CAN IMAGE GENERATION BENEFIT VISUAL REASONING?

Anonymous authors
Paper under double-blind review

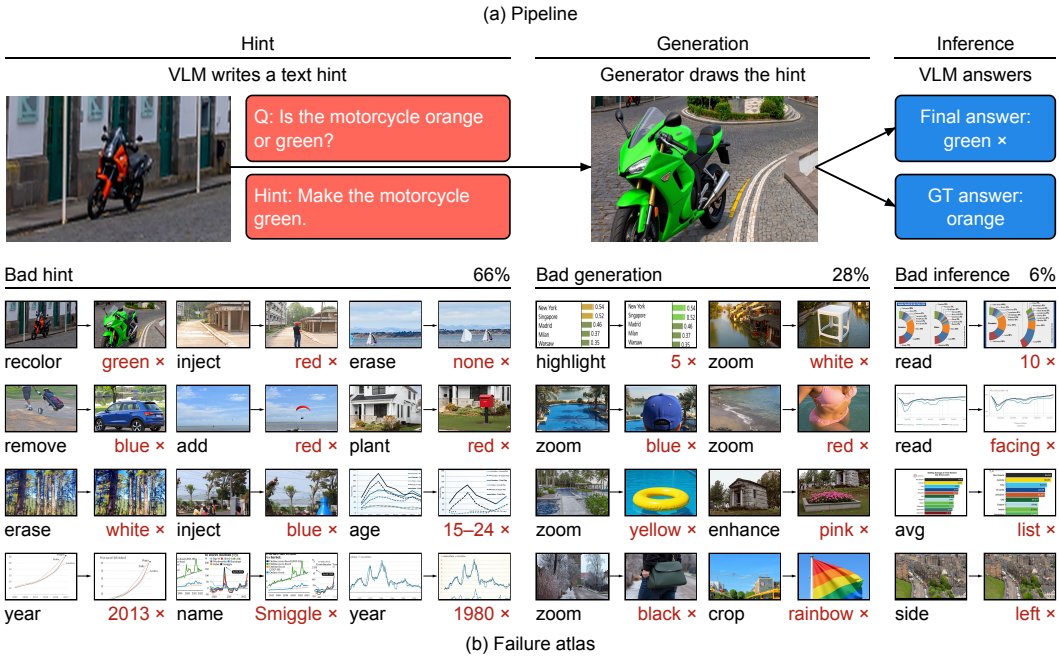


Figure 1: **Where does generation-augmented reasoning fail? (a)** The VLM first proposes a textual hint, the generator renders it as visual evidence, and the VLM then answers from the generated image. The example illustrates how an intermediate visual edit can alter task-relevant evidence and derail the final prediction. **(b)** Across annotated failures, errors are dominated by bad hints (66%) and bad generations (28%), while final inference errors are rare (6%). The atlas shows representative failures spanning incorrect edits, and wrong image attributes, and misread generated evidence.

ABSTRACT

If language models can "think in words," can multimodal models "think in images"? Recent advances in visual generation raise this possibility, but whether generated images can reliably support multimodal reasoning remains largely untested. We conduct a systematic study of visual generation for multimodal reasoning, spanning both modular VLM–generator pipelines and unified models across diverse benchmarks. We find that, in common settings, incorporating visual generation does not improve over text-only reasoning, with gains confined to limited conditions. We find two main limitations. First, current models struggle to use generated images as effective intermediate reasoning steps: visual information that appears useful does not always translate into better downstream reasoning. Second, test-time scaling mainly benefits text reasoning, while scaling visual generation brings little improvement. Overall, current VLMs remain much stronger at reasoning in text than reasoning with generated images as intermediate evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

When humans face a hard visual problem (a geometry diagram, a cluttered scene, a spatial puzzle), we instinctively sketch, annotate, or imagine alternatives to think it through. This ability to construct visual intermediate steps is a central part of how humans reason about visual information. Today's vision-language models (VLMs) (Liu et al., 2023; Wang et al., 2024b; Bai et al., 2025b) can articulate elaborate textual rationales *about* images, yet they cannot reason *in* them: even when scaled with chain-of-thought (CoT) (Shao et al., 2024; Zhang et al., 2024b; Qi et al., 2024) or reinforcement learning (RL) (Zheng et al., 2025; DeepSeek-AI, 2025; Shen et al., 2025b; Wu et al., 2025c), the reasoning chain runs entirely in text and consumes the visual input only once, yielding a brittle, single-shot trajectory that misses fine-grained semantics and environment dynamics.

Complex visual understanding remains difficult, particularly for tasks that rely on multi-step derivation, temporal and spatial dynamics, or goal execution. Addressing these difficult visual questions may require models to think in images, not just about them. A recent wave of work tries to close this gap with a *think with images* paradigm, where the visual channel is no longer a frozen input but an active participant in the reasoning chain (Xu et al., 2025b; Li et al., 2025b; Liu et al., 2025; Zhou et al., 2025). Models learn to produce and consume visual cues mid-reasoning (Bigverdi et al., 2025; Su et al., 2025b), calling on external tools (Hu et al., 2024; Su et al., 2025a; Liu et al., 2025) or image transformations (Fan et al., 2026; Zheng et al., 2025) to extract finer-grained, task-relevant perceptual signals from the existing inputs. Yet these operations only crop, zoom, or annotate existing inputs; they cannot synthesize entirely new visual hypotheses or unseen counterfactual scenes.

A few efforts (Li et al., 2025b; Xu et al., 2025b; Wu et al., 2025b; Lu et al., 2025) attempt visual generation as auxiliary evidence. Some demonstrate that generation models can directly solve some reasoning tasks (Wiedemer et al., 2025; Cai et al., 2025; Tong et al., 2025), and others (Li et al., 2025b; Wu et al., 2025b; Yang et al., 2025b) prove that given high-quality training data in simplified settings, visual-augmented reasoning is feasible and may outperform single-modality training. To examine generation-augmented multimodal reasoning more thoroughly and its ability to scale to complex scenarios, we ask: can image generation act as a unified form of visual thought in reasoning chains? And does this form of thought benefit general multimodal reasoning?

To this end, we introduce *Reasoning with Imagination*, a framework that structures inference as visual-augmented reasoning traces where each visual artifact is a generated visual thought guided by a textual hint from a reasoner model. The generated visuals encode hypotheses about the scene, diagram, or configuration, and are reinterpreted to refine subsequent textual reasoning. To our knowledge, this is among the first systematic studies of whether image generation can serve as beneficial visual thoughts for general reasoning, and whether a single model can perform end-to-end visual-augmented reasoning without external tool scaffolding. The most closely related works are MVoT (Li et al., 2025b), GoViG (Wu et al., 2025b), and ThinkMorph (Gu et al., 2025), which also present visual-augmented chains but are trained on specific and limited multimodal data. We use visual-augmented reasoning to probe whether current VLMs can effectively invoke visual generation and whether the generator produces useful evidence to support reasoning.

Concretely, we compare different multimodal reasoning patterns spanning multiple facets of QA tasks: charts, STEM, embodied understanding, logic, grounding, and fine-grained visual perception. Our study covers a diverse set of models, encompassing both modular pipelines and unified multimodal models. Our empirical findings can be summarized as follows:

- Visual generation hurts more than it helps: it degrades accuracy on most benchmarks across both modular and unified models, with reliable gains largely confined to a narrow set of spatial tasks (and an isolated STEM gain for unified models).
- Failures are cross-modal: VLMs produce under-specified hints and generators render imprecise visuals, with the dominant source of failure varying across different task types.
- Text-side scaling yields stronger overall benefits: additional reasoning improves two of three tasks, while additional visual generation fails to recover baseline accuracy.

