VIG: Visual Information Gain as a Reward Signal for Multimodal Chain-of-Thought Compression

Wen Luo[1*] **Xiaohan Yi**[2*] **Xiaotao Huang**[1†] **Liqun Huang**[1†]
[1]School of Software Engineering, Huazhong University of Science and Technology
[2]Tsinghua University
{wen_chaser, huangliqun}@hust.edu.cn huangxt@mail.hust.edu.cn
yxh24@mails.tsinghua.edu.cn

Abstract

Multimodal large reasoning models often rely on long Chain-of-Thought (CoT) traces in which a substantial fraction of tokens, such as repeated visual descriptions, self-reflection, and other visually-disengaged filler, inflate inference cost without contributing to the answer. Existing CoT compression methods optimize output length but never measure whether a reasoning token is actually grounded in the image. We propose **VIG** (Visual Information Gain), an information-theoretic GRPO reward that scores each reasoning token by how much the image reduces its predictive uncertainty. VIG is computed online from two forward passes of the same policy, one with and one without the image, so no reference chains, external annotations, or auxiliary reward models are needed. Across six main multimodal reasoning benchmarks and three Qwen3-VL-Thinking model sizes (2B/4B/8B), plus an additional R1-Onevision-Bench evaluation on 8B, VIG consistently improves the accuracy–efficiency trade-off, supporting our central claim: *efficient multimodal reasoning emerges from raising visual information density, where every reasoning token earns its place by anchoring to the image, rather than from imposing a length budget.* Our source code is available at https://github.com/chaser682/vig.

1 Introduction

A modern Multimodal Large Reasoning Model (MLRM) may emit over a thousand tokens of "thinking" to answer a single geometry question—yet a substantial fraction of those tokens never actually consult the image. They restate the question, narrate the solver's own metacognition, and rephrase visual relationships already established earlier in the chain. The result is a reasoning trace that is both wasteful to compute and, as we will show, often *less* accurate than a shorter, more visually grounded reasoning chain.

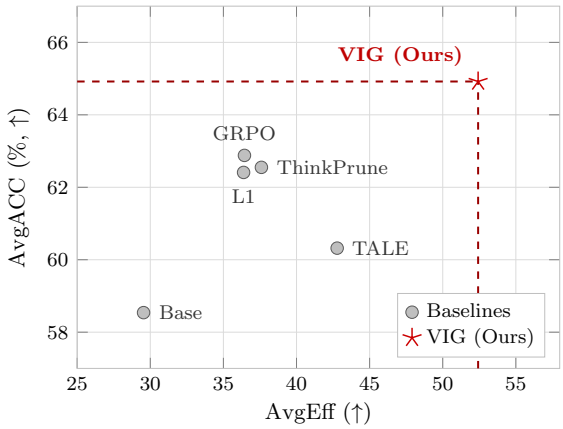


Figure 1: **Accuracy–efficiency Pareto frontier on Qwen3-VL-8B-Thinking.** Each point reports the average accuracy (AvgACC, %) and efficiency (AvgEff, ACC per 1k tokens) of one method across the six multimodal benchmarks in Table 1. VIG (★) Pareto-dominates the evaluated baselines (Base, GRPO, L1, ThinkPrune, TALE) on both axes, defining a new accuracy–efficiency Pareto frontier (red dashed lines).

This phenomenon is a direct consequence of how deep reasoning is currently trained. Inspired by OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), modern MLRMs (Bai et al., 2023; Liu et al., 2023a; Chen et al., 2024b; Bai et al., 2025) produce long Chain-of-Thought (CoT) within <think> blocks, articulating visual details, spatial relationships, and cross-modal alignments before committing to an answer. Longer chains often correlate with higher accuracy on text-only benchmarks, but in multimodal scenarios this paradigm comes with two costs. First, the inflated token count drives up inference latency, restricting deployment in interactive settings. More surprisingly—and more importantly—over-reasoning actively harms correctness: redundant steps inject misleading context, small perceptual

[*] Equal contribution.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 4574–4593
October 24-29, 2026 ©2026 Association for Computational Linguistics

errors accumulate across hundreds of tokens, and models that initially read the image correctly are often talked out of correct conclusions by their own elaboration.

Figure 2 makes this concrete on an elementary geometry problem (rectangle of area $10\,\text{cm}^2$ with one side labeled $5\,\text{cm}$): Qwen3-VL-8B-Thinking (Bai et al., 2025) spends 823 tokens and still answers "no correct answer", because its long deliberation second-guesses the figure's plain label and talks itself out of the obvious division. L1's length penalty (Aggarwal and Welleck, 2025) cuts the chain to 412 tokens and recovers the correct answer, but still uses nearly $4\times$ the tokens of our method—it truncates self-doubt without teaching the model to trust the image. Our VIG model finishes in 105 tokens by simply reading $5\,\text{cm}$ from the figure and applying area $=$ length$\times$width. This is the blind spot shared by current CoT compression methods—prompting (Han et al., 2025; Team et al., 2025), distillation on compressed traces (Li et al., 2026; Xia et al., 2025; Ma et al., 2025), and RL with length, token-limit, or preference penalties (Aggarwal and Welleck, 2025; Hou et al., 2025; Hong et al., 2025)—and recent multimodal reasoners (Meng et al., 2025; Huang et al., 2025b,a; Yang et al., 2025) inherit it: every compression signal in the literature is purely textual, unable to tell a visually-essential observation from linguistic filler. What is missing is a per-token signal that measures *how much each token actually depends on the image.*

For each reasoning token, we ask exactly that: how much does the image reduce the model's uncertainty about it. Tokens that require the image—a direct reading of a value, a spatial relationship, an identification of a visual entity—receive high scores; tokens derivable from language context alone—meta-cognitive openers, problem restatements, generic algebra—receive scores near zero. We call this quantity Visual Information Gain (VIG); it coincides with the conditional mutual information between the token and the image (Cover and Thomas, 2006). Computing it requires only two forward passes of the same policy—one with the visual tokens, one without—whose per-token entropy difference, averaged over the `<think>` block, becomes a scalar reward plugged directly into GRPO (Shao et al., 2024). No external annotator, reference chain, step segmentation, or auxiliary model is needed. Trained with VIG, the model learns to make every token earn its place by anchoring to the image, with chain shortening emerging as a byproduct of optimizing visual information density rather than of a length constraint.

Empirically, this principle yields consistent gains. Across six multimodal reasoning benchmarks and three Qwen3-VL-Thinking model sizes, VIG sets a new accuracy–efficiency Pareto frontier. As shown in Figure 1, on Qwen3-VL-8B-Thinking VIG Pareto-dominates the evaluated baselines—Base, plain GRPO (Shao et al., 2024), L1 (Aggarwal and Welleck, 2025), ThinkPrune (Hou et al., 2025), and TALE (Han et al., 2025)—in both accuracy and efficiency, reaching 64.92% average accuracy and an efficiency (ACC per 1k tokens) of 52.42 and beating the strongest compression baseline (ThinkPrune) by $+2.37\%$ in accuracy and $+39.4\%$ in efficiency. The gains extend to R1-Onevision-Bench (Yang et al., 2025), where VIG ranks first overall and first or tied-first on three of five subject categories, confirming that the benefit of visual-density-aware optimization is not specific to a particular benchmark family.

Contributions.

1. We introduce **Visual Information Gain**, a per-token, information-theoretic measure of how much an MLRM's reasoning token actually relies on the visual input, reframing CoT compression as increasing visual density rather than reducing length.
2. We turn VIG into a fully online RL reward computable from two forward passes of the policy itself—no external models, no annotated traces, no step segmentation—dropping cleanly into GRPO with no additional supervision.
3. On Qwen3-VL-8B-Thinking, VIG delivers $+39.4\%$ efficiency and $+2.37\%$ accuracy over the strongest compression baseline; across all three Qwen3-VL-Thinking model sizes it lies on the main-benchmark Pareto frontier, and on R1-Onevision-Bench it ranks first overall and first or tied-first on three of five subject categories.

2 Related Work

2.1 Multimodal Large Reasoning Models

RL-based deep reasoning in LLMs (Guo et al., 2025; Jaech et al., 2024) has quickly been extended to multimodal settings. Early systems such as LLaVA-CoT (Xu et al., 2025) and Mulberry (Yao et al., 2024) introduced structured visual reasoning via stage-wise decomposition and search. Recent

