PRIMEFOCUS: HARNESSING INTRINSIC REGULARITIES FOR TEST-TIME POLICY OPTIMIZATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time reinforcement learning (RL) adapts vision-language models (VLMs) on unlabeled test data by rewarding agreement with a majority-voted pseudo-label. This recipe has two weaknesses. First, the reward depends only on the final textual answer, so a consensus that arises from language priors is reinforced as strongly as one supported by the image. Second, training runs for a fixed number of steps, although the best stopping point differs across test sets and no labeled validation set is available. We propose PrimeFocus, a test-time policy optimization method that adds intrinsic signals of the policy to this recipe. For the reward, *consensus-anchored attention sharpening* derives three annotation-free terms from the attention of answer tokens to image tokens: the share of attention on the image (visual enhancement), the concentration of that attention (visual focus), and the concentration of the attention map averaged over all responses that give the same answer (group visual consensus). For stopping, we track the cumulative deviation of the policy's mean token entropy from its exponential moving average, a statistic motivated by prequential estimates of epiplexity, and terminate training when it plateaus. On three visual recognition and four visual question answering benchmarks with InternVL3-2B and InternVL2.5-4B, PrimeFocus improves the mean accuracy of TTRV under the same backbone by 2.27 and 2.42 points. Ablations and diagnostics examine each reward term, attention-level failure modes, and the behavior of the stopping rule.

1 INTRODUCTION

Reinforcement learning (RL) has substantially improved vision-language models (VLMs) (Ma et al., 2026; Feng et al., 2025; Wang et al., 2024; Schoepp et al., 2024). With high-quality annotated data and policy optimization, VLMs reach strong accuracy on visual question answering (VQA) (Lyu, 2025; Jian et al., 2024; Weng et al., 2025) and visual recognition (He et al., 2025; Hu et al., 2025; Lu & Wang, 2025). Large-scale annotation is costly, however, and a model trained on one distribution often degrades on test data drawn from another. Test-time RL addresses this by training on the unlabeled test set itself (Zuo et al., 2026; Singh et al., 2026). Current test-time RL for VLMs builds on Group Relative Policy Optimization (GRPO) (Shao et al., 2024): for each query it samples several responses, takes the most frequent answer as a pseudo-label, and rewards responses by their agreement with it (Singh et al., 2026).

This recipe has two limitations (Figure 1). First, the reward is computed only from the final textual answer. Agreement among sampled answers does not imply that the answers rely on the image: when the policy answers from language priors, majority voting rewards the resulting consensus as strongly as a visually supported one, and training can drift toward text-driven answering, a failure mode also documented for VLM decoding (Huang et al., 2024; Leng et al., 2024; Liu et al., 2024). Second, existing methods train for a fixed number of steps. Models start from different priors on different test sets, and unsupervised RL with intrinsic rewards typically improves and then degrades (He et al., 2026a), so a fixed schedule underfits some datasets and overfits others. Without labels, the stopping point cannot be chosen on a validation set.

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.

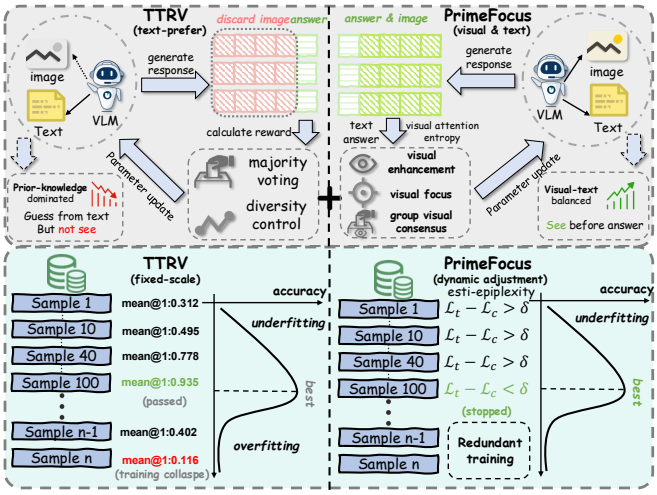

Figure 1: Two limitations of test-time RL for VLMs. **Visual-text imbalance** (top): methods such as TTRV reward agreement among textual answers and ignore intrinsic visual signals, so a consensus formed from language priors is reinforced. **Fixed training scale** (bottom): models start from different priors on different test sets, so a single fixed schedule underfits some datasets and overfits others.

We propose PrimeFocus, a test-time policy optimization method that uses intrinsic signals of the policy for both problems. For the reward, *consensus-anchored attention sharpening* reads, during the log-probability forward pass, the causal attention from response tokens to image tokens in the final decoder layer, and derives three terms from it. **Visual enhancement** rewards the share of attention placed on image tokens, which counteracts answering from language priors. Because this term alone can be increased by spreading attention uniformly over the image, **visual focus** penalizes the entropy of the attention distribution over image patches. Because focus alone can be satisfied by collapsing onto a single patch, the two terms are used jointly. **Group visual consensus** targets the weakness of answer-level voting: for each set of responses that give the same answer, we average their attention maps and penalize the entropy of the average. Responses that agree because they attend to the same image regions produce a concentrated average, whereas agreement that is not accompanied by shared attention produces a diffuse one.

For stopping, we accumulate the absolute deviations of the policy's mean token entropy on its own rollouts from an exponential moving average (EMA) of its past values, and stop training when this cumulative quantity plateaus. The construction is motivated by, but distinct from, the prequential estimate of epiplexity (Finzi et al., 2026), which measures the structural information a bounded learner extracts from data as the area between the online loss curve and the final loss (Section 3.3).