Together, these findings suggest that stronger models and better integration alone are insufficient: the core challenge is generating precise, task-relevant visual evidence for downstream reasoning. We hope this study guides better planning, generation fidelity, and cross-modal evaluation.

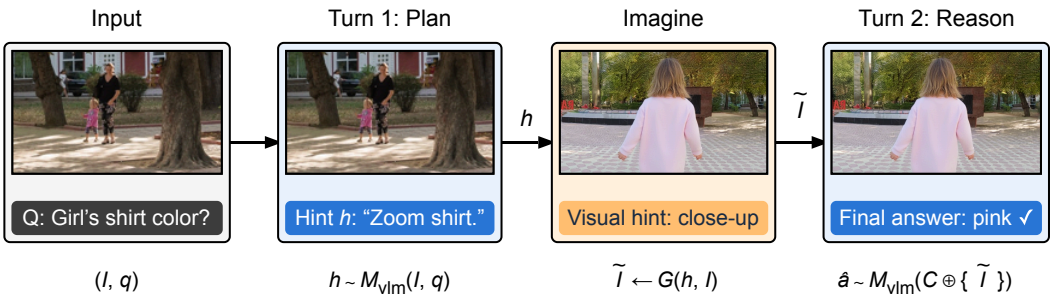


Figure 2: **Reasoning with Imagination interleaves textual planning with model-generated visual evidence to support multimodal inference.** Given input images and a question, the model first produces a textual hint describing what visual content would help. The hint conditions an external generator in the modular setting or the model's own generator in the unified setting to synthesize a helper image. In Turn 2, the model reasons over the original inputs and generated image to answer.

2 Reasoning With Imagination

We introduce *Reasoning with Imagination*, a unified formulation that incorporates auxiliary image generation into the reasoning pipeline. We present both a modular instantiation with a VLM and an external generator (Section 2.1) and a unified formulation (Section 2.2), as illustrated in Figure 2. Formally, given an input $(\mathcal{I}, q)$ with images $\mathcal{I} = \{I_1, \ldots, I_m\}$ and question q, the system uses a vision-language model $\mathcal{M}_{\text{vlm}}$ and an image generator $\mathcal{G}$ to generate and reason over visual evidence.

2.1 VLM planner + external generator

We first present a modular setting where the understanding module $\mathcal{M}_{\text{vlm}}$ and generation module $\mathcal{G}$ are two separate, single-purpose models. This setup lets us examine how textual planning and visual synthesis interact during reasoning, and isolates the contribution of generated visual evidence.

Turn 1 (planning / hint generation). We prompt $\mathcal{M}_{\text{vlm}}$ with all input images and the question to produce a short textual hint h, as in Equation (1), where h is constrained to be concise and actionable (e.g., "draw a clean bar chart highlighting category X vs Y"; "zoom into the top-left panel and annotate the legend"; "rotate the first image to the right by 30 degrees"). Prompts are in Appendix B.

$$h \sim \mathcal{M}_{\text{vlm}}(\, \mathcal{I},\, q\,), \tag{1}$$

Tool call (image generation). The generator uses the hint and input images to synthesize $\tilde{I}$:

$$\tilde{I} \leftarrow \mathcal{G}(\, h,\, \mathcal{I}\,). \tag{2}$$

In our implementation, $\mathcal{G}$ can be any image-to-image or conditional generator; the key requirement is that it returns an image artifact that can be re-inserted into the VLM context. Unless otherwise stated, we use a single generated image; experiments with multiple images are reported in Section 4.2.

Turn 2 (answering with generated evidence). Let $\mathcal{C}$ denote the multimodal conversation including the question, input images, and generation hint produced by the VLM. We append the generated image to the conversation, and the VLM reasons over the augmented input to produce the final answer $\hat{a}$:

$$\hat{a} \sim \mathcal{M}_{\text{vlm}}\Big(\, \mathcal{C} \oplus \{\tilde{I}\}\, \Big). \tag{3}$$

2.2 Visual-Augmented Unified Model

In the unified setting, a single model $\mathcal{M}$ serves as both $\mathcal{M}_{\text{vlm}}$ and $\mathcal{G}$. The pipeline of Section 2.1 is unchanged, except that the model generates internally in place of the external tool call in (2):

$$z \sim \mathcal{M}(\, h,\, \mathcal{I}\,), \qquad \tilde{I} = \mathcal{D}(z), \tag{4}$$

where z is a sequence of image tokens (or a latent representation) decoded by a fixed renderer $\mathcal{D}$. Planning (1) and answering (3) remain identical, but are now executed by the same backbone, removing cross-model coordination and enabling end-to-end training over when and how to imagine.

VLM + Generator	ChartQAPro	VisuLogic	VisualPuzzles	VSTAR	Rover-TG
Gemini	**66.4**	**36.4**	**46.2**	**88.0**	**41.9**
+ *Nano-Banana*	58.1 (−8.3)	34.1 (−2.3)	34.9 (−11.3)	71.7 (−16.3)	22.1 (−19.8)
Qwen3-VL 4B	**35.6**	**26.3**	22.1	**81.3**	**31.9**
+ *SD3.5*	33.6 (−2.0)	25.4 (−0.9)	23.3 (+1.2)	63.4 (−17.9)	18.3 (−13.6)
+ *FLUX.1*	35.5 (−0.1)	26.0 (−0.3)	22.7 (+0.6)	77.5 (−3.8)	18.3 (−13.6)
+ *Qwen-Image-Edit*	33.2 (−2.4)	24.7 (−1.6)	21.9 (−0.2)	50.8 (−30.5)	16.0 (−15.9)
+ *Nano-Banana*	35.2 (−0.4)	24.9 (−1.4)	**23.4** (+1.3)	59.7 (−21.6)	21.5 (−10.4)
Qwen3-VL 8B	**40.8**	25.6	23.1	**81.6**	22.9
+ *SD3.5*	38.6 (−2.2)	24.4 (−1.2)	23.4 (+0.3)	54.5 (−27.1)	17.4 (−5.5)
+ *FLUX.1*	38.7 (−2.1)	25.1 (−0.5)	22.4 (−0.7)	69.6 (−12.0)	18.9 (−4.0)
+ *Qwen-Image-Edit*	38.3 (−2.5)	25.4 (−0.2)	23.4 (+0.3)	50.8 (−30.8)	22.1 (−0.8)
+ *Nano-Banana*	40.4 (−0.4)	**26.4** (+0.8)	**25.0** (+1.9)	58.3 (−23.3)	**24.1** (+1.2)
GLM-4.6V Flash	**60.5**	**25.9**	**37.0**	**86.2**	**33.4**
+ *SD3.5*	56.9 (−3.6)	24.8 (−1.1)	34.9 (−2.1)	60.7 (−25.5)	28.8 (−4.6)
+ *FLUX.1*	58.9 (−1.6)	25.1 (−0.8)	34.2 (−2.8)	82.2 (−4.0)	29.1 (−4.3)
+ *Qwen-Image-Edit*	56.7 (−3.8)	25.0 (−0.9)	33.9 (−3.1)	54.9 (−31.3)	32.1 (−1.3)
+ *Nano-Banana*	58.9 (−1.6)	25.4 (−0.5)	35.1 (−1.9)	57.9 (−28.3)	32.5 (−0.9)
Unified Model					
BAGEL	**29.4**	**41.3**	**26.2**	**67.0**	**25.6**
+ *Generation*	26.0 (−3.4)	39.3 (−2.0)	25.2 (−1.0)	47.6 (−19.4)	20.6 (−5.0)

Table 1: Visual generation generally degrades performance across chart, logic, visual-search, and cross-modal reasoning tasks. Across modular VLM–generator pipelines, the degradation is particularly consistent on ChartQA-Pro, VSTAR, and Rover-TG, while only a few small gains appear on VisuLogic and VisualPuzzles. The unified BAGEL model exhibits the same overall trend. Bold: group-best accuracy (ties included). Parentheses: changes from the corresponding text-only baseline in percentage points, with increases/decreases in green/red.