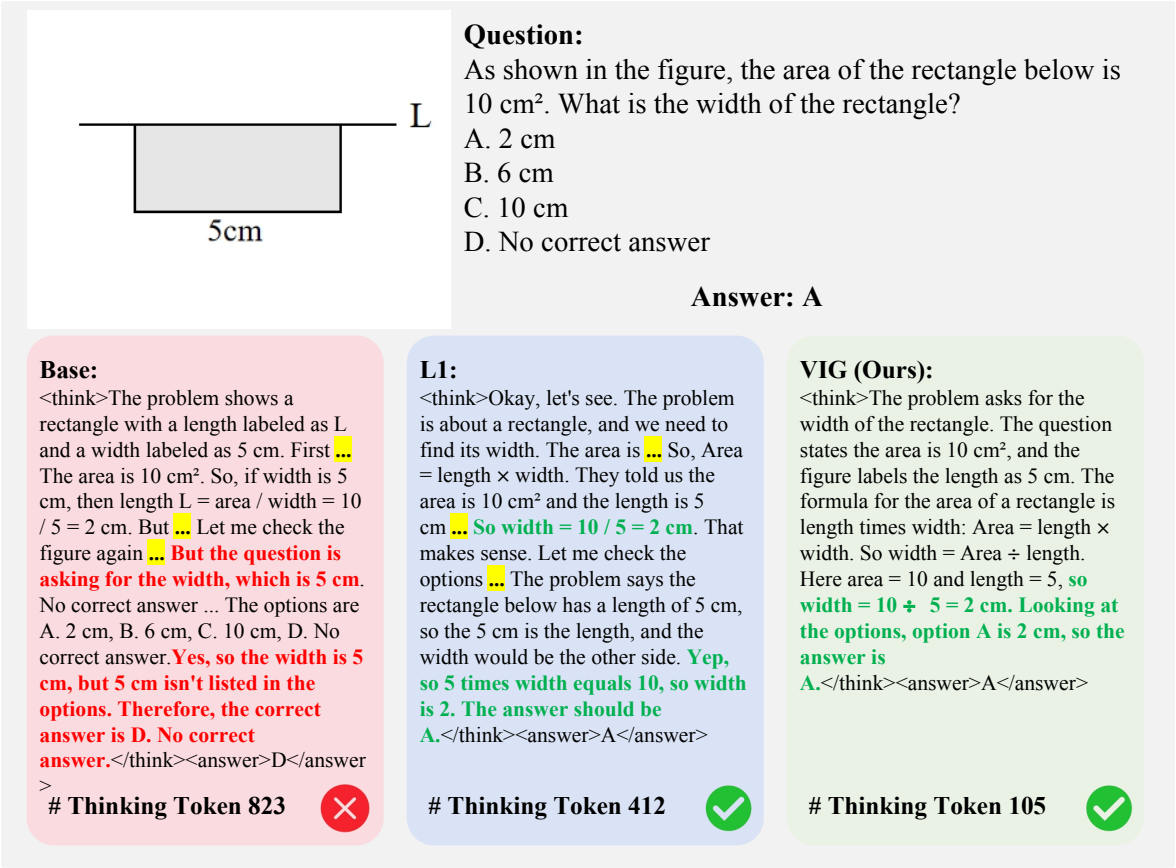


Figure 2: **Length penalties trim the chain; only visually grounded reasoning removes the redundancy at its source.** Token counts and final answers from Base, L1, and our VIG model on the same geometry question (gold answer A, 2 cm).

RL-based MLRMs further scale long-CoT reasoning in vision-language tasks: VisualThinker-R1-Zero (Zhou et al., 2025) and MM-Eureka (Meng et al., 2025) study R1-style reasoning and "aha-moment" behavior, while OThink-MR1 (Liu et al., 2025), ThinkLite-VL (Wang et al., 2025a), Vision-R1 (Huang et al., 2025b), LMM-R1 (Peng et al., 2025), R1-Omni (Zhao et al., 2025a), Insight-V (Dong et al., 2024), and R1-OneVision (Yang et al., 2025) improve training recipes and broaden the scope of multimodal reasoning. However, stronger reasoning often comes with substantially longer CoTs, including repeated visual descriptions and unnecessary self-reflection. Closest to ours is VPPO (Huang et al., 2025a), which reweights trajectory advantages by overall visual dependency; in contrast, VIG targets explicit CoT compression by turning a with-/without-image entropy difference into a per-token reward.

2.2 Chain-of-Thought Compression

Existing CoT compression methods fall into three broad families distinguished by where the compression signal originates. *Prompt-based* methods control CoT length through input instructions without modifying model parameters: TALE (Han et al., 2025) estimates a per-question token budget and embeds it in the system prompt, and Kimi k1.5 (Team et al., 2025) toggles between long and short reasoning via a budget flag—training-free but compressing by truncation rather than by selection. *Training-based* methods fold the compression signal into a supervised or RL objective. At the token level, TokenSkip (Xia et al., 2025) ranks tokens by semantic importance for controllable skipping, while CoT-Valve (Ma et al., 2025) identifies a parameter-space direction along which chain length elastically varies; at the step level, StepEntropy (Li et al., 2026) and Prune-on-Logic (Zhao et al., 2025b) segment the chain and prune low-utility steps before SFT distillation. RL-based variants—L1 (Aggarwal and Welleck, 2025), ThinkPrune (Hou et al., 2025), and small-scale preference optimization (Hong et al., 2025)—reward shorter outputs through length-target rewards, iteratively tightened token limits, or preference-based

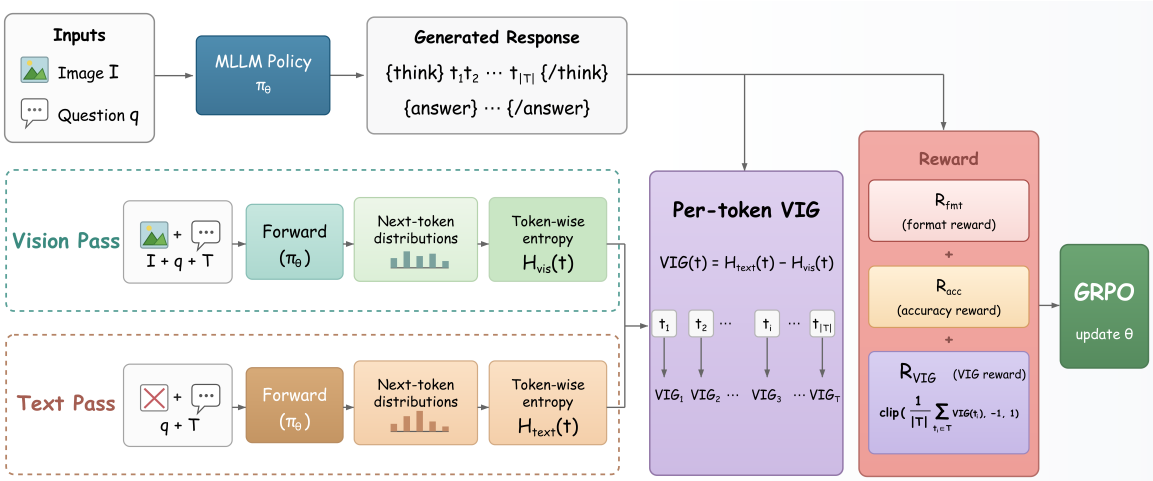


Figure 3: **VIG framework overview.** Two forward passes over the generated reasoning chain—one with the image and one with visual tokens removed—yield a token-level entropy difference, which is aggregated into R_{VIG} and used as a GRPO reward.

penalties. *Latent and architectural* alternatives bypass tokenized CoT entirely: Coconut (Hao et al., 2024) treats hidden states as continuous thoughts, LightThinker (Zhang et al., 2025) compresses thought steps into gist tokens via specialized attention masks, the Markovian Thinker (Aghajohari et al., 2025) chunks reasoning with explicit carry-over state, and R1-Compress (Wang et al., 2025b) performs chunk-level inference-time compression—typically at the cost of architectural changes or the interpretability of explicit reasoning.

All these approaches share a blind spot for the multimodal setting: their compression signal is purely textual—token count, text-only entropy, or text-internal importance—and none measures how much each reasoning token actually depends on the image. Information-theoretic objectives are well established in representation learning (Belghazi et al., 2018; Oord et al., 2018) and have been applied through the information-bottleneck lens to attribute and prune redundant visual features (Tishby et al., 2000; Alemi et al., 2017; Schulz et al., 2020), but always at the level of the visual encoder, not the reasoning chain that consumes it. VIG transfers this IB intuition from "compressing the encoder" to "compressing the chain of thought" by using visual–textual conditional mutual information as a per-token, online RL reward.

3 Method

Our framework, summarized in Figure 3, scores every reasoning token by its VIG (Eq. (2)), aggregates these per-token scores into a sequence-level reward, and combines it with standard format and accuracy rewards inside GRPO. The reward is computed online from two forward passes of the same policy—one with the image and one with visual tokens removed—so no reference chains, external annotators, or auxiliary models are required. Crucially, VIG never rewards shorter outputs directly: it pushes the policy toward chains in which every token is anchored to the image, and the chain shortening observed in our experiments emerges as a *byproduct* of this density-aware optimization. The remainder of this section defines VIG (§3.1) and details the reward design and two-pass computation (§3.2); we adopt GRPO (Shao et al., 2024) as the policy-optimization framework and refer the reader to Appendix A for the standard background.

3.1 Visual Information Gain

With this optimization background in place, we next define the central quantity used to score multimodal reasoning tokens. Specifically, we ask how much the image reduces the model's uncertainty about a given token in the reasoning chain. Consider a single token t_τ generated at position τ, conditioned on the question q and the preceding context $\text{ctx}_{<t} = t_{<\tau}$. By the chain rule of conditional entropy, the predictive uncertainty of t_τ admits the orthogonal decomposition

$$H(t_\tau \mid q, \text{ctx}_{<t}) = \underbrace{H(t_\tau \mid q, I, \text{ctx}_{<t})}_{H_{\text{vis}}(t_\tau)} + \underbrace{I(t_\tau; I \mid q, \text{ctx}_{<t})}_{\text{VIG}(t_\tau)}, \quad (1)$$

Table 1: Main results on six benchmarks. **Bold** and underline denote the best and second-best results within each model block; tied values share the same marking.

