Do MLLMs Really See It: Reinforcing Visual Attention in Multimodal LLMs

Anonymous ACL submission

Abstract

While chain-of-thought (CoT) reasoning has substantially improved multimodal large language models (MLLMs) on complex reasoning tasks, existing approaches largely rely on long textual reasoning trajectories and provide limited mechanisms for learning stable visual attention policies. Our analysis shows that current MLLMs exhibit weak visual focus: early-stage visual misalignment is rarely corrected during subsequent reasoning, leading to error propagation and failed inferences. We argue that this limitation stems from inadequate credit assignment for visual attention during training. To address this issue, we propose SAYO, a visual reasoning model trained with a reinforcement learning (RL) framework that introduces a region-level visual attention–based reward. This reward explicitly aligns optimization signals with visually grounded reasoning steps, enabling the model to learn more reliable attention behaviors. Extensive experiments across multiple multimodal benchmarks demonstrate that SAYO consistently improves performance on diverse reasoning and perception tasks.

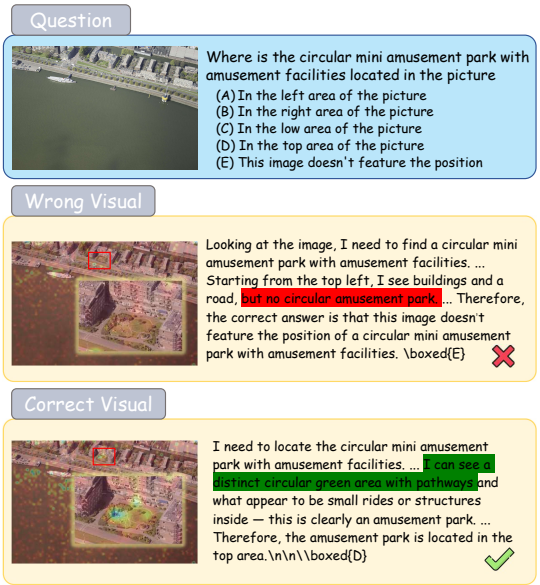


Figure 1: In the reasoning process of CoT, initial visual focusing errors can mislead the inference processing. We emphasize enhancing the model's visual capabilities to ensure MLLMs can proactively focus on text-relevant visual regions. The resolution of image is 2752x1824. The attention map displays the average attention of all generated tokens.

1 Introduction

Recent advances in multimodal large language models (MLLMs) (Bai et al., 2025; Wang et al., 2025c; Huang et al., 2026) have significantly improved performance on complex reasoning tasks that require integrating visual and linguistic information. In particular, chain-of-thought (CoT) reasoning has emerged as an effective mechanism for decomposing complex problems into sequential inference steps (Yang et al., 2023; Xu et al., 2025). Despite these successes, existing MLLMs continue to exhibit fundamental limitations when reasoning over visually complex inputs, especially in scenarios requiring precise and sustained visual grounding across long reasoning trajectories. Most prior approaches emphasize textual reasoning processes or rely on heuristic visual prompt engineering, such as programmatic region highlighting or prompt reflection mechanisms (Yang et al., 2025a). While these methods can indirectly influence visual perception, they do not explicitly address how visual attention behaviors are learned during training. Prior works have begun to probe this problem that current MLLMs often develop unstable visual attention policies, particularly when confronted with complex multi-object scenes or information-dense documents (Liu et al., 2025b; Yang et al., 2025b). These findings suggest that MLLMs may struggle to focus their visual attention and become distracted by an excess of other visual signals. Although techniques such as ViP (Cai et al., 2024) introduce visual cues to mark regions of interest, their effectiveness ultimately depends on the model's

pre-existing attention behavior, which remains insufficiently optimized.

Previous studies (Tong et al., 2024; Verma et al., 2024) have revealed a consistent pattern: when an MLLM attends to incorrect visual regions at early inference stages, this misalignment is rarely corrected during subsequent reasoning. Instead, erroneous visual assumptions propagate through the chain of thought, leading to systematic inference failures. This phenomenon is especially pronounced in long reasoning sequences, where early errors exert a disproportionate influence on final predictions. We argue that this limitation reflects a broader optimization deficiency, rather than a lack of representational capacity. Specifically, existing training objectives fail to provide effective credit assignment signals for learning reliable visual attention behaviors during multimodal reasoning.

To investigate this issue, we evaluate representative MLLMs on challenging visual question answering benchmarks involving complex multi-object scenes. As shown in Figure1, analysis indicates that incorrect predictions are frequently accompanied by misplaced attention or inaccurate spatial localization. Quantitatively, we observe a strong correlation between visual attention accuracy and overall task performance, supporting the hypothesis that attention misalignment is a primary driver of reasoning errors.

Motivated by these findings, we propose Entropy-Based Target Attention Reward, a novel reinforcement learning–based framework designed to improve visual information learning in MLLMs. Rather than relying on external visual prompts or tools, it introduces an attention-aware reward that directly aligns optimization signals with visually grounded reasoning steps. The reward is selectively applied based on token-level entropy, encouraging the model to prioritize visual information at uncertain decision points. This design provides a principled mechanism for addressing credit assignment in multimodal reasoning. As illustrated in Figure 1, models trained with this reward exhibit significantly improved attention alignment during inference, consistently focusing on task-relevant visual regions at critical reasoning steps. Importantly, this improvement emerges without requiring explicit visual prompts, special tokens, or iterative prompt refinement at inference time. The resulting model is able to maintain stable visual grounding throughout long reasoning trajectories.

In summary, this paper makes the following contributions:

- **An optimization-centric analysis of visual attention in MLLMs.** We identify visual attention misalignment as a credit assignment failure and demonstrate that chains-of-thought struggle to recover from early attention errors.
- **A reinforcement learning framework for visual attention learning.** We propose Entropy-Based Target Attention Reward, which introduces an entropy-selective, attention-based reward to enable stable and sustained visual grounding during multimodal reasoning.
- **A strong and generalizable visual reasoning model.** Using the proposed framework, we train SAYO and show consistent improvements across diverse multimodal benchmarks, demonstrating strong generalization in both reasoning and perception tasks.

2 Related Works

2.1 MultiModal Large Language Models

Multimodal Large Language Models (Huang et al., 2026; Team et al., 2025)have made remarkable progress by integrating various modalities—such as text, images, and video—into a unified framework for understanding and reasoning. In this framework, different modality encoders project inputs into a shared semantic space, which is then processed by a language model to generate responses. However, although most existing MLLMs possess powerful reasoning capabilities, they struggle to pinpoint the truly relevant parts within complex visual information accurately (Liu et al., 2025b). This prevents the reasoning capabilities of MLLMs from being effectively utilized.

