# Bridging Visual Representation and Reinforcement Learning from Verifiable Rewards in Large Vision-Language Models

Yuhang Han[1,3], Yuyang Wu[1], Zhengbo Jiao[1], Yiyu Wang[1,3], Xuyang Liu[1], Shaobo Wang[1], Hanlin Xu[2], Xuming Hu[3], and Linfeng Zhang[1⋆]

[1] EPIC Lab, Shanghai Jiao Tong University, Shanghai, China
[2] Huawei, China
[3] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

**Abstract.** Reinforcement Learning from Verifiable Rewards (RLVR) has substantially enhanced the reasoning capabilities of large language models in abstract reasoning tasks. However, its application to Large Vision-Language Models (LVLMs) remains constrained by a structural representational bottleneck. Existing approaches generally lack explicit modeling and effective utilization of visual information, preventing visual representations from being tightly coupled with the reinforcement learning optimization process and thereby limiting further improvements in multimodal reasoning performance. To address this limitation, we propose `KAWHI` (**K**ey-Region **A**ligned **W**eighted **H**armonic **I**ncentive), a plug-and-play reward reweighting mechanism that explicitly incorporates structured visual information into uniform reward policy optimization methods (*e.g.*, GRPO and GSPO). The method adaptively localizes semantically salient regions through hierarchical geometric aggregation, identifies vision-critical attention heads via structured attribution, and performs paragraph-level credit reallocation to align spatial visual evidence with semantically decisive reasoning steps. Extensive empirical evaluations on diverse reasoning benchmarks substantiate `KAWHI` as a general-purpose enhancement module, consistently improving the performance of various uniform reward optimization methods. Code is available at `https://github.com/kawhiiiileo/KAWHI`.

**Keywords:** Reinforcement Learning · Large Vision-Language Models · Harmonic Incentive

## 1 Introduction

Reinforcement Learning from Verifiable Rewards (RLVR), particularly as instantiated through online training paradigms such as Group Relative Policy Optimization (GRPO), has substantially advanced the logical deduction capabilities of Large Language Models (LLMs) within abstract reasoning scenarios

⋆ Corresponding author.

confined to unimodal textual contexts [7,29,37,42,53,66,72]. Recently, this training paradigm has progressively extended into multimodal cognitive domains, thereby stimulating sustained scholarly efforts toward its cross-modal adaptation and theoretical reconfiguration within Large Vision-Language Models (LVLMs). Existing research predominantly advances along three principal technical trajectories: systematic curation of training corpora [8, 19, 24, 25, 28, 36, 63], principled modeling of reward functions [17, 43, 46, 50, 57, 58, 64], and structural reformulation of underlying optimization algorithms [47, 71]. However, although prior studies have achieved certain performance gains in enhancing the reasoning capabilities of LVLMs, they have ***failed to establish a principled coupling between visual representations and the RLVR optimization paradigm***, resulting in suboptimal performance improvements.

Recent LVLMs [4, 5, 55] have exhibited impressive Chain-of-Thought (CoT) reasoning capabilities in complex tasks such as mathematical problem solving and chart interpretation. However, their effectiveness remains constrained by a fundamental representational bottleneck: the implicit assumption that dense tokenization suffices to achieve exhaustive and homogeneous visual encoding. This premise overlooks ***the intrinsically sparse spatial distribution and task-contingent relevance that characterize visual scenes***, thereby preventing models from accurately isolating the discriminative visual evidence essential to downstream reasoning. Such a perceptual bottleneck consequently biases attention allocation in RLVR-trained LVLMs, diminishing sensitivity to visually salient cues and ultimately constraining overall performance.

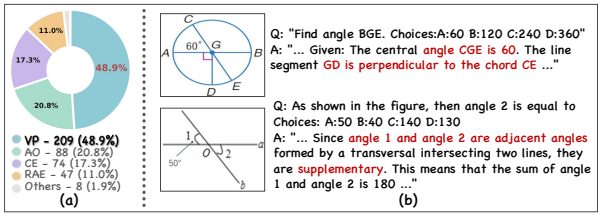


**Fig. 1:** (a) Error distribution across five categories: VP (visual perception error), AO (answer only), CE (calculation error), RAE (rule application error), and Others. (b) Two VP cases from MathVerse. Red highlights indicate visual misinterpretation.

To empirically substantiate this claim, we conduct a dual diagnostic analysis (Fig. 1) on Qwen2.5-VL-7B-Instruct [5] under the GRPO method. At the quantitative level (Fig. 1(a)), we employ Qwen3-VL-30B-A3B [4] as an arbiter to perform fine-grained attribution analysis of erroneous responses on MathVerse [69]. The results indicate that VP errors constitute 48.9% of failures, substantially exceeding other categories such as CE and RAE (see Fig. 1 caption for definitions). This disparity underscores that **deficiencies in visual encoding**, rather than peripheral noise factors, represent the primary bottleneck constraining overall model performance. Complementing these quantitative findings, qualitative case analysis (Fig. 1(b)) further reveals the underlying mechanism. Specifically, **systematic misinterpretation of discriminative visual evidence** (see the

red-highlighted region in Fig. 1(b)) precipitates cascading failures along the reasoning chain, substantiating the proposition that misalignment between sparsely distributed visual signals and dense encoding strategies structurally distorts multimodal reasoning dynamics.

Motivated by this diagnosis, recent efforts [18, 20] have sought to alleviate the bottleneck through fine-grained reward modeling. However, these approaches remain constrained by two fundamental limitations: **(i) Semantic Indirection**: VPPO [18] approximates visual dependency via indirect statistical surrogates (*e.g.*, divergence-based masking), without explicitly grounding reward signals in structured visual semantics. **(ii) Architectural Incompatibility**: AT-RL [20] requires explicit extraction of cross-modal attention weights, which conflicts with efficiency-oriented operators such as flashattention [11], thereby undermining scalable training.

Based on the above analysis, we propose the **K**ey-Region **A**ligned **W**eighted **H**armonic **I**ncentive (`KAWHI`), a plug-and-play, visually-perceived reward reweighting mechanism integrable into uniform reward policy optimization methods [13, 42, 66, 72]. `KAWHI` adaptively localizes semantically salient regions through hierarchical geometric aggregation and local structural modeling, thereby establishing spatial priors oriented toward visual representation to effectively mitigate information sparsity and foreground-background distributional imbalance. Furthermore, based on MME [12] evaluation, we identify vision-correlated attention heads as anchors. The mechanism then leverages image-critical states as semantic identifiers and response query states as information probes to construct passage-level weighting metrics, enabling fine-grained credit assignment across reasoning steps. In summary, our contributions are threefold:

1. **Structural Bottleneck Identification:** We systematically identify critical bottlenecks in applying RLVR to LVLMs and, for the first time, explicitly incorporate visual-geometric information into reward modeling, substantially enhancing multimodal grounding capabilities.
2. **Vision-Aware Reward Reweighting:** We propose `KAWHI`, a plug-and-play reward reweighting mechanism compatible with uniform reward policy optimization methods, enabling fine-grained, vision-aware credit assignment without architectural modifications.
3. **Consistent Performance Gains:** Through extensive evaluations on mathematical and chart-based benchmarks, we demonstrate significant improvements in reasoning accuracy while mitigating hallucination, validating the effectiveness and generalizability of our approach.