3 Visual Generation Augmented Reasoning Falls Short

3.1 Experimental Setup

Model Selection. We evaluate Qwen3-VL (Bai et al., 2025a) at 4B and 8B, GLM-4.6V Flash (GLM-V Team et al., 2025), and Gemini 3.0 Flash (Google DeepMind, 2025b) as VLM backbones, paired with Stable Diffusion 3.5 Large (Esser et al., 2024), FLUX.1-Kontext Dev (Labs, 2024), Qwen-Image-Edit-2511 (Wu et al., 2025a), and Nano-Banana (Google DeepMind, 2025a) as generators (covering open-source and proprietary paradigms). We additionally include BAGEL (Deng et al., 2025) as a representative unified model that jointly handles perception and generation. All models use non-thinking mode unless otherwise specified.

Benchmark Selection. We evaluate chart understanding (ChartQAPro (Masry et al., 2025)), visual search (VSTAR (Wu & Xie, 2024)), visual logic (VisuLogic (Xu et al., 2025a), VisualPuzzles (Song et al., 2025b)), embodied and spatial reasoning (SAT (Ray et al., 2024), RoboSpatial (Song et al., 2025a), ERQA (Team et al., 2025)), STEM problem solving (MathVista-mini (Lu et al., 2024)), and cross-modality reasoning (Rover-TG (Liang et al., 2025)). We exclude Rover-TG's spatial-context task and ChartQAPro's paragraph-context task.

Evaluation Protocol. We report final accuracy under unified prompts (Appendix B), using rule-based matching with official criteria for deterministic and multiple-choice formats, and Gemini-2.5-Flash (Comanici et al., 2025) as an LLM judge for extended reasoning outputs not amenable to rule-based matching. Implementation and judge details are in Appendix A and Appendix B.

VLM + Generator	MathVista	SAT	ERQA	RS-Compat.	RS-Config.
Gemini	**78.6**	**79.3**	**66.8**	**76.2**	**82.9**
+ *Nano-Banana*	74.5 (−4.1)	76.7 (−2.6)	64.3 (−2.5)	**76.2** (0.0)	81.0 (−1.9)
Qwen3-VL 4B	**62.3**	**66.3**	**44.3**	45.7	**80.8**
+ *SD3.5*	53.3 (−9.0)	64.7 (−1.6)	41.8 (−2.5)	68.6 (+22.9)	67.5 (−13.3)
+ *FLUX.1*	58.3 (−4.0)	65.3 (−1.0)	43.5 (−0.8)	65.7 (+20.0)	68.3 (−12.5)
+ *Qwen-Image-Edit*	48.8 (−13.5)	60.0 (−6.3)	38.5 (−5.8)	**69.5** (+23.8)	75.6 (−5.2)
+ *Nano-Banana*	57.6 (−4.7)	61.4 (−4.9)	40.0 (−4.3)	**69.5** (+23.8)	75.8 (−5.0)
Qwen3-VL 8B	**61.2**	**70.1**	**44.3**	61.6	80.4
+ *SD3.5*	52.4 (−8.8)	64.3 (−5.8)	40.0 (−4.3)	**77.1** (+15.5)	77.2 (−3.2)
+ *FLUX.1*	57.3 (−3.9)	65.4 (−4.7)	42.8 (−1.5)	70.5 (+8.9)	**81.3** (+0.9)
+ *Qwen-Image-Edit*	56.9 (−4.3)	61.6 (−8.5)	38.5 (−5.8)	**77.1** (+15.5)	75.6 (−4.8)
+ *Nano-Banana*	60.7 (−0.5)	61.8 (−8.3)	40.0 (−4.3)	72.4 (+10.8)	79.4 (−1.0)
GLM-4.6V Flash	**83.3**	65.8	**47.8**	66.7	**79.8**
+ *SD3.5*	79.9 (−3.4)	65.3 (−0.5)	37.3 (−10.5)	**79.1** (+12.4)	76.4 (−3.4)
+ *FLUX.1*	80.1 (−3.2)	64.0 (−1.8)	43.0 (−4.8)	78.1 (+11.4)	79.7 (−0.1)
+ *Qwen-Image-Edit*	77.0 (−6.3)	**68.0** (+2.2)	43.0 (−4.8)	67.6 (+0.9)	78.8 (−1.0)
+ *Nano-Banana*	80.1 (−3.2)	**68.0** (+2.2)	43.2 (−4.6)	71.4 (+4.7)	76.4 (−3.4)
Unified Model					
BAGEL	41.8	**63.3**	**43.2**	46.7	**82.1**
+ *Generation*	**49.9** (+8.1)	54.7 (−8.6)	40.8 (−2.4)	**72.4** (+25.7)	73.2 (−8.9)

Table 2: Visual generation improves only a narrow subset of reasoning settings. Across modular VLM–generator pipelines, generation generally reduces accuracy on MathVista, SAT, ERQA, and RoboSpatial-Configuration, while RoboSpatial-Compatibility shows the most consistent gains. The unified BAGEL model exhibits a similar pattern, with additional improvement on MathVista. Bold: group-best accuracy (ties included). Parentheses: changes from the corresponding text-only baseline in percentage points, with increases/decreases in green/red.

3.2 Results

Table 1 and Table 2 summarize results. For modular VLM-generator pipelines, generation degrades the majority of benchmarks across all four backbones (Qwen3-VL 4B/8B (Bai et al., 2025a), GLM-4.6V Flash (GLM-V Team et al., 2025), Gemini 3.0 Flash (Google DeepMind, 2025b)) and all four generators (SD3.5 (Esser et al., 2024), FLUX.1 (Labs, 2024), Qwen-Image-Edit (Wu et al., 2025a), Nano-Banana (Google DeepMind, 2025a)) in the tested pairings. The failure is systematic rather than incidental. Degradation is sharpest where fine-grained or structured visual information is required: VSTAR (Wu & Xie, 2024) drops by up to 31 points, MathVista (Lu et al., 2024) by up to 13, and ChartQAPro (Masry et al., 2025) by up to 8; VisuLogic (Xu et al., 2025a) and Rover-TG (Liang et al., 2025) also predominantly decline. The proprietary Nano-Banana exhibits the same overall degradation, reinforcing that the failure is not generator-specific. Broadly consistent gains are confined to *RoboSpatial-Compatibility* (Song et al., 2025a) (up to +23.8 points across modular pipelines), where qualitative spatial structure tolerates approximate imagery.

Unified models show a mixed pattern. BAGEL (Deng et al., 2025) gains on MathVista (+8.1) and RoboSpatial-Compatibility (+25.7), but drops on VSTAR (−19.4), SAT (−8.6), RoboSpatial-Configuration (−8.9), Rover-TG (−5.0), and 1–3.4 points elsewhere. The MathVista gain (where all modular pipelines fail) suggests co-trained generation can aid STEM reasoning, yet persistent drops on spatial and visual-search tasks show architectural integration alone cannot produce visuals precise enough for fine-grained perception. Qualitative examples appear in Figure 4.