2.2 Enhance Visual Reasoning

A significant recent trend in research involves explicitly applying visual processing to images using external tools (e.g., Python programs) before inputting them into models, such as resizing images or adding bounding boxes around target objects. Studies indicate that such methods can significantly impact the performance of MLLMs, particularly in tasks like visual localization. Recent studies, such as Visual SketchPad (Hu et al., 2024), ReFocus (Fu et al., 2025), and ControlMLLM (Wu et al., 2025) have explored the performance of visual cueing frameworks across various visual comprehension tasks. However, these methods rely on

where MLLMs locate visual prompts for processing. BLINK (Fu et al., 2024) indicates that most open-source multimodal language models struggle to comprehend visual prompts, and incorrect localization undermines their effectiveness. Furthermore, constrained by the fixed nature of the procedures, the aforementioned methods face challenges in transferring to new visual reasoning tasks.

2.3 Strengthen Visual Focus

Consistent with our proposed visual focus, recent research (Yang et al., 2025b; Zhang et al., 2025a) emphasizes the importance of enhancing attention to visual cues in long-chain reasoning. (Chen et al., 2025) discovered that the model's overreliance on prior language led to neglect of visual details in dense information tasks. Look-Back (Yang et al., 2025a) enhances the focus of the thinking process on images by incorporating look-back labels into long-term reasoning chains. Building upon this foundation, Reflection-V (Jian et al., 2025) further introduces an attention reward mechanism. By rewarding the overall attention of the thought chain text toward the image, it promotes the discovery of visual information. Unlike the aforementioned methods, we leverage data annotated with visual bounding boxes to enhance the visual attention capabilities of RL-trained MLLMs. The resulting MLLMs maintain visual focus and reasoning on target objects without requiring lengthy textual inference and reflection processes.

3 Do MLLMs Know where to focus?

Recent studies suggest that multimodal large language models (MLLMs) frequently fail to attend to the correct visual regions during reasoning, resulting in inference errors. This observation raises two key questions: (i) whether current MLLMs are able to accurately localize target objects in complex multi-object scenes, and (ii) how visual attention misalignment affects downstream reasoning performance. To answer these questions, we conduct a diagnostic analysis on the GQA (Hudson and Manning, 2019) dataset, focusing on the relationship between visual attention and model accuracy.

As shown in Figure 1, attention alignment plays a critical role in inference quality: directing attention to the correct visual regions significantly improves prediction accuracy. To quantitatively characterize this effect, we introduce a target attention ratio, which measures the extent to which a model allocates visual attention to the target object relative to irrelevant image regions.

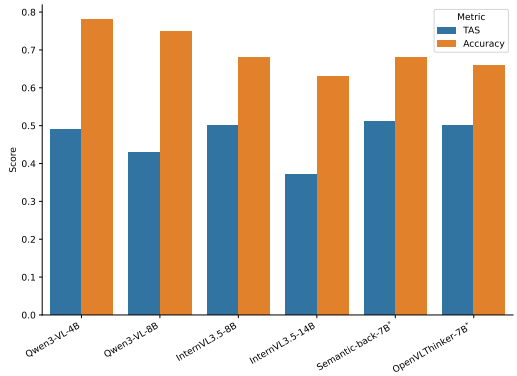


Figure 2: Comparison of target attention score (TAS) and accuracy across Models on a part of GQA dataset. The displayed score and accuracy represent the average across all samples. * denotes models based on the Qwen2.5-7B series.

We extract attention weights from the final transformer layer, where multimodal fusion is fully realized. Let $\alpha^{(h)}_{t_i,t_g}$ denote the attention weight from generated token t_g to image token p at attention head h. We denote by $\mathcal{T}_{\text{target}}$ the set of image tokens corresponding to the target region and by $\mathcal{T}_{\text{all}}$ those corresponding to entire image regions. For each generated token t_g, we compute the average attention mass assigned to the target and entire image regions, then we average across all attention heads to obtain the overall target attention score a and entire image attention score v, respectively:

$$a = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{|\mathcal{T}_{\text{target}}|}\sum_{t_i\in\mathcal{T}_{\text{target}}}\alpha^{(h)}_{t_i,t_g}, \quad (1)$$

$$v = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{|\mathcal{T}_{\text{all}}|}\sum_{t_i\in\mathcal{T}_{\text{all}}}\alpha^{(h)}_{t_i,t_g}. \quad (2)$$

Based on these quantities, we define an attention advantage score to quantify visual focus:

$$R_a = \frac{1}{2}\left(1+\tanh\left(\log\frac{a+\varepsilon}{v+\varepsilon}\right)\right) \quad (3)$$

where ε is a small constant used for stability.

As shown in Figure 2, experiments across multiple visual models reveal a positive correlation between target attention scores and response accuracy. For models within the same series, higher attention weights on the target visual region yield better inference performance. Notably, despite recent advances in reinforcement learning–based op-

timization for MLLMs, all evaluated models exhibit consistently low target attention scores. This suggests that while existing RL methods improve textual reasoning trajectories, they fail to provide effective learning signals for precise visual focus. Consequently, models may develop strong reasoning capabilities without reliably grounding their inferences in the correct visual evidence, fundamentally limiting their visual reasoning performance.

4 Method

In the analysis behind, we observe that current models exhibit insufficient attention to target visual information. This deficiency in visual information utilization severely limits their powerful reasoning capabilities when addressing visual reasoning problems. To address this issue, we propose a reinforcement learning reward training strategy based on visual attention. First, we constructed training data emphasizing visual attention focus using visual reasoning datasets with detailed object annotation information. Subsequently, we employed GRPO to incentivize attention toward target visual signals through reward function.

4.1 Construction of Data with Visual Focus

Existing large visual language model training paradigms typically focus solely on the accuracy and formatting compliance of model responses and evaluation accordingly. The neglect of visual attention shifts during reasoning processes prevents these data from demonstrating the visual attention focus we propose. Inspired by recent studies, we employ a precisely designed toolchain that aligns textual questions with visual token information to accomplish data construction tasks. The data construction process is detailed in the following.

As shown in Figure 3, in the first stage, we extract the final target object text from the question-answer pairs and match it with image segmentation information to obtain bounding box coordinates. Subsequently, based on the model's image processor, we convert the bounding boxes into corresponding visual token ranges. This process yields the objects eligible for visual attention rewards.

4.2 Visual Attention Based Reward

Following previous studies, we employ the reinforcement learning algorithm GRPO to enhance the perception, localization, and reasoning capabilities of visual models. Building upon the original format reward, we have introduced a new reward mechanism based on target visual attention. This aims to incentivize the model to accurately focus on the correct visual tokens. Based on the analyses in Section 3, model's attention to target visual regions is positively correlated with accuracy.

