# REINFORCED ATTENTION LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Post-training via Reinforcement Learning (RL) has catalyzed significant reasoning gains in Large Language Models (LLMs) by scaling test-time computation. While recent efforts attempt to extend this paradigm to Multimodal Large Language Models (MLLMs) by incentivizing verbose rationales, we demonstrate that such gains are marginal for perception; indeed, excessive verbalization can inadvertently degrade performance. We propose **Reinforced Attention Learning** (**RAL**), a policy-gradient framework that treats internal attention distributions as an optimization target. Unlike conventional RL that focuses on high-reward token sequences, **RAL** explicitly explores and rewards optimal information allocation patterns during the generation process. By shifting optimization from "what to generate" to "where to attend", **RAL** enhances the model's ability to isolate salient signals within complex multimodal contexts. Extensive experiments across diverse image and video benchmarks show that **RAL** consistently outperforms standard GRPO and other baselines. We further extend this approach to On-Policy Attention Distillation, proving that transferring latent attention behaviors to student models yields superior cross-modal alignment compared to standard knowledge distillation. Our results establish attention policies as a principled, generic alternative for multimodal post-training. We will release our code and data upon publication.

## 1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable proficiency in complex domains such as mathematics and programming. Beyond massive unsupervised pre-training Brown et al. (2020); Gemini Team (2025), post-training has emerged as a critical technology for eliciting long-form Chain-of-Thought (CoT) reasoning. Current paradigms predominantly employ Reinforcement Learning (RL) to optimize the model's policy, utilizing rewards derived from learned models or programmatic verifiers to favor high-utility token sequences. Fundamentally, these policy gradient methods refine the next-token distribution to maximize expected rewards. Recent research has further highlighted a robust correlation between reasoning length and task accuracy—a phenomenon central to the concept of test-time scaling.

To extend these gains to multimodal LLMs (MLLMs), recent studies have attempted to incorporate "thinking processes" into Visual Question Answering (VQA) tasks. Under this paradigm, models are incentivized to generate exhaustive textual descriptions of visual inputs as a CoT prior to providing a final answer. However, our results reveal that for core perception tasks—such as image or video question answering—extensive textual reasoning provides only marginal gains and may even degrade the model's fundamental perceptual capabilities.

We attribute this limitation to the insufficiency of next-token prediction as the fundamental policy objective in MLLM post-training. In typical MLLM architectures, visual inputs are encoded as tokens and projected into the textual embedding space to serve as context for generation. Accurately answering fine-grained questions requires the model to precisely identify and attend to task-relevant visual information. This process is governed by the Transformer's attention mechanism, which must learn to assign high weights to salient multimodal tokens. Standard RLHF, however, optimizes for the result (the token) rather than the process (the internal information allocation).

Inspired by this observation, we reformulate the post-training policy for MLLMs to operate directly on the attention distribution during generation. This yields **Reinforced Attention Learning**



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

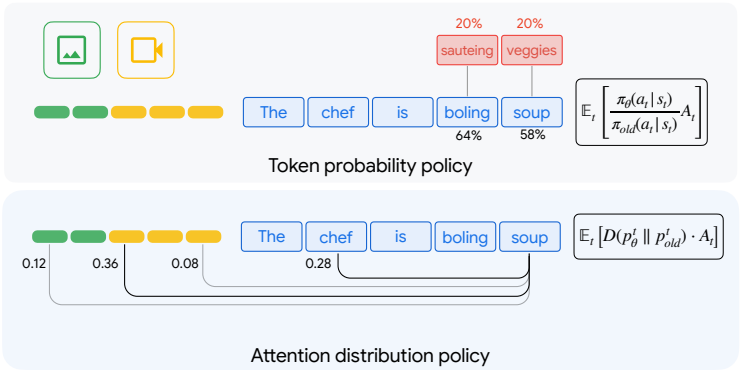


Figure 1: **`Reinforced Attention Learning` formulates internal attention distributions as a policy.** Unlike traditional methods that optimize next-token probabilities ("what to generate"), our approach prioritizes the selective allocation of information focus ("where to look"). Optimizing for the advantage, the model explores a high-rewarding attention policy that effectively isolates salient information from dense contexts.

(**RAL**), an algorithm designed to optimize the model toward high-utility attention trajectories. Unlike traditional methods, **RAL** treats the attention pattern itself as the policy: when a response receives a high reward, the algorithm encourages the underlying attention distribution by minimizing the divergence between the current and reference policies. Conversely, for low-reward responses, the model is penalized by increasing the divergence from those sub-optimal patterns. By shifting the optimization target from token likelihood to attention-based allocation, **RAL** fine-tunes MLLMs more directly for multimodal alignment. Our results indicate that **RAL** consistently outperforms Group Relative Policy Optimization (GRPO) across video and image benchmarks, particularly on perception-intensive tasks.