Method	WeMath		MMMU		MathVision		DynaMath		MMK12		Geo3K		Summary		
	ACC	TLen	ACC	TLen	ACC	TLen	ACC	TLen	ACC	TLen	ACC	TLen	AvgACC	AvgTLen	AvgEff
Qwen3-VL-2B-Thinking	51.72	3120	36.78	2855	15.39	3928	51.60	2252	33.25	3683	33.78	3087	37.09	3154.2	12.71
GRPO	57.36	2850	**45.00**	2614	17.17	3825	54.27	2046	36.70	3592	38.60	2929	41.52	2975.9	15.29
L1	**61.67**	2241	43.33	2561	17.14	3668	53.65	1926	34.30	3429	35.94	2670	41.01	2749.0	16.74
TALE	53.39	3033	40.00	2667	17.24	3872	54.01	2096	35.20	3595	36.11	3040	39.32	3050.5	14.08
ThinkPrune	60.80	2438	42.56	**2372**	17.53	3708	55.53	1863	**39.70**	3413	37.44	2679	42.26	2745.5	17.17
VIG (Ours)	60.98	**2159**	43.33	2395	**18.72**	**3639**	**56.55**	**1728**	37.20	**3317**	**40.10**	**2550**	**42.81**	**2631.3**	**18.52**
Qwen3-VL-4B-Thinking	66.84	2650	53.00	2432	24.38	3657	65.33	2012	52.05	3230	54.08	2635	52.61	2769.6	20.46
GRPO	73.79	2095	57.78	2142	30.39	3378	70.30	1714	58.00	2949	61.56	2336	58.64	2435.5	26.37
L1	75.06	1876	56.78	2157	**31.71**	3305	70.70	1636	60.35	2738	61.73	2114	59.39	2304.4	28.40
TALE	71.90	2152	57.22	2002	29.21	3389	69.40	1569	54.15	3027	55.41	2506	56.22	2440.7	25.81
ThinkPrune	74.94	1944	**58.89**	2082	31.51	3276	71.10	1614	59.90	2817	58.07	2230	59.07	2327.2	27.97
VIG (Ours)	**79.31**	**850**	**58.89**	**1416**	30.95	**2415**	**73.59**	**772**	**65.90**	**1795**	**64.23**	**1310**	**62.15**	**1426.5**	**54.79**
Qwen3-VL-8B-Thinking	73.56	1796	60.89	1955	28.85	3332	69.04	1464	59.85	2868	59.07	2075	58.54	2248.5	29.54
GRPO	79.20	1465	63.11	1779	30.63	3099	73.87	1319	66.60	2369	63.89	1821	62.88	1975.4	36.44
L1	78.51	1436	63.67	1795	30.20	3022	72.40	1339	66.75	2360	62.90	1759	62.41	1951.8	36.38
TALE	76.78	1123	59.67	1557	32.50	3004	70.00	1013	63.40	2163	59.57	**1460**	60.32	1720.2	42.78
ThinkPrune	79.43	1367	62.56	1690	**32.57**	3038	72.10	1263	65.40	2366	63.23	1804	62.55	1921.4	37.60
VIG (Ours)	**81.26**	**947**	**64.67**	**1364**	32.20	**2735**	**75.31**	**892**	**70.70**	**1728**	**65.39**	1480	**64.92**	**1524.3**	**52.42**

where the first term $H_{\text{vis}}(t)$ is the residual uncertainty that remains *after* observing the image, and the second term is the additional uncertainty *removed by* the image, i.e., the conditional mutual information between the token and the visual input. We refer to the latter as the token's *visual information gain*.

Definition 1 (Visual Information Gain)**.** *For a token t_τ at position τ in the reasoning chain T, the visual information gain is defined as*

$$VIG(t_\tau) = H_{txt}(t_\tau) - H_{vis}(t_\tau), \quad (2)$$

where $H_{vis}(t_\tau) = H(t_\tau \mid q, I, t_{<\tau})$ and $H_{txt}(t_\tau) = H(t_\tau \mid q, \emptyset, t_{<\tau})$ are the Shannon entropies of the policy's next-token distribution with *and* without *the visual input.*

By the data processing inequality (Cover and Thomas, 2006), $H_{\text{vis}}(t_\tau) \leq H_{\text{txt}}(t_\tau)$ and $\text{VIG}(t_\tau) \geq 0$ almost surely (Appendix C); Eq. (1) further identifies $\text{VIG}(t_\tau)$ with the per-token conditional mutual information $I(t_\tau;\, I \mid q, t_{<\tau})$. Intuitively, a *high-VIG* token is confident with the image but uncertain without it—the signature of a visually grounded observation—whereas a *near-zero VIG* token is derivable from the textual context alone. VIG thus gives a principled per-token measure of how much each reasoning token depends on the visual evidence.

3.2 Reward Design

Having defined VIG at the token level, we now show how it is converted into a sequence-level reward that can be used during policy optimization. Our reward design proceeds in three stages: aggregating token-level VIG over the `<think>` block, computing the resulting score online with a two-pass forward procedure, and combining it with format and accuracy rewards.

3.2.1 Token-Level Aggregation

We aggregate VIG over all $|T|$ tokens in the `<think>` block by the clipped token-level mean

$$R_{\text{VIG}} = \text{clip}\left(\frac{1}{|T|}\sum_{t\in T}\text{VIG}(t),\ -1,\ 1\right). \quad (3)$$

This choice is principled: by the chain rule of mutual information, $I(T; I \mid q) = \sum_{t\in T}\text{VIG}(t)$, so the unclipped mean is a length-normalized estimate of the total visual information carried by the reasoning chain. In practice, this simple aggregation already provides an effective and stable reward signal for RL training.

3.2.2 Two-Pass Forward Computation

R_{VIG} is computed *online* during the GRPO rollout via two forward passes of the same policy π_θ: a *vision pass* on the full multimodal input (I, q, s_T) yields $H_{\text{vis}}(t)$, and a *text pass* that removes the visual tokens (the `<vision>`$\cdots$`</vision>` spans) from the same input yields $H_{\text{txt}}(t)$. The token-level difference, restricted to the `<think>` block via a binary mask, is averaged and clipped to $[-1, +1]$. Both passes share the same weights—no auxiliary parameters or gradients are introduced—making VIG a drop-in scalar reward. In practice, the second

Table 2: Performance comparison on R1-Onevision-Bench. **Bold** and underline denote the best and second-best results.

Method	Physics (278)		Mathematics (327)		Biology (134)		Chemistry (105)		Deduction (98)		Overall		
	ACC	TLen	ACC	TLen	ACC	TLen	ACC	TLen	ACC	TLen	AvgACC	AvgTLen	AvgEff
Qwen3-VL-8B-Thinking	53.96	2793	45.26	2894	56.72	2121	60.95	2315	17.35	3953	48.30	2799.8	18.48
GRPO	61.87	2444	47.09	2775	**62.69**	1845	64.76	2125	24.49	3836	53.29	2583.1	22.62
L1	61.51	2338	48.32	2708	59.70	1844	**68.57**	2101	21.43	3676	53.29	2508.9	23.00
TALE	56.83	2149	48.62	2565	61.94	1621	**68.57**	1647	23.47	**3444**	52.55	2297.1	26.41
ThinkPrune	59.35	2416	46.48	2808	58.21	1671	63.81	1993	**26.53**	3797	51.80	2542.8	22.99
VIG (Ours)	**65.47**	**1974**	**54.13**	**2473**	60.45	**1446**	**68.57**	**1619**	23.47	3488	**56.79**	**2190.0**	**29.19**

forward operates on a shorter, vision-stripped input and does not require gradients, so per-step training time grows by only ∼17% over plain GRPO (Appendix G). The full procedure is given as Algorithm 1 in Appendix B.

3.2.3 Full Reward Function

The final reward returned to GRPO combines three signals: format compliance, answer correctness, and sequence-level visual information gain:

$$R(o) = w_f\, R_{\text{format}}(o) + w_a\, R_{\text{acc}}(o, a^*) + w_v\, R_{\text{VIG}}(o). \tag{4}$$

We use w_f=0.1, w_a=1.0, and w_v=1.0 by default; a sensitivity analysis on w_v is reported in Section 4.3. Below, we describe the three terms in turn.

Format reward (R_{format}). This term encourages the model to preserve the desired <think>/<answer> structure:

$$R_{\text{format}}(o) = \begin{cases} 0 & \text{if both tags are present,} \\ -0.5 & \text{if only <answer> is present,} \\ -1 & \text{otherwise.} \end{cases} \tag{5}$$

Accuracy reward (R_{acc}). This term reflects whether the extracted final answer $A(o)$ matches the ground truth a^*:

$$R_{\text{acc}}(o, a^*) = \begin{cases} 1 & \text{if } A(o) = a^*, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

VIG reward (R_{VIG}). This term is the clipped mean token-level visual information gain over the <think> block, as defined in Eq. (3) and computed by Algorithm 1. Larger values indicate that a greater fraction of the reasoning chain is supported by visual evidence rather than by text-only continuation.

Taken together, these three terms guide the model toward responses that are well-formed, correct, and visually grounded. In particular, R_{VIG} shapes the reasoning style by rewarding chains whose tokens remain visually informative on average. As a result, the policy learns to retain tokens that directly read or relate visual evidence, while suppressing meta-cognitive openers, repeated problem restatements, and other visually disengaged filler. This is how VIG produces shorter and denser reasoning chains without introducing an explicit length penalty.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate VIG on Qwen3-VL-2B-Thinking, Qwen3-VL-4B-Thinking, and Qwen3-VL-8B-Thinking (Bai et al., 2025). We compare against five baselines: **Base**, the original thinking model without RL fine-tuning; **GRPO**, trained with only format and accuracy rewards; **L1** (Aggarwal and Welleck, 2025), which uses a length-target RL reward; **TALE** (Han et al., 2025), a dynamic budget prompting method; and **ThinkPrune** (Hou et al., 2025), which applies RL under an iteratively tightened token limit.