## 2 Related Work

**Visual Token Selection.** Visual token selection research offers valuable insights into identifying critical visual regions. ToMe [6] merges similar tokens to accelerate ViT inference. FastV [9] provides a training-free framework for visual token pruning. Through one-to-many token aggregation, FiCoCo [15] develops

distinct compression methods for visual and linguistic representations. [54] critically examine token pruning approaches in multimodal large language models. DivPrune [3] introduces diversity-aware pruning for vision-language models. SparseVLM [70] focuses on visual token sparsification for efficient inference. VFlowOpt [61] guides token pruning via visual information flow. These studies collectively indicate that not all visual tokens contribute equally, and identifying high-information-density regions is key to efficiency. However, current strategies are primarily adapted for the texture-dense characteristics of natural images and are mainly employed for **inference acceleration** [14, 16, 30, 62] in LVLMs, without accounting for the pixel-level sparsity and geometric structural priors inherent in structured visual scenarios (e.g., charts and geometric diagrams).

**Visual-Text Alignment.** Visual-text alignment research reveals intrinsic mechanisms within multimodal reasoning processes. RefCOCO [65] establishes benchmarks for modeling context in referring expressions. Attention Rollout [1] quantifies attention flow in transformers. GLIP [22] unifies grounded language-image pre-training. GroundingDINO [27] marries DINO with grounded pre-training for open-set detection. SparseMM [48] analyzes sparse attention heads in multimodal language models. Structured attention for document understanding [26] highlights pattern regularity in document-level tasks. VISTA [23] enhances alignment via mutual information maximization. Attention Re-Alignment [10] addresses suppressed visual grounding through intermediate layer guidance. Spatial-MLLM [56] designs frameworks specifically for spatial reasoning. Spatial Preference Rewarding [39] constructs reward functions for spatial understanding tasks. These findings highlight that disconnection between text and image is a core bottleneck constraining multimodal reasoning performance. However, existing alignment insights are mostly utilized for post-hoc interpretation or static supervision in SFT, lacking research on **integrating visual alignment quality as a dynamic signal** into the optimization process.

**RL Optimization Methods.** RL optimization methods have increasingly focused on improving credit assignment efficiency in complex reasoning tasks. PPO [41] establishes a stable foundation for proximal policy optimization through clipped surrogate objectives. DPO [40] simplifies the alignment pipeline by directly optimizing preferences without explicit reward modeling. GRPO [42] removes the critic network and performs group-relative comparisons to enhance training efficiency. VinePPO [21] introduces fine-grained credit assignment with value networks to better capture token-level contributions. DAPO [66] improves stability via decoupled clipping and dynamic sampling, while GSPO [72] further stabilizes updates through group-level clipping mechanisms. SAPO [13] proposes a soft adaptive policy optimization strategy tailored for language models. Beyond the 80/20 Rule [51] emphasizes the importance of high-entropy minority tokens in driving effective learning. VPPO [18] reweights advantages through visually grounded signals, and Step-wise GRPO [67] refines advantage estimation at the step level. AT-RL [20] leverages cross-modality connectivity to enable more precise reinforcement learning. Overall, these studies indicate that, compared with uniform signal allocation strategies, fine-grained signal modulation

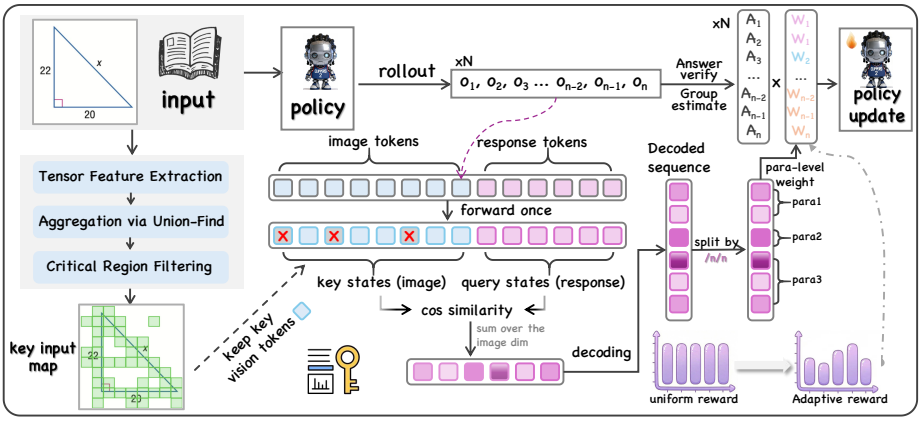


**Fig. 2:** Overview of the KAWHI mechanism. Critical regions are selected using the SGUF algorithm (see section 3.2 for details). The decoded sequence denotes the textual output generated from response tokens, segmented by the paragraph delimiter \n\n.

can more effectively enhance the quality and stability of policy updates. However, most existing approaches do not explicitly model or fully exploit visual modality information, thereby limiting their potential in multimodal reasoning scenarios. Moreover, they generally lack an **explicit spatial credit assignment mechanism** that incorporates geometric spatial priors, making it difficult to capture higher-level structured dependencies.

## 3 Method

In this paper, we propose KAWHI (Fig. 2). This section details its core components. We first review GRPO [42] and the uniform reward mechanism (Sec. 3.1), then introduce the visual token selection algorithm (Sec. 3.2) and the evaluation metrics for critical visual tokens (Sec. 3.3). Finally, we describe how these metrics are integrated into the reasoning process (Sec. 3.4).

### 3.1 Preliminary: Group-Relative Policy Optimization (GRPO)

Given a multimodal prompt $x = (I, q)$, GRPO performs critic-free policy optimization by constructing a baseline from a within-prompt rollout pool. Let $\pi_{\theta_{\text{old}}}$ sample $G$ completions $\{y_g\}_{g=1}^{G}$, where $y_g = (y_{g,1}, \ldots, y_{g,T_g})$ and $T_g$ is its length. A verifier assigns a sequence-level reward $R_g \in \mathbb{R}$ (typically $R_g \in \{0, 1\}$).

Rewards are standardized within the group to obtain a relative advantage:

$$\mu = \frac{1}{G}\sum_{j=1}^{G} R_j, \quad \sigma = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \mu)^2}, \quad A_g = \frac{R_g - \mu}{\sigma + \epsilon}, \tag{1}$$

where $\epsilon > 0$ ensures numerical stability. Since rewards are sequence-level, the advantage is broadcast across timesteps: $A_{g,t} = A_g$.