The efficacy of optimizing attention distributions naturally extends to **On-Policy Distillation**. While traditional distillation focuses on token-level probability alignment, we propose a dual-distillation approach that transfers knowledge via both token and attention distribution alignment. Our experiments indicate that the inclusion of attention distillation provides significant additional performance gains.

In summary, this paper introduces a novel post-training paradigm for MLLMs. Our contributions are as follows:

- **`Reinforced Attention Learning`**: We propose **RAL**, a policy-gradient method that shifts the optimization objective from next-token prediction to attention-distribution alignment, enabling direct reinforcement of visual grounding rather than indirect supervision through textual outputs.
- **On-Policy Attention Distillation**: We further extend this framework to an On-Policy Distillation setting, which substantially improves a student model's ability to inherit fine-grained perceptual and grounding behaviors from a teacher.
- **Empirical Validation**: Extensive experiments demonstrate that **RAL** consistently improves upon GRPO across diverse visual question answering benchmarks that require fine-grained visual understanding and perception.

## 2 Related Work

### 2.1 Post training LLMs through Reinforcement Learning

Post-training is now the standard for aligning Large Language Models (LLMs) with human intent Ouyang et al. (2022). The traditional RLHF pipeline involves Supervised Fine-Tuning (SFT), training a Reward Model (RM) to mimic human preferences, and optimizing the policy via Reinforcement Learning (RL) Christiano et al. (2017). While early RLHF methods significantly improved model safety and helpfulness Bai et al. (2022), they relied heavily on Proximal Policy Optimization (PPO) Schulman et al. (2017). However, PPO's actor-critic framework is memory-intensive due to the

auxiliary critic model. Group Relative Policy Optimization (GRPO) Shao et al. (2024) addresses this by replacing the critic with group-averaged reward estimates. This shift reduces computational overhead while maintaining high performance, particularly in verifiable domains like reasoning and code Guo et al. (2025), leading to the domain of RL with Verifiable Rewards (RLVR).

Extending post-training to multimodal LLMs (MLLMs) introduces challenges beyond text-only alignment, including visual hallucination and robust cross-modal grounding Liu et al. (2023); Li et al. (2023); Zhu et al. (2023). Recent methods adapt RLHF, RLVR, or Direct Preference Optimization (DPO) to improve visual grounding and reduce hallucinations Bai et al. (2025); Zhu et al. (2025); Huang et al. (2025).

A persistent issue is *modality bias*, where the model over-relies on linguistic priors or, conversely, overfits to superficial visual cues Cai et al. (2025). To address this, recent work designs reward functions and training signals that discourage text-only shortcuts, penalize spurious visual correlations, and promote faithful, evidence-based responses Xia et al. (2025); Wang et al. (2025).

Our approach targets the same goal by leveraging a fundamental information-selection mechanism: attention. Since cross-modal reasoning depends on identifying salient evidence in both modalities, directly shaping attention weights provides a principled way to control the cross-modal reasoning policy in our method, rather than relying solely on text-token-level policies.

### 2.2 Distilling knowledge and beyond from teacher to student models

Knowledge Distillation (KD) transfers knowledge from a high-capacity teacher to a student by matching softened output distributions rather than hard labels Hinton et al. (2015). By providing richer supervisory signals, KD has been widely adopted for model compression, domain adaptation, and efficient deployment Romero et al. (2015); Zagoruyko & Komodakis (2016). In the context of large language models, distillation has been extended beyond output logits to intermediate representations, attention maps, and hidden states, enabling improved preservation of structural and reasoning behaviors Sun et al. (2019); Jiao et al. (2020); Wang et al. (2020).

More recent work has explored *on-policy* distillation Agarwal et al. (2024), in which the student generates responses under its own policy and receives supervision from teacher evaluations along these trajectories. Compared to offline KD on static datasets, on-policy distillation mitigates exposure bias and better aligns the student's generation distribution with deployment-time behavior. This paradigm is closely related to RL–based post-training, yet retains the stability and efficiency of supervised learning objectives.

In this work, we investigate knowledge distillation and attention distillation as alternative mechanisms for learning effective attention distributions. Incorporating attention-level supervision within an on-policy distillation framework provides a principled means of regularizing the student's internal information allocation while maintaining policy alignment. This motivates our experimental study of on-policy attention distillation as a complementary approach to purely reward-driven optimization.

## 3 Reinforced Attention Learning

Traditional reinforcement learning (RL) for Large Language Models (LLMs) typically optimizes a policy by maximizing the expected return over the output token distribution. Modern off-policy algorithms like PPO and GRPO utilize a surrogate objective based on importance sampling, which we write as a loss to be minimized:

$$\mathcal{L}_{\text{RL}} = -\,\mathbb{E}_t \left[ \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)} A_t \right] \tag{1}$$

where $A_t$ denotes the advantage estimate and the negative sign converts the maximization of expected advantage into a minimization, so that all objectives in this section are consistently expressed as losses. This formulation explicitly optimizes the **divergence between the current and the behavioral policy's token distributions**. While effective for maximizing immediate rewards, this token-level optimization often precipitates **diversity collapse**. In such cases, the model overfits to

specific high-reward surface forms, thereby weakening its generalization across diverse reasoning patterns and leading to "reward hacking" of the linguistic structure rather than the underlying logic.