Furthermore, prior research (Wang et al., 2025b,a) has shown that a small subset of high-entropy tokens contains disproportionately rich information. Selecting these information-dense tokens as optimization targets can therefore yield stronger performance. Importantly, in SAYO, high-entropy token selection is used only for reward construction, rather than directly participating in the loss computation. We formalize the reasoning process as a trajectory $\tau = (v, q, t_1, t_2, ..., t_T)$. For a high-entropy token $t_k \in \mathcal{Q}high$, the model exhibits high epistemic uncertainty, often caused by insufficient grounding in the visual context v. Standard Next-Token Prediction minimizes $-\log p(t_k|v, t< k)$, which allows the model to bypass visual verification and rely instead on linguistic priors, leading to hallucination. In contrast, SAYO does not directly optimize these tokens through the training loss; instead, it uses them to identify uncertain reasoning states for reward assignment. Specifically, we encourage higher visual focus at these high-entropy tokens through the reinforcement learning reward:

$$\nabla_t = \mathbb{I}(t \in Q_{high}) \cdot r_v(a_t) \tag{4}$$

This mechanism effectively suppresses "blind" reasoning. We argue that this "Look-to-Verify" policy constitutes a domain-agnostic meta-skill. Once the model learns to ground its attention in complex natural scenes (dense objects) and charts (structured elements), this attention-sharpening capability naturally transfers to other visual domains, such as geometric diagrams, enabling the pre-trained mathematical reasoning engine to operate on correctly perceived visual primitives.

Based on these findings, we designed the following reward rules: Let $\mathcal{Q}_{\text{high}}$ denote the set of generated tokens whose entropies rank in the top 30% among all generated tokens. Using attention weights from the final transformer layer (Jian et al., 2025), we compute the average attention mass from the selected tokens to the target visual region and

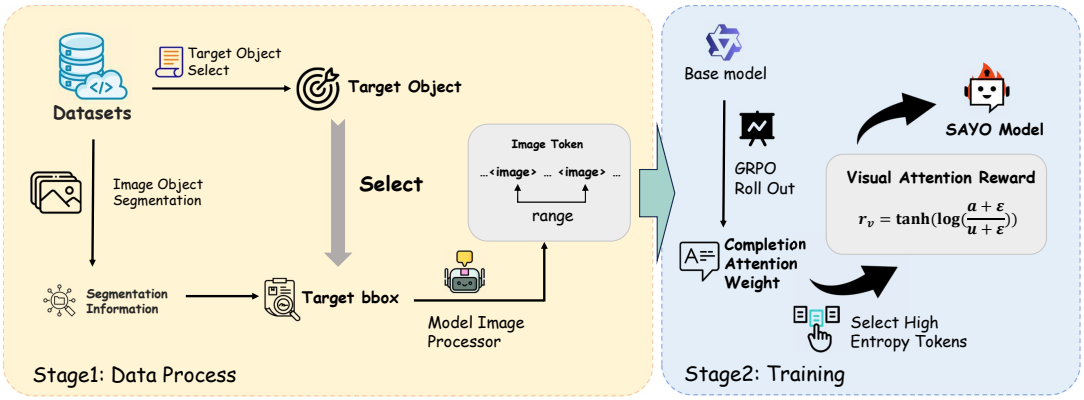


Figure 3: The workflow for our method, including constructing reasoning data with region visual information and training with region attention reward

to all visual tokens, respectively:

$$a_q = \frac{1}{|\mathcal{Q}_{\text{high}}|} \sum_{t_g \in \mathcal{Q}_{\text{high}}} \frac{1}{H} \sum_{h=1}^{H} \frac{1}{|\mathcal{T}_{\text{target}}|} \sum_{t_i \in \mathcal{T}_{\text{target}}} \alpha_{t_i,t_g}^{(h)}, \tag{5}$$

$$v_q = \frac{1}{|\mathcal{Q}_{\text{high}}|} \sum_{t_g \in \mathcal{Q}_{\text{high}}} \frac{1}{H} \sum_{h=1}^{H} \frac{1}{|\mathcal{T}_{\text{all}}|} \sum_{t_i \in \mathcal{T}_{\text{all}}} \alpha_{t_i,t_g}^{(h)}. \tag{6}$$

Therefore, the visual attention-based reward is given by:

$$r_v = \tanh\left(\log \frac{a_q + \varepsilon}{v_q + \varepsilon}\right), \tag{7}$$

where ε is a small constant for numerical stability. The reward $r \in (-1, 1)$ reflects whether the model allocates relatively more attention to the target region than to the visual context as a whole. The overall reward r_o in GRPO is the weighted sum of the region visual attention-based reward r_v and format reward r_f (Shao et al., 2024), the equation is given as: $r_o = r_v + r_f$

5 Experiments

5.1 Experimental Setup

Implementations. To evaluate our method, we use Qwen3-VL (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025c) as the base model. During the training stage, we trained the model using GRPO with entropy and visual attention-based reward. The detailed composition of training data is shown in Appendix C. Training details are provided in Appendix A.

Benchmarks. We conducted detailed analytical experiments to evaluate how our approach enhances the model's visual reasoning capabilities. To ensure thoroughness in our assessment, we selected multiple widely recognized visual understanding benchmarks across various domains. These benchmarks encompass structured image reasoning, mathematical reasoning, and general visual reasoning. For evaluating mathematical reasoning, we adopt We-Math (Qiao et al., 2025) and MathVision (Wang et al., 2024a). To evaluate performance across general visual reasoning, we utilize M3CoT (Chen et al., 2024b), V*Bench (Wu and Xie, 2024), MMStar (Chen et al., 2024a), and MME-RealWorld-Lite (Zhang et al., 2025b). Furthermore, ChartQA (Masry et al., 2022), CharXiv (Wang et al., 2024b), and AI2D (Kembhavi et al., 2016) are used to assess structured image reasoning ability, as they cover a broad range of chart understanding questions. Additionally, we compared SAYO against various of baselines: (i) close-source MLLMs such as GPT-4o (OpenAI) and Gemini 2.5 pro (Comanici et al., 2025); (ii) open-source general MLLMs, such as Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025c), and Kimi-VL-16B (Team et al., 2025); (iii) open-source reasoning MLLMs, such as OpenVLThinker-7B (Deng et al., 2025), Semantic-back-7B (Yang et al., 2025a), ViGoRL (Sarch et al., 2025), and NoisyRollout-Geo3k-7B (Liu et al., 2025a).