Evaluation benchmarks. We evaluate on seven widely-used multimodal reasoning benchmarks: WeMath (Qiao et al., 2024), MMMU (Yue et al., 2023), MathVision (Wang et al., 2024), DynaMath (Zou et al., 2024), MMK12 (Meng et al., 2025), Geo3K (Lu et al., 2021), and R1-Onevision-Bench (Yang et al., 2025); the last one provides subject-level evaluation across Physics, Mathematics, Biology, Chemistry, and Deduction, and is reported separately to highlight per-subject performance.

Metrics. We report accuracy (ACC ↑), average reasoning-chain length in tokens (TLen ↓), and efficiency $\text{Eff} = \text{ACC}/\text{TLen} \times 1000$ (↑), which measures accuracy per kilotoken. AvgACC and AvgTLen are averaged over the six main benchmarks, while AvgEff is the mean of per-benchmark efficiency scores. All numbers we report are from

Table 3: Evaluation results on non-mathematical visual tasks with Qwen3-VL-4B-Thinking. All models use the same checkpoints as the 4B block of Table 1. **Bold** and underline denote the best and second-best results.

Method	RealWorldQA		OCRBench	
	ACC	TLen	ACC	TLen
Qwen3-VL-4B-Thinking	67.84	784	77.70	587
GRPO	68.10	635	77.30	483
L1	67.19	1116	76.00	779
ThinkPrune	68.10	453	77.50	367
VIG (Ours)	**68.63**	**374**	**78.30**	**201**

Table 4: Ablation results on Qwen3-VL-8B-Thinking. All values are averaged over the six benchmarks of Table 1.

Factor	Variant	AvgACC	AvgTLen	AvgEff
Reward	w/o VIG	62.88	1975.4	36.44
	w/ VIG	**64.92**	**1524.3**	**52.42**
Vision pass	masked image	62.98	1885.0	38.70
	no image	**64.92**	**1524.3**	**52.42**
w_v	0.5	63.06	1676.1	45.04
	1.0	**64.92**	**1524.3**	**52.42**
	2.0	64.30	1546.0	50.76

a *single* deterministic (greedy, temperature 0) decoding pass per checkpoint rather than a mean over sampled runs, so they carry no decoding variance; where we compare two methods and the margin matters, we report paired-bootstrap confidence intervals over items instead (§4.2, Appendix F).

Training setup. All VIG models are trained directly with GRPO, without an SFT warm-up. The training set contains 2,948 samples from MM-Star (Chen et al., 2024a), MathVista (Lu et al., 2023), and LogicVista (Xiao et al., 2024). Unless otherwise specified, VIG uses the no-image comparison pass and $w_v = 1.0$. Detailed dataset statistics, hyperparameters, and implementation settings are provided in Appendix D.

4.2 Main Results

We compare VIG against the base model and four representative compression baselines—plain GRPO, L1, TALE, and ThinkPrune—across three Qwen3-VL-Thinking model sizes and six multimodal reasoning benchmarks (Table 1). Two patterns emerge. First, length-budget baselines do shorten reasoning chains relative to the base model, but only marginally improve, and sometimes degrade, average accuracy: on the 8B backbone, L1 trims AvgTLen from 2249 to 1952 while AvgACC moves only from 58.54 to 62.41, and TALE actually drops accuracy to 60.32. Second, VIG behaves qualitatively differently—rather than trading accuracy for length, it improves both simultaneously. On Qwen3-VL-8B-Thinking, VIG attains AvgACC = 64.92% and AvgEff = 52.42, surpassing the strongest compression baseline ThinkPrune (Hou et al., 2025) by $+2.37\%$ in accuracy and $+39.4\%$ in efficiency, and shortening AvgTLen by an additional 397 tokens. Compared with plain GRPO (Shao et al., 2024), which leaves visually-disengaged elaboration unchecked, VIG cuts the average chain from 1975 to 1524 tokens (-23%) while *raising* accuracy by $+2.04\%$, indicating that the gains come from removing tokens that did not help answer the question rather than from stricter length optimization. The same pattern holds on 4B, where VIG ranks first on AvgACC, AvgTLen, and AvgEff. VIG also ranks first on all three metrics at 2B, but we state that accuracy claim more narrowly: the margin over the strongest baseline is only $+0.55$ points and is not significant under a unified step-400 protocol ($+0.74$), because 53% of 2B generations hit the 4,096-token limit (vs. 31.5% at 8B) with a high repetition rate, so this block is confounded by runaway-length behaviour (Appendix F). We therefore read 2B as consistently best efficiency with accuracy on par with the strongest baseline, and treat 4B/8B as carrying the accuracy claim.

We additionally evaluate VIG on R1-Onevision-Bench (Yang et al., 2025), which provides subject-level evaluation across Physics, Mathematics, Biology, Chemistry, and Deduction (Table 2). On Qwen3-VL-8B-Thinking, VIG attains the best overall accuracy (56.79) and the best AvgEff (29.19), ranks first on Physics and Mathematics, and ties for first on Chemistry while using fewer tokens than every baseline on four of the five subjects. For instance, VIG outperforms the strongest baseline by $+3.60\%$ on Physics at 80.8% of its tokens, and by $+5.51\%$ on Mathematics at $\sim 96\%$ of its chain length. TALE compresses aggressively but trails VIG by 4.24% on overall accuracy, confirming that the joint accuracy–efficiency improvement is driven by the visual-density signal rather than by length budgeting.

Beyond mathematics. Our benchmark suite is already broader than pure mathematics—MMMU spans 30 subjects across six disciplines, and R1-Onevision covers physics, chemistry, and biology—

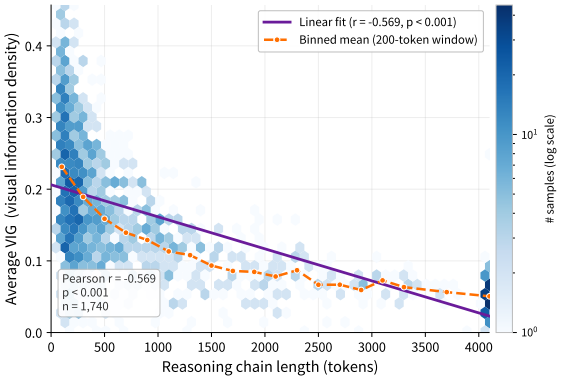


Figure 4: Average per-token VIG decreases as reasoning chains lengthen ($r = -0.569$, $p < 0.001$): long chains are dominated by visually-disengaged tokens, exactly the redundancy VIG suppresses.

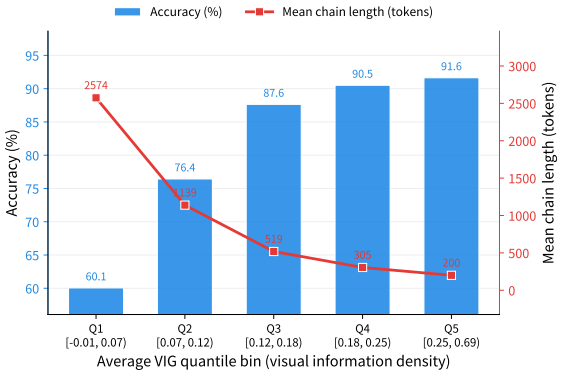


Figure 5: Accuracy and mean chain length stratified by per-sample average VIG (WeMath, $n = 1{,}740$): higher visual density co-occurs with both higher accuracy and shorter chains. This association is descriptive; it is partly mediated by chain length and item difficulty (§4.4).

yet perception-centric tasks such as visual question answering, scene understanding, and document understanding remain under-represented. To test whether the benefit of visual-density-aware optimization carries over to such tasks, we evaluate the same 4B checkpoints on RealWorldQA (xAI, 2024), which targets real-world scene understanding, and OCRBench (Liu et al., 2023b), which targets text and document recognition, under an identical decoding protocol (Table 3). The efficiency effect transfers cleanly: VIG produces the shortest chains on both benchmarks, reducing output tokens by 52% and 66% relative to the base model, while attaining the highest accuracy of all methods. We read the accuracy result conservatively, since the gaps among the stronger methods fall within noise (paired-bootstrap intervals in Appendix F); the claim these results support is that VIG preserves accuracy at a substantially lower token cost outside mathematical reasoning, rather than that it improves accuracy there.

4.3 Ablation Studies

We ablate the three design choices of VIG on Qwen3-VL-8B-Thinking, averaging across the six benchmarks of Table 1 (Table 4). *(i) The VIG reward.* Removing the VIG term from the reward reduces the method to plain GRPO, which collapses AvgEff from 52.42 to 36.44 (-30%) and lowers AvgACC by 2.04 points; conversely, adding the VIG term shortens chains by 451 tokens *while* improving accuracy. This is the central observation of the ablation: increasing visual information density and shortening reasoning are not in tension—the former drives the latter. *(ii) The text-only contrast pass.* Replacing the text-only forward pass with a 50% masked-image pass ($H_{\text{vis}}^{\text{mask}} - H_{\text{vis}}$) weakens the information-theoretic contrast, costing 1.94 points of accuracy and 13.72 points of efficiency, which validates the conditional-mutual-information formulation: removing visual context entirely produces a sharper signal than partially masking it. *(iii) Reward weight w_v.* A sweep over $w_v \in \{0.5, 1.0, 2.0\}$ traces an inverted-U on AvgEff: too small under-rewards visual grounding, while too large over-emphasizes the visual-density signal at the expense of the accuracy reward—accuracy drops by 0.62 points without producing any meaningful further chain shortening, since TLen stays around 1.5k once visually-disengaged tokens have been suppressed. $w_v = 1.0$ is the joint optimum and is used throughout the rest of the paper.

