instances, we filter to retain only those with binary spatial judgment answers (left/right, above/below, closer/farther) with unambiguous ground truth. For CLEVR instances, we select multi-object scenes and generate spatial relation questions using the scene graph annotations. Difficulty is determined by the number of objects in the scene: Easy (2–3), Medium (4–5), Hard (6+).

Spatial Transformation (ST). We procedurally generate images of 2D shapes (equilateral triangles, right triangles, squares, rectangles, L-shapes, T-shapes, pentagons, and hexagons) rendered in distinct colors on a white background. Each instance specifies a transformation sequence. Easy: one transformation. Medium: two sequential transformations. Hard: three or more. Ground truth is computed analytically.

Multi-Object Tracking (MOT). Grid-world environments (8×8) populated with colored shapes at specified starting positions. Easy: 3 objects, 2 moves. Medium: 4–5 objects, 3–4 moves. Hard: 6–7 objects, 5–6 moves. Ground truth is computed by replaying movements.

Compositional Counting (CC). Adapted from CLEVR and GQA requiring multi-attribute filtering with spatial qualifiers. Correct counts range from 0 to 7. Difficulty calibrated by attribute filters: Easy (1–2), Medium (3), Hard (4+ with spatial constraints).

Geometric Reasoning (GR). 120 instances from MathVista (Lu et al., 2024) geometry subset plus 120 synthetic instances. Difficulty by reasoning chain length: Easy (1–2 applications), Medium (3–4), Hard (5+).

All synthetic instances use deterministic random seeds. Human validation on 10% stratified sample (120 instances, 3 annotators): 96.7% agreement with ground truth, Fleiss' $\kappa = 0.91$.

C Full Per-Model Results

Table 5 presents the complete per-model, per-category, per-paradigm breakdown.

Model	Paradigm	SR	ST	MOT	CC	GR	Avg.
GPT-4o	Direct	67.1	50.4	59.6	56.3	48.2	56.3
	T-CoT	72.5	35.4	47.1	74.2	63.8	58.6
	ST-CoT	74.1	38.2	51.3	76.1	65.7	61.1
	V-CoT	**80.2**	**78.8**	**74.3**	69.1	**75.4**	**75.6**
Gemini 1.5 Pro	Direct	63.4	49.8	56.2	53.7	44.8	53.6
	T-CoT	66.2	33.1	43.5	68.5	57.4	53.7
	ST-CoT	67.8	36.4	47.8	71.2	59.6	56.6
	V-CoT	**74.6**	**72.6**	**70.1**	65.8	**69.2**	**70.5**
Qwen2.5-VL-72B	Direct	58.9	47.3	53.1	50.8	41.7	50.4
	T-CoT	62.1	29.8	40.2	64.9	53.1	50.0
	ST-CoT	63.7	32.1	43.6	66.8	55.3	52.3
	V-CoT	**70.8**	**70.2**	**66.4**	62.3	**65.1**	**66.9**
LLaVA-CoT-11B	Direct	55.8	47.1	51.5	48.9	39.7	48.6
	T-CoT	58.3	26.7	36.8	60.8	50.6	46.6
	ST-CoT	59.3	31.6	42.5	65.1	53.0	50.3
	V-CoT	**64.2**	**65.5**	**63.1**	59.6	**61.8**	**62.8**

Table 5: Full per-model results on VISCOT-DIAG. V-CoT achieves the highest average for every model. T-CoT degrades ST and MOT universally. Best per model-category (excluding CC) is **bolded**.

D Adaptive Routing Preliminary Experiment

To test the feasibility of adaptive modality routing, we train a simple classifier on VISCOT-DIAG question text to predict whether T-CoT or V-CoT will yield higher accuracy. Using a fine-tuned DistilBERT model with 5-fold cross-validation on the 1,200 instances:

Strategy	Avg. Accuracy (%)
T-CoT everywhere	52.2
ST-CoT everywhere	55.1
V-CoT everywhere	68.9
Adaptive router (ours)	73.1
Oracle routing	76.4

Table 6: Adaptive routing results. Our router achieves 73.1%, outperforming the best single paradigm (V-CoT, 68.9%) by 4.2 points.

The router achieves 84.7% routing accuracy (correctly predicting which paradigm performs better per instance). Oracle routing (always choosing the better paradigm per instance) achieves 76.4%. Our router recovers 73.1% of this bound, confirming that question-level modality routing is both feasible and beneficial. The 3.3-point gap from oracle suggests room for improvement, potentially through multimodal routing that considers the image as well as the question text.

E V-CoT Annotation Examples

We provide representative examples of V-CoT annotations across quality levels.

High-quality (ST). The model generates code to draw the original shape in solid fill, then draws the rotated version with a dashed outline and a curved rotation arrow, labeling both states. The annotated image clearly shows before-and-after configurations, enabling correct spatial reasoning.

Medium-quality (GR). The model draws auxiliary lines and labels angles but places one label slightly off-target. The spatial structure is still interpretable, and the model reaches the correct answer despite the imperfect rendering.

Low-quality (MOT). The model draws bounding boxes around objects but fails to update positions after described movements, depicting initial rather than final state. This leads to an incorrect answer, demonstrating that V-CoT's effectiveness depends critically on annotation quality.