The token-wise importance ratio is

$$r_{g,t}(\theta) = \frac{\pi_\theta(y_{g,t} \mid x, y_{g,<t})}{\pi_{\theta_{\text{old}}}(y_{g,t} \mid x, y_{g,<t})}. \tag{2}$$

GRPO maximizes a PPO-style clipped surrogate:

$$\ell_{g,t}(\theta) = \min\Big(r_{g,t}(\theta)A_{g,t},\ \text{clip}(r_{g,t}(\theta), 1-\varepsilon, 1+\varepsilon)\, A_{g,t}\Big), \tag{3}$$

and optimizes

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_g\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{T_g}\sum_{t}\ell_{g,t}(\theta)\right], \tag{4}$$

where $\varepsilon$ is the clipping parameter.

### 3.2 Structure-Guided Union-Find (SGUF)

*What distinguishes structured visual inputs from natural imagery?* Unlike natural images characterized by rich textures and high semantic density, many vision-language reasoning tasks involve **structured visual content** such as mathematical expressions, geometric diagrams, and analytical charts. These inputs exhibit pronounced **spatial sparsity** and **foreground–background imbalance**: informative signals are concentrated in thin strokes, symbols, axes, or curve regions, while large portions of the image remain semantically redundant. Such structural properties introduce substantial redundancy into patch token sequences produced by vision encoders. As a result, computational resources are allocated to visually inactive regions, diluting optimization signals and hindering accurate estimation of policy gradients or advantages.

To address this, we propose **SGUF**, a geometry-aware token filtering mechanism that adaptively identifies informative visual regions. By modeling local structural patterns, SGUF provides **spatial priors** for downstream RL optimization, enabling criticality metrics to concentrate on semantically meaningful regions rather than redundant background areas.

**Structure Tensor Feature Extraction.** Given an input image $\mathcal{I}$, we first convert it to the CIELab [33] perceptually uniform color space and extract the luminance channel $L \in [0, 100]$. After applying Gaussian smoothing to $L$, we compute the gradient field $\nabla L = (\partial_x L, \partial_y L)^\top$ using Sobel operators. For each image patch $\mathcal{P}_{i,j}$, we characterize its local geometry through the second-order structure tensor, defined as the accumulation of gradient outer products over the patch domain:

$$\mathbf{S}_{i,j} = \sum_{(x,y)\in\mathcal{P}_{i,j}} (\nabla L)(\nabla L)^\top = \begin{pmatrix} \sum g_x^2 & \sum g_x g_y \\ \sum g_x g_y & \sum g_y^2 \end{pmatrix}. \tag{5}$$

In practice, we normalize the tensor entries by the patch area $|\mathcal{P}_{i,j}|$ to reduce sensitivity to patch size:

$$\mathbf{S}_{i,j} \leftarrow \frac{1}{|\mathcal{P}_{i,j}|}\mathbf{S}_{i,j}. \tag{6}$$

Performing eigendecomposition $\mathbf{S}_{i,j} = \mathbf{R}^\top \boldsymbol{\Lambda}\mathbf{R}$ yields eigenvalues $\lambda^{i,j}_{\max} \geq \lambda^{i,j}_{\min} \geq 0$, which provide rotation-invariant descriptors of local structure. Specifically, the condition $\lambda_{\max} \gg \lambda_{\min}$ indicates anisotropic regions (e.g., stroke edges), whereas $\lambda_{\max} \approx \lambda_{\min} \approx 0$ corresponds to flat background regions. Crucially, these eigenvalues remain invariant under rotation and illumination variations, ensuring robustness against document skew and shadow artifacts.

**Hierarchical Aggregation via Union-Find.** We cast region aggregation as graph partitioning over the patch grid. Each $\mathcal{P}_{i,j}$ is a node in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ under 4-connectivity. We define a heterogeneous similarity metric combining structural and luminance coherence:

$$d(\mathcal{P}_i, \mathcal{P}_j) = \underbrace{\frac{\|\boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_j\|_F}{\max(\lambda^i_{\max}, \lambda^j_{\max}, \epsilon)}}_{\text{structure}} + \alpha \cdot \underbrace{\frac{|L_i - L_j|}{100}}_{\text{luminance}}, \tag{7}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{\max}, \lambda_{\min})$, $\|\cdot\|_F$ is the Frobenius norm, and $\alpha$ controls the trade-off. While $d(\cdot,\cdot)$ provides a unified dissimilarity score, our implementation uses a *dual-threshold* merge rule to ensure conservative aggregation: two nodes are merged into component $\mathcal{C}_k$ via Union-Find if

$$\frac{\|\boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_j\|_F}{\max(\lambda^i_{\max}, \lambda^j_{\max}, \lambda^i_{\min}, \lambda^j_{\min}, \epsilon)} < \delta_s \quad \text{and} \quad |L_i - L_j| < \delta_l, \tag{8}$$

where $\delta_s$ and $\delta_l$ are the structure and luminance thresholds, respectively. This enforces transitive aggregation of structurally coherent stroke regions while avoiding merges between chromatically similar but geometrically distinct areas (e.g., white backgrounds vs. white formula interiors).

**Key Region Identification and Adaptive Sampling.** Given connected components $\{\mathcal{C}_k\}$, we measure structural saliency by the average trace of their structure tensors:

$$E(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|}\sum_{\mathcal{P}\in\mathcal{C}_k} \mathrm{tr}(\mathbf{S}) = \frac{1}{|\mathcal{C}_k|}\sum_{\mathcal{P}\in\mathcal{C}_k} (\lambda_{\max} + \lambda_{\min}), \tag{9}$$

where $E(\mathcal{C}_k)$ reflects the aggregated gradient magnitude within each region. We set the energy threshold $\tau$ using a median-based statistic over $\{E(\mathcal{C}_k)\}$:

$$\tau = \beta \cdot \mathrm{median}\big(\{E(\mathcal{C}_k)\}\big), \tag{10}$$

where $\beta$ is a scaling factor. This partitions regions into key regions ($E(\mathcal{C}_k) > \tau$) and background regions ($E(\mathcal{C}_k) \leq \tau$).

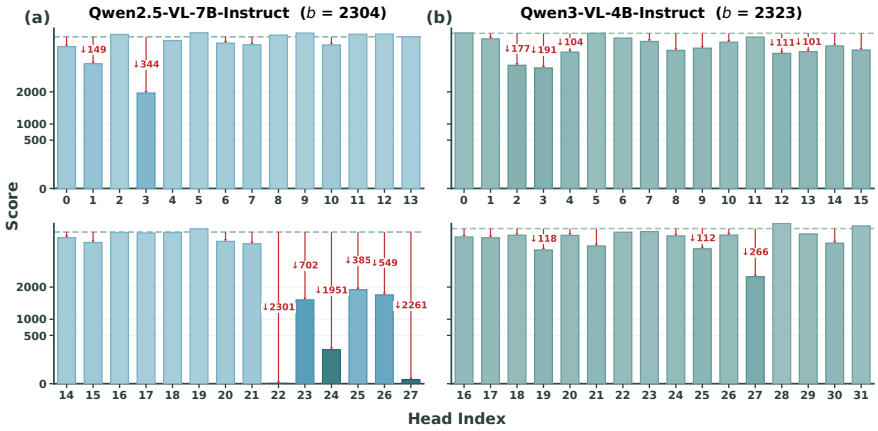


**Fig. 3:** Vision-Critical head identification via global ablation on the MME benchmark. Here, b denotes the performance score of the baseline models (Qwen2.5-VL-7B-Instruct [5] and Qwen3-VL-4B-Instruct [4]).