To mitigate this, we propose shifting the optimization target from the external output distribution to the **internal attention distributions**. By supervising how the model allocates its computational focus over contextual information, we provide a robust form of structural regularization. We frame the model's aggregate attention over an input prompt as an **information-gathering policy**.

### 3.1 Aggregated causal Attention Distribution Policy

We posit that the model's internal attention mechanism constitutes an alternative, latent policy space. By reinforcing these internal distributions, we guide the model's reasoning process—specifically, how it integrates both the original context and its own generated rationale—without strictly constraining the output tokens to a narrow, brittle distribution.

Let the total sequence be $S = (x_1, \ldots, x_T)$, where $x_1, \ldots, x_P$ represents the prompt and $x_{P+1}, \ldots, x_T$ represents the generated response. Let $\alpha_{t,i}$ represent the attention weight from the generated token at position $t$ to any preceding token at position $i$ ($i < t$), extracted from the final layer and averaged across all heads. For each generated token $t \in [P+1, T]$, we define the causal Attention Distribution Policy $p_\theta^t$ as the distribution over all preceding positions:

$$p_\theta^t(i) = \frac{\alpha_{t,i}}{\sum_{j=1}^{t-1} \alpha_{t,j}}, \quad \forall i \in [1, t-1] \tag{2}$$

This formulation captures how the model attends to its own emerging reasoning in addition to the initial instructions and visual input.

### 3.2 Advantage-Weighted Attention Divergence

To ensure the model retains an effective **information-gathering policy**, we derive a per-token loss function. Drawing inspiration from the importance sampling ratios in PPO/GRPO, this objective encourages attention patterns that correlate with high rewards. The objective is defined as the expected divergence of the current attention policy and the old attention policy:

$$L_{\text{AttnRL}} = \mathbb{E}_t \left[ A_t \cdot D(p_\theta^t \parallel p_{\text{old}}^t) \right] \tag{3}$$

where $A_t$ is the sequence-level advantage and $D(\cdot)$ is a symmetric, bounded divergence measure, such as Jensen-Shannon Divergence (JSD). Using JSD ensures training stability and behaves according to the sign of the advantage:

- If $A_t > 0$, minimizing $L_{\text{AttnRL}}$ pulls the current policy $p_\theta$ toward the successful strategy $p_{\text{old}}$.
- If $A_t < 0$, the objective "pushes" $p_\theta$ away from the suboptimal strategy.

This per-token granularity prevents the "vanishing gradient" effect often encountered when averaging attention over long sequences, ensuring that even late-stage tokens in a long response are supervised by the internal attention objective.

### 3.3 Combined Optimization Objective

The final training objective integrates the standard token-level policy gradient with our internal attention regularizer:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{attn}} L_{\text{AttnRL}} \tag{4}$$

where $\lambda_{\text{attn}}$ is a hyperparameter balancing explicit output maximization with internal attention-level exploration. This dual-objective approach ensures the model remains linguistically flexible while maintaining a structured and reward-aligned reasoning process. Importantly, RAL is architecture- and algorithm-agnostic. The formulation requires only causal attention weights and a per-trajectory

advantage estimate, both of which are readily available in any standard Transformer-based MLLM and any policy-gradient algorithm (e.g., PPO, GRPO, DAPO). RAL can therefore be plugged into existing RL pipelines as an auxiliary objective with no changes to the underlying policy or reward model.

### 3.4 Gradient Derivation

The gradient of $\mathcal{L}_{\text{AttnRL}}$ with respect to the attention logits is derived via the chain rule. Recall from Section 3.1 that the attention policy is averaged over the $H$ heads of the final layer. Let $e^h_{t,i}$ denote the pre-softmax attention logit of head $h$ from position $t$ to position $i$, and $p^{h,t}_\theta = \mathrm{softmax}(e^h_{t,\cdot})$ the corresponding per-head distribution, so that $p^t_\theta = \frac{1}{H}\sum_{h=1}^{H} p^{h,t}_\theta$. Let $J_t = \mathrm{JSD}(p^t_\theta \| p^t_{\text{old}})$.