4.4 Visualization Analysis

To understand *why* VIG simultaneously improves accuracy and shortens reasoning, we examine the relationship between visual information density and chain behavior on a held-out subset of WeMath generations.

4.4.1 Visual Density vs. Chain Length

We first investigate how per-token visual information density evolves as reasoning chains grow (Figure 4, $\sim$1.7k VIG-trained generations). A clear negative trend emerges (Pearson $r = -0.569$, $p < 0.001$): as chains grow longer, the average per-token visual information gain decreases monotonically. In other words, the additional tokens that lengthen a chain are disproportionately tokens

whose distributions can be predicted from linguistic context alone—meta-cognitive openers, problem restatements, self-verification, generic algebra—rather than tokens grounded in the image. This is precisely the failure mode that uniform length penalties cannot target, and it is the reason VIG's reward, which directly upweights visually-grounded tokens, induces shorter chains as a *byproduct* of optimizing for visual density rather than as an explicit length constraint.

4.4.2 Visual Density vs. Accuracy

We next stratify the same generations into VIG quintiles and examine how accuracy and chain length co-vary with visual density (Figure 5). Both quantities move monotonically with visual density: bins with higher average VIG attain substantially higher accuracy (Q1 $\rightarrow$ Q5: 60.1% $\rightarrow$ 91.7%) and use markedly fewer tokens (Q1 $\rightarrow$ Q5: 2574 $\rightarrow$ 200; intermediate quintiles are 76.4%/1139, 87.6%/519, and 90.5%/305).

We read this analysis as *descriptive rather than causal*: a partial-correlation check shows the sample-level ACC–VIG association is largely mediated by chain length and item difficulty (partial $r \approx 0$ for correctness after controlling both, while the VIG–length relation survives difficulty control at partial $r = -0.63$). The causal evidence that raising visual density *produces* this behaviour is therefore carried by the controlled ablations of §4.3 and Appendix F, not by this correlation. With that qualification, the two analyses indicate that VIG acts not as a length budget but as a mechanism that discards visually-irrelevant elaboration, yielding the chains documented in Tables 1–4.

5 Conclusion

We introduced **VIG** (Visual Information Gain), an information-theoretic reward for multimodal chain-of-thought compression that measures how much each reasoning token depends on the image. Unlike existing compression methods that optimize for shorter outputs directly, VIG trains the model to concentrate its reasoning on visually informative content, thereby improving efficiency by increasing visual density rather than by imposing a length budget. The reward is fully online, requires no reference chains or external annotators, and integrates naturally with GRPO through a two-pass forward computation. Across six main multimodal reasoning benchmarks and three Qwen3-VL-Thinking model sizes, plus an additional R1-Onevision-Bench evaluation on 8B, VIG consistently improves the accuracy–efficiency trade-off and ranks first overall and first or tied-first on three of five R1-Onevision subject categories. Overall, our results suggest that efficient multimodal CoT generation requires not merely thinking less, but making each token rely more on visual evidence.

Limitations

VIG has several important limitations. Computing the reward requires an additional text-only forward pass during RL training, which raises optimization cost relative to standard GRPO, and it depends on a clean separation between visual and textual tokens, so its implementation is somewhat model-specific.

Most importantly, VIG measures *visual dependence*, which is neither *visual correctness* nor *counterfactual necessity*. A score is high whenever the image substantially shifts the policy's next-token distribution, so a confidently hallucinated reading can score as high as genuinely grounded perception; the accuracy reward counteracts this only at the trajectory level, and combining VIG with explicit grounding signals is a natural extension. Nor does high VIG imply indispensability: as Appendix F shows, deleting low-VIG sentences from an already-compressed chain hurts more than deleting high-VIG ones, because what survives at low VIG is largely load-bearing computation. VIG is therefore suited to shaping the output distribution during training, but not to post-hoc pruning.

Finally, our study covers static image reasoning with Qwen3-VL-Thinking models, and we regard cross-architecture generality as *unresolved*: a rigorous cross-family study would require matched thinking-mode backbones and per-family re-tuning of the RL stack, and a single under-tuned transfer run would be weak evidence either way. We release our vision-span abstraction, which lets other families be plugged in without changing the reward, to facilitate such verification.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Milad Aghajohari, Kamran Chitsaz, Amirhossein Kazemnejad, Sarath Chandar, Alessandro Sordoni, Aaron Courville, and Siva Reddy. 2025. The marko-

vian thinker: Architecture-agnostic linear scaling of reasoning. *arXiv preprint arXiv:2510.06557*.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv e-prints*, pages arXiv–1801.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2024. Insight-V: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24842–24855.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Bin Hong, Jiayu Liu, Kai Zhang, Jianwen Sun, Mengdi Zhang, and Zhenya Huang. 2025. Pruning long chain-of-thought of large reasoning models via small-scale preference optimization. *arXiv preprint arXiv:2508.10164*.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. 2025. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning. *arXiv preprint arXiv:2504.01296*.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2025a. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2025b. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. 2026. Making slow thinking faster: Compressing llm chain-of-thought via step entropy. In *The Fourteenth International Conference on Learning Representations*.

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

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023b. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. 2025. OThink-MR1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*.

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

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6774–6786.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. Cot-valve: Length-compressible chain-of-thought tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6035.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. LMM-R1: Empowering 3b LMMs with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025a. SoTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yibo Wang, Haotian Luo, Huanjin Yao, Tiansheng Huang, Haiying He, Rui Liu, Naiqiang Tan, Jiaxing Huang, Xiaochun Cao, Dacheng Tao, and 1 others. 2025b. R1-compress: Long chain-of-thought compression via chunk compression and search. *arXiv preprint arXiv:2505.16838*.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and 1 others. 2025c. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

xAI. 2024. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v. RealWorldQA benchmark released at https://huggingface.co/datasets/xai-org/RealworldQA.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

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.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025. Lightthinker: Thinking step-by-step compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13318–13339.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025a. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Shangziqi Zhao, Jiahao Yuan, Jinyang Wu, Zhenglin Wang, Guisong Yang, and Usman Naseem. 2025b. Can pruning improve reasoning? revisiting long-cot compression with capability in mind for better reasoning. *arXiv preprint arXiv:2505.14582*.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-Zero's "aha moment" in visual reasoning on a 2b non-SFT model. *arXiv preprint arXiv:2503.05132*.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.

A Preliminaries

A.1 Multimodal CoT

Given a visual question (I, q) composed of an image I and a textual query q, a modern Multimodal Large Reasoning Model produces a structured response of the form

$$\underbrace{\langle \texttt{think} \rangle\ t_1 \cdots t_{|T|}\ \langle \texttt{/think} \rangle}_{\text{reasoning chain } T}\ \underbrace{\langle \texttt{answer} \rangle\ a\ \langle \texttt{/answer} \rangle}_{\text{final answer}}$$

where the <think> block contains an ordered sequence of reasoning tokens $T = (t_1, \ldots, t_{|T|})$, and a is the final answer. VIG operates exclusively on the reasoning chain T, leaving the format and answer portions unaffected.

A.2 GRPO

To optimize the policy, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which replaces a learned value baseline with intra-group reward normalization. For each prompt (I, q), the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_1, \ldots, o_G\}$ with scalar rewards $\{r_1, \ldots, r_G\}$. The group-relative advantage assigned to the i-th sample is

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \tag{7}$$

Let $\rho_{i,t}(\theta) = \pi_\theta(o_{i,t} \mid q, I, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, I, o_{i,<t})$ be the importance ratio at token t. GRPO then maximizes the clipped surrogate objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,I),\,\{o_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \mathcal{L}_{i,t} \right], \tag{8}$$

where the token-level term is

$$\begin{aligned} \mathcal{L}_{i,t} = \ & \min \Big(\rho_{i,t} \hat{A}_i, \text{clip}(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon) \hat{A}_i \Big) \\ & - \beta\, D_{\text{KL}} \big(\pi_\theta \,\|\, \pi_{\text{ref}} \big). \end{aligned} \tag{9}$$

Here $(q, I) \sim \mathcal{D}$ and $\{o_i\} \sim \pi_{\theta_{\text{old}}}$, while ε is the PPO-style clipping range and β controls the KL regularization strength. Under this framework, all task-specific supervision enters through the scalar reward r_i, which is exactly where our VIG-based reward is introduced.

B Two-Pass VIG Computation

Algorithm 1 details how R_{VIG} is computed online during each GRPO rollout. For every sampled response $o = (s_T, a)$ produced by π_θ, the controller assembles two parallel inputs from the same prompt and chain: `vis_inputs`, the standard multimodal input that includes the visual token spans <vision>$\cdots$</vision>; and `txt_inputs`, identical except that all visual spans have been stripped out so the model conditions purely on textual context. Both inputs are batched and passed through π_θ in a single forward call, so the cost of computing R_{VIG} is two forward passes per rollout sample—no extra weights, no backward pass, and no auxiliary models.