We evaluate PrimeFocus on seven benchmarks with InternVL3-2B and InternVL2.5-4B. Under the same backbone and protocol, PrimeFocus improves the mean accuracy of TTRV by 2.27 and 2.42 points. Our contributions are:

- We identify that frequency-based test-time rewards for VLMs are blind to whether a consensus relies on the image, and propose three annotation-free attention-derived reward terms, including a group-level consensus term defined over answer clusters.
- We propose a label-free stopping rule based on the plateau of an EMA-residual entropy statistic, and relate it to prequential epiplexity through a bound on the EMA approximation.
- We report same-backbone comparisons with TTRV on seven benchmarks, component ablations, attention failure-mode diagnostics, correlation analyses, and stopping analyses.

2 Related Work

Vision-language models. VLMs such as InternVL (Chen et al., 2024; Zhu et al., 2025; Wang et al., 2025), Qwen-VL (Bai et al., 2025) and GLM-V (Team et al., 2026) couple a vision encoder with a

large language model, address tasks from video understanding (Tang et al., 2023; Zhang et al., 2024b; 2023) and GUI grounding (Qin et al., 2025; Li et al., 2025) to captioning (Jung et al., 2025; Tang et al., 2025) and VQA (Lyu, 2025; Jian et al., 2024), and increasingly use RL in post-training.

Test-time reinforcement learning. RL with verifiable rewards (RLVR) improves reasoning but requires ground-truth labels (Zhang et al., 2025; Shao et al., 2025; Yu et al., 2025). Test-time RL removes this requirement by training on unlabeled test data with self-generated rewards; unlike test-time adaptation of CLIP-style classifiers via entropy minimization (Shu et al., 2022) or CLIP rewards (Zhao et al., 2024), it adapts generative VLMs by policy optimization. TTRL (Zuo et al., 2026) derives a pseudo-label for each question by majority voting and rewards agreement with it. TTRV (Singh et al., 2026) transfers this paradigm to VLMs with a frequency-based reward and a diversity-control term on the entropy of the answer distribution. SCOPE (Wang et al., 2026b) weights votes by step-wise confidence and derives local consensus from subgroups to obtain denser and more reliable rewards. SCRL (Yan et al., 2026) handles unreliable consensus by selectively combining positive and complementary supervision. ETTRL (Liu et al., 2025) uses token entropy to branch rollouts and to reshape advantages, TTVS (Bai et al., 2026) synthesizes semantically equivalent variants of test queries, TTPO (Wang et al., 2026a) treats agreeing and disagreeing rollouts asymmetrically to tolerate wrong pseudo-labels, and Vakada et al. (2026) restrict test-time RL updates to bias parameters. All of these methods improve the reliability of the *textual* vote. PrimeFocus is complementary: it adds a signal about whether the vote is accompanied by visual attention, and can be combined with confidence- or selectivity-based vote corrections. He et al. (2026a) show that unsupervised RL with intrinsic rewards follows a rise-then-fall trajectory, which motivates our stopping rule.

Visual signals as rewards and grounding cues. Several works exploit attention or perception signals in VLMs. OPERA (Huang et al., 2024) penalizes over-trust patterns in attention during decoding, VCD (Leng et al., 2024) contrasts outputs under distorted images, and PAI (Liu et al., 2024) amplifies attention to image tokens at inference time. VADE (Prabhakaran et al., 2025) derives grounding signals from visual attention maps to detect and remove hallucinations, and Min et al. (2026) train perception-centric process reward models that provide token-level rewards for VLM RL. These methods either act at inference time without updating the policy or require annotated data or a trained reward model. Label-free intrinsic rewards have also been used for test-time RL of vision-language-action policies, e.g., trajectory confidence (Chen et al., 2026). Concurrently, He et al. (2026b) measure image and question dependence by teacher-forcing each response under ablated inputs, which needs extra forward passes; our signals come from the log-probability pass GRPO already performs. Unlike these methods, PrimeFocus turns attention statistics into annotation-free rewards inside unlabeled test-time RL, and introduces a group-level term that evaluates the attention shared by responses in the same answer cluster.

3 METHODOLOGY

Figure 2 gives an overview. Section 3.1 fixes the test-time RL objective, Section 3.2 defines the attention-derived rewards, and Section 3.3 defines the stopping rule. Algorithm 1 in Appendix B summarizes the training loop.

3.1 PRELIMINARIES

Test-time RL for VLMs. The VLM is a policy π_θ that generates a response o for a query q (image and text). RL maximizes $\mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)}[R(o,q)]$. Following Shao et al. (2024), GRPO samples a group $\{o_1,\dots,o_G\}$ from $\pi_{\theta_{\text{old}}}$ and maximizes

$$\mathcal{J}(\theta)=\mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G}\min\big(\rho_i(\theta)A_i,\ \text{clip}(\rho_i(\theta),1-\epsilon,1+\epsilon)A_i\big)\Big],\qquad A_i=\frac{r_i-\text{mean}(\{r_j\})}{\text{std}(\{r_j\})+\delta},\ (1)$$

where $\rho_i(\theta)=\pi_\theta(o_i\mid q)/\pi_{\theta_{\text{old}}}(o_i\mid q)$ is the importance ratio, the clipped surrogate follows PPO (Schulman et al., 2017), and δ is a small constant.

Without labels, the reward is built from the policy's own samples. Following TTRV (Singh et al., 2026), we sample N responses per query, extract their final answers $a_1,\dots,a_N$, and group them into