The selected token set $\mathcal{S}$ is constructed via hybrid sampling: retaining all tokens from key regions and sparsely sampling background regions at rate $r_{\text{skip}}$:

$$\mathcal{S} = \bigcup_{E(\mathcal{C}_k)>\tau} \mathcal{T}_k \cup \bigcup_{E(\mathcal{C}_k)\leq\tau} \text{Sample}(\mathcal{T}_k, 1 - r_{\text{skip}}), \tag{11}$$

where $\mathcal{T}_k$ denotes tokens corresponding to $\mathcal{C}_k$. This preserves geometrically informative regions while compressing low-information areas, focusing computation on reasoning-relevant content.

### 3.3 Metrics for Spatially-Aware Credit Assignment

To integrate the key regions identified by SGUF with the RL rollout process for spatially-aware credit assignment, we revisit the functional differentiation of Query, Key, and Value vectors in the self-attention mechanism [45]:

- **Query (Q):** Serves as an **information probe**, representing the information retrieval demands of the current generation token regarding visual content.
- **Key (K):** Acts as a **semantic identifier**, carrying the semantic identity of visual tokens to respond to and match external queries.
- **Value (V):** Functions as an **information payload**, containing the actual representations transmitted downstream after attention weighting.

Building upon this functional differentiation, we instantiate the generated response tokens from LVLMs as **Q** (information probes), and the visual tokens within SGUF-selected key regions as **K** (semantic identifiers). This establishes attentional coupling between generated content and critical visual regions via **Q-K** matching, thereby quantifying the spatial saliency of the reasoning process.

Nevertheless, the fidelity of this coupling depends on the functional specialization of individual attention heads, which exhibit substantial heterogeneity in their visual contributions [48]. To identify the most critical computational units, we perform structured attribution analysis to localize vision-specific heads.

We perform global head ablation on MME [12], which contains 14 subtasks spanning recognition and reasoning (Fig. 3). Heads with identical indices are masked across all decoder layers to measure their marginal impact on visual performance. Heads whose removal causes an aggregated MME drop exceeding **100** are defined as *vision-critical*. Using this criterion, we identify attention head subsets in Qwen2.5-VL-7B-Instruct [5] and Qwen3-VL-4B-Instruct [4] that are critical for visual perception.

Leveraging these vision-critical heads, we establish the spatial alignment mechanism between generated responses and SGUF-selected regions. Let $\mathcal{R}$ denote the response token indices and $\mathcal{S} \subseteq \mathcal{I}$select denote the SGUF-selected key region indices. As query states are not retained in the decoding cache, we conduct an additional forward pass based on the hidden states of the final decoder layer to recover the corresponding query and key states. For Grouped Query Attention [2] with $H_q = g \cdot H_k$ heads, we align the key states with query states by repeating each key head $g$ times. The spatial alignment between response token $t \in \mathcal{R}$ and critical visual token $v \in \mathcal{S}$ is measured by the cosine similarity between $\ell_2$-normalized query and key vectors restricted to $\mathcal{H}$critical:

$$s_{t,v}^{(h)} = \frac{\mathbf{q}_t^{(h)} \cdot \mathbf{k}_v^{(h)}}{\|\mathbf{q}_t^{(h)}\| \|\mathbf{k}_v^{(h)}\|}, \quad h \in \mathcal{H}_{\text{critical}}, \tag{12}$$

where $\mathbf{q}_t^{(h)} \in \mathbb{R}^d$ is the query vector and $\mathbf{k}_v^{(h)}$ is the corresponding expanded key vector. The aggregated spatial attention score for token $t$ is then computed as:

$$\alpha_t = \frac{1}{|\mathcal{S}| \cdot |\mathcal{H}_{\text{critical}}|} \sum_{h \in \mathcal{H}_{\text{critical}}} \sum_{v \in \mathcal{S}} s_{t,v}^{(h)}. \tag{13}$$

This score $\alpha_t \in [-1, 1]$ quantifies the degree to which the generation process focuses on the SGUF-identified critical visual regions through the lens of vision-critical heads, serving as a spatial prior for downstream credit assignment.

### 3.4 Semantic-Preserving Spatial Weighting

Having established spatially-aware credit assignment metrics, the remaining challenge is to couple these signals effectively with reward allocation over generated responses. The segmentation granularity critically determines the stability and validity of credit attribution. Excessively fine-grained segmentation (e.g., token- or sentence-level) fragments coherent semantic units and disrupts contextual dependencies within the reasoning chain, thereby amplifying noise and destabilizing credit signals. In such cases, rewards become misaligned with semantically coherent reasoning steps and are diffusely assigned to isolated lexical fragments, undermining long-range dependencies and logical consistency. As a result, the

spatial saliency metric $\alpha_t$ cannot form a stable correspondence with overall reasoning quality.

To address this issue, we adopt a **paragraph-level coarse-grained segmentation** strategy. Using paragraph delimiters (`\n\n`) as boundaries, each segment preserves a self-contained reasoning step with intact semantic structure. This design maintains the natural organization of cot reasoning while enabling stable and semantically consistent credit allocation.

Formally, let the response be partitioned into $M$ paragraphs $\{P_j\}_{j=1}^{M}$, where each paragraph $P_j$ comprises a set of token indices $\mathcal{S}_j$. We first aggregate the spatial saliency scores within each paragraph via mean pooling:

$$\bar{\alpha}_j = \frac{1}{|\mathcal{S}_j|} \sum_{t \in \mathcal{S}_j} \alpha_t. \tag{14}$$

Subsequently, these paragraph-level scores are normalized into weights $w_j \in [w_{\min}, w_{\max}]$ through temperature-scaled softmax followed by uniform mixing:

$$\tilde{w}_j = (1-\lambda)\frac{\exp(\bar{\alpha}_j/\tau)}{\sum_{k=1}^{M}\exp(\bar{\alpha}_k/\tau)} + \frac{\lambda}{M}, \quad w_j = w_{\min} + (w_{\max} - w_{\min}) \cdot \tilde{w}_j, \tag{15}$$

where $\tau$ denotes the temperature parameter controlling the sharpness of the weight distribution, and $\lambda$ is the smoothing coefficient ensuring minimum credit allocation to all paragraphs.

To integrate these spatial priors into the policy optimization objective, we modulate the group-wise baseline advantage by the paragraph-level weights. For a response group $g$ with total reward $R_g$, the baseline advantage is first computed via group-wise standardization:

$$A_g = \frac{R_g - \mu_g}{\sigma_g + \epsilon}, \tag{16}$$

where $\mu_g$ and $\sigma_g$ denote the mean and standard deviation of rewards within group $g$, and $\epsilon$ is a numerical stability constant. This group-level scalar is then broadcast to all response tokens $t$, yielding the token-level advantage $A_{g,t} = A_g \cdot \mathbb{I}(t \in \mathcal{R}_g)$, where $\mathbb{I}(\cdot)$ is the indicator function for response positions.