From the resulting logits we compute per-token entropies $H_{\text{vis}}(t) = -\sum_v p_v^{\text{vis}} \log p_v^{\text{vis}}$ and $H_{\text{txt}}(t) = -\sum_v p_v^{\text{txt}} \log p_v^{\text{txt}}$, then take their position-wise difference. A binary mask restricts the difference to positions inside the <think> block; positions corresponding to format tags, the <answer> segment, and any padding are excluded so that the reward reflects only the reasoning chain. The masked sequence is averaged across positions and clipped to $[-1, +1]$ to absorb the rare floating-point negative values discussed in Appendix C. The resulting scalar enters the composite reward of Eq. (4) without any gradient flowing through the two forward passes, so VIG behaves as a pure reward signal rather than a differentiable auxiliary loss.

Algorithm 1 Computing R_{VIG} for a sampled response

Input: model π_θ, processor, image I, question q, think text s_T
Output: scalar $R_{\text{VIG}} \in [-1, +1]$
[Vision pass] $\leftarrow$ build `vis_inputs` from (I, q, s_T)
  Run forward: $\text{logits}_{\text{vis}} \leftarrow \pi_\theta(\text{vis_inputs})$
  $H_{\text{vis}}(t) \leftarrow -\sum_v p_v \log p_v$ (per token)
[Text pass] Remove all `<vision>`$\cdots$`</vision>` tokens
  Run forward: $\text{logits}_{\text{txt}} \leftarrow \pi_\theta(\text{txt_inputs})$
  $H_{\text{txt}}(t) \leftarrow -\sum_v p_v \log p_v$ (per token)
[Aggregate] mask $\leftarrow$ positions in $\langle$`think`$\rangle$ block
  $\text{VIG_seq} \leftarrow (H_{\text{txt}} - H_{\text{vis}})[\text{mask}]$
  $R_{\text{VIG}} \leftarrow \text{clip}(\text{mean}(\text{VIG_seq}),\ -1,\ +1)$

C Proof of VIG Non-Negativity

Population statement. Let p denote the true joint distribution over $(t_\tau, q, I, \text{ctx}_{<t})$. The conditional mutual information satisfies $I(t_\tau;\ I \mid q,\ \text{ctx}_{<t}) \geq 0$ with equality iff t_τ is conditionally independent of I given $(q,\ \text{ctx}_{<t})$ (Cover and Thomas, 2006). Equivalently, conditioning on additional information cannot increase entropy:

$$H(t_\tau \mid q,\ I,\ \text{ctx}_{<t}) \ \leq\ H(t_\tau \mid q,\ \text{ctx}_{<t}), \quad (10)$$

so $H_{\text{vis}}(t_\tau) \leq H_{\text{txt}}(t_\tau)$ and $\text{VIG}(t_\tau) \geq 0$ at the population level.

Estimator and sources of negative values. In practice H_{vis} and H_{txt} are not population entropies but *plug-in estimates* obtained from the policy π_θ: $\hat{H}_{\text{vis}}(t_\tau) = -\sum_v \pi_\theta(v \mid q, I, \text{ctx}_{<t}) \log \pi_\theta(v \mid q, I, \text{ctx}_{<t})$ and analogously for $\hat{H}_{\text{txt}}$ with the visual tokens stripped from the input. Because π_θ is not the true conditional—and, crucially, because the text-only pass evaluates the same multimodal-trained network on an out-of-distribution input—the empirical entropies need not satisfy the population inequality above. We identify three sources of occasional negative empirical VIG:

1. *Policy mis-specification.* π_θ is a learned model, not the data-generating distribution; the data-processing inequality is a statement about p, not about π_θ, so noisy estimates can flip sign on individual tokens.
2. *Out-of-distribution text-only pass.* The shared weights encode multimodal training; removing the visual tokens places the network outside its training distribution, occasionally inducing higher-entropy or attention-leakage behaviour (*phantom* attention to absent vision positions).
3. *Numerical noise.* Computing entropy in float16/bf16 with `F.log_softmax` introduces $\mathcal{O}(10^{-3})$ rounding error; in well-grounded tokens this can manifest as small negative VIG.

Empirically, negative values are rare and concentrated near zero (less than $\sim$1% of think-block tokens, $|\text{VIG}| < 0.05$ for the vast majority of the affected tokens). We therefore clip R_{VIG} to $[-1, +1]$ and average over the `<think>` block, which renders the reward stable while preserving the information-theoretic interpretation of Eq. (1) at the population level.

D Data Details

To comprehensively evaluate multimodal CoT compression, we use seven reasoning benchmarks in total: six benchmarks for the main comparison plus one additional benchmark with subject-level evaluation. For RL training, we construct the training mixture from three datasets: MMStar, MathVista, and LogicVista. Below we summarize the role and characteristics of each dataset used in this paper.

Evaluation benchmarks.

- **WeMath** (Qiao et al., 2024): a multimodal mathematics benchmark designed to study how large multimodal models solve visual math problems rather than only measuring end-to-end performance. We evaluate on 1,740 examples.
- **MMMU** (Yue et al., 2023): a college-level multimodal benchmark spanning diverse disciplines and requiring broad knowledge together with strong visual understanding. We evaluate on 900 examples.
- **MathVision** (Wang et al., 2024): a benchmark for multimodal mathematical reasoning over diagrams, formulas, charts, and geometry figures. We evaluate on 3,040 examples.
- **DynaMath** (Zou et al., 2024): a visual mathematical reasoning benchmark designed to test robustness under dynamic perturbations and distribution shifts. We evaluate on 5,010 examples.
- **MMK12** (Meng et al., 2025): a multi-subject K–12 multimodal mathematics benchmark with human-verified answers, introduced alongside MM-Eureka. We evaluate on the 2,000-example test split released by Wang et al. (2025c).
- **Geo3K** (Lu et al., 2021): a geometry benchmark centered on diagram understanding and symbolic reasoning. We evaluate on 601 examples.
- **R1-Onevision-Bench** (Yang et al., 2025): a comprehensive multimodal reasoning benchmark covering mathematics, physics, chemistry,

biology, and deduction. We use all 942 examples for subject-level evaluation.

RL training data.

- **MMStar** (Chen et al., 2024a): a challenging multimodal benchmark emphasizing fine-grained perception, spatial reasoning, and image–text alignment. We sample from its validation split for RL training.
- **MathVista** (Lu et al., 2023): a diverse multimodal mathematics benchmark covering charts, geometry, symbolic reasoning, and visual word problems. We use its testmini split for RL training.
- **LogicVista** (Xiao et al., 2024): a benchmark for multimodal logical reasoning in visual contexts, containing problems that require multi-step deduction grounded in images. We use its test split for RL training.

In total, the RL training mixture contains 2,948 examples.

E Implementation Details

Baseline implementation. **Base** directly uses the original Qwen3-VL-Thinking checkpoints without additional post-training. **GRPO** (Shao et al., 2024) uses the same training pipeline as VIG but removes the VIG reward, retaining only the format and accuracy terms. **L1** (Aggarwal and Welleck, 2025) follows the same GRPO framework with a length-oriented reward, **TALE** (Han et al., 2025) is implemented as a training-free prompting baseline with a dynamic token budget, and **ThinkPrune** (Hou et al., 2025) uses a pruning-oriented RL signal based on reasoning-length control. For fair comparison, all RL-based baselines use the same training data mixture, model backbones, and evaluation protocol as VIG.

VIG implementation. For each sampled response, VIG computes token-level visual information gain by running the same policy twice: once with the image and once with the visual tokens removed. In Qwen3-VL, visual tokens are enclosed by `<|vision_start|>` and `<|vision_end|>` markers; the text-only pass removes the full span between these markers. The resulting entropy difference is restricted to the `<think>` block, averaged over tokens, and clipped to obtain R_{VIG}. Unless otherwise specified, all main experiments use token-level aggregation, the no-image comparison pass, and reward weight $w_v = 1.0$.

Training configuration. All models are trained with full-parameter GRPO using group size $G = 8$, temperature 1.0, KL coefficient 0.04, and at most 500 optimization steps. We use batch size 1 per device, gradient accumulation 4, learning rate 10^{-6}, warmup steps 10, gradient checkpointing, and DeepSpeed ZeRO-2. The optimal checkpoint for VIG is selected at step 400 in the main experiments; baselines use step 500 for Qwen3-VL-2B-Thinking and Qwen3-VL-8B-Thinking, and step 400 for Qwen3-VL-4B-Thinking.

Numerical details. Entropy is computed in nats using `F.log_softmax` for numerical stability. In practice, VIG values are clipped to $[-1, 1]$ before being used as rewards.

F Signal-Integrity Controls, Deletion Probe, and Scope

This appendix reports the controls that establish *why* the VIG term produces the gains of §4.3, a probe that delimits what the signal measures, and the analyses that bound the scope of our claims.

F.1 Protocols

Confidence-only arm. This arm shares the pipeline, data, seed, and step count of VIG, with the reward's visual term replaced by $-H_{\text{vis}}$ computed over the same `<think>` mask. It therefore rewards low predictive entropy *with* the image and is by construction insensitive to the visual contrast, which isolates whether the gain could come from entropy or brevity shaping alone.

Stratification. For every evaluation item we run the untrained base model with the image removed under otherwise identical decoding settings. Items it still answers correctly form the *text-solvable* stratum and the remainder the *vision-required* stratum; VIG and GRPO are then re-scored within each stratum, with confidence intervals from a 10k paired bootstrap over items.