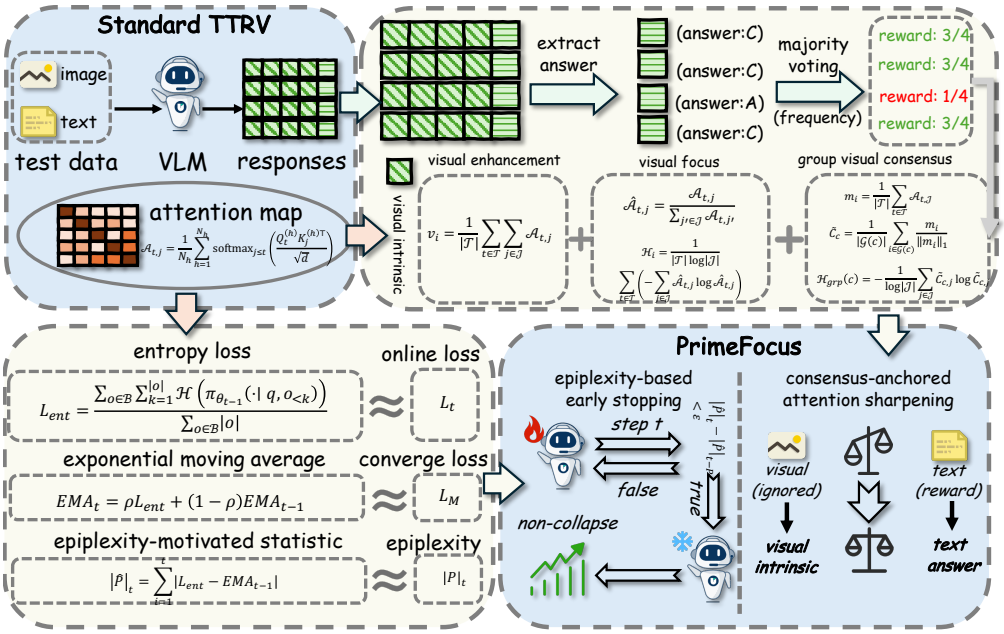

Figure 2: **Overview of PrimeFocus.** The causal response-to-image attention of the final decoder layer is converted into three intrinsic reward terms (visual enhancement, visual focus, group visual consensus) that are added to the TTRV frequency reward. In parallel, the mean token entropy L_t of the policy on its rollouts is compared with its EMA; the accumulated absolute deviations $|\hat{P}|_t$ are monitored, and training stops when their growth over a window falls below a threshold.

answer clusters $\mathcal{G}(c) = \{i \mid a_i = c\}$. Each response receives the *frequency reward*

$$r_i^{\text{freq}} = \frac{|\mathcal{G}(a_i)|}{N} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(a_j = a_i), \tag{2}$$

so responses that agree with the majority answer receive the largest reward (e.g., $3/4$ versus $1/4$ in Figure 2). Rewards are computed from all N votes, and G of the N responses of each query are retained for the update. TTRV additionally subtracts a diversity-control term $\lambda \bar{H}_{\text{out}}$, where $\bar{H}_{\text{out}}$ is the entropy of the empirical answer distribution normalized by the logarithm of the number of distinct answers. This term is identical for all responses to a query and therefore cancels in the group-wise standardization of Eq. (1); it does not affect the policy update. We keep it in the implementation for parity with TTRV but omit it from the notation below. With all three attention coefficients set to zero, PrimeFocus reduces exactly to the TTRV update, which we use as the base configuration in all comparisons.

Epiplexity. Epiplexity (Finzi et al., 2026) separates the information in data that a computationally bounded observer can extract (structure) from what remains unpredictable to it (time-bounded entropy). For a random variable X and the class $\mathcal{P}_T$ of models that run within time T, let $\mathtt{P}^\star = \arg\min_{\mathtt{P}\in\mathcal{P}_T}\{|\mathtt{P}| + \mathbb{E}[\log 1/\mathtt{P}(X)]\}$. The epiplexity is $S_T(X) = |\mathtt{P}^\star|$ and the time-bounded entropy is $H_T(X) = \mathbb{E}[\log 1/\mathtt{P}^\star(X)]$. In practice, S_T is estimated by prequential coding: when a network is trained on a data stream $z_1, z_2, \ldots$, the structural information is approximated by the area between the online loss curve and the final loss,

$$|P| = \sum_{t=1}^{T} \big(L_t(z_t) - L_M(z_t)\big), \tag{3}$$

where L_t is the loss of the model at step t on the incoming example and L_M the loss of the final model.

3.2 CONSENSUS-ANCHORED ATTENTION SHARPENING

Under the frequency reward, the visual pathway receives no direct signal. Contrastive visual-text alignment would require positive and negative pairs, which are unavailable at test time. We therefore derive annotation-free signals from the policy's own attention.

Answer-to-image attention. During the log-probability forward pass of the rollout policy $\pi_{\theta_{\text{old}}}$ on each sampled response (teacher forcing), a forward hook on the final decoder layer captures its query and key states after the rotary position embedding. Let $\mathcal{T}$ be the positions of the generated response tokens, $\mathcal{J}$ the positions of image tokens, and L the sequence length. We use the head-averaged causal attention of this layer,

$$\mathcal{A}_{tj} = \frac{1}{N_h}\sum_{h=1}^{N_h} \text{softmax}_{j\le t}\Big(\frac{Q_t^{(h)} K_j^{(h)\top}}{\sqrt{d}}\Big), \qquad t \in \mathcal{T}, \tag{4}$$

with $\mathcal{A}_{tj} = 0$ for $j > t$, where d is the head dimension. These are exactly the attention weights the decoder applies at position t: later response tokens do not enter the normalization, and since all image tokens precede the response, every response token can attend to the full image. All reward terms are computed without gradients and affect the update only through the scalar reward in Eq. (1).

Visual enhancement. We define the visual share of response i as the fraction of the causal attention of response tokens that falls on image tokens,

$$v_i = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}} \mathcal{A}_{tj}. \tag{5}$$

Each row of $\mathcal{A}$ sums to one, so $v_i \in [0, 1]$. Rewarding v_i with coefficient $\alpha > 0$ counteracts answering from language priors.

Visual focus. v_i alone can be increased by spreading attention uniformly over all patches. We therefore renormalize attention over image tokens and penalize its normalized entropy with coefficient β:

$$\hat{\mathcal{A}}_{tj} = \frac{\mathcal{A}_{tj}}{\sum_{j'\in\mathcal{J}} \mathcal{A}_{tj'}}, \qquad \mathcal{H}_i = \frac{1}{|\mathcal{T}|\log|\mathcal{J}|}\sum_{t\in\mathcal{T}}\Big(-\sum_{j\in\mathcal{J}} \hat{\mathcal{A}}_{tj}\log\hat{\mathcal{A}}_{tj}\Big) \in [0,1]. \tag{6}$$

The two terms constrain each other: enhancement without focus admits uniform smearing, whereas focus without enhancement admits collapse onto a single, possibly irrelevant, patch while the image receives little attention overall. Section 4.3 measures both failure modes.