**Gradient w.r.t. Distribution.** The partial derivative of JSD with respect to the head-averaged distribution $p^t_\theta$ is:

$$\nabla_{p^t_\theta} J_t = \frac{1}{2} \ln\left(\frac{2p^t_\theta}{p^t_\theta + p^t_{\text{old}}}\right)$$

**Gradient w.r.t. Logits.** Because $p^t_\theta$ is a mixture of $H$ softmax distributions rather than a single softmax, the softmax Jacobian $\frac{\partial p^h_j}{\partial e^h_i} = p^h_i(\delta_{ij} - p^h_j)$ must be applied per head, using that head's own distribution, and scaled by the averaging factor $1/H$. The gradient for a specific logit $e^h_{t,i}$ is therefore:

$$\nabla_{e^h_{t,i}} J_t = \frac{1}{H}\, p^{h,t}_\theta(i) \left(\nabla_{p^t_\theta(i)} J_t - \sum_j p^{h,t}_\theta(j)\, \nabla_{p^t_\theta(j)} J_t\right)$$

Note that the outer gradient $\nabla_{p^t_\theta} J_t$ is taken with respect to the head-averaged distribution, while the inner Jacobian is head-specific; the two are coupled only through the shared divergence term.

**Total Parameter Update.** The final gradient for the attention loss is the advantage-weighted accumulation across the sequence and across heads:

$$\nabla_\theta \mathcal{L}_{\text{AttnRL}} = \mathbb{E}_\tau \left[A_\tau \sum_{t=P+1}^{T} \sum_{h=1}^{H} \sum_{i=1}^{t-1} \left(\nabla_{e^h_{t,i}} J_t\right) \nabla_\theta e^h_{t,i}\right]$$

When $A_\tau > 0$, the update minimizes divergence from successful patterns. When $A_\tau < 0$, it pushes the model to explore alternative attention fragments, penalizing the specific reasoning path that led to a low reward.

### 3.5 On-Policy Attention Distillation

Beyond reward-driven optimization, the attention policy framework can be extended to an on-policy distillation setting where the student model $\pi_\theta$ aims to inherit the attention distribution, as the structural reasoning patterns, from a teacher model $\pi_\phi$. In this regime, we provide dense supervision by minimizing the divergence between the student's and teacher's internal attention distributions over the student's own generated trajectories.

**Teacher-Student Alignment.** For each token $t$ in a trajectory $\tau$ sampled from the student policy $\pi_\theta$, we define the distillation loss as the sum of divergences between the student attention policy $p^t_\theta$ and the teacher attention policy $p^t_\phi$:

$$\mathcal{L}_{\text{AttnDistill}} = \mathbb{E}\tau \sim \pi_\theta \left[\sum_{t=P+1}^{T} \mathrm{JSD}(p^t_\theta | p^t_\phi)\right] \tag{5}$$

Crucially, this objective does not involve an advantage term $A_t$. The goal is pure structural imitation, ensuring that for any generated token, the student utilizes the same contextual evidence as the teacher. This provides a denser gradient signal than token-level KL divergence alone.

**Unified Distillation Objective.** The final objective for on-policy distillation combines the RL objective, the generalized knowledge distillation (GKD) on output logits, and our proposed attention alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \mu\mathcal{L}_{\text{GKD}} + \gamma_{\text{attn}}\mathcal{L}_{\text{AttnDistill}} \tag{6}$$

where $\mathcal{L}_{\text{GKD}}$ typically represents the reverse Kullback-Leibler divergence between the output distributions $\pi_\theta(a_t|s_t)$ and $\pi_\phi(a_t|s_t)$. $\mu$ controls the strength of the distillation and $\gamma_{\text{attn}}$ balances the strength of the attention distillation.

**Gradient Flow.** The gradient of $\mathcal{L}_{\text{AttnDistill}}$ with respect to student parameters $\theta$ is derived similarly to 3.4, but is purely aligning toward the teacher's distribution:

$$\nabla_\theta \mathcal{L}_{\text{AttnDistill}} = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=P+1}^{T}\sum_{h=1}^{H}\sum_{i=1}^{t-1}\left(\nabla_{e^h_{t,i}} J_t\right)\nabla_\theta e^h_{t,i}\right] \tag{7}$$

where $J_t = \text{JSD}(p^t_\theta \| p^t_\phi)$. By sampling trajectories from the student ($\tau \sim \pi_\theta$), the model learns to maintain "teacher-like" attention patterns even when navigating states it would not encounter under the teacher's original distribution, thereby mitigating exposure bias in internal representations.

# 4 Experiments

In this section, we empirically evaluate **RAL** for post-training MLLMs. Our experiments investigate both the standard **RAL** objective and the on-policy attention distillation variant, implemented on top of the Group Relative Policy Optimization (GRPO) framework.

## 4.1 Experimental Setup

**Model Configurations** We adopt **Qwen-2.5-VL-7B** as our foundation MLLM. For the on-policy attention distillation experiments, we utilize **Qwen-2.5-VL-32B** as the teacher model. In all experimental settings, the visual encoder and multimodal projector are kept frozen, with gradients updated only for the language model backbone. RAL imposes minimal infrastructure cost: we patch only the final attention block of the policy to eager attention so its weights are materialized, while the rollout engine (e.g., vLLM) and the rest of the model remain untouched. The JSD-based attention reward is computed at training time rather than during rollout, so RAL preserves rollout throughput and remains fully compatible with standard distributed RL setups.