Sentence typing. Sentences are labelled *visual* when they contain image-referential expressions, *algebra* when they contain arithmetic or symbolic manipulation, *meta* for metacognitive openers and self-checks, and *other* otherwise, with *visual* taking precedence when a sentence qualifies for both. Over the 70,754 sentences of VIG-8B

WeMath outputs, mean VIG by type is 0.153 (visual, n=10,821), 0.106 (meta, 6,780), 0.088 (other, 32,524), and 0.065 (algebra, 20,629).

Deletion probe. Sentences of a completed chain are ranked by VIG, the stated fraction is removed, and the surviving text is supplied as a forced prefix from which the answer is re-decoded greedily. We report two conditions, with the image still in context and with the image also removed at re-decoding. The VIG-chain probe covers rates $\{10, 20, 30, 40, 50\}\%$ and the GRPO-chain probe $\{25, 50\}\%$.

F.2 Interpreting the Deletion Asymmetry

Table 6 shows that on VIG-trained chains, deleting the lowest-VIG sentences is the most damaging condition, the opposite of what a token-importance reading of VIG would predict. Two mechanisms combine to produce this.

The first is content type. Low-VIG sentences are enriched in algebra and derivation, which is visually ungrounded yet computation-bearing, so removing it breaks the arithmetic. A high-VIG sentence, by contrast, is typically a readout of a value that remains recoverable at re-decoding time because the image is still in context; consistent with this, high-VIG deletion costs only 1.67 points even when the image is *also* removed.

The second is chain state. The asymmetry is specific to chains produced by a policy already trained for visual density. On chains from a GRPO-trained policy the ordering reverses, so the reward does track answer-critical visual content when that content has not yet been compressed away. One confound is worth stating explicitly: low-VIG sentences are somewhat longer on average, so deleting a fixed fraction of *sentences* removes more tokens in that condition, although the strict ordering holds at every rate.

F.3 Scope of the Comparison

Trajectory-level reweighting (VPPO). VPPO (Huang et al., 2025a) is the closest contemporaneous method, reweighting trajectory advantages by an overall visual-dependency measure rather than scoring individual tokens. Under our protocol, VIG reaches 71.37 AvgACC at 1,297 tokens versus 63.39 at 1,100 tokens for the released VPPO-8B checkpoint: VPPO produces slightly shorter chains and therefore a higher efficiency ratio, while VIG retains substantially higher accuracy. We treat this as *indicative only* rather than an apples-to-apples comparison, because the available checkpoint is built on a different backbone (Qwen3-VL-8B-Instruct rather than the Thinking variant) and trained on different data; a fully matched reproduction is left to future work.

Table 5: Signal-integrity controls on Qwen3-VL-8B-Thinking. All arms share identical data, seed, and step count. **(a)** Reward ablation, 5-benchmark macro-average. **(b)** VIG vs. GRPO split by whether the base model solves the item without the image, 6-benchmark macro-average.

(a) Reward ablation

Reward	AvgACC	AvgTLen
GRPO (no visual term)	69.33	1751
Confidence-only ($-H_{\text{vis}}$)	68.21	1441
VIG (ours)	**71.37**	**1297**

(b) Stratified accuracy

Stratum	VIG	GRPO	Δ
Text-solvable	83.18	81.69	+1.49
Vision-required	56.01	53.19	**+2.82**

Failure cases. MathVision, the hardest and longest-chain benchmark in our suite, is the one benchmark where VIG does not lead accuracy: ThinkPrune (32.57) and TALE (32.50) edge past VIG (32.20) at 8B, and L1 leads at 4B (31.71 vs. 30.95). The margins lie within noise (≤ 0.8 points) but are consistent across scales, while VIG retains by far the shortest chains (2,735 vs. $\geq$3,004 tokens at 8B). Our reading is that when a problem genuinely requires long exploration, the density signal has less redundancy to remove and mild compression can trim useful search. Two further limits are worth stating. On 2B, 53% of generations hit the 4,096-token limit (versus 31.5% at 8B) with a high repetition rate, so accuracy differences in that block are dominated by runaway-length behaviour. And because R_{VIG} is a chain-level mean, a chain that hinges on one small visual detail can be under-rewarded relative to one with many mildly visual sentences, as illustrated in Figure 7(b). At more generous decoding budgets (8k tokens, no truncation pressure), accuracy differences among RL recipes narrow while VIG's efficiency advantage persists.

Table 6: Quantile-deletion probe (WeMath, $n = 1{,}740$). Sentences are ranked by VIG, a fraction is deleted, and the answer is re-decoded from the surviving prefix with the image still in context. **Bold** marks the most damaging condition per block.

(a) VIG-8B chains

Deleted	10%	20%	30%	40%	50%
lowest-VIG	80.92	79.94	78.22	75.17	**72.36**
random	81.84	80.92	80.06	79.25	77.41
highest-VIG	82.07	81.95	81.78	81.32	80.75

(b) GRPO-8B chains

Deleted	25%	50%
lowest-VIG	81.21	80.00
random	80.57	79.48
highest-VIG	80.46	**75.92**

G Training Cost

VIG requires a second, text-only forward pass per rollout sample, so its theoretical worst-case overhead is $2\times$ the cost of plain GRPO. In practice the gap is substantially smaller because (i) the text pass runs on a shorter sequence (the visual-token spans are removed before the forward), (ii) only logits are needed—no backward pass and no optimizer update for the second pass, and (iii) loading the rollout, computing advantages, and stepping the optimizer all amortize across the two passes.

We report end-to-end training cost on Qwen3-VL-8B-Thinking, all jobs trained on a single $8\times$ H20 node for 500 GRPO steps with the same batch size, rollout length, and dataset.

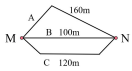
Table 7: Training cost on Qwen3-VL-8B-Thinking ($8\times$H20, 500 GRPO steps). *Rel.* is the relative wall-clock cost vs. plain GRPO.

Method	**s/it**	**Wall-clock**	**GPU-h**	**Rel.**
GRPO	31.8	4h 25m	35.3	1.00×
L1	33.5	4h 39m	37.2	1.05×
ThinkPrune	32.3	4h 29m	35.8	1.01×
VIG (Ours)	37.2	5h 10m	41.3	1.17×

VIG adds only $\sim$17% to per-step wall-clock relative to plain GRPO, and remains comparable to other length-based RL baselines such as L1 ($1.05\times$) and ThinkPrune ($1.01\times$), because the second forward operates on a shorter, vision-stripped input and does not require gradient computation. We did not observe sample-efficiency degradation: VIG converges within the same 500-step budget as the other baselines and reaches its final accuracy without additional rollouts. We therefore view the cost as a modest, predictable overhead that buys an information-theoretic reward signal otherwise unavailable through purely textual compression objectives.

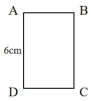
sentence VIG: <0 ≤ 0.1 ≤ 0.25 ≤ 0.5 >0.5

(a) *There are three paths from point M to point N, one of which is the shortest. This path is ( ).*
Options: A. A; B. B; C. C; D. Cannot be determined; E. No correct answer | gold **B**, model **B**



VIG	Reasoning sentence
+0.14	So, let's see.
+0.04	The problem is about finding the shortest path from M to N.
+0.04	There are three paths: A, B, C.
+0.55	Let's check their lengths.
+1.16	Path A: The label says 160m.
+0.44	Path B: 100m.
+0.28	Path C: 120m.
+0.13	So comparing 100, 120, 160, the shortest is B.
+0.45	So the answer should be B.

(b) *The perimeter of rectangle ABCD is 20 cm. What is the length of AB in cm?*
Options: A. 4; B. 3; C. 2; D. 5; E. No correct answer | gold **A**, model **A**



VIG	Reasoning sentence
-0.10	So, let's solve this.
+0.73	First, the rectangle has sides AD and BC as 6 cm (since . . .
+0.21	The perimeter of a rectangle is 2*(length + width).
+0.29	Let's denote AB as the length, which is the top side, and . . .
+0.09	The perimeter is 20 cm.
+0.10	So formula: 2*(AB + AD) = 20.
+0.05	We know AD is 6, so plug that in: 2*(AB + 6) = 20.
+0.01	Divide both sides by 2: AB + 6 = 10.
+0.10	Then AB = 10 - 6 = 4.
+0.98	So the answer should be A.

Figure 6: Sentence-level VIG within complete VIG-8B chains, shown beside the input image. Sentences that read a quantity off the figure score highest ($+1.16$, $+0.73$), while symbolic manipulation scores near zero even when answer-bearing ("Divide both sides by 2", $+0.01$).

When is the overhead repaid? The overhead is a one-time training cost, whereas the token savings recur at every inference request, so it is useful to state the break-even point explicitly. Training VIG costs $41.3 - 35.3 = 6.0$ additional GPU-hours. At inference under our unified 4,096-token protocol, VIG emits 1,297 output tokens per query versus 1,751 for plain GRPO, a saving of $\sim$454 tokens per request. Taking a deliberately conservative aggregate decode throughput of 1,000 output tokens per second per GPU for batched serv-

ing of an 8B model on this hardware, 6.0 GPU-hours corresponds to $\sim$21.6M generated tokens, so the extra training cost is repaid after roughly $21.6\text{M}/454 \approx 48\text{K}$ inference requests. Beyond that point VIG is strictly cheaper end-to-end than plain GRPO while also being more accurate. The estimate scales linearly with the assumed throughput, so a faster serving stack shortens the break-even point proportionally; the qualitative conclusion is that the overhead is amortized within a small fraction of a deployed model's lifetime.