Finally, we apply the paragraph-specific weights to modulate the token-level advantages. For token $t$ belonging to paragraph $P_j$ (i.e., $t \in \mathcal{S}_j$), the weighted advantage is computed as:

$$\hat{A}_{g,t} = A_{g,t} \cdot w_j. \tag{17}$$

This formulation ensures that tokens within spatially salient paragraphs (high $w_j$) receive amplified gradient signals, while maintaining stable credit allocation across semantically coherent reasoning blocks. The resulting weighted advantages $\hat{A}_{g,t}$ serve as the final surrogate objectives for policy gradient updates, effectively coupling spatial visual attention with linguistic reasoning quality.

# 4 Experiments

## 4.1 Experimental Setup

**Model, training data and evluation** We conduct experiments on Qwen2.5-VL-7B-Instruct [5] and Qwen3-VL-4B-Instruct [4]. For training, Geo3K [32] is used for mathematical reasoning, while a 20K subset of ChartQA [35] is constructed for chart understanding (details in Appendix). Evaluation is performed on MathVista [31], MathVerse [69], MathVision [49], and WeMath [38] for mathematical reasoning, and on ChartXivDesc [52], ChartXivRea [52], ChartQA [35], ChartQA-Pro [34], and ChartMimic [59] for chart comprehension. Following prior work, we adopt an LLM-as-a-judge protocol with Qwen3-VL-30B-A3B [4] for MathVista, MathVerse, and MathVision, and Exact Match for WeMath. All experiments are conducted under a unified setup using the LMMs-Eval [68] framework, and reinforcement learning is performed without any intermediate supervised fine-tuning (SFT).

**Table 1:** Main results on **Qwen2.5-VL-7B-Instruct** and **Qwen3-VL-4B-Instruct**. $+\Delta/-\Delta$ denote changes vs. baselines; **(+X.XX)** shows gains over base RL methods. Bold = best per model size; gray rows = ours.

| Method | MathVista | MathVerse | MathVision | WeMath | Avg |
|---|---|---|---|---|---|
| ***Qwen2.5-VL-7B-Instruct*** | | | | | |
| Base | 67.47 | 45.56 | 23.36 | 54.32 | 47.68 |
| *fine-grained reward* | | | | | |
| Step-GRPO | 66.87$_{-0.60}$ | 44.32$_{-1.24}$ | 24.76$_{+1.40}$ | 55.32$_{+1.00}$ | 47.82$_{+0.14}$ |
| FT-RL | 67.68$_{+0.21}$ | 46.88$_{+1.32}$ | 26.12$_{+2.76}$ | 56.78$_{+2.46}$ | 49.37$_{+1.69}$ |
| VPPO | 68.42$_{+0.95}$ | 47.21$_{+1.65}$ | 28.42$_{+5.06}$ | 57.12$_{+2.80}$ | 50.29$_{+2.61}$ |
| *uniform reward* | | | | | |
| GRPO | 68.37$_{+0.90}$ | 45.94$_{+0.38}$ | 28.29$_{+4.93}$ | 58.10$_{+3.78}$ | 50.18$_{+2.50}$ |
| +KAWHI (Ours) | 69.30$_{+1.83}$ | 47.56$_{+2.00}$ | 29.28$_{+5.92}$ | 57.84$_{+3.52}$ | 51.00$_{+3.32}^{(+0.82)}$ |
| DAPO | 69.20$_{+1.73}$ | 47.75$_{+2.19}$ | 28.21$_{+4.85}$ | 57.98$_{+3.66}$ | 50.79$_{+3.11}$ |
| +KAWHI (Ours) | 69.08$_{+1.61}$ | 48.20$_{+2.64}$ | 31.91$_{+8.55}$ | 58.21$_{+3.89}$ | 51.85$_{+4.17}^{(+1.06)}$ |
| GSPO | **69.42**$_{+1.95}$ | 49.87$_{+4.31}$ | 27.30$_{+3.94}$ | 59.02$_{+4.70}$ | 51.40$_{+3.72}$ |
| +KAWHI (Ours) | 69.37$_{+1.90}$ | **50.10**$_{+4.54}$ | **32.57**$_{+9.21}$ | **59.71**$_{+5.39}$ | **52.94**$_{+5.26}^{(+1.54)}$ |
| ***Qwen3-VL-4B-Instruct*** | | | | | |
| Base | 70.43 | 40.86 | 21.38 | 69.43 | 50.53 |
| *uniform reward* | | | | | |
| GRPO | 71.07$_{+0.64}$ | 41.12$_{+0.26}$ | 22.16$_{+0.78}$ | 70.31$_{+0.88}$ | 51.17$_{+0.64}$ |
| +KAWHI (Ours) | 71.80$_{+1.37}$ | 42.10$_{+1.24}$ | 23.03$_{+1.65}$ | 70.98$_{+1.55}$ | 51.98$_{+1.45}^{(+0.81)}$ |
| DAPO | 69.40$_{-1.03}$ | 40.48$_{-0.38}$ | 27.63$_{+6.25}$ | 69.67$_{+0.24}$ | 51.80$_{+1.27}$ |
| +KAWHI (Ours) | 71.21$_{+0.78}$ | 41.36$_{+0.50}$ | 28.43$_{+7.05}$ | 71.23$_{+1.80}$ | 53.06$_{+2.53}^{(+1.26)}$ |
| GSPO | 70.50$_{+0.07}$ | 41.62$_{+0.76}$ | 32.89$_{+11.51}$ | 71.55$_{+2.12}$ | 54.14$_{+3.61}$ |
| +KAWHI (Ours) | **72.10**$_{+1.67}$ | **46.82**$_{+5.96}$ | 31.09$_{+9.71}$ | **72.15**$_{+2.72}$ | **55.54**$_{+5.01}^{(+1.40)}$ |

**Comparison Methods** We first report the experimental results of fine-grained reward allocation methods, including Step-GRPO [67], FT-RL [51], and VPPO [18]. For uniform reward allocation approaches, we present the baseline results of GRPO [42], DAPO [66], and GSPO [72], followed by the performance obtained after incorporating our proposed method. All experiments are conducted on 8 NVIDIA H200 GPUs, with details in the Appendix.

### 4.2 Main Results

Table 1 presents comprehensive experiments on Qwen2.5-VL-7B-Instruct [5] and Qwen3-VL-4B-Instruct [4], integrating `KAWHI` into the uniform reward-based GRPO [42], DAPO [66], and GSPO [72] frameworks. The results show consistent and substantial improvements in average performance across four datasets after incorporating `KAWHI`. From a methodological perspective, `KAWHI` explicitly captures critical visual information and adaptively assigns weights across different stages of the reasoning process, thereby enhancing the modeling of key signals in vision–language interactions. This structured attention adjustment reduces irrelevant interference while improving cross-modal alignment, enabling the model to focus more precisely on task-relevant regions during complex reasoning and ultimately enhancing overall perception and reasoning performance.