5.2 Main Results

We evaluate the performance of our model, SAYO, on various of visual reasoning benchmarks across three categories: visual math problems, structured image problems, and general reasoning. As shown in Table 1, the results indicate that our model significantly outperforms base models and other open-source models of a similar scale in terms of reason-

Model	General				Math		Chart			Avg
	MMERealWorld	M3CoT	V*	MMStar	MathVision	We-Math	ChartQA	AI2D	CharXiv	
Close-Source Models										
GPT4o	73.06	74.20†	-	64.70†	30.40†	69.00†	75.32	84.60	48.90†	-
Gemini 2.5 Pro	-	-	83.80†	77.50†	-	80.60†	83.30†	90.90†	-	-
Open-Source General Models										
Qwen3-VL-4B	56.80	65.10	79.06	60.80	21.64	55.63	80.12	75.16	38.00	59.15
Qwen3-VL-8B	56.23	64.71	81.15	62.60	22.20	52.64	78.96	75.55	42.70	59.64
InternVL3_5-8B	45.34	54.70	73.82	60.27	13.82	20.06	79.12	73.22	34.10	50.49
InternVL3_5-14B	48.57	51.86	70.16	55.93	13.85	33.10	81.76	69.43	38.90	51.51
InternVL3_5-38B	54.56	53.75	79.58	54.20	13.36	24.20	81.60	75.10	**44.40**	53.42
Kimi-VL-16B	42.00	54.01	75.39	63.47	17.50	42.01	82.08	77.04	31.30	53.87
Open-Source Reasoning Models										
ViGoRL	55.34	68.72	60.73	61.53	22.07	62.01	59.64	79.70	30.40	55.57
OpenVLThinker-7B	48.31	63.29	80.10	63.07	25.59	64.48	75.44	82.61	35.20	59.79
Semantic-back-7B	46.43	67.77	78.01	63.07	27.57	66.61	**82.32**	81.74	38.10	61.29
NoisyRollout-7B	51.69	67.90	76.96	63.67	22.01	70.57	82.16	80.73	40.50	61.80
Ours Trained Model										
SAYO-Qwen-4B	57.63	64.32	**83.25**	63.53	22.47	63.97	81.96	80.51	41.90	62.17
SAYO-Qwen-8B	**65.87**	**68.81**	82.20	**67.60**	**27.99**	**73.51**	81.16	**83.48**	42.50	65.70
SAYO-Intern-8B	50.03	57.20	72.77	62.73	13.62	25.23	82.28	76.88	43.20	53.77

Table 1: Performance of SAYO across various visual reasoning benchmarks with different tasks. † indicates results are taken from the respective models' official reports. The best results of each benchmark among open-source models are **bold** and the secondary results are underlined.

ing capability. Compared to models trained using other methods, SAYO also demonstrates significant performance advantages. In contrast to existing reasoning MLLMs, which show improved math reasoning but a decline in general reasoning capabilities, SAYO yields significant improvements across a variety of reasoning tasks. Besides, experimental results show that our proposed method is effective across different models and various scales.

A counter-intuitive finding in Table 1 is the significant improvement on We-Math and MathVision, despite the exclusion of mathematical datasets from our training phase. We attribute this cross-domain generalization to the decoupling of visual parsing and logical reasoning. Current state-of-the-art base models already possess strong latent mathematical reasoning capabilities derived from massive textual pre-training. However, their performance is often bottlenecked by visual misalignment which propagates errors into the reasoning chain. By training with visual attention focus, SAYO learns a robust structure-aware attention policy to correcting the input signal and effectively eliminating the garbage in phase. This allows the model's pre-existing mathematical engine to process valid visual premises, thereby unlocking performance gains without explicit domain-specific training.

5.3 Ablation Study

Effectiveness of attention reward. To systematically assess the contribution of the proposed regional visual attention–based reward, we conduct a series of controlled ablation experiments designed to isolate its effect from other reward components. Specifically, we consider the following variants: (i) Accuracy-only reward, where the attention-based reward is replaced by a conventional answer accuracy reward; (ii) Attention-only reward, where optimization relies solely on the proposed visual attention reward; (iii) Combined reward, where both attention-based and accuracy rewards are applied.

The results, summarized in Table 2, demonstrate that incorporating a visual attention–based reward leads to substantial and consistent performance improvements across benchmarks. Notably, models trained using only visual attention reward achieve performance comparable to those trained with the combined rewards, whereas models trained with accuracy reward alone exhibit only marginal gains.

These findings suggest that deficiencies in current MLLMs stem less from limited reasoning capacity and more from insufficient visual perception and localization. Basically, while existing models are capable of performing complex reasoning once relevant information is identified, they often fail to

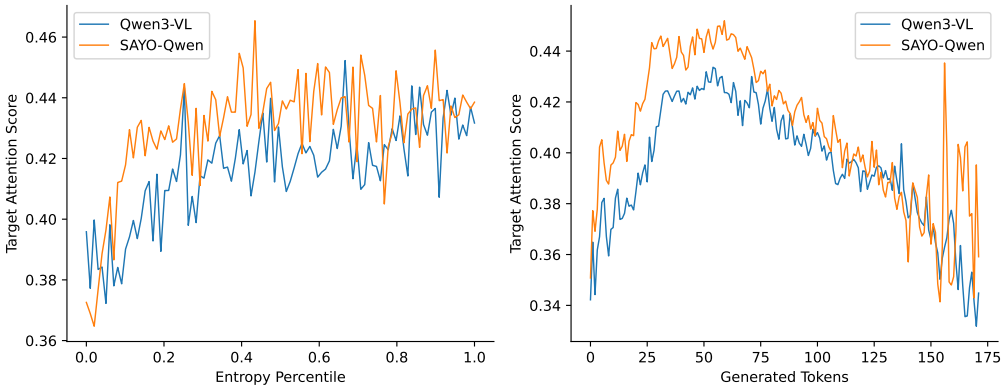


Figure 4: Attention weights of model-generated tokens to target visual tokens (last layer) and attention weights of tokens with different entropy values to target visual tokens. The entropy values shown have been normalized across samples, and the displayed attention weights represent the average across all samples.

Model	General				Math		Chart			Avg
	MMERealWorld	M3CoT	Vstar	MMStar	MathVision	We-Math	ChartQA	AI2D	CharXiv	
Qwen3-VL-8B	56.23	64.71	81.15	62.60	22.20	52.64	78.96	75.55	42.70	59.64
+ Attn. Reward	62.85	67.86	82.20	65.27	25.26	64.83	81.84	83.06	42.50	63.96
+ Acc. Reward	57.01	65.96	78.01	65.20	23.59	56.61	82.52	78.01	41.40	60.92
Acc. + Attn.	61.59	67.77	82.20	66.00	25.95	66.15	82.56	83.03	43.20	64.27
InternVL3.5-8B	45.34	54.70	73.82	60.27	13.82	20.06	79.12	73.22	34.10	50.49
+ Attn. Reward	50.03	57.20	72.77	62.73	13.62	25.23	82.28	76.88	43.20	53.77
+ Acc. Reward	47.94	57.16	74.35	62.20	14.05	24.31	82.68	75.62	41.60	53.32
Acc. + Attn.	49.04	56.77	74.35	63.73	14.14	27.24	82.04	76.81	42.80	54.10

Table 2: Ablation results for different combinations of rewards and visual attention-based reward on performance improvement. Attn. Reward denotes visual attention-based reward, and Acc. Reward denotes accuracy reward. All combinations include format rewards.

reliably extract and attend to the necessary visual evidence. By explicitly incentivizing correct visual focus, the proposed attention reward effectively unlocks the model's latent reasoning capabilities, leading to more accurate and robust inference.