Model / Variant	Jigsaw	Spatial	Maze	Multi-hop
Qwen3-VL 8B (Text)	47.3	42.1	44.7	65.4
+ *Controlled Prompt*	48.0	44.3	–	–
+ *Oracle Visual*	49.8	–	49.8	67.5
+ *Visual-Augmented*	46.2	37.8	40.0	63.3

(a) Controlled and oracle interventions

Condition	ChartQA Pro	MathVista	ERQA
Text-only	40.8	61.2	44.3
+ *Visual Aug.*	38.7	57.3	42.8
Text-only + Thinking	56.6	63.7	36.7
+ *Visual Aug.*	51.2	60.8	28.8

(b) Thinking ablation

Table 3: **(a)** Controlled and oracle interventions reveal positive headroom, whereas standard visual augmentation underperforms text-only inference. We compare Qwen3-VL 8B under text-only inference, controlled generation prompts, oracle helper images, and the standard visual-augmented pipeline. Controlled or oracle interventions improve performance where available, while Visual-Augmented degrades performance on all four tasks. Accuracy (%); generator: Nano-Banana. **(b)** Thinking does not eliminate the visual-augmentation gap. Visual augmentation reduces accuracy both with and without explicit thinking, with the gap widening on ChartQAPro and ERQA under thinking. Accuracy (%) is reported for Qwen3-VL 8B with FLUX.1-Kontext. Red cells indicate degradation relative to the corresponding text-only condition.

4 Why Does Generation Fail to Help?

We conduct two targeted analyses: Section 4.1 probes a potential upper bound under idealized conditions, and Section 4.2 contrasts test-time scaling along text and vision axes.

4.1 Benefits Under Oracle Conditions

To probe whether visual augmentation *can* help once alignment issues are removed, we evaluate two oracle conditions on controlled benchmarks (Ray et al., 2024; Liang et al., 2025; Li et al., 2025a; Hao et al., 2025; Shi et al., 2025b) that supply ground-truth visual targets and accurate textual descriptions: (i) *oracle visual*, injecting ground-truth helper images directly to remove generation noise; and (ii) *oracle hint*, replacing model hints with task-specific hard-coded prompts that drive the generator. Benchmark choice and prompt designs are detailed in Appendix A.

Table 3a (Qwen3-VL 8B with Nano-Banana) shows both oracle conditions consistently improve over the text-only baseline, by roughly $+2\%$ with controlled prompts and $+3\%$ with ground-truth visuals. The potential is therefore real: when intermediate visual content is accurate or textual signals are well aligned, multimodal reasoning beats text-only. The bottleneck of current pipelines lies in the cross-modal interaction itself. Textual and visual contents produced by current models are often poorly aligned and not mutually interpretable.

4.2 Text Scales; Vision Does Not

A second explanation is that extra test-time compute allocated to language reasoning yields more reliable accuracy gains than scaling visual generation does. To test this, we compare two complementary scaling axes: (i) *text scaling*, increasing reasoning tokens via standard chain-of-thought prompting, and (ii) *visual scaling*, increasing generator compute by allowing multiple generation calls per instance, with all other components and decoding hyperparameters held fixed.

Figure 3 shows that text-side scaling yields stronger benefits overall: additional reasoning improves ChartQAPro and MathVista, although it reduces accuracy on ERQA. Increasing visual generation from one to two calls, meanwhile, leaves all three benchmarks below the text-only baseline. Together with the analysis in Section 4.1, these results indicate that the evaluated pipelines benefit more from added textual reasoning than from added visual generation: simply allocating more compute to the generation pipeline is insufficient to systematically improve multimodal reasoning performance.

4.2.1 Does Thinking Help?

We test whether enabling explicit *thinking* mode (extended chain-of-thought before answering) closes the visual-augmentation gap, comparing four conditions on Qwen3-VL 8B: text-only and visual-augmented, each with and without thinking, across all three benchmarks in this study.

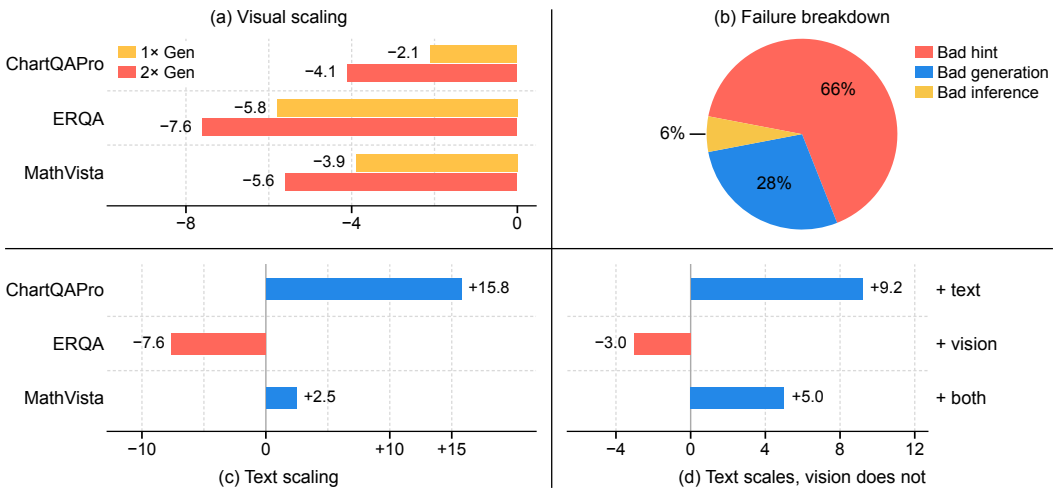


Figure 3: Comparison between text-side and visual-side test-time scaling. Increasing visual generation compute consistently reduces accuracy across all three benchmarks, while additional text reasoning yields task-dependent gains. Values in panels (a), (c), and (d) show changes in accuracy relative to the text-only baseline, measured in percentage points (pp).

Table 3b shows that thinking helps text-only performance on tasks admitting step-by-step deliberation (ChartQAPro, MathVista) but hurts perception-driven ERQA. Crucially, thinking does not eliminate the visual-augmentation penalty on any of the three benchmarks: the gap widens on ChartQAPro and ERQA, while narrowing on MathVista. This persistent penalty indicates that stronger textual deliberation alone is insufficient to make generated images consistently useful, pointing to cross-modal alignment as a remaining obstacle. Details are in Appendix C.6.

5 Additional Analysis

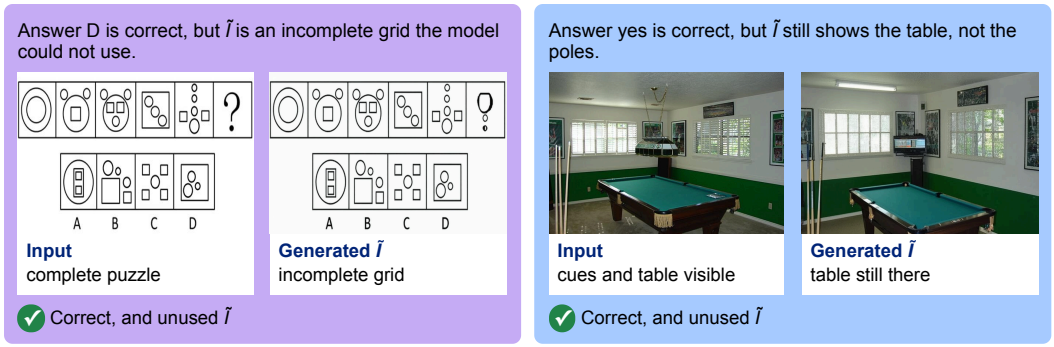
5.1 Forced vs. Selective Generation

To distinguish extra-token effects from a deeper reasoning mismatch, we compare two prompting variants: *Forced-Gen* always triggers generation, while *Selective-Gen* lets the VLM decide whether an auxiliary image would help. We also report the selective invocation rate as a proxy for perceived usefulness, holding all other components fixed.