H Supplementary Visualization Examples

This appendix contains two kinds of qualitative material: additional sentence-level VIG colourings (Figure 7), and side-by-side output comparisons against baselines (Figures 8–9).

To complement the motivating example in Figure 2, we also provide two case studies comparing the base model, L1, and VIG. These examples illustrate the same qualitative pattern observed throughout the paper: the base model tends to over-elaborate and drift away from decisive visual evidence, L1 shortens the chain but does not always improve visual grounding, and VIG produces shorter reasoning that remains tightly anchored to the image.

I Datasets and Licenses

All evaluation benchmarks used in this work are publicly released for research use. WeMath (Qiao et al., 2024), MathVision (Wang et al., 2024), DynaMath (Zou et al., 2024), Geo3K (Lu et al., 2021), MMMU (Yue et al., 2023), MMK12 (Meng et al., 2025; Wang et al., 2025c), and R1-Onevision-Bench (Yang et al., 2025) are distributed under research-permissive licenses (Apache-2.0, MIT, or CC-BY-NC-4.0 variants); our use is restricted to non-commercial academic evaluation, consistent with their stated intended use. The two additional perception-centric benchmarks used in §4.2 are also publicly available: RealWorldQA (xAI, 2024), released by xAI alongside the Grok-1.5V announcement and distributed on Hugging Face under CC-BY-ND-4.0, and OCRBench (Liu et al., 2023b), released with the MultimodalOCR project. We use the public splits without modification and report only aggregate evaluation metrics, consistent with their stated intended use. The Qwen3-VL-Thinking 2B/4B/8B base models (Bai et al., 2025) are released under the Qwen license and used here

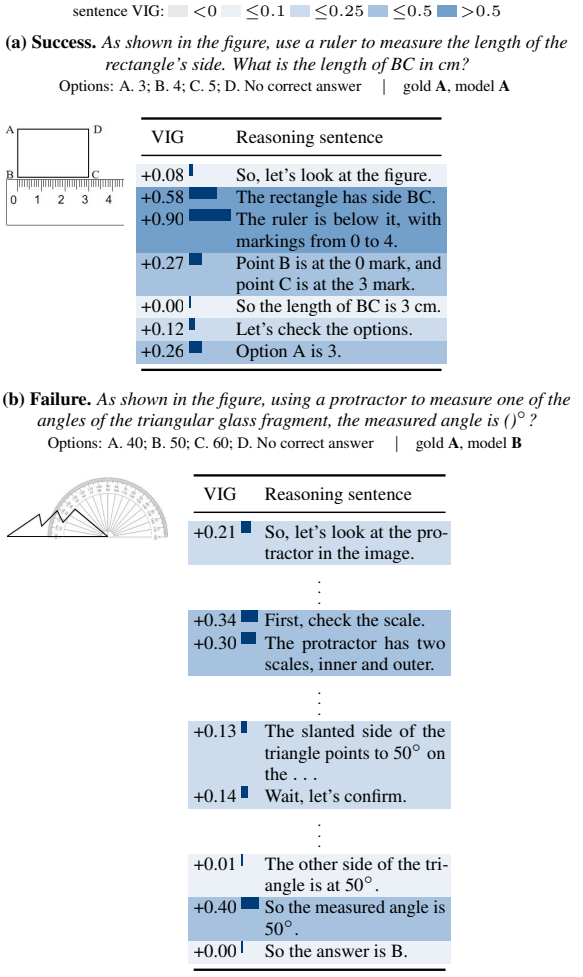


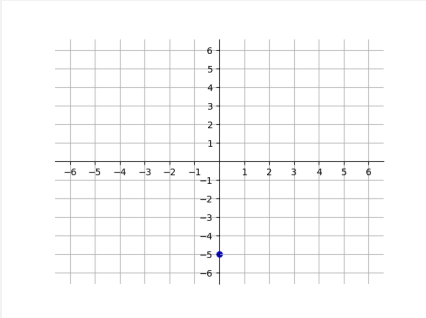
Figure 7: Additional cases (⋮ omitted sentences). (a) Reading the ruler scores $+0.90$, the answer-bearing arithmetic $+0.004$. (b) A failure case: every sentence is visually dependent, yet the model misreads the protractor as 50° instead of 40°. High VIG certifies image-driven generation, not a correct reading.

only for research. None of the benchmarks contain personally identifying information or offensive content: all items are mathematical, scientific, or geometric reasoning problems in English.

J Use of AI Assistants

We used AI assistants (ChatGPT and Claude) in a strictly limited capacity. Their use was confined to (i) language polishing of the manuscript—improving grammar, phrasing, and consistency of terminology—and (ii) minor coding assistance for non-research utilities such as plotting scripts, LaTeX formatting, and small data-processing snippets. All research ideas, the VIG formulation, the algorithmic design, the training pipeline, the experimental protocol, and the analysis are the authors' own. No AI assistant was used to generate experimental

results, write the information-theoretic derivations,
or produce any scientific claim in this paper.



Question:
If the blue point is reflected about the origin, what are the coordinates of its image? Please answer in the form (_, _)

Answer: (0, 5)

Base:
<think>Got it, let's see. The problem is about reflecting a blue point about the origin. First, I need to find **...** Wait, let's check. The x-coordinate **...** So the original point is (-1, -5). Now, reflecting about the origin. The rule for reflection over the origin is **...** So (1, 5) is in the first quadrant, which makes sense. Let's confirm the coordinates again. The blue dot: x is -1 (left of y-axis), y is -5 (below x-axis). **So (-1, -5). Reflecting over origin: (1, 5). Yep, that seems right.**</think><answer>(1, 5)\n</answer>

 # Thinking Token 695

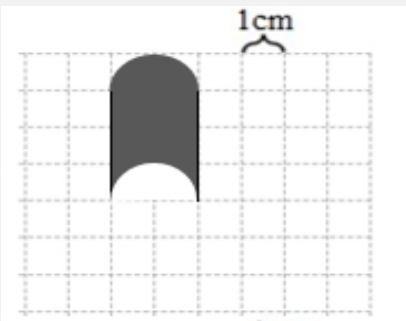
L1:
<think>Got it, let's see. The problem is about reflecting a blue point about the origin. First, I need to find **...** So if the original point is (a, b), the reflected point is (-a, -b). So here, a is 0, b is -5. So the reflected point would be (-0, -(-5)) which is (0, 5). Wait, let me confirm **...** Let me check the grid **...** that's the same distance from the origin but in the opposite direction. So from (0, -5), moving 5 units up along the y-axis (since origin is (0,0)), **we get to (0, 5). that's correct.**</think><answer>(0, 5)</answer>

 # Thinking Token 511

VIG (Ours):
<think>So, let's see. The blue point is on the grid. First, find its current coordinates. The x-axis is horizontal, y-axis vertical. The point is at x = 0 (since it's on the y-axis) and y = -5 (because it's 5 units down from the origin). So original coordinates are (0, -5). Reflecting about the origin means changing both x and y signs. So (0, -5) becomes (0, 5). Wait, let me check. Origin reflection: (x, y) → (-x, -y). So 0 becomes 0 (since -0 is 0), and -5 becomes 5. **So the image is (0, 5).**</think><answer>(0, 5)</answer>

 # Thinking Token 168

Figure 8: An additional qualitative comparison between Base, L1, and VIG. VIG produces a shorter reasoning chain while preserving the key visual evidence needed for the final answer.



Question:
Move and calculate, what is the area of the shaded part in the figure below?
A. 5 cm²
B. 6 cm²
C. 8 cm²
D. No correct answer

Answer: B

Base:
<think>The figure shows a grid with a shaded shape. The shape appears to be a rectangle with a semicircle on top and a semicircle at the bottom, but the bottom semicircle is hollow (white) **...** So shaded area = 8 + 0.125π - 0.125π = 8 cm².Wait, but the options include 8 cm² as option C. But let me check again. Maybe **...** But if the top and bottom semicircles have the same radius, then they cancel out, so the shaded area is just the area of the rectangle **...** **So the answer should be 8 cm², which is option C.**</think><answer>C</answer>

 # Thinking Token 841

L1:
<think>Okay, let's see. I need to find the area of the shaded part in the figure. The problem says to move and calculate, so **...** If I take the top semicircle and move it down to the bottom, would they fit together? Wait **...** But wait, the bottom semicircle is a hole, so **...** But since the top and bottom semicircles are the same size, they cancel out. So the area is just the area of the rectangle. Wait, the rectangle is 3cm wide and 2cm tall. **So 3*2=6cm². So the answer is B. 6 cm².**</think><answer>B</answer>

 # Thinking Token 968

VIG (Ours):
<think>The task is to calculate the area of the shaded region. Looking at the figure, the shaded region resembles a \"capsule\" shape — semicircles on both the top and bottom. Let me count the grid squares for the length and width **...** The top and bottom semicircles can be fitted together, so the total shaded area equals the area of the middle rectangle. Counting: length = 3 cm (3 horizontal squares), width = 2 cm (2 vertical squares). **Area = 3 × 2 = 6 cm². Option B is 6 cm².**</think><answer>B</answer>

 # Thinking Token 174

Figure 9: A second qualitative example comparing Base, L1, and VIG. Compared with the baseline methods, VIG better concentrates its reasoning on the visually relevant information and avoids unnecessary self-reflection.