Design an Effective Attention Reward Beyond validating the necessity of the attention-based reward, we further investigate how different design choices affect its effectiveness. We evaluate several reward configurations that vary in reward token selection strategy, as reported in Table 3. Across all settings, introducing visual attention–based supervision consistently improves performance, confirming the general effectiveness of this approach.

A key observation is that selectively rewarding a subset of high-information tokens yields significantly larger performance gains than applying rewards across all tokens. Further analysis of the training process reveals that entropy-selective token rewards lead to more stable and faster convergence, as evidenced by reduced variance in training rewards and improved consistency. By focusing the reward on tokens that correspond to critical decision points, model receives clearer supervision. Additional analyses of training stability and behavior are provided in Appendix D and E.

5.4 Results Analysis

In Section 3, we demonstrate that existing MLLMs exhibit shortcomings in terms of accurate visual perception capabilities through attention advantage score R_a. Based on this metric, we will analyze in this section whether the performance improvement of SAYO truly stems from the training strategy targeting regional visual attention, and what changes this training strategy brings to the model.

We compared the target visual attention score R_a between Qwen3-VL-8B-Instruct and SAYO-Qwen-8B on the same task. As shown in Figure 4, we found that throughout the generated token sequence, SAYO consistently exhibited significantly higher visual attention weights toward the target region than Qwen3-VL, and significantly enhanced attention to the target region of the image during the later stages of inference when generating answers.

Model	General				Math		Chart			Avg
	MMERealWorld	M3CoT	Vstar	MMStar	MathVision	We-Math	ChartQA	AI2D	CharXiv	
Qwen3-VL-8B	56.23	64.71	81.15	62.60	22.20	52.64	78.96	75.55	42.70	59.64
key tokens	62.85	67.86	82.20	65.27	25.26	64.83	81.84	83.06	42.50	63.96
all tokens	56.59	66.01	75.92	64.07	22.57	58.39	82.00	77.43	41.60	60.51
InternVL3.5-8B	45.34	54.70	73.82	60.27	13.82	20.06	79.12	73.22	34.10	50.49
key tokens	50.03	57.20	72.77	62.73	13.62	25.23	82.28	76.88	43.20	53.77
all tokens	46.64	55.22	72.77	63.40	13.75	24.20	81.80	74.79	41.90	52.72

Table 3: Comparative results of visual attention reward for all generated tokens and key generated tokens on reasoning performance improvement.

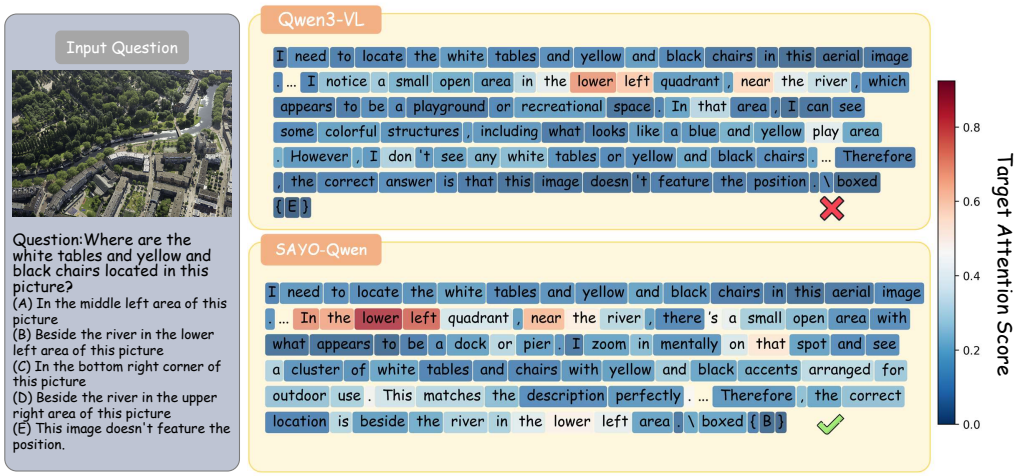


Figure 5: An example demonstrating how areas of visual attention shift during the reasoning process. The background color of tokens in the figure indicates the magnitude of the target visual attention score.

Additionally, to verify whether this method can effectively improve the attention of high-entropy tokens toward target visual regions, we compared the attention weights assigned to these regions by SAYO and Qwen3-VL when generating sequences with tokens of varying entropy levels. As shown in Figure 4, compared to the baseline model, SAYO maintains consistently high visual attention on most tokens with higher entropy.

As discussed in previous sections, model itself possesses sufficient reasoning capabilities, with performance limitations primarily stemming from deficiencies in focusing on relevant visual information. Our method enhances reasoning performance by strengthening the model's attention to effective visual information. Figure 5 presents a comprehensive example demonstrating that these gains are indeed due to the model's visual focus ability. In this example, our model correctly identified and focused on the object mentioned in the problem during the early stages of reasoning, and consistently maintained high visual attention on that region at critical junctures of the reasoning process. This demonstrates the key mechanism for visual reasoning. As shown in Figure 5, sustained and accurate visual attention can continuously guide the reasoning process away from erroneous visual information, thereby arriving at the correct conclusion. In summary, our method demonstrates a more precise focus on visual information.

6 Conclusion

In this paper, we propose that the key factor influencing the reasoning performance of visual models is their ability to perceive and focus on critical visual information. Through quantitative research, we reveal that existing models exhibit poor attention to visual regions containing critical information for inference. We analyze the relationship between visual attention and inference accuracy: visual attention significantly impacts a model's accuracy. To address this bottleneck, we propose a region-level visual attention enhancement training strategy: integrating token-level visual attention rewards into RL training. This strategy enhances model reasoning performance across multiple benchmarks, validating that precise visual attention effectively improves model capabilities.

Limitation

While our method significantly outperforms other methods under multiple tasks, the performance gains are different among various models, especially the weaker models gain less than the stronger models. This discrepancy suggests that benefits of visual attention reward are more pronounced in models with a higher baseline capacity, highlighting its potential to enhance the performance of more powerful architectures more effectively. Moving forward, we plan to explore strategies for improving omni modal applicability, aiming to achieve more excellent performance across a broader range of modality.

Ethic Consideration

Our training data is sampled from open-source datasets, and the training framework is open-source code, edited by our own proprietary code. All models used are open-source, and their sources are clearly credited. The entire process follows transparent and ethical guidelines, ensuring there are no ethical concerns or issues with the data generation. We are committed to maintaining high standards of integrity and transparency in our work.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Huanqia Cai, Yijun Yang, and Winston Hu. 2025. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *Preprint*, arXiv:2502.00698.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *Preprint*, arXiv:2503.06486.

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? *Preprint*, arXiv:2403.20330.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *Preprint*, arXiv:2503.17352.