Figure 5 shows that *Selective-Gen* invokes generation sparingly (roughly 2–12% across benchmarks), with accuracy close to the text-only baseline. *Forced-Gen* leaves ChartQAPro near-flat but degrades VSTAR by up to 12 points. Low voluntary invocation suggests VLMs recognize when generation is unlikely to help. Consistent with Section 4, the bottleneck lies in the mismatch between generated content and task requirements, rather than invocation policy.

5.2 Qualitative Case Studies

Figure 4 summarizes representative successes and failures. For success cases, despite the correctness of the final answer, intermediate visual content is often not the primary source of reasoning success and gains plausibly come from longer conversations or textual randomness. Failures arise on both sides: VLMs often produce confusing generation hints that drift the reasoning, while generators fail in two regimes, (i) *complex manipulation* requiring precise transformations under constraints, and (ii) *structured tasks* requiring visually consistent diagrams or exact configurations. These examples support the conclusion that current VLMs are text-centric and that general cross-modality reasoning requires joint advances in both understanding and generation. Full discussion in Appendix C.5.



(a) Correct answers, unused helpers. When the final answer is right, $\tilde{I}$ is still imprecise. The generated image is not what the model needed to answer.



(b) Wrong answers, unusable helpers. When the chain fails, the generated image is the thing that is wrong: it misses the asked region or destroys the evidence that was already in the input.

Figure 4: **Case study of generation-augmented reasoning.** **(a)** Examples where such reasoning chains produce accurate final answers (row 1). **(b)** Examples where they fail due to either text prompts or visual generation (rows 2 and 3).

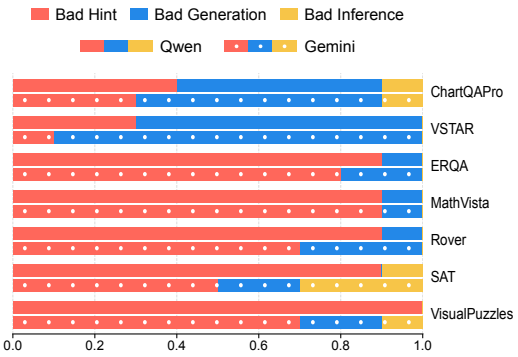
5.3 Quantitative Failure Attribution

We decompose pipeline errors into (i) *bad hint*, (ii) *bad generation*, and (iii) *bad inference*, manually annotating 10–20 failed instances per benchmark. Figure 5 reveals a task-level split: chart and visual-search tasks (ChartQAPro, VSTAR) produce plausible hints but the generator fails to preserve fine-grained details, shifting errors toward generation; spatial, embodied, puzzle, and STEM tasks (ERQA, MathVista, SAT, VisualPuzzles, Rover) are dominated by planning failures, with generic hints like *"crop"* or *"zoom."* Pure inference failures are negligible.

Compact the Figure 5-to-text gap locally through the conclusion. Improving either component alone is insufficient: precise hints and faithful generations are both required, with their relative importance varying by task. Per-model statistics are in Appendix C.4. We also study task difficulty (Appendix C.1) and multi-turn vs. KV-cache inference (Appendix C.3).

Model / Variant	ChartQAPro		VSTAR	
	Acc.	Call (%)	Acc.	Call (%)
Qwen3-VL 4B (Text)	35.6	0.0	81.3	0.0
+ *Forced-Gen*	35.5	100.0	77.5	100.0
+ *Selective-Gen*	35.6	**9.8**	80.6	**4.0**
Qwen3-VL 8B (Text)	40.8	0.0	81.6	0.0
+ *Forced-Gen*	38.7	100.0	69.6	100.0
+ *Selective-Gen*	40.2	**11.5**	81.6	**2.0**

(a) Generation policy ablation



(b) Failure attribution

Figure 5: **Generation is rarely selected and fails for complementary reasons.** **(a)** Qwen3-VL invokes the generator on fewer than 12% of examples under *Selective-Gen*, while *Forced-Gen* consistently reduces accuracy, especially on VSTAR. **(b)** Failures arise from bad hints (coral), bad generations (blue), and bad inference (yellow); planning errors dominate spatial and embodied tasks, while generation errors are more prominent on chart and perception tasks.

6 Related Work

Multimodal reasoning. Following advances in LLMs and large reasoning models (Wei et al., 2023; DeepSeek-AI, 2025; Shen et al., 2025b), VLMs acquire reasoning ability via CoT decomposition (Shao et al., 2024; Wang et al., 2025b) and reinforcement finetuning (Shen et al., 2025a; Peng et al., 2025; Sarch et al., 2025). Recent work (Xu et al., 2025b; Li et al., 2025b; Wu et al., 2025b) treats images as active reasoning components, extracting task-centric signals such as bounding boxes, depth, and masks (Hu et al., 2024; Shao et al., 2024; Liu et al., 2025; Fu et al., 2025) via SFT (Zhang et al., 2024a) or RFT (Zheng et al., 2025; Su et al., 2025b). Visual thoughts are typically sourced from tool usage (Fan et al., 2026; Hu et al., 2024; Shao et al., 2024; Zheng et al., 2025; Su et al., 2025b) or code execution (Vinker et al., 2024; Zhang et al., 2024a; Liu et al., 2025; Fu et al., 2025), both limited by tool coverage and the original input. Only a few works (Xu et al., 2025b; Li et al., 2025b; Wu et al., 2025b; Gu et al., 2025) explore image generation as an alternative source, and only on narrow domains; whether models can reliably invoke and consume generated visuals at scale remains open.

Unified understanding and generation. Unified multimodal models jointly handle understanding and generation in a single architecture (Tong et al., 2024; Wu et al., 2024b; Ma et al., 2025; Chen et al., 2025), spanning autoregressive (Dong et al., 2023; Team, 2024; Chern et al., 2024; Wu et al., 2024a; Wang et al., 2024a; Pan et al., 2025; Xie et al., 2025), diffusion-based (Yang et al., 2025a; Wang et al., 2025a; Swerdlow et al., 2025; Shi et al., 2025a), and hybrid (Zhou et al., 2024; Xie et al., 2024; Ma et al., 2024; Deng et al., 2025; Liao et al., 2025) designs. Building on LLM post-training (DeepSeek-AI, 2025), recent efforts apply CoT and RL to unified models for image editing and reasoning generation (Duan et al., 2025; Chern et al., 2025; Jiang et al., 2025; Xiao et al., 2025), showing gains when thinking in text before generating in vision. ThinkMorph (Gu et al., 2025) extends this to visual-augmented reasoning via specialized data, yet such data is hard to curate at scale, and whether generation aids general perception and understanding remains unknown.

7 Conclusion

We presented the first systematic evaluation of generation-augmented multimodal reasoning, identifying two barriers: *cross-modal incompatibility*, where planners and generators fail to produce mutually usable signals, and *language bias*, where added textual reasoning yields stronger overall benefits than added visual generation. Beyond static-image QA, future directions include video and world models for dynamics and counterfactuals; generated evidence for planning, agentic execution, and open-ended generation; and reasoning-specialist VLMs with longer text–image trajectories.