Group visual consensus. Majority voting rewards agreement among textual answers regardless of whether the agreeing responses attend to common visual evidence, so a confidently wrong majority is amplified. For each answer cluster $\mathcal{G}(c)$ we average the normalized per-response image attention maps with equal weights and measure the normalized entropy of the average:

$$m_i = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathcal{A}_{t,\mathcal{J}}, \qquad \tilde{C}_c = \frac{1}{|\mathcal{G}(c)|}\sum_{i\in\mathcal{G}(c)} \frac{m_i}{\|m_i\|_1}, \qquad \mathcal{H}_{\text{grp}}(c) = -\frac{1}{\log|\mathcal{J}|}\sum_{j\in\mathcal{J}} \tilde{C}_{cj}\log\tilde{C}_{cj}. \tag{7}$$

Normalizing each map first gives every member the same weight, so the statistic reflects where the members attend rather than how much attention any single response places on the image. If the members attend to different regions, the average is diffuse and $\mathcal{H}_{\text{grp}}$ is high; if they attend to the same regions, it stays concentrated. Clusters are formed over all N votes. Consensus is undefined for a singleton, so a singleton cluster is assigned the mean $\mathcal{H}_{\text{grp}}$ of the multi-member clusters of the same query, a neutral value that neither rewards nor penalizes it relative to them (the term is dropped if no multi-member cluster exists). Dividing by $\log|\mathcal{J}|$ makes both entropies lie in $[0, 1]$ independently of the number of image tokens, which varies with dynamic tiling.

Total reward. The complete reward of response i is

$$r_i = r_i^{\text{freq}} + \alpha v_i - \beta\mathcal{H}_i - \gamma\,\frac{|\mathcal{G}(a_i)|}{N}\,\mathcal{H}_{\text{grp}}(a_i), \tag{8}$$

and enters GRPO through Eq. (1). Setting $\alpha = \beta = \gamma = 0$ recovers the TTRV update. Since $|\mathcal{G}(a_i)|/N = r_i^{\text{freq}}$, the first and last terms equal $r_i^{\text{freq}}(1 - \gamma\mathcal{H}_{\text{grp}}(a_i))$: visual consensus rescales the frequency reward by a factor in $[1-\gamma, 1]$, so a rare answer cannot overtake a clear majority through the group term, while clusters of similar size are ranked by the concentration of their shared attention.

3.3 Entropy-Plateau Early Stopping

Without a validation set, existing methods train every test set for the same number of steps. Since intrinsic-reward RL tends to rise and then fall (He et al., 2026a), we need a label-free signal of when further updates stop being useful. We first define the statistic operationally and then discuss its relation to epiplexity.

Statistic. Let $\mathcal{B}_t$ be the rollout batch at step t. We compute the mean token entropy of the policy on its own rollouts,

$$L_t = \frac{\sum_{o\in\mathcal{B}_t}\sum_{k=1}^{|o|} H\big(\pi_{\theta_{t-1}}(\cdot \mid q, o_{<k})\big)}{\sum_{o\in\mathcal{B}_t}|o|}, \qquad H(p) = -\sum_v p_v \log p_v, \tag{9}$$

maintain its EMA, and accumulate the absolute deviation of each new value from the EMA of the previous values:

$$\text{EMA}_t = \rho L_t + (1-\rho)\text{EMA}_{t-1}, \qquad \delta_t = \big|L_t - \text{EMA}_{t-1}\big|, \qquad |\hat{P}|_t = \sum_{i=1}^{t}\delta_i, \tag{10}$$

with $\text{EMA}_0 = L_1$, so that $\delta_1 = 0$ and $|\hat{P}|_1 = 0$, and $\rho \in (0, 1]$. A plateau counter c_p is incremented when the growth $|\hat{P}|_t - |\hat{P}|_{t-P}$ over a window of P steps falls below a threshold ε and is decremented (floored at zero) otherwise; training stops when c_p reaches P. Intuitively, $|\hat{P}|_t$ stops growing once the policy's output distribution stops changing.

Relation to prequential epiplexity. Eq. (10) has the form of Eq. (3) with two substitutions. The online loss is replaced by L_t: on its own rollouts, the expected negative log-likelihood of the sampling policy equals its mean token entropy, so L_t is the prequential loss of the policy on the self-generated stream. The final loss L_M, which is unknown during training, is replaced by the EMA. The following approximation argument supports the second substitution under idealized assumptions.

Proposition 1. *Consider a sequence $(L_t)_{t\geq 1}$ with limit L_M such that (A1) $|L_t - L_M| \leq c\,\eta^t$ for some $c > 0$, $0 < \eta < 1$, and (A2) $L_t \geq L_M$ for all t. Let $|P|_T = \sum_{t=1}^{T}(L_t - L_M)$ and $|\hat{P}|_T$ as in Eq. (10). Then for every $\kappa \in (\max(1-\rho, \eta), 1)$ there is a constant c' depending only on c, η, ρ, κ such that $\big||\hat{P}|_T - |P|_T\big| \leq c'/(1-\kappa)$ for all T.*

The proof (Appendix C) unrolls the EMA recursion and bounds $|\text{EMA}_t - L_M| \leq c'\kappa^t$. Since a constant offset alone does not align plateau steps, Corollary 1 bounds the quantity the rule actually tests: the window growths of $|\hat{P}|$ and $|P|$ differ by at most $e_t = c'\kappa^{t-P}/(1-\kappa)$, so beyond a finite step the detector fires if the prequential-style sum grows by less than $\varepsilon - e_t$ and only if it grows by less than $\varepsilon + e_t$. This is an idealized argument under (A1)–(A2), not a guarantee for non-stationary GRPO dynamics.