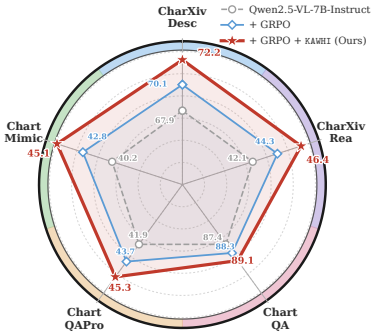


**Fig. 4:** Radar chart comparison of GRPO and GRPO+`KAWHI` on five chart benchmarks.

Furthermore, we conduct additional validation on chart understanding using Qwen2.5-VL-7B-Instruct with a 20K-sample subset from ChartQA under the GRPO method. As shown in Fig. 4, incorporating `KAWHI` leads to performance gains of **2.1%**, **2.1%**, **0.8%**, **1.6%**, and **2.3%** on ChartXivDesc [52], ChartXivRea [52], ChartQA [35], ChartQA-Pro [34], and ChartMimic [59], respectively. These results indicate that `KAWHI` generalizes well to structured visual tasks such as chart comprehension and consistently provides stable improvements.

### 4.3 Ablation Studies

To evaluate the effectiveness of `KAWHI`, we conduct five ablation studies (Table 2) on Qwen-2.5-VL-7B-Instruct [5], using GRPO as the baseline and reporting results on MathVerse and MathVision. The ablations cover the following components: ① *Region Selection*, ② *Reward Metric*, ③ *Response Granularity*, ④ *Critical Visual Head*, and ⑤ *Positional Encoding*.

① ***Region Selection.*** Under the *Region Selection* configuration, we evaluate whether SGUF effectively identifies critical visual tokens that contribute to performance gains. We examine three settings: **All vision tokens**, **Random**, and

**Table 2: Ablation on Qwen2.5-VL-7B-Instruct.** Results on MathVerse and MathVision. ↓X.XX indicates drop vs. GRPO+ KAWHI.

| Ablation | Experiment | MathVerse (%)↑ | MathVision (%)↑ | Average |
|---|---|---|---|---|
| – | GRPO | $45.94_{\downarrow 1.62}$ | $28.29_{\downarrow 0.99}$ | 37.11 |
| – | GRPO+KAWHI (Ours) | **47.56** | **29.28** | **38.42** |
| ① *Region Selection* | All vision tokens | $46.58_{\downarrow 0.98}$ | $28.45_{\downarrow 0.83}$ | 37.52 |
| | Random | $46.34_{\downarrow 1.22}$ | $28.98_{\downarrow 0.30}$ | 37.66 |
| | Inverse selection | $45.74_{\downarrow 1.82}$ | $27.82_{\downarrow 1.46}$ | 36.78 |
| ② *Reward Metric* | Key–Key | $47.08_{\downarrow 0.48}$ | $28.95_{\downarrow 0.33}$ | 38.02 |
| ③ *Response Granularity* | Sentence-level | $47.11_{\downarrow 0.45}$ | $26.64_{\downarrow 2.64}$ | 36.88 |
| | Token-level | $46.21_{\downarrow 1.35}$ | $26.64_{\downarrow 2.64}$ | 36.42 |
| ④ *Critical Visual Head* | w/o Critical Head | $46.57_{\downarrow 0.99}$ | $28.32_{\downarrow 0.96}$ | 37.45 |
| ⑤ *Positional Encoding* | before RoPE | $47.08_{\downarrow 0.48}$ | $28.62_{\downarrow 0.66}$ | 37.85 |

**Inverse selection**. **All vision tokens** improves performance over GRPO by 0.64% on MathVerse and 0.16% on MathVision, indicating that although retaining complete visual information avoids missing important regions, redundant tokens dilute informative signals. **Random** achieves gains of 0.4% and 0.7%, suggesting that random sampling may occasionally capture critical regions but lacks stability. In contrast, **Inverse selection** underperforms GRPO on both benchmarks, demonstrating that attending to visually irrelevant regions introduces noise and degrades performance.

② ***Reward Metric.*** We compare different similarity computation schemes for reward modeling. Computing similarity between *image (key states)* and *response (key states)* results in performance drops of 0.48% and 0.33% relative to KAWHI. Although the *Key–Key* formulation preserves basic evaluation ability, it is less discriminative than the *Key (image)–Query (response)* scheme, as it relies on symmetric static matching. In contrast, *Key–Query* supports asymmetric, task-conditioned alignment, enabling more precise cross-modal modeling.

③ ***Response Granularity.*** We further investigate finer-grained response partition strategies, including sentence-level and token-level segmentation. Compared to KAWHI, sentence-level segmentation yields performance drops of 0.45% and 2.64%, while token-level segmentation results in larger decreases of 1.35% and 2.64%. These results suggest that preserving semantic structural integrity is critical, as excessively fine-grained segmentation disrupts contextual coherence and weakens reasoning performance.

④ ***Critical Visual Head.*** When the metric is computed over all attention heads instead of restricting it to vision-related heads, performance drops by 0.99% and 0.96%. This suggests that concentrating on critical visual heads suppresses irrelevant signals and enhances discriminative stability.

⑤ ***Positional Encoding.*** Using Key and Query states prior to positional encoding leads to performance decreases of 0.48% and 0.66%. This indicates that positional encoding incorporates spatial information into the representations, enabling more accurate visual alignment and more reliable reward computation.

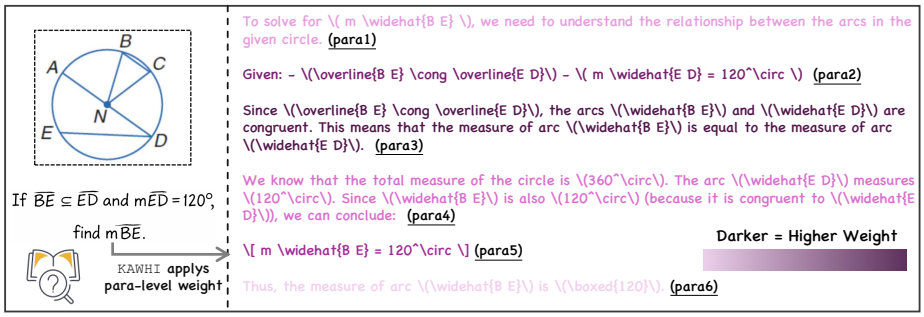


**Fig. 6:** Qualitative Analysis of Weight Allocation in KAWHI. Results are illustrated using cases from Geo3K [32]. Color intensity represents weight magnitude, where darker colors indicate larger weights, illustrating the allocation behavior of KAWHI.

### 4.4 A Viable Alternative to SGUF

To examine the transferability of our framework, we replace SGUF with VisionZip [60] (50% vision tokens) to obtain salient visual tokens. As shown in Fig. 5, the performance drops on MathVista [31], MathVerse [69], and MathVision [49] are only 0.13%, 0.22%, and 0.30%, respectively, indicating that attention-based token selection [6, 15, 60] remains compatible with KAWHI and confirming the transferability of KAWHI.