AI USE STATEMENT

We used AI to assist with manuscript revision, reference checking, figure and table format adjustment, and associated plotting and LaTeX code. It also provided feedback on methodological descriptions and the interpretation of reported results. Separately, generative models serve as experimental components of our framework, and Gemini-2.5-Flash is used for LLM-based evaluation, as described in Appendix A, Appendix B, and Appendix C.2.

ETHICS STATEMENT

This study evaluates existing models on publicly available benchmarks. Dataset sources and available licensing information are documented in Appendix F. A central concern is that generated images can alter or fabricate task-relevant evidence while remaining visually plausible, potentially misleading downstream reasoning. Our findings highlight the need to verify generated visual evidence before relying on it, particularly in safety-critical applications. We discuss these risks and broader implications in Appendix E, and the scope and limitations of our evaluation in Appendix D.

REPRODUCIBILITY STATEMENT

We describe the modular and unified inference pipelines in Section 2.1 and Section 2.2, and the evaluated models, benchmarks, and evaluation protocol in Section 3.1. Implementation details, decoding settings, and benchmark adjustments are provided in Appendix A, with evaluation and generation prompts in Appendix B. Additional results and methodological details are reported in Appendix C, including judge reliability, inference implementations, failure attribution, and thinking-mode comparisons. Dataset and model sources are listed in Appendix F.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025.

Zefan Cai, Haoyi Qiu, Tianyi Ma, Haozhe Zhao, Gengze Zhou, Kung-Hsiang Huang, Parisa Kordjamshidi, Minjia Zhang, Wen Xiao, Jiuxiang Gu, Nanyun Peng, and Junjie Hu. MMGR: Multimodal generative reasoning, 2025. URL https://arxiv.org/abs/2512.14691v2.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images, 2026. URL https://arxiv.org/abs/2505.15879.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006. Model card: https://huggingface.co/zai-org/GLM-4.6V-Flash.

Google DeepMind. Gemini 2.5 Flash: Model card, December 2025a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf. Native audio and image capabilities.

Google DeepMind. Gemini 3 Flash: Model card, December 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

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

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

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

Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation. *arXiv preprint arXiv:2511.01163*, 2025.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025.

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

TaiMing Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Genex: Generating an explorable world. In *ICLR*, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024.

Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2403.16999.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025a.

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search, 2025b. URL https://arxiv.org/abs/2502.02508.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv:2505.23606*, 2025a.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Chaoyou Fu, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark, 2025b. URL https://arxiv.org/abs/2509.24897.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. Oral Presentation.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025b.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv:2503.20853*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*, 2025.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. Sketchagent: Language-driven sequential sketch generation. *arXiv preprint arXiv:2411.17673*, 2024.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024a.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *arXiv:2505.20147*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Fengyi Wu, Yifei Dong, Zhi-Qi Cheng, Yilong Dai, Guangyu Chen, Hang Wang, Qi Dai, and Alexander G Hauptmann. Govig: Goal-conditioned visual navigation instruction generation. *arXiv preprint arXiv:2508.09547*, 2025b.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025c.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024a.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *arXiv preprint arXiv:2505.13031*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

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

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

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025a.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025b.

Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Silvio Savarese, et al. Provision: Programmatically scaling vision-centric instruction data for multimodal language models. *arXiv preprint arXiv:2412.07012*, 2024a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024b. URL https://arxiv.org/abs/2302.00923.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Yejin Choi, James Zou, Chaorui Deng, Shen Yan, Haoqi Fan, Cihang Xie, Huaxiu Yao, and Qinghao Ye. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought, 2025. URL https://arxiv.org/abs/2511.02779.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

APPENDIX

This appendix provides additional methodology details, evaluation protocols, and per-benchmark analyses supporting the main paper:

- § A details the modular VLM–generator pipeline and unified-model inference setup, including decoding hyperparameters and benchmark adjustments.
- § B lists all prompts used for text-only inference, generation-hint planning, visual-augmented reasoning, and the LLM judge.
- § C reports additional experiments: generator capability stratified by difficulty, judge reliability against rule-based and human evaluation, multi-turn vs. KV-cache inference, and detailed breakdowns of failure attribution, qualitative case studies, and thinking-mode results.
- § D discusses the limitations of our study and adjacent paradigms left to future work.
- § E considers broader societal impacts of generation-augmented multimodal reasoning.
- § F lists all dataset licenses, source URLs, and terms of use.

A Implementation Details
B Evaluation Prompts
C Additional Experiments
C.1 Generator Capability
C.2 Judge Reliability
C.3 Multi-Turn vs. KV-Cache Inference
C.4 Detailed Failure Attribution
C.5 Detailed Qualitative Case Studies
C.6 Detailed Thinking-Mode Results
D Limitations
E Broader Impact
F Dataset Licenses and Asset Credits

A IMPLEMENTATION DETAILS

This section summarizes evaluation settings for both modular VLM–generator pipelines and unified multimodal models.

Text-only and unified model inference. For all VLM evaluations, we fix the random seed for reproducibility. Unless otherwise specified, decoding uses a temperature of 0.1, top-p sampling with $p = 1.0$, and a maximum token budget of 8192, which is sufficient to accommodate the longer reasoning chains required by benchmarks such as MathVista and VisualPuzzles without truncation.

Reasoning with Imagination Framework. In the modular VLM–generator setting, inference proceeds in two stages. In **Turn 1**, the VLM produces a generation hint describing the desired auxiliary visual content. We use a temperature of 0.1, top-$p = 1.0$, and a token budget of 2048 for this stage. The hint is then passed to an external image generator to produce a helper image.

For Stable Diffusion 3.5, we use 28 denoising steps, guidance scale 5.0, and strength 0.6 for image-to-image generation. For FLUX.1-Kontext, we use 28 denoising steps, guidance scale 2.5, and CFG scale 1.0. In **Turn 2**, the VLM receives both the original inputs and the generated helper image and produces the final answer using the same decoding settings as in text-only inference. Detailed hyper-parameter settings are provided in Table 4.

For unified models, we follow the default inference hyperparameters provided in their official repositories.

Component	Parameter	Value
Text Inference (VLM / Unified)	Temperature	0.1
	Top-p	1.0
	Max tokens	8192
Turn 1 Hint Generation	Temperature	0.1
	Top-p	1.0
	Max tokens	2048
Stable Diffusion 3.5	Denoising steps	28
	Guidance scale	5.0
	Strength	0.6
FLUX.1-Kontext	Denoising steps	28
	Guidance scale	2.5
	CFG scale	1.0

Table 4: Inference hyperparameters used in our evaluation.

Benchmark adjustments. To reduce ambiguity in evaluation and focus on reasoning-centric tasks, we use the Compatibility and Configuration splits from RoboSpatial and exclude the embodied-interaction tasks from Rover-Bench, as these settings require environment dynamics or action feedback that are not compatible with our static generation framework.

Oracle-Conditions Setup (Section 4.1). We use a small set of controlled benchmarks covering visual perception, spatial relations, maze navigation, and multi-hop counting (Ray et al., 2024; Liang et al., 2025; Li et al., 2025a; Hao et al., 2025; Shi et al., 2025b), chosen because they expose ground-truth visual targets and accurate textual descriptions (either provided by the benchmark (Shi et al., 2025b; Liang et al., 2025; Zou et al., 2025) or constructed task-specifically). The two settings are: (i) injecting ground-truth helper images directly into the VLM input for maze navigation and multi-hop reasoning, isolating reasoning capability when generation noise is removed; and (ii) replacing model-generated hints with task-specific, hard-coded prompts that accurately describe the required visual content, controlling for prompt quality while retaining the generation step.