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

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

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 139348–139379. Curran Associates, Inc.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, Jingcheng Hu, Kangheng Lin, Liang Zhao, Mitt Huang, Song Yuan, Wenwen Qu, Xiangfeng Wang, Yanlin Lai, Yingxiu Zhao, and 74 others. 2026. Step3-vl-10b technical report. *Preprint*, arXiv:2601.09668.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. 2025. Look again, think slowly: Enhancing visual reflection in vision-language models. *Preprint*, arXiv:2509.12132.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *Preprint*, arXiv:2504.13055.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. 2025b. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *Preprint*, arXiv:2510.17771.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. Hello gpt-4o.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

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

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

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. 2024. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. *Preprint*, arXiv:2402.16832.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. 2025a. Sparsemm: Head sparsity emerges from visual concept responses in mllms. *Preprint*, arXiv:2506.05344.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pages 95095–95169. Curran Associates, Inc.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Preprint*, arXiv:2506.01939.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025c. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2025. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Preprint*, arXiv:2407.21534.

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

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025a. Look-back: Implicit visual re-focusing in mllm reasoning. *Preprint*, arXiv:2507.03019.

Siqi Yang, Zilve Gao, Haibo Qiu, Fanfan Liu, Peng Shi, Zhixiong Zeng, Qingmin Liao, and Lin Ma. 2025b. Learning when to look: A disentangled curriculum for strategic perception in multimodal reasoning. *Preprint*, arXiv:2512.17227.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *Preprint*, arXiv:2303.11381.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *Preprint*, arXiv:2502.17422.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *Preprint*, arXiv:2408.13257.

A Implementation of Training Details and Hyperparameters

During the Group Relative Policy Optimization (GRPO) phase, we limit max responses to 110 tokens and apply KL divergence with a coefficient of $1e^{-3}$. We use 6 GPUs with 4 epochs of training on TRL (von Werra et al., 2020) framework. More details and hyperparameters are shown in Table 4

Parameters	Value
Epochs	4
Per Device Batch Size	64
Warmup	False
Rollout	16
Rollout Temperature	1.0
Rollout Top-P	0.9
KL divergence coefficient	1×10^{-3}
Learning rate	5×10^{-6}
Weight Decay	1×10^{-2}
Max Grad Norm	0.8
Optimizer	AdamW

Table 4: The hyperparameters used during GRPO with visual attention based reward.

B Prompts in Training and Evaluation

Prompt Template in GRPO

A conversation between a user and an assistant. The user asks a question, and the assistant must answer it.
The assistant first conducts internal reasoning, followed immediately by the final answer enclosed within <answer>...</answer> tags.
Requirements:
1. The reasoning outside <answer></answer> should be concise, minimal, and essential.
2. The final answer must be correct, complete, and directly address the user's question.
<question>

Prompt Template in Evaluation

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in
boxed{}.

Prompt Template using for ablation study

A conversation between a user and an assistant. The user asks a question, and the assistant must answer it.
The assistant first conducts internal reasoning, followed immediately by the final answer enclosed within <answer>...</answer> tags.
Requirements:
1. The reasoning outside <answer></answer> should be concise, minimal, and essential.
2. The final answer must be correct, complete, and directly address the user's question.
3. When reasoning, focus on the areas of aim object in the image.
<question>

C Training Data Source

We collected data from the dataset in the field of real-world visual reasoning and the structured graph reasoning for RL training. The specific composition is shown in Table 5.

Datasets	Size
GQA (Hudson and Manning, 2019)	$\sim$ 16k
Refocus_Data (Fu et al., 2025)	$\sim$ 4k

Table 5: Detailed composition of the datasets used for RL training stage.

D Experiments Results

Detailed Study of Visual Attention Shifts We further evaluated SAYO's optimization of visual attention on the test dataset. As shown in Figures 6 and 7, SAYO maintains higher attention to both the target region and the entire image throughout the generation process compared to baseline models. When comparing tokens with different entropies, SAYO significantly enhances visual attention for all tokens except those with extremely low entropy.

Detailed Study of Entropy Sensitivity The result in 5.3 aligns with the intuition that not all tokens contribute equally to visual grounding during multimodal reasoning. Many tokens serve primarily syntactic or connective roles and do not require direct visual attention. Including such low-information tokens in reward computation introduces noise, weakening the learning signal and potentially destabilizing policy optimization. To further investigate the impact of entropy selection range on attention rewards, we trained multiple entropy-based reward token selection ranges. As shown in Table 6, selecting fewer tokens will exclude some tokens carrying important information, leading to reduced training effectiveness. Conversely, selecting more tokens introduces additional training noise, thereby diminishing training outcomes.

Improvement in Model Inference Length Figure 8 illustrates the relationship between model output length and accuracy across different benchmarks. Thanks to improved visual attention, SAYO achieves higher reasoning performance with shorter reasoning lengths. This demonstrates that accurate visual perception can effectively reduce the extra inference required by the model due to ambiguous visual information.

Evaluation on abstract tasks To more broadly evaluate the performance of our model on abstract reasoning tasks, we conducted evaluation experiments on ScienceQA (Lu et al., 2022) and MMIQ (Cai et al., 2025) benchmark. Table 7 demonstrates that our method can effectively enhance the model's reasoning capabilities even for tasks without explicit visual targets, highlighting its strong generalization ability.

E Trainging Behaviors

Figure 9 shows the training details for selecting top token rewards versus full token rewards. After excluding low-entropy tokens, the reward values can increase normally during training due to reduced noise signals. However, if the range of tokens selected for rewards is too narrow, it will actually increase training instability. We speculate that an overly narrow selection range prevents some tokens containing critical information from receiving

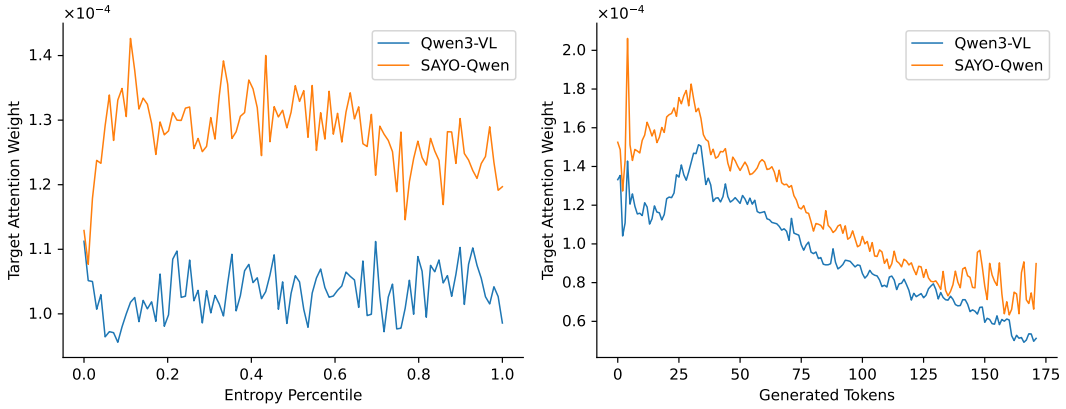


Figure 6: Attention weight (last layer) on target visual token in Qwen3-VL-8B-Instruct and SAYO-8B. The entropy values shown have been normalized across samples, and the displayed attention weights represent the average across all samples.