### 4.5 Qualitative Experimental Analysis

Fig. 6 visualizes the paragraph-level weight distribution induced by KAWHI under GRPO (Qwen2.5-VL-7B-Instruct). Splitting the solution into para1–para6 by \n\n reveals clear functional roles along the reasoning chain. **Para2** states the key constraints (congruent chords and the given arc measure), providing the factual basis for inference. **Para3** applies the decisive rule—congruent chords subtend congruent arcs—to equate the target arc with the known arc. Together, **para2**–**para3** constitute constraint injection and rule application. In contrast, **para1** is introductory and **para6** restates the answer, contributing minimally to the derivation. **Para4** mentions the circle's total measure (360°) but is unnecessary and adds limited information. Accordingly, KAWHI assigns substantially higher weights to **para2** and **para3** while down-weighting auxiliary or redundant segments. This distribution suggests a positive alignment between paragraph weight and structural importance, supporting the effectiveness of KAWHI.

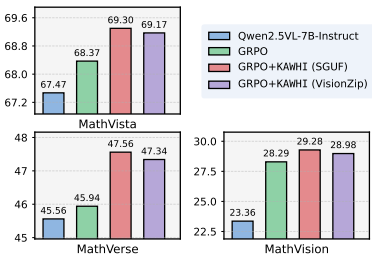


**Fig. 5:** Validation of visionzip as an effective alternative to SGUF under 50% visual tokens.

### 4.6 Efficiency Analysis

To assess the computational efficiency of the proposed method, we conduct rigorous ablation studies measuring per-step training latency under identical hardware configurations and training pipelines (As shown in Fig. 7). Specifically, within the Qwen2.5-VL-7B-Instruct [5] and VERL [44] framework, we compare executions with (R1) and without (R2) KAWHI, yielding per-step durations of 521.0s and 540.6s, respectively. The induced temporal overhead of 19.6s, which is attributable to the integrated weighting mechanism, constitutes merely 3.6% relative to the baseline, demonstrating that our approach achieves substantial performance gains while incurring only marginal computational costs.

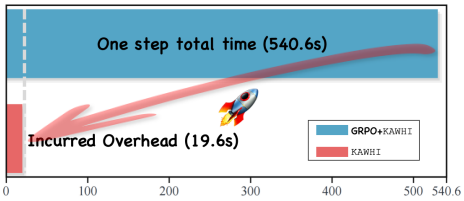


**Fig. 7:** Computational Efficiency Analysis of KAWHI

## 5 Conclusions

This work identifies a structural bottleneck in applying RLVR to LVLMs, arising from the mismatch between sparse, structured visual signals and dense visual token representations. Through quantitative and qualitative analyses, we show that visual perception errors constitute the primary failure mode in multimodal reasoning, indicating that existing reward mechanisms underutilize spatial visual information. To address this limitation, we introduce KAWHI, a plug-and-play reward reweighting module that integrates structured visual region alignment and paragraph-level credit assignment into uniform reward optimization. Experiments on mathematical and chart-based benchmarks demonstrate that KAWHI improves reasoning accuracy while reducing hallucination, highlighting the value of structurally grounded reward modeling in multimodal reinforcement learning.

## References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 4190–4197 (2020)
2. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints (2023), https://arxiv.org/abs/2305.13245
3. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: DivPrune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)

4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
6. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
7. Cai, X.Q., Wang, W., Liu, F., Liu, T., Niu, G., Sugiyama, M.: Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers (2025), https://arxiv.org/abs/2510.00915
8. Chen, L., Gao, H., Liu, T., Huang, Z., Sung, F., Zhou, X., Wu, Y., Chang, B.: G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning (2025), https://arxiv.org/abs/2505.13426
9. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large Vision-Language Models. In: European Conference on Computer Vision. Springer (2024)
10. Chen, Y., Wang, P., Qin, G., Wu, W., Chen, M., Hao, Y.: Attention re-alignment in multimodal large language models via intermediate-layer guidance (2025)
11. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
12. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394
13. Gao, C., Zheng, C., Chen, X.H., Dang, K., Liu, S., Yu, B., Yang, A., Bai, S., Zhou, J., Lin, J.: Soft adaptive policy optimization (2025), https://arxiv.org/abs/2511.20347
14. Han, Y., Liu, X., Zhang, Z., Ding, P., Chen, J., Chen, H., Wang, D., Yan, Q., Huang, S.: Filter, correlate, compress: Training-free token reduction for mllm acceleration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 4601–4609 (2026)
15. Han, Y., Liu, X., Zhang, Z., Ding, P., Chen, J., Wang, D., Chen, H., Yan, Q., Huang, S.: Filter, correlate, compress: Training-free token reduction for mllm acceleration (2025), https://arxiv.org/abs/2411.17686
16. Han, Y., Yang, W., Chen, Y., Jin, X., Zhang, Y., Huang, S., Zhang, L.: Star-kv: Spatio-temporal adaptive re-weighting for kv cache compression in gui vision-language models. arXiv preprint arXiv:2606.01790 (2026)
17. He, Z., Qu, X., Li, Y., Huang, S., Liu, D., Cheng, Y.: Framethinker: Learning to think with long videos via multi-turn frame spotlighting (2025), https://arxiv.org/abs/2509.24304
18. Huang, S., Qu, X., Li, Y., Luo, Y., He, Z., Liu, D., Cheng, Y.: Spotlight on token perception for multimodal reinforcement learning (2025), https://arxiv.org/abs/2510.09285

19. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models (2026), https://arxiv.org/abs/2503.06749
20. Jiao, Z., Wang, S., Zhang, Z., Wang, W., Zhao, B., Wei, H., Zhang, L.: Credit where it is due: Cross-modality connectivity drives precise reinforcement learning for mllm reasoning (2026), https://arxiv.org/abs/2602.11455
21. Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A., Roux, N.L.: Vineppo: Refining credit assignment in rl training of llms (2025), https://arxiv.org/abs/2410.01679
22. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Gao, J.: GLIP: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
23. Li, M., Su, N., Qu, F., Zhong, Z., Chen, Z., Li, Y., Tu, Z., Li, X.: VISTA: Enhancing vision-text alignment in MLLMs via cross-modal mutual information maximization. arXiv preprint arXiv:2505.10917 (2025)
24. Li, S., Xu, X., Deng, K., Wang, L., Shen, H.T., Shen, F.: Truth in the few: High-value data selection for efficient multi-modal reasoning (2026), https://arxiv.org/abs/2506.04755
25. Liang, Y., Qiu, J., Ding, W., Liu, Z., Tompkin, J., Xu, M., Xia, M., Tu, Z., Shi, L., Zhu, J.: Modomodo: Multi-domain data mixtures for multimodal llm reinforcement learning (2025), https://arxiv.org/abs/2505.24871
26. Liu, C., Chen, H., Cai, Y., Wu, H., Ye, Q., Yang, M.H., Wang, Y.: Structured attention matters to multimodal LLMs in document understanding. arXiv preprint arXiv:2506.21600 (2025)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
28. Liu, X., Ni, J., Wu, Z., Du, C., Dou, L., Wang, H., Pang, T., Shieh, M.Q.: Noisyrollout: Reinforcing visual reasoning with data augmentation (2025), https://arxiv.org/abs/2504.13055
29. Liu, X., Liang, T., He, Z., Xu, J., Wang, W., He, P., Tu, Z., Mi, H., Yu, D.: Trust, but verify: A self-verification approach to reinforcement learning with verifiable rewards (2025), https://arxiv.org/abs/2505.13445
30. Liu, X., Wang, Z., Chen, J., Han, Y., Wang, Y., Yuan, J., Song, J., Huang, S., Chen, H.: Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7350–7358 (2026)
31. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts (2024), https://arxiv.org/abs/2310.02255
32. Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning (2021), https://arxiv.org/abs/2105.04165
33. Luo, M.R.: Cielab. In: Encyclopedia of color science and technology, pp. 251–257. Springer (2023)
34. Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., Thakkar, M., Parvez, M.R., Hoque, E., Joty, S.: Chartqapro: A more diverse and challenging benchmark for chart question answering (2025), https://arxiv.org/abs/2504.05506

35. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022), https://arxiv.org/abs/2203.10244
36. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., Zhang, K., Luo, P., Qiao, Y., Zhang, Q., Shao, W.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning (2025), https://arxiv.org/abs/2503.07365
37. Mroueh, Y.: Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification (2025), https://arxiv.org/abs/2503.06639
38. Qiao, R., Tan, Q., Dong, G., Wu, M., Sun, C., Song, X., GongQue, Z., Lei, S., Wei, Z., Zhang, M., Qiao, R., Zhang, Y., Zong, X., Xu, Y., Diao, M., Bao, Z., Li, C., Zhang, H.: We-math: Does your large multimodal model achieve human-like mathematical reasoning? (2024), https://arxiv.org/abs/2407.01284
39. Qiu, H., Gao, P., Lu, L., Zhang, X., Shao, L., Lu, S.: Spatial Preference Rewarding for MLLMs spatial understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 720–730 (2025)
40. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (2024), https://arxiv.org/abs/2305.18290
41. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), https://arxiv.org/abs/1707.06347
42. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
43. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., Xu, R., Zhao, T.: Vlm-r1: A stable and generalizable r1-style large vision-language model (2025), https://arxiv.org/abs/2504.07615
44. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. In: Proceedings of the Twentieth European Conference on Computer Systems. p. 1279–1297. EuroSys '25, ACM (Mar 2025). https://doi.org/10.1145/3689031.3696075, http://dx.doi.org/10.1145/3689031.3696075
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
46. Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., Shen, H., Xiong, J., Xin, Y., Jiang, Y., Tao, C., He, Y., Zhang, M., Yan, S.: Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning (2025), https://arxiv.org/abs/2506.01713
47. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning (2025), https://arxiv.org/abs/2504.08837
48. Wang, J., Liu, Z., Rao, Y., Lu, J.: Sparsemm: Head sparsity emerges from visual concept responses in mllms (2025), https://arxiv.org/abs/2506.05344
49. Wang, K., Pan, J., Shi, W., Lu, Z., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset (2024), https://arxiv.org/abs/2402.14804
50. Wang, P., Wei, Y., Peng, Y., Wang, X., Qiu, W., Shen, W., Xie, T., Pei, J., Zhang, J., Hao, Y., Song, X., Liu, Y., Zhou, Y.: Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning (2025), https://arxiv.org/abs/2504.16656

51. Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., Dang, K., Chen, X., Yang, J., Zhang, Z., Liu, Y., Yang, A., Zhao, A., Yue, Y., Song, S., Yu, B., Huang, G., Lin, J.: Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning (2025), https://arxiv.org/abs/2506.01939
52. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., Chevalier, A., Arora, S., Chen, D.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms (2024), https://arxiv.org/abs/2406.18521
53. Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., Bian, J., Yang, M.: Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms (2025), https://arxiv.org/abs/2506.14245
54. Wen, Z., Gao, Y., Li, W., He, C., Zhang, L.: Token pruning in Multimodal Large Language Models: Are we solving the right problem? In: Findings of the Association for Computational Linguistics: ACL 2025 (2025)
55. Wen, Z., Yang, B., Chen, S., Zhang, Y., Han, Y., Ke, J., Wang, C., Fu, Y., Zhao, J., Yao, J., et al.: Innovator-vl: A multimodal large language model for scientific discovery. arXiv preprint arXiv:2601.19325 (2026)
56. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence (2025), https://arxiv.org/abs/2505.23747
57. Xia, J., Zang, Y., Gao, P., Li, S., Zhou, K.: Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning (2025), https://arxiv.org/abs/2505.14677
58. Xiao, T., Xu, X., Huang, Z., Gao, H., Liu, Q., Liu, Q., Chen, E.: Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward (2025), https://arxiv.org/abs/2506.07218
59. Yang, C., Shi, C., Liu, Y., Shui, B., Wang, J., Jing, M., Xu, L., Zhu, X., Li, S., Zhang, Y., Liu, G., Nie, X., Cai, D., Yang, Y.: Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation (2025), https://arxiv.org/abs/2406.09961
60. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models (2024), https://arxiv.org/abs/2412.04467
61. Yang, S., Xu, R., Cui, C., Wang, T., Lin, D., Pang, J.: VFlowOpt: A token pruning framework for LMMs with visual information flow-guided optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
62. Yang, Z., Xu, D., Pang, W., Yuan, Y.: Script: Graph-structured and query-conditioned semantic token pruning for multimodal large language models (2025), https://arxiv.org/abs/2512.01949
63. Yao, H., Yin, Q., Zhang, J., Yang, M., Wang, Y., Wu, W., Su, F., Shen, L., Qiu, M., Tao, D., Huang, J.: R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo (2025), https://arxiv.org/abs/2505.16673
64. Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., Wei, H., Sun, J., Han, C., Ge, Z., Zhang, X., Jiang, D., Wang, J., Tao, W.: Perception-r1: Pioneering perception policy with reinforcement learning (2025), https://arxiv.org/abs/2504.07954
65. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85. Springer (2016)
66. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang,

M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Qiao, M., Wu, Y., Wang, M.: Dapo: An open-source llm reinforcement learning system at scale (2025), https://arxiv.org/abs/2503.14476

67. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization (2025), https://arxiv.org/abs/2503.12937
68. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. p. 881–916. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.findings-naacl.51, http://dx.doi.org/10.18653/v1/2025.findings-naacl.51
69. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., Li, H.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? (2024), https://arxiv.org/abs/2403.14624
70. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., Zhang, S.: SparseVLM: Visual token sparsification for efficient vision-language model inference. In: Proceedings of the International Conference on Machine Learning (2025)
71. Zhao, A., Wu, Y., Yue, Y., Wu, T., Xu, Q., Yue, Y., Lin, M., Wang, S., Wu, Q., Zheng, Z., Huang, G.: Absolute zero: Reinforced self-play reasoning with zero data (2025), https://arxiv.org/abs/2505.03335
72. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., Lin, J.: Group sequence policy optimization (2025), https://arxiv.org/abs/2507.18071