**Training Pipeline** Our training pipeline utilizes the Video-R1 (Feng et al., 2025) dataset and consists of two primary stages, implemented using the `veRL` infrastructure (Sheng et al., 2025):

- **Supervised Fine-Tuning (SFT):** To align the model with our target reasoning schema, we perform SFT using the Video-R1-COT-165k subset. This dataset provides 165k instances of video-question pairs paired with structured Chain-of-Thought reasoning. Following the format in Figure 3, the "thinking" process and final answer are enclosed in <think> and <answer> tags, respectively. The computational overhead for this stage is roughly 10 hours on a cluster of 8× NVIDIA H100s.
- **Reinforcement Learning (RL):** We subsample the first 51.2k instances from the Video-R1-260k dataset and perform training for a single epoch. For each input, the policy generates $G = 8$ rollouts to facilitate advantage estimation. This RL phase requires approximately 120 hours of compute on a cluster of 8× NVIDIA H100 GPU cluster.

### 4.2 Implementation Details

**Visual Processing** For the visual encoder, we set the maximum resolution of images to $5120 \times 28 \times 28$ pixels and the minimum to $128 \times 28 \times 28$ pixels. Video frames are sampled at 1 frame per second (fps), with the total number of frames capped at 128. Each individual frame is constrained to a maximum of $128 \times 28 \times 28$ pixels.

**Hyperparameters** We extract attention weights from the last Transformer layer of the LLM by patching its attention implementation with eager attention, and average them across all attention heads. The final layer is chosen because later layers in Transformer LMs are known to encode higher-level, task-relevant semantics, while head-averaging captures the aggregate information-routing behavior and suppresses noise from individual head specialization. For RL training, we maintain a fixed sampling temperature of $\tau = 0.9$ and a KL coefficient of $\beta = 0.04$. The learning rate is set to $5 \times 10^{-6}$ for SFT and $1 \times 10^{-5}$ for RL. Detailed hyperparameter ranges for **RAL** are provided in Table 3.

**Baselines.** We primarily compare our method against the **Group Relative Policy Optimization (GRPO)** algorithm, which eliminates the need for a separate value model by computing the relative advantage of each response within a group of $G$ sampled outputs for the same query. For a given response $i$, the advantage $\hat{A}_i$ is calculated as:

$$\hat{A}_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\text{std}(r_1, \ldots, r_G)} \tag{8}$$

where $r_i$ represents the total reward for response $i$. Our reward system is entirely rule-based, comprising two components: **Accuracy Reward ($r_{acc}$):** We extract the content within the `<answer>...</answer>` tags and compare it against the ground truth. The model receives a reward of 1.0 for an exact match and 0.0 otherwise. **Format Reward ($r_{fmt}$):** A regular expression-based verifier checks if the response strictly adheres to the template: `<think>...</think><answer>...</answer>`. A reward of 1.0 is granted for perfect formatting and 0.0 otherwise. The final reward $r_i$ is a weighted combination: $r_i = 0.9 \cdot r_{acc} + 0.1 \cdot r_{fmt}$. These settings remain consistent across all experiments, including **RAL**, on-policy attention distillation, the baseline GRPO and the GRPO-based on-policy knowledge distillation.

We also include Video-R1-7B for comparison, which was introduced alongside the dataset. This model is trained using Temporal-GRPO, a proposed variant of the standard GRPO framework that incorporates a temporal coherence reward. Specifically, this method perturbs the chronological ordering of video frames and provides an auxiliary reward when the model demonstrates higher accuracy on correctly sequenced inputs compared to shuffled ones. Since Video-R1-7B utilizes the same Qwen-2.5-VL-7B backbone as our other configurations, it facilitates a strictly controlled and fair comparison across our evaluation suite.

### 4.3 Evaluation Benchmarks

We evaluate **RAL** across a broad spectrum of vision-centric benchmarks to assess spatial, temporal, and reasoning-intensive capabilities.

**Image QA Tasks.** For static image understanding, we utilize:(1) $V^*$ Bench (Wu & Xie, 2023) for fine-grained visual search and spatial reasoning;(2) MMMU Pro (Yue et al., 2024) for multi-step expert-level knowledge;(3) MME (Fu et al., 2025a) for basic perception and high-level cognition;(4) MuirBench (Wang et al., 2024a) for robustness across diverse image types;(5) ChartQA (Masry et al., 2022) for complex data extraction;(6) VizWiz (Gurari et al., 2018) for real-world visual grounding;(7) Blink (Fu et al., 2024) for foundational perception tasks; and (8) CVBench (Tong et al., 2024) for benchmarking core computer vision capabilities.