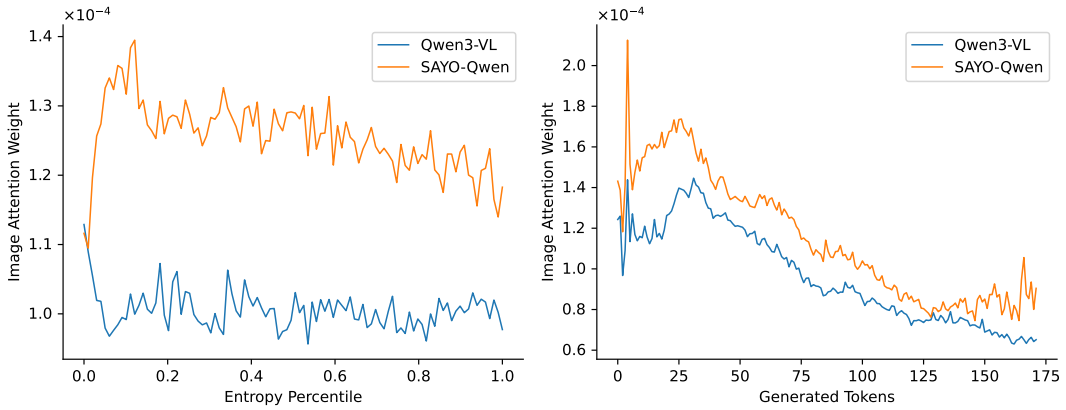


Figure 7: Attention weight (last layer) on all visual tokens in Qwen3-VL-8B-Instruct and SAYO-8B. The entropy values shown have been normalized across samples, and the displayed attention weights represent the average across all samples.

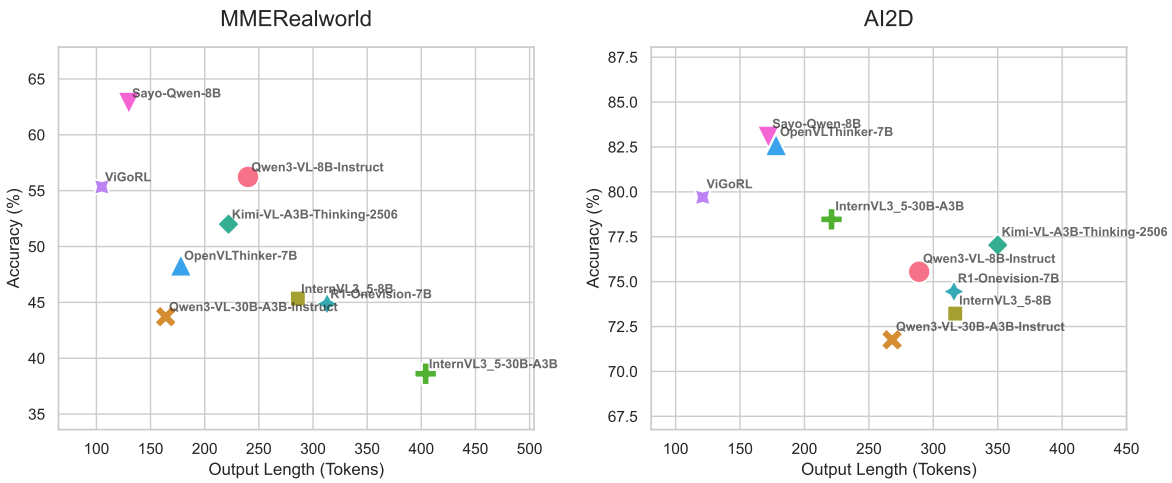


Figure 8: Comparative analysis of model accuracy and average output length across MMERealworld and AI2D.

Range	General		Math		Chart	
	MMERealWorld	M3CoT	MathVision	We-Math	AI2D	CharXiv
Sayo-Qwen-8B	62.85	68.46	25.26	64.83	83.06	42.50
Top 20% tokens	58.36	65.19	17.80	59.08	79.40	40.90
Top 40% tokens	62.48	66.18	20.13	66.84	82.93	39.30

Table 6: Ablation results for different ranges of entropy select strategies in visual attention-based reward.

Model	MMIQ	ScienceQA
Qwen3-VL-8B-Instruct	16.19	88.27
Sayo-Qwen-8B	19.93	89.46

Table 7: Extend evaluation results on abstract reasoning tasks.

rewards, thereby destabilizing the training process. In addition, reducing the range of choices lowers the computational cost of calculating visual attention weights and increases training speed.

F Case Study

Figure 10, 11, 12 show examples of performance on benchmarks in different domains. The erroneous visual information obtained by the baseline model significantly affects the model's performance during inference, whereas SAYO ensures the accuracy of inference by improving the model's perception of visual information.

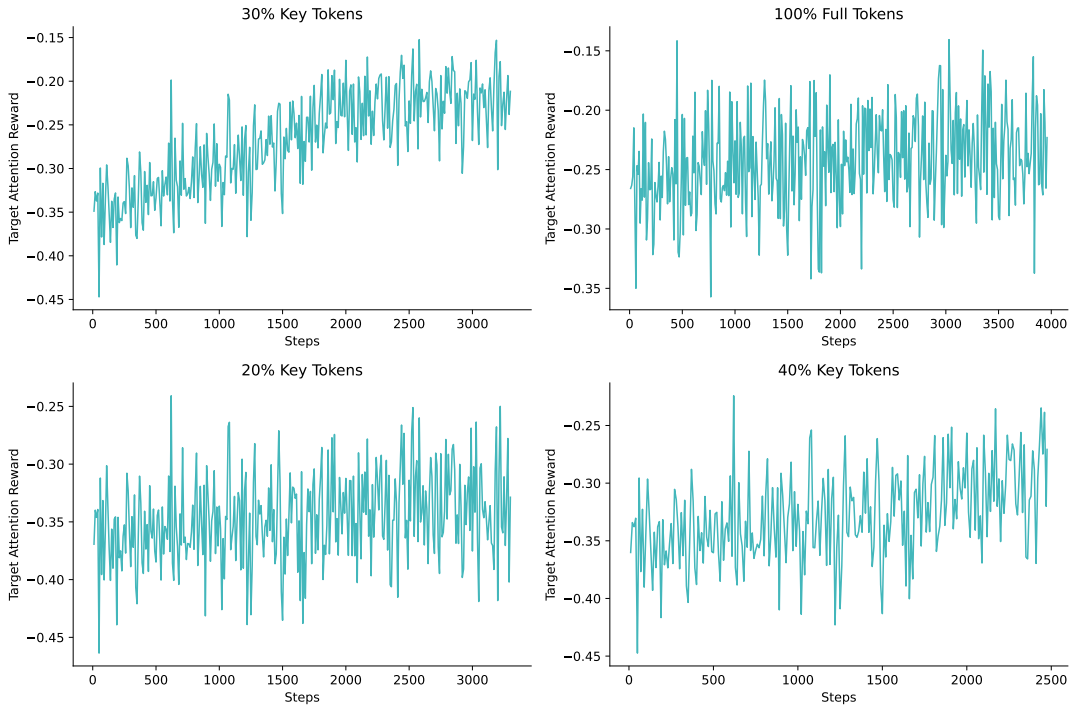
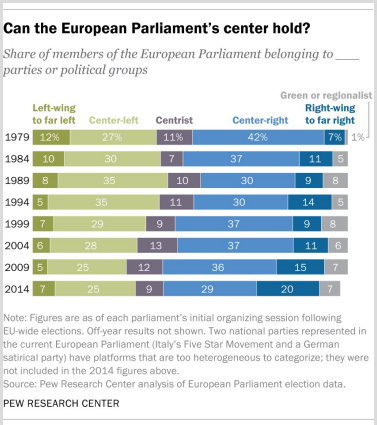