Scope. The proposition concerns the entropy stream, not the test set. Because L_t is measured on responses sampled from the current policy, it decreases whenever generation becomes sharper, including when the policy becomes confidently wrong. $|\hat{P}|$ is therefore a smoothed output-entropy stabilization signal that indicates when the policy has stopped changing, rather than an estimator of the epiplexity of the test set. The epiplexity view motivates its cumulative-residual form, which reacts to the total amount of change rather than to the instantaneous entropy level; Section 4.3 evaluates the rule empirically.

Model	DTD	ImageNet-R	ImageNet-A	MathVerse	MathVista	RealWorldQA	AI2D	Mean
Pretrained models without test-time adaptation (reference only)								
LLaMA-3.2-11B	87.54	72.26	83.07	19.36	35.40	41.18	59.54	56.91
LLaVA-1.5-7B	76.68	96.46	94.39	26.02	34.29	59.87	47.49	62.17
Phi-3.5-vision	88.26	96.05	85.78	36.02	51.22	57.25	75.55	70.02
InternVL3-2B								
Base	37.24	66.01	67.92	44.10	58.26	63.47	39.68	53.81
TTRV	89.73	**96.89**	96.31	48.51	66.11	64.29	67.75	75.66
PrimeFocus	**93.85**	96.02	**97.88**	**50.86**	**67.14**	**66.83**	**72.91**	**77.93**
Δ vs. TTRV	+4.12	−0.87	+1.57	+2.35	+1.03	+2.54	+5.16	+2.27
InternVL2.5-4B								
Base	47.33	79.53	90.67	51.69	65.49	65.25	51.55	64.50
TTRV	81.98	88.21	96.00	53.02	66.94	66.00	61.09	73.32
PrimeFocus	**84.62**	**90.63**	**96.41**	**56.35**	**68.21**	**69.52**	**64.47**	**75.74**
Δ vs. TTRV	+2.64	+2.42	+0.41	+3.33	+1.27	+3.52	+3.38	+2.42

Table 1: Top-1 accuracy (%) on seven benchmarks. TTRV results are reported from Singh et al. (2026) (fixed schedule); PrimeFocus results are averaged over three seeds (standard deviation below 1.58 points for every entry) and use early stopping. Bold marks the better of TTRV and PrimeFocus.

4 Experiments

4.1 Experimental Setup

Benchmarks and baselines. We use seven benchmarks: DTD (Cimpoi et al., 2013), ImageNet-R (Hendrycks et al., 2020) and ImageNet-A (Hendrycks et al., 2019) for visual recognition, and AI2D (Kembhavi et al., 2016), MathVerse (Zhang et al., 2024a), RealWorldQA (xAI, 2024) and MathVista (Lu et al., 2023) for VQA. Following the protocol of TTRV, the model adapts on twenty unlabeled samples per dataset and is evaluated on the full test split. Our main comparison is TTRV (Singh et al., 2026); we follow its backbones and adaptation protocol and report its published results. For context we also list LLaMA-3.2-11B (Grattafiori et al., 2024), LLaVA-1.5-7B (Liu et al., 2023) and Phi-3.5-vision (Abdin et al., 2024) as pretrained models without test-time adaptation.

Models and metrics. We use InternVL3-2B (Zhu et al., 2025) and InternVL2.5-4B (Chen et al., 2024) and report top-1 accuracy of PrimeFocus averaged over three runs with different data-shuffling seeds. All runs use one node with eight NVIDIA RTX 4090 GPUs.

Hyperparameters. We sample $N = 32$ responses per query for voting and retain $G = 16$ for the policy update. The learning rate is 5×10^{-7}, and the maximum training budget is three epochs. The reward coefficients are $\alpha = 0.4$, $\beta = 0.3$ and $\gamma = 0.3$. For stopping, $\rho = 0.05$, $P = 5$ and $\varepsilon = 10^{-4}$, and L_t is the token-mean entropy of the rollout policy. Attention is taken from the final decoder layer and averaged over heads. PrimeFocus uses early stopping, whereas the TTRV results follow its original fixed schedule. The same hyperparameters are used throughout (Appendix A).

4.2 Main Results

Table 1 compares PrimeFocus with TTRV under the same backbone and adaptation protocol. PrimeFocus improves the mean accuracy from 75.66 to 77.93 on InternVL3-2B and from 73.32 to 75.74 on InternVL2.5-4B. It improves on six of seven benchmarks for InternVL3-2B, the exception being ImageNet-R (−0.87), where TTRV is already close to saturation, and on all seven for InternVL2.5-4B. The largest gains occur on benchmarks whose answers depend on reading the image, such as AI2D (+5.16 and +3.38) and RealWorldQA (+2.54 and +3.52), while the gain on the nearly saturated ImageNet-A is smaller (+1.57 and +0.41).

Configuration	DTD	MathVerse	AI2D
TTRV (fixed schedule)	89.73	48.51	67.75
TTRV + early stop	90.20	48.93	69.18
+ enhancement (α)	90.46	49.37	69.24
+ focus (β)	90.28	48.90	68.83
+ group consensus (γ)	91.15	50.84	71.66
full PrimeFocus	**93.85**	**50.86**	**72.91**

Table 2: Component ablation on InternVL3-2B. The first row is TTRV with its fixed schedule (Table 1); all other rows use early stopping, and each term is added to TTRV + early stop.

v_i quintile	Q1	Q2	Q3	Q4	Q5
Acc. (%)	24.6	32.4	39.5	46.9	55.0

Table 3: Accuracy by visual-share quintile on AI2D with the base InternVL3-2B.

Component	Time (s)	Rel.	Mem. (GB)
TTRV	26.50	1.00×	16.79
+ attn. capture	28.34	1.07×	18.60
+ rewards	29.80	1.12×	18.60

Table 4: Cost per training step on InternVL3-2B.