**Video QA Tasks.** To evaluate temporal reasoning and long-context integration, we adopt:(1) LongVideoBench (Wu et al., 2024) for long-range referring reasoning;(2) NExT-QA (Xiao et al., 2021) for causal and temporal action explanation;(3) Video-MME (Fu et al., 2025b) for comprehensive multi-domain evaluation;(4) Video-MMMU (Hu et al., 2025) for expert-level knowledge acquisition;(5) LVBench (Wang et al., 2024b) for extreme long-form comprehension;(6) MVBench (Li et al., 2024) for multi-task temporal perception; and(7) TempCompass (Liu et al., 2024) for sensitivity to motion and temporal order.

Table 1: **Performance comparison across long-video benchmarks.** **RAL** and Attention Distillation (in blue) are evaluated against baselines; **RAL**-zero (Section 4.6) ablates the explicit thinking process to isolate the attention-policy signal. Optimizing internal attention distributions provides a complementary signal to token-based RL, consistently improving multimodal understanding.

| Method | LongVideoBench | NExTQA | VideoMME | VideoMMMU | LVBench | MVBench | TempCompass |
|---|---|---|---|---|---|---|---|
| Qwen-2.5-VL-7B | 57.0 | 73.7 | 61.6 | 47.6 | 40.5 | 63.4 | 69.5 |
| *Reinforcement Learning* | | | | | | | |
| Video-R1 | 56.5 | 65.3 | 62.5 | 44.7 | 43.5 | 62.0 | 68.0 |
| GRPO | 57.9 | 70.7 | 62.0 | **49.7** | 43.9 | 64.0 | 68.3 |
| **RAL** | **60.1** | **74.1** | **63.4** | 48.6 | **44.2** | **65.5** | **70.0** |
| **RAL**-zero | 58.8 | 76.2 | 65.1 | 49.2 | 45.9 | 62.7 | 68.5 |
| *On-Policy Distillation* | | | | | | | |
| Knowledge Distillation | 59.7 | 70.9 | 61.3 | 47.3 | 43.7 | 65.5 | 69.2 |
| + Attention Distillation | **59.7** | **75.3** | **63.9** | **48.5** | **44.7** | **65.5** | **70.3** |

Table 2: **Performance comparison across image VQA benchmarks.**

| Method | Vstar | MMMUpro | MME | MuirBench | ChartQA | VizWiz | Blink | CVBench |
|---|---|---|---|---|---|---|---|---|
| Qwen-2.5-VL-7B | 70.7 | 36.4 | 2309.3 | 44.9 | 84.0 | 71.2 | 56.3 | 77.6 |
| *Reinforcement Learning* | | | | | | | | |
| Video-R1 | 66.5 | 36.0 | 2266.2 | 40.9 | 84.2 | 63.5 | 47.4 | 72.3 |
| GRPO | 68.6 | 36.8 | 2258.7 | 43.9 | 81.7 | 67.9 | 54.9 | 78.1 |
| **RAL** | **73.3** | 37.8 | **2352.8** | **47.4** | **86.4** | 71.7 | **57.2** | **79.0** |
| **RAL**-zero | 72.3 | **38.4** | 2306.2 | 43.4 | 82.7 | **71.9** | 55.4 | 78.4 |
| *On-Policy Distillation* | | | | | | | | |
| On-policy Distillation | 68.1 | 37.8 | 2344.7 | 39.9 | 81.7 | 71.2 | 56.3 | **78.7** |
| + Attention Distillation | **72.3** | **38.4** | **2345.1** | **43.4** | **83.6** | **72.1** | **57.2** | 78.4 |

### 4.4 Main Results

Table 2 and Table 1 compare **RAL** with GRPO and standard on-policy distillation across image and video VQA benchmarks. Overall, **RAL** consistently outperforms both GRPO and the base Qwen-2.5-VL-7B, suggesting that optimizing attention distributions provides a more stable and effective signal than token-level policy gradients.

**Image VQA.** Under RL, **RAL** surpasses GRPO on all eight benchmarks, with notable gains on $V^*$ (+4.7), MME (+94.1), ChartQA (+4.7), and VizWiz (+3.8). It not only avoids the degradation often seen with GRPO but also exceeds the base model across all metrics, indicating improved generalization.

Under on-policy distillation, attention supervision improves results on 7/8 benchmarks, especially on $V^*$ (+4.2) and MuirBench (+3.5), highlighting the value of modeling where the teacher attends. Strong performance on $V^*$ further demonstrates improved fine-grained visual reasoning.

**Video VQA.** On long-video tasks, **RAL** outperforms GRPO on 6/7 datasets, with largest gains on LongVideoBench (+2.2), NExTQA (+3.4), and MVBench (+1.5), indicating better temporal and multi-hop reasoning. It remains competitive on VideoMMMU.

With distillation, attention alignment improves or matches performance across all benchmarks, with strong gains on NExTQA (+4.4) and VideoMME (+2.6), suggesting benefits for long-context reasoning and event localization.

**Implications.** Across settings, attention-centric learning yields consistent gains, unlike GRPO's benchmark-specific trade-offs. This supports the view that supervising internal attention provides a more stable and generalizable signal. Attention distillation further shows that attention patterns encode transferable reasoning behavior.