Test-Time Scaling Setup (Section 4.2). *Text scaling* increases reasoning tokens by applying standard chain-of-thought at inference time. *Visual scaling* increases vision-side compute by allowing multiple generator calls per instance. The underlying VLM, evaluation protocol, and decoding hyperparameters are kept identical across both axes; we vary only the scaling dimension under study.

B Evaluation Prompts

Text-only Direct prompting. We do not add additional instructions unless otherwise specified. A typical prompt follows the format:

> You are given a visual reasoning problem based on the given image(s).
> Question: {Question}
> options: {Options}(if exists)
> Return only the final answer and Wrap your final answer in $boxed\{\}$.

Text-only Chain-of-Thought prompting. For the thinking and scaling experiments in Section 4.2.1 and Section 4.2, we use the following standard chain-of-thought prompt:

> You are given a visual reasoning problem based on the given image(s).
> Question: {Question}
> options: {Options}(if exists)
> Think step by step, carefully reasoning through the problem before giving your answer.
> Wrap your final answer in $boxed\{\}$.

Visual-augmented reasoning: Turn 1 (generation hint) for Flux.1 and SD3.5. The VLM first produces a short instruction to guide the external image generator. Since Flux.1 and SD3.5 are built for dealing with only one input image, we specifically prompt the edit instruction on the first input image:

> You will be given a question and one or more input images. Do NOT answer the question yet.
> An external tool can generate ONE helper image, but it can ONLY take the FIRST image as the init image.
> So your instruction must describe how to edit the FIRST image into a helper image. If there are multiple images, you may use them as reference, but the helper must be generated from the FIRST image.
> Write a SHORT instruction (1-2 sentences) for generating the helper image.
> Wrap your instruction in ¡instruction¿ tags:¡instruction¿Your image generation instruction here.¡/instruction¿
> Do NOT answer the question yet.

Visual-augmented reasoning: Turn 1 (generation hint) for Qwen-Image-Edit and Nano-banana. Qwen-Image-Edit-2511 and Nano-banana can take in multiple images for input:

> You will be given a question and one or more input images. Do NOT answer the question yet.
> You can call an external tool that will generate ONE helper image.
> The tool can see ALL the original images as context, so your instruction may reference any of them.
> Write a SHORT instruction (1-2 sentences) for generating the helper image.
> Wrap your instruction in ¡instruction¿ tags: ¡instruction¿Your image generation instruction here.¡/instruction¿
> Do NOT answer the question yet.

Visual-augmented reasoning: Turn 1 (generation hint) CoT prompt. For the additional experiment in Section 4.2.1, we add a thinking prompt to the hint-generation turn. The final-answer turn uses the text-only CoT prompt.

> Before writing your instruction, think step by step about what visual information is missing or hard to extract from the original image(s), and how a generated helper image could make answering the question easier.

Visual-augmented reasoning: Turn 2 (final answer). After the helper image is generated, the VLM receives both original and generated images:

> You will see (A) the ORIGINAL input image(s) and (B) an OPTIONAL GENERATED helper image.
> Answer using evidence from all provided image(s).

LLM-as-Judge evaluation prompt. For benchmarks without rule-based evaluation, we use the following prompt to assess answer correctness:

> You are a strict evaluation judge for QA tasks.
> Goal: decide if the model prediction should be counted as correct for the given sample.
> Rules:
> - If options are provided, the prediction is correct iff it clearly selects the same option as the ground-truth answer.
> - Accept either option letter (A, B, C...) or the option text, as long as it maps unambiguously.
> - Be robust to minor formatting differences (case, punctuation, whitespace).
> - For open-ended answers without options, mark correct only if the prediction is semantically equivalent to the ground truth.
> - If prediction is ambiguous, incomplete, or picks multiple conflicting answers, mark incorrect.
> - Do NOT be lenient: if you are unsure, mark incorrect.
> Return ONLY JSON with keys:
> correct: boolean
> matched_option_index: integer or null (0-based if options exist)

Setting	Jigsaw		Math-Geo	
	Easy	Hard	Easy	Hard
Text-only	76.8	34.2	55.2	23.1
+ *Generation*	77.1 (+0.3)	31.0 (-3.2)	54.9 (-0.3)	19.8 (-3.3)

Table 5: Generation degrades performance primarily on hard, structured instances. On easy examples, generation leaves accuracy nearly unchanged, whereas performance drops by over 3 points on the hard subsets of both Jigsaw and Math-Geo. Accuracy (%) is reported for Qwen3-VL 8B with Nano-Banana; parentheses show changes relative to text-only inference.

Judge	ChartQAPro	VSTAR	Audit Set
Rule-based	40.8	81.6	–
Gemini-2.5-Flash	41.2 (+0.4)	81.6 (+0.0)	37.0 (-2.0)
Human	–	–	39.0

Table 6: LLM-based evaluation closely agrees with rule-based and human scoring. Gemini-2.5-Flash differs from rule-based accuracy by at most 0.4 points on ChartQAPro and VSTAR, and by 2.0 points from human annotation on a 100-instance audit set sampled across benchmarks. Results are based on Qwen3-VL 8B text-only predictions. The cross-run standard deviation of the LLM judge is 0.97 points.

```
normalized_pred: string
normalized_answer: string
rationale: string (one short sentence)
```

C Additional Experiments

C.1 Generator Capability

We further investigate whether the limited gains of visual generation stem from the generator's capability rather than the reasoning model. Concretely, we perform a focused case study that stratifies tasks by difficulty and contrasts scenarios where image generation is expected to be sufficient (simple perceptual edits) against those requiring complex manipulations or structured visual outputs. Unless otherwise specified, experiments in this subsection use Qwen3-VL 8B (Bai et al., 2025a) as the reasoning model and Nano-Banana (Google DeepMind, 2025a) as the image generator. We chose two tasks, a perception-heavy jigsaw task and a structured math-geometry task, and control the generation hint to be minimal short prompts without any reasoning involved.

Table 5 shows difficulty-stratified results. Generation provides marginal benefit on easy perceptual instances (e.g., Jigsaw Easy: 76.8→77.1) but this benefit diminishes or reverses on hard and structured instances (Math-Geo Hard: 23.1→19.8), consistent with the failure modes identified in Section 5.3.

C.2 Judge Reliability

We use Gemini-2.5-Flash as an LLM judge for benchmarks where rule-based matching is not applicable. To assess reliability, we compare it against (1) rule-based evaluation on benchmarks with deterministic answer formats, and (2) human annotations on a randomly sampled set of 100 instances drawn from all datasets. All results in Table 6 use Qwen3-VL 8B text inference.

Consistency with rule-based evaluation. The two evaluators yield closely matching aggregate accuracies, with identical scores on VSTAR and a difference of only 0.4 percentage points on ChartQAPro.

Strategy	VSTAR	ERQA	SAT
Text-only	81.6	44.3	70.1
+ *Gen (KV-Cache)*	32.4	18.7	45.8
+ *Gen (Multi-Turn)*	50.8	38.5	61.1

Table 7: Multi-turn inference substantially outperforms KV-cache reuse, but still fails to recover the text-only baseline. Re-encoding the generated image in a new turn improves accuracy over KV-cache reuse by 15–20 points across all three benchmarks. Accuracy (%) is reported for Qwen3-VL 8B with Qwen-Image-Edit.