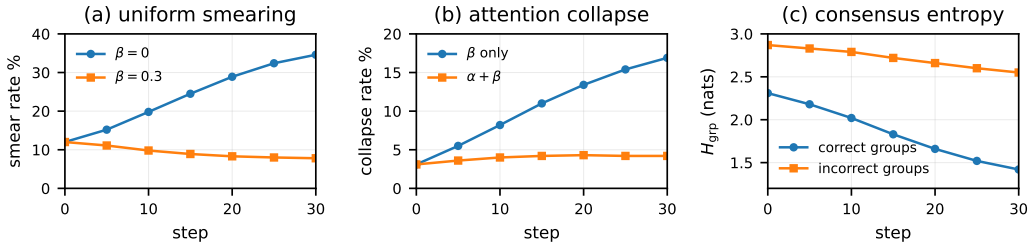

Figure 3: Attention failure-mode diagnostics on AI2D with InternVL3-2B. (a) Smear rate with and without the focus penalty. (b) Collapse rate with the focus penalty alone and with visual enhancement. (c) Mean consensus entropy of the majority cluster (in nats, i.e., $\log|\mathcal{J}| \cdot \mathcal{H}_{\text{grp}}$) for queries whose majority answer is correct or incorrect. Rates are defined in Appendix D.

4.3 Ablation and Analysis

Components. On InternVL3-2B (Table 2), our stopping rule improves TTRV by 0.47, 0.42 and 1.43 points, and the attention rewards add a further 3.65, 1.93 and 3.73 points. Among the individual terms, group visual consensus gives the largest gain (+2.48 on AI2D), visual enhancement improves on all three datasets, and visual focus alone is roughly neutral ($+0.08$, -0.03, -0.35), consistent with its role as a constraint against the smearing induced by enhancement (Figure 3a). The full combination is best on all three datasets.

Attention failure modes. Figure 3 tracks the two failure modes of Section 3.2 on AI2D with InternVL3-2B; exact definitions of the rates are given in Appendix D. Without the focus penalty ($\beta = 0$), the smear rate (high visual share but near-uniform attention) rises from 12% to about 35%, whereas with $\beta = 0.3$ it decreases to about 8%. With the focus penalty alone, the collapse rate (attention on a single patch) rises to about 17%; adding visual enhancement keeps it near 4%. Panel (c) tracks $\mathcal{H}_{\text{grp}}$ of the majority cluster, split by whether the majority answer is correct (labels are used only for this analysis): it decreases much more for correct than for incorrect majorities.

Early stopping. Figure 4 shows one training trajectory on AI2D with InternVL3-2B. $|\hat{P}|$ saturates before accuracy peaks, and the plateau counter stops training at step 27, three steps after the oracle peak at step 24 and within 0.1 points of its accuracy. Training to the end of the schedule ends at 71.03%, below the stopped checkpoint. Table 5 repeats the analysis on RealWorldQA and MathVerse, where the fixed schedule is considerably longer: the rule stops after 68 and 40 steps instead of 120 and 70, and the stopped checkpoints reach 66.25 and 50.93 compared with 63.10 and 49.78 at the end of the fixed schedule. In all three cases the rule avoids late degradation with fewer steps.

Do attention signals carry information about correctness? Before any training, on AI2D with the base InternVL3-2B, the visual share v_i of a response correlates with its correctness (Spearman $\rho = 0.38$, $p < 10^{-4}$), and accuracy increases monotonically from 24.6% in the lowest v_i quintile

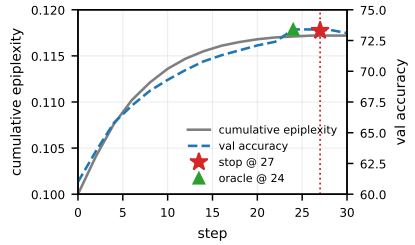

Figure 4: Stopping statistic and accuracy on AI2D with InternVL3-2B (single seed). For display, $|\hat{P}|_t$ is shifted by a constant offset; the rule depends only on its growth.

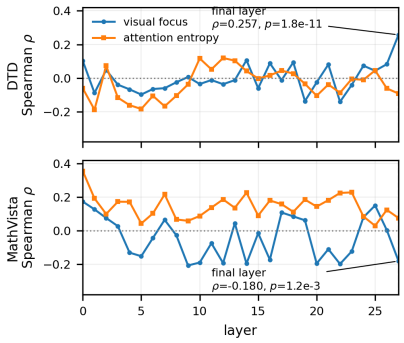

Figure 5: Layer-wise Spearman correlation of the visual share v_i ("visual focus" in the legend) and the attention entropy $\mathcal{H}_i$ with correctness on the base InternVL3-2B.

Dataset	Schedule	Step	Acc.	Oracle step	Oracle acc.
RealWorldQA	early stop	68	**66.25**	62	68.47
	fixed	120	63.10	65	69.23
MathVerse	early stop	40	**50.93**	23	51.12
	fixed	70	49.78	20	50.14

Table 5: Early stopping versus the fixed schedule on InternVL3-2B (single seed; TTRV reward with group visual consensus). "Step" is the stopping step or the last step of the fixed schedule; oracle values use test labels for analysis only.

Dataset	Q1	Q2	Q3	Q4	Q5	Spearman ρ
DTD	31.6	11.3	47.8	48.9	38.8	0.16
SEED	26.3	38.8	39.6	27.8	41.5	0.05
MathVista	65.7	58.7	48.9	62.1	44.4	−0.10
MathVerse	49.2	50.0	50.0	50.8	39.4	−0.07

Table 6: Accuracy (%) by visual-share quintile and Spearman correlation of the visual share with correctness on further datasets (base InternVL3-2B).

to 55.0% in the highest (Table 3). The partial correlation given output entropy is 0.31, so attention adds information beyond output uncertainty. Across further datasets (Table 6 and Appendix E), the attention entropy is negatively correlated with correctness on every dataset, while the relation of the visual share is task-dependent. This supports pairing the visual share with the focus penalty.