In summary, attention-based optimization is an effective complement to RL and distillation for multimodal post-training, particularly for perception-heavy and long-horizon tasks.

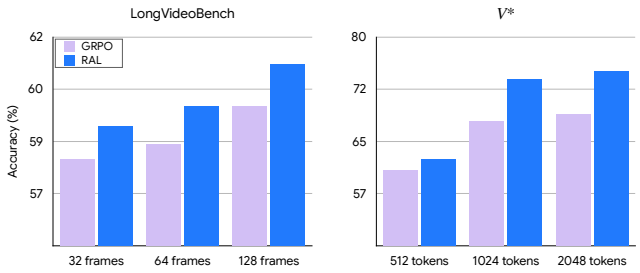


Figure 2: **RAL improves GRPO along the increasing video frames or image resolution.**

### 4.5 Qualitative Analysis: Where Does RAL Attend?

Beyond aggregate benchmark gains, we examine whether **RAL** changes *where* the model attends. Figure 4 compares last-layer attention heatmaps for Base (Qwen-2.5-VL-7B), GRPO, and **RAL** on three perception-heavy questions. Across all examples, Base and GRPO produce diffuse attention spread across the image and answer incorrectly on two of three questions, while **RAL** concentrates attention on the task-relevant region and answers correctly. This indicates that **RAL** does not act as a generic regularizer; it actively shifts the model's information-gathering toward the visual evidence required to answer the question.

### 4.6 Ablation Studies

**RAL yields consistent improvements across visual resolutions and frame rates.** We compare **RAL** against GRPO across video sampling lengths and image resolutions using LongVideoBench (sparse retrieval over long video) and $V^*$ (high-resolution scene reasoning). We sweep maximum frame counts of $\{32, 64, 128\}$ on LongVideoBench and per-image token budgets of $\{512, 1024, 2048\}$ on $V^*$ (where one $28 \times 28$ patch = 1 token, so 1024 tokens $\approx 896 \times 896$). As shown in Figure 2, **RAL** outperforms GRPO at every temporal scale, and the $V^*$ margin widens with resolution from $+1.6$ at 512 tokens to $+6.3$ at 2048 tokens, indicating that attention-based policies scale better as visual information becomes more granular.

**Is an explicit thinking process necessary for VQA?** To isolate attention-level supervision, we introduce **RAL**-zero, which removes the thinking process from both SFT and RL: the `<think>` blocks are stripped from the SFT data (Figure 3) and the RL format reward enforces an `<answer>...</answer>`-only output. All other hyper-parameters and data volumes match the main experiments. On long-video benchmarks (Table 1), **RAL**-zero beats the base model on 5/7 and GRPO on 5/7, achieving the best scores on NExTQA, VideoMME, and LVBench. On image benchmarks (Table 2), it improves over the base on 4/8 and tops all RL methods on MMMU-Pro and VizWiz. These results show that policy-gradient updates applied directly to attention can drive cross-modal reasoning gains even without verbalized rationales.

## 5 Conclusion

We introduced **Reinforced Attention Learning**, an MLLM post-training paradigm that shifts optimization from output token distributions to internal attention distributions. By treating attention as a policy, **RAL** directly reinforces visual grounding, supervising the cross-modal reasoning process that outcome-based RL leaves untouched. Across image and long-video benchmarks, **RAL** consistently outperforms both the base Qwen-2.5-VL-7B and GRPO, with more uniform gains than token-level RL, which can degrade base performance. The same perspective extends to on-policy distillation, where transferring "where to look" from a teacher provides a signal that output imitation alone lacks. Because **RAL** requires only causal attention weights and an advantage estimate, it is a drop-in auxiliary objective for any Transformer-based MLLM and any policy-gradient algorithm.

More broadly, this work establishes attention distributions as a first-class optimization target for multimodal alignment, and points toward fine-grained credit assignment and the optimization of other internal structures, such as MoE routing or cross-modal fusion.

## AI USE STATEMENT

In this work, we did not use generative AI tools for any task requiring disclosure: the research hypotheses, the method and its derivations, the experimental design, the implementation, the data processing, and the interpretation of results are entirely the authors' own work; assistance with synthetic data generation, translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, for tasks with recommended disclosure, we used large language models to edit the manuscript for readability and concision, to convert it to the ICLR 2027 LaTeX template and fix formatting, to proofread the notation and equations for consistency, to cross-check that the numbers reported in the text agree with those in the tables, and to format references.