Consistency with human annotation. On the 100-instance sample set, Gemini-2.5-Flash scores 37.0 versus human evaluation at 39.0, a 2-point difference well within the margin of inter-annotator variation. We additionally rerun Gemini-2.5-Flash three times on ChartQAPro and VSTAR, observing a standard deviation of 0.97 percentage points across runs. Together, these checks confirm that our evaluation judge is robust and that findings are not an artifact of judge design.

C.3 Multi-Turn vs. KV-Cache Inference

Once a helper image $\tilde{I}$ has been produced, we compare two implementations for feeding it back to the VLM. *Multi-turn* inference adds $\tilde{I}$ in a new conversational turn and re-encodes the full context, including the question, original images, hint, and helper image. *KV-cache* inference instead reuses the cached prefix and processes the appended image tokens. We compare these implementations on VSTAR, ERQA, and SAT using Qwen3-VL 8B paired with Qwen-Image-Edit.

Across all three benchmarks, the multi-turn implementation outperforms the KV-cache implementation: 50.8 vs. 32.4 on VSTAR (+18.4), 38.5 vs. 18.7 on ERQA (+19.8), and 61.1 vs. 45.8 on SAT (+15.3). We therefore use multi-turn inference for all main experiments. These results compare the two evaluated implementations; they do not establish an inherent limitation of KV caching. Under causal attention, recomputing an unchanged prefix does not let earlier tokens attend to a later helper image. The source of the observed gap remains unresolved.

Even with multi-turn inference, accuracy remains below the text-only baseline (81.6 / 44.3 / 70.1). Thus, the main finding holds under the better-performing implementation: generated helper images still reduce accuracy on these three benchmarks.

C.4 Detailed Failure Attribution

Aggregated across benchmarks, **76%** of Qwen3-VL failures and **57%** of Gemini failures stem from bad hints, while **21%** (Qwen) and **36%** (Gemini) come from bad generation; pure inference failures are negligible. Per-task ranges sharpen the dichotomy reported in Section 5.3: chart and visual-search tasks (ChartQAPro, VSTAR) reach 50–90% bad generation, while spatial, embodied, puzzle, and STEM tasks (ERQA, MathVista, SAT, VisualPuzzles, Rover) reach 70–100% bad hints. Stylistically, Gemini more frequently produces tool-like hints (e.g., Python-style editing instructions) while Qwen3-VL produces shorter, under-specified ones; despite this stylistic difference, both models exhibit the same task-dependent split.

C.5 Detailed Qualitative Case Studies

For success cases, despite the correctness in the final answer, we observe that intermediate visual content is not guaranteed to be useful—improvement is more likely to come from longer conversations or randomness of textual reasoning. For failure cases, we analyze the failure in text understanding and visual generation separately. We find that VLMs do not fully understand what visual information could be helpful and often produce confusing generation hints, leading to drifted reasoning. On the visual generation side, failures become common in two regimes: (i) *complex manipulation*, where the edit requires precise transformations with constraint satisfaction, and (ii) *structured tasks*, where correctness depends on producing visually consistent diagrams or exact configurations. These examples support the conclusion in Section 4.2 that current VLMs are text-centric

and that achieving general cross-modality reasoning requires joint advances in both understanding and generation abilities.

C.6 Detailed Thinking-Mode Results

For text-only inputs, thinking yields ChartQAPro +15.8 (40.8→56.6) and MathVista +2.5 (61.2→63.7) but hurts ERQA (−7.6, 44.3→36.7). Adding visual augmentation without thinking already degrades all three (FLUX.1: ChartQAPro −2.1, MathVista −3.9, ERQA −5.8), and enabling thinking does not recover the gap and sometimes widens it (−5.4, −2.9, −7.9 respectively). On harder reasoning tasks, longer reasoning over the generated image may even amplify the noise the synthetic visual injects.

D Limitations

Our study evaluates static image generators and omits two adjacent paradigms that may unlock cases where stills fall short: **video generators** and **world-model / simulator generators**, which can render dynamics, action consequences, and counterfactual rollouts that single-frame edits cannot. Whether these temporally grounded generators sidestep the planning and language-bias bottlenecks identified here is an open question we leave to future work. Our evaluation is also confined to **question-answering benchmarks**; visual generation may play a different and possibly more load-bearing role in non-QA settings such as long-horizon planning, agentic execution, dialog, or open-ended generation, where intermediate visual artifacts are themselves part of the output rather than an aid to a single answer. Extending the framework to instruction-tuned reasoning-specialist VLMs and to longer multi-turn text–image–text–image trajectories is another natural direction.

E Broader Impact

This work studies the integration of visual generation into multimodal reasoning systems and finds that generated visual evidence does not consistently improve reasoning and may even degrade performance in structured tasks. These insights can inform the design of more reliable systems and guide when generation should be used.

At the same time, incorporating generated images introduces risks. Generated content may contain hallucinated or misleading details, which could be interpreted as valid evidence, especially in high-stakes settings such as healthcare or autonomous systems. There is also potential for misuse, such as generating visually plausible but incorrect evidence.

We encourage future work on improving the reliability of visual generation, incorporating verification mechanisms, and explicitly modeling uncertainty in multimodal reasoning systems.

F Dataset Licenses and Asset Credits

We use the following publicly released benchmarks under their respective licenses. All datasets are used only for academic research and we comply with their terms.

- **ChartQAPro** (Masry et al., 2025) – MIT License (https://github.com/vis-nlp/ChartQAPro).
- **V*Bench / VSTAR** (Wu & Xie, 2024) – MIT License (https://github.com/penghao-wu/vstar).
- **VisuLogic** (Xu et al., 2025a) – Apache 2.0 (https://github.com/VisuLogic-Benchmark/VisuLogic-Eval).
- **VisualPuzzles** (Song et al., 2025b) – MIT License (https://github.com/neulab/VisualPuzzles).
- **SAT** (Ray et al., 2024) – MIT License (https://huggingface.co/datasets/array/SAT).

- **RoboSpatial** (Song et al., 2025a) – NVIDIA Source Code License, non-commercial research use only (https://github.com/NVlabs/RoboSpatial/blob/main/LICENSE).
- **ERQA** (Team et al., 2025) – CC-BY 4.0 (https://github.com/embodiedreasoning/ERQA).
- **MathVista** (Lu et al., 2024) – CC-BY-SA 4.0 (https://huggingface.co/datasets/AI4Math/MathVista).
- **Rover-Bench** (Liang et al., 2025) – MIT License (https://github.com/cheryyunl/ROVER).
- **Zebra-CoT** (Li et al., 2025a) – CC-BY-NC 4.0, non-commercial research use only (https://huggingface.co/datasets/multimodal-reasoning-lab/Zebra-CoT).
- **EMMA-mini** (Hao et al., 2025) – dataset license not specified in the public dataset card (https://huggingface.co/datasets/luckychao/EMMA-mini).
- **RealUnify** (Shi et al., 2025b) – Apache 2.0 (code) with academic-research-only restriction on dataset usage (https://github.com/FrankYang-17/RealUnify).

We also use the following models under their official releases: Qwen3-VL (Bai et al., 2025a), GLM-4.6V (GLM-V Team et al., 2025), Gemini 3.0 Flash (Google DeepMind, 2025b) and Nano-Banana (Google DeepMind, 2025a), Stable Diffusion 3.5 (Esser et al., 2024), FLUX.1-Kontext (Labs, 2024), Qwen-Image-Edit (Wu et al., 2025a), and BAGEL (Deng et al., 2025).