Choice of attention layer. Figure 5 correlates the attention signals of every layer with correctness. On DTD, the visual-share correlation is concentrated in the final layer, while other layers stay below $|\rho| = 0.15$; on the reasoning-dominant MathVista, no layer gives a positive signal (Appendix E). Although representations align best in middle layers (Appendix F), the predictive attention signal on visually dependent tasks resides in the final layer, which motivates our choice.

Computational cost. The attention rewards add about 12% step time and 1.81 GB peak memory over TTRV (Table 4), while early stopping reduces the number of steps (Table 5).

5 Limitations

The attention terms are attention-based proxies for visual reliance rather than verified grounding; the stopping rule detects entropy stabilization and cannot distinguish a correct from a confidently wrong policy; the group term may under-reward multi-region evidence; and we evaluate only InternVL backbones.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. Details are described in the paper.

ETHICS STATEMENT

This work adapts publicly available vision-language models on public benchmarks. It involves no human subjects or private data. Test-time adaptation can reinforce a model's own errors when its consensus is wrong; the attention terms and the stopping rule reduce but do not remove this risk, and adapted models should be validated before deployment in high-stakes settings.

REPRODUCIBILITY STATEMENT

Section 3 specifies the rewards and the stopping rule, and Algorithm 1 in Appendix B gives the full training loop. All hyperparameters are listed in Section 4.1 and Table 7, and are identical across datasets and backbones. Appendix C contains the proof of Proposition 1 with all assumptions, and Appendix D defines the diagnostic metrics. Anonymized code will be provided as supplementary material.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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

Sikai Bai, Haoxi Li, Jie Zhang, Yongjiang Liu, and Song Guo. Ttvs: Boosting self-exploring reinforcement learning via test-time variational synthesis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32651–32665, 2026.

Siyao Chen, Jiakang Yuan, Jiaxin Wang, and Tao Chen. Trust your instincts: Confidence-driven test-time RL for vision-language-action models. *arXiv preprint arXiv:2606.29892*, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2013. URL https://api.semanticscholar.org/CorpusID:4309276.

Lang Feng, Weihao Tan, Zhiyi Lyu, Longtao Zheng, Haiyang Xu, Mingshi Yan, Fei Huang, and Bo An. Towards efficient online tuning of vlm agents via counterfactual soft reinforcement learning. *ArXiv*, abs/2505.03792, 2025. URL https://api.semanticscholar.org/CorpusID:278368088.

Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J. Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *ArXiv*, abs/2601.03220, 2026. URL https://api.semanticscholar.org/CorpusID:284512934.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, et al. How far can unsupervised rlvr scale llm training? In *The Fourteenth International Conference on Learning Representations*, 2026a.

Chiyuan He, Zihuan Qiu, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Desclip: Robust continual learning via general attribute descriptions for vlm-based visual recognition. *IEEE Transactions on Multimedia*, 28:5021–5035, 2025. URL https://api.semanticscholar.org/CorpusID:276094247.

Xinrui He, Ting-Wei Li, Junting Wang, Mengting Ai, Xinyu He, Hanghang Tong, and Jingrui He. Harnessing image question dependence for better VLM test-time reinforcement learning. *arXiv preprint arXiv:2609.13296*, 2026b.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15257–15266, 2019. URL https://api.semanticscholar.org/CorpusID:196831327.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2020. URL https://api.semanticscholar.org/CorpusID:220250257.

Zhi Hu, Liang Liao, and Weisi Lin. Internvl-vpr: Hierarchical zero-shot visual place recognition with vlm-driven re-ranking. *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2412–2417, 2025. URL https://api.semanticscholar.org/CorpusID:283355607.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:273901553.

Mingi Jung, Saehuyng Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. *ArXiv*, abs/2502.01419, 2025. URL https://api.semanticscholar.org/CorpusID:276107107.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. URL https://api.semanticscholar.org/CorpusID:2682274.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. URL https://api.semanticscholar.org/CorpusID:277740982.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. URL https://api.semanticscholar.org/CorpusID:258179774.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:264491155.

Yinbin Lu and Alan Wang. Integrating language into medical visual recognition and reasoning: A survey. *Medical image analysis*, 102:103514, 2025. URL https://api.semanticscholar.org/CorpusID:276682037.

Tiange Lyu. A prompt that can stimulate the self-correction ability of the vlm, improving the performance on vqa task. *Applied and Computational Engineering*, 2025. URL https://api.semanticscholar.org/CorpusID:277485194.

Weiyu Ma, Yongcheng Zeng, Yan Song, Xinyu Cui, Jianzhe Zhao, Xu-Hui Liu, and Mohamed Elhoseiny. Freshness-aware prioritized experience replay for llm/vlm reinforcement learning. *ArXiv*, abs/2604.16918, 2026. URL https://api.semanticscholar.org/CorpusID:287618079.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33099–33109, 2026.

Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. VADE: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14949–14965, 2025.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haolin Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents. *ArXiv*, abs/2501.12326, 2025. URL https://api.semanticscholar.org/CorpusID:275788034.

Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R. Zaiane, and Matthew E. Taylor. The evolving landscape of llm- and vlm-integrated reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2024. URL https://api.semanticscholar.org/CorpusID:276558335.

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

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hanna Hajishirzi, Pang Wei Koh, and Luke S. Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr. *ArXiv*, abs/2506.10947, 2025. URL https://api.semanticscholar.org/CorpusID:279318769.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33153–33163, 2026.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *ArXiv*, abs/2506.15220, 2025. URL https://api.semanticscholar.org/CorpusID:279447296.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36:1355–1376, 2023. URL https://api.semanticscholar.org/CorpusID:266690572.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

Naveen Vakada, Mingyuan Li, and Shaoxiong Ji. Label-free steering: Compressing test-time reinforcement learning into bias-only subspaces. *arXiv preprint arXiv:2609.18587*, 2026.