Figure 9: Attention rewards vary with training step. The left figure shows rewards covering the top 30% of high-entropy tokens, while the right figure covers all tokens.

Question: Is the value of light green bar 35 in 1989?



Qwen3VL

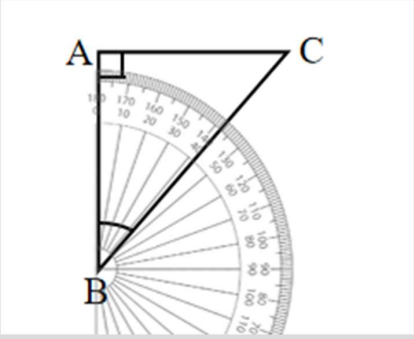
To determine if the value of the light green bar is 35 in 1989, I need to examine the provided bar chart.
1. **Locate the year 1989:** I find the row labeled "1989" in the chart.
2. **Identify the light green bar:** The legend at the top indicates that the "Left-wing to far left" category is represented by the light green bar.
3. **Read the value for 1989:** Looking at the "1989" row, the light green bar (Left-wing to far left) has the value "8" written inside it.
The value shown for the light green bar in 1989 is 8, not 35. The value 35 is in the same row but corresponds to the "Center-left" category, which is represented by a darker green bar.
Therefore, the statement is false.
\boxed{No}

Sayo-Qwen

I need to examine the chart to find the light green bar corresponding to the year 1989. The light green bar represents the "Center-left" political group. Looking at the row for 1989, I see that the light green bar has a value of 35, which is clearly labeled in the chart. This value is consistent with the trend shown in the chart for the center-left group, which was relatively strong in the 1980s and early 1990s before declining slightly in later years. The value is also visually distinct from the other bars in that row.
\boxed{Yes}

Figure 10: Case 1 of Sayo-Qwen-8B in ChartQA.

Question: As shown in the figure, the protractor measures the degree of ∠B as ()°? Options: A. 42; B. 50; C. 60; D. No correct answer



Qwen3VL

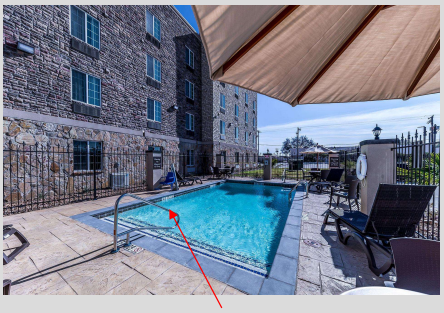
Looking at the figure, I see a right triangle ABC with the right angle at A. There is a protractor centered at point B, and it's measuring angle ABC, which is angle B. The protractor has two scales: the outer one going from 0° to 180° clockwise, and the inner one going from 0° to 180° counterclockwise. Since the angle is at the vertex B, and the rays go out to A and C, I need to see which scale to read. Looking at the protractor, the ray BA is aligned with the 0° mark on the outer scale (which is the one going from 0° to 180° clockwise). The ray BC is pointing to the 50° mark on the outer scale. Therefore, the measure of angle ABC is 50°. Let me double-check. The protractor is centered at B. The baseline (0°) is along BA. The ray BC is at 50° on the outer scale. Since the angle is inside the triangle and the protractor is oriented such that the 0° is on BA and the 180° is on the opposite side, the angle between BA and BC is indeed 50°. So, the answer should be 50°. \boxed{B}

Sayo-Qwen

I need to determine the measure of angle B using the protractor shown in the figure. The protractor is centered at point B, and the angle is formed by rays BA and BC. The protractor has two scales: the outer one going from 0° to 180° clockwise, and the inner one going counterclockwise. Since point A is at the 0° mark on the outer scale, and point C is somewhere along the arc, I need to read the angle from the outer scale. Looking at the protractor, the ray BC intersects the outer scale at approximately 42 degrees. The inner scale would read 138 degrees, but since the angle is measured from the 0° mark at A, we use the outer scale. The arc from A to C clearly aligns with the 42° mark on the outer scale. The right angle at A confirms that this is a right triangle, but the question is specifically about angle B. Therefore, the measure of angle B is 42 degrees. \boxed{A}

Figure 11: Case 2 of Sayo-Qwen-8B in We-Math

Question: What's the text on the pool by the blue chair? Choices:(A) 4FT12, (B) 7FT12, (C) 4FT13, (D) 4FT11, (E) The image doesn't feature the content.



Qwen3VL

I need to examine the image closely to find the text on the pool near the blue chair. The blue chair is on the left side of the pool, next to a metal handrail. There is a small white sign with black text on the pool deck right next to the handrail, near the blue chair. Looking at the sign, I can clearly read the text: "4FT11". This appears to be a depth marker indicating the water depth at that point is 4 feet 11 inches. Now, I will compare this to the given choices:- (A) 4FT12- (B) 7FT12- (C) 4FT13- (D) 4FT11- (E) The image doesn't feature the content. The text on the sign matches option (D) 4FT11. \boxed{D}

Sayo-Qwen

I need to examine the text on the pool near the blue chair. The blue chair is on the left side of the pool, next to the metal handrail. Looking closely at the pool's edge near that area, there's a small white sign with black lettering. The sign is partially visible and appears to read "4FT12" — indicating a depth of 4 feet 12 inches. This matches the depth marker commonly placed on pool edges for safety. The other options don't align with the visible text. The image clearly shows the text, so the correct choice is not "the image doesn't feature the content." \boxed{(A) 4FT12}

Figure 12: Case 3 of Sayo-Qwen-8B in MME-Realworld