We have reviewed all AI-assisted work. All AI-edited text was read and revised by the authors to ensure that it accurately reflects our method, experiments, and results, and all experimental numbers were produced by our own training and evaluation pipeline. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

We have taken the following steps to facilitate reproducibility. The **RAL** objective and its gradient are fully specified in Section 3, including the definition of the head-averaged attention policy (Section 3.1), the advantage-weighted JSD loss, the combined objective, and the on-policy attention distillation variant. All experiments build on publicly available models: the Qwen-2.5-VL-7B and Qwen-2.5-VL-32B models, the VIDEO-R1 datasets (Feng et al., 2025), and the open-source `veRL` framework (Sheng et al., 2025). Section 4 describes the model configuration, the two-stage SFT and RL training pipeline, the data subsets and their sizes, the visual pre-processing settings, the reward functions, and the compute budget. The complete hyperparameter configuration, including the sampling temperature, KL coefficient, learning rates, group size, and the swept ranges of the attention-loss weights $\lambda_{\text{attn}}$ and $\gamma_{\text{attn}}$, is listed in Appendix B, and the training data format is illustrated in Appendix A. The implementation requires only patching the final attention block of the policy to eager attention so that attention weights are materialized; no changes to the rollout engine, reward model, or underlying policy-gradient algorithm are needed.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott

Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS Datasets and Benchmarks Track*, 2025a.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025b.

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

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

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

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. URL https://arxiv.org/abs/2501.13826.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv: 2403.00476*, 2024.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL https://arxiv.org/abs/1412.6550.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019. URL https://arxiv.org/abs/1908.09355.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024b. URL https://arxiv.org/abs/2406.08035.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

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

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

# APPENDIX

## A TRAINING DATA FORMAT

We use the VIDEO-R1-COT-165K subset for SFT and the first 51.2k examples of VIDEO-R1-260K for RL. Both stages share the structured `<think>`/`<answer>` schema illustrated in Figure 3: SFT supervises the full chain-of-thought, while RL applies a verifiable reward to the final answer and a format reward on the schema.

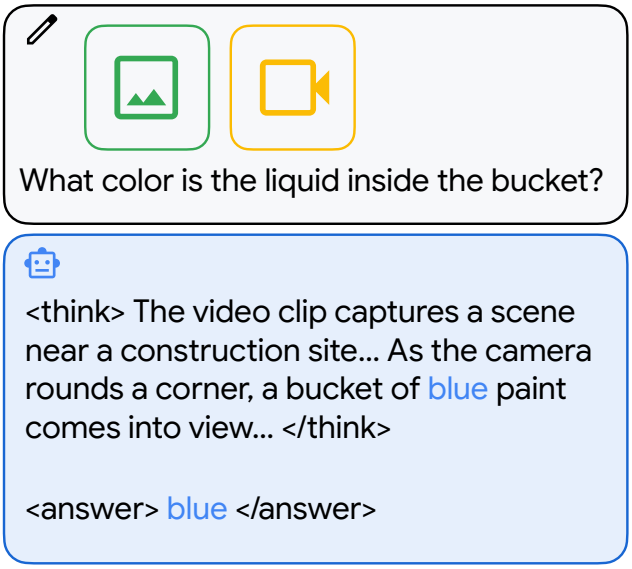


Figure 3: **Sample data for the SFT and RL training stages.** SFT adapts the model to a "think-and-answer" paradigm, and RL applies a reward function that verifies both the format and the correctness of rollout responses.

## B HYPERPARAMETERS

Table 3 lists the full hyperparameter configuration used in our experiments. The attention-loss weights $\lambda_{\text{attn}}$ and $\gamma_{\text{attn}}$ were swept over the listed ranges; remaining settings were kept fixed across all runs.

Table 3: Hyperparameters for **RAL** experiments.

| Parameter | Value |
|---|---|
| Rollout batch size | 8 |
| KL coefficient ($\beta$) | 0.04 |
| Sampling temperature ($\tau$) | 0.9 |
| Learning rate (SFT) | $5 \times 10^{-6}$ |
| Learning rate (RL) | $1 \times 10^{-5}$ |
| Group size $G$ (GRPO) | 8 |
| $\lambda_{\text{attn}}$ (RL) | $\{0.5, 1, 5\}$ |
| $\gamma_{\text{attn}}$ (distillation) | $\{0.05, 0.5, 1\}$ |

## C QUALITATIVE ATTENTION VISUALIZATIONS

To complement the quantitative results in Section 4, we visualize the last-layer attention maps of the base model, GRPO, and **RAL** on perception-heavy questions from $V^*$. Across all three examples in

Figure 4, Base and GRPO distribute attention diffusely across the scene and answer incorrectly on two of three questions, while **RAL** concentrates attention on the task-relevant region (the umbrella, the scarf, the poster) and answers correctly. This provides direct evidence that **RAL** actively redirects the model's information-gathering toward the visual evidence required to answer the question, rather than acting as a generic regularizer.




Figure 4: **Last-layer attention heatmaps for Base (Qwen-2.5-VL-7B), GRPO, and RAL on perception-heavy questions.** Base and GRPO produce diffuse attention and answer incorrectly on two of three examples, while **RAL** concentrates attention on the task-relevant region and answers correctly.