Aozhe Wang, Zhengxi Lu, Jianze Wang, Shangke Lv, Ying Liu, Weiming Lu, Jun Xiao, Yueting Zhuang, Hua Yang, Qianglong Chen, and Yongliang Shen. TTPO: Test-time policy optimization. *arXiv preprint arXiv:2608.27448*, 2026a.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37251–37265, 2026b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:267499679.

Zihan Weng, Lucas Gomez, Taylor Webb, and Pouya Bashivan. Caption this, reason that: Vlms caught in the middle. *ArXiv*, abs/2505.21538, 2025. URL https://api.semanticscholar.org/CorpusID:278959992.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, 2024.

Dong Yan, Jian Liang, Yanbo Wang, Shuo Lu, Ran He, and Tieniu Tan. What if consensus lies? selective-complementary reinforcement learning at test time. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28957–28970, 2026.

Tianyu Yu, Bo Ji, Shouli Wang, Shunyu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Rlpr: Extrapolating rlvr to general domains without verifiers. *ArXiv*, abs/2506.18254, 2025. URL https://api.semanticscholar.org/CorpusID:279999943.

Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:4496–4506, 2023. URL https://api.semanticscholar.org/CorpusID:258812691.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 2024a. URL https://api.semanticscholar.org/CorpusID:268554279.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *ArXiv*, abs/2502.19655, 2025. URL https://api.semanticscholar.org/CorpusID:276647659.

Yunzuo Zhang, Yameng Liu, Weili Kang, and Ran Tao. Vss-net: Visual semantic self-mining network for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:2775–2788, 2024b. URL https://api.semanticscholar.org/CorpusID:261573135.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *International Conference on Learning Representations*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

A Implementation Details

Table 7 lists the full configuration of PrimeFocus. The same values are used for all datasets and both backbones, and TTRV uses the same settings with $\alpha = \beta = \gamma = 0$.

Rollouts and answer extraction. Responses are sampled with temperature 1.0 and top-$p = 1.0$, with at most 1024 new tokens. For each query, all $N = 32$ responses are used for voting and for the answer clusters of Eq. (7); the first $G = 16$ responses in sampling order, i.e., a uniformly random subset, are retained for the policy update, and Eq. (1) standardizes rewards over these G responses. Final answers are extracted with the rule-based answer parser of TTRL (Zuo et al., 2026) and verified with its rule-based matcher.

Optimization. We update the language model and the vision-language projector with full-parameter training under FSDP; the vision encoder is frozen. Each step uses eight prompts, one PPO epoch, clipping $\epsilon = 0.2$, $\delta = 10^{-6}$ in Eq. (1), and no KL or entropy regularization. We use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01) with a cosine schedule and 3% linear warm-up.

B Algorithm

C Proof of Proposition 1

Proof. Unrolling the recursion $\mathrm{EMA}_t = \rho L_t + (1-\rho)\mathrm{EMA}_{t-1}$ from $\mathrm{EMA}_0 = L_1$ gives, for $t \geq 1$,

$$\mathrm{EMA}_t - L_M = \rho \sum_{i=1}^{t} (1-\rho)^{t-i} \big(L_i - L_M\big) + (1-\rho)^t \big(\mathrm{EMA}_0 - L_M\big). \tag{11}$$

Hyperparameter	Value
Responses per query for voting N	32
Responses retained for the update G	16 (first G in sampling order)
Sampling temperature / top-p / max. new tokens	1.0 / 1.0 / 1024
Prompts per step / PPO epochs	8 / 1
Clipping ϵ / δ in Eq. (1)	0.2 / 10^{-6}
KL / entropy regularization	0 / 0
Optimizer	AdamW, cosine, 3% warm-up
Trainable parameters	LLM + projector (vision encoder frozen)
Learning rate	5×10^{-7}
Maximum training budget	3 epochs
Visual enhancement coefficient α	0.4
Visual focus coefficient β	0.3
Group consensus coefficient γ	0.3
TTRV diversity coefficient λ (no effect)	0.75
EMA rate ρ	0.05
Plateau window P	5
Plateau threshold ε	10^{-4}
Attention	final layer, causal, post-RoPE, head-avg.
Stopping signal L_t	token-mean rollout entropy
Hardware	8× NVIDIA RTX 4090

Table 7: Hyperparameters of PrimeFocus.

Algorithm 1 PrimeFocus

Require: unlabeled test samples $\mathcal{D}$, policy π_θ, votes N, retained responses G, coefficients α, β, γ, EMA rate ρ, window P, threshold ε

$|\hat{P}|_0 \leftarrow 0$, $c_p \leftarrow 0$
for each training step $t = 1, 2, \ldots$ until the maximum budget **do**
  Sample N responses $\{o_i\}$ from $\pi_{\theta_{\text{old}}}(\cdot \mid q)$ for each query q in the batch; extract answers $\{a_i\}$
  Compute r_i^{freq} (Eq. 2)
  In the log-probability forward pass, capture Q, K of the final layer and compute $\mathcal{A}$ (Eq. 4)
  Compute v_i, $\mathcal{H}_i$ and $\mathcal{H}_{\text{grp}}(a_i)$ (Eqs. 5–7) without gradients
  Form r_i (Eq. 8) for the G retained responses, the group-wise advantages, and update θ with Eq. (1)
  Compute L_t (Eq. 9); if $t = 1$ set $\text{EMA}_0 \leftarrow L_1$; update δ_t, $|\hat{P}|_t$, EMA_t (Eq. 10)
  if $t > P$ and $|\hat{P}|_t - |\hat{P}|_{t-P} < \varepsilon$ **then**
    $c_p \leftarrow c_p + 1$
  else
    $c_p \leftarrow \max(c_p - 1, 0)$
  end if
  if $c_p = P$ **then**
    break
  end if
end for
return adapted policy π_θ

(For $t = 1$ the right-hand side equals $\rho(L_1 - L_M) + (1-\rho)(L_1 - L_M) = L_1 - L_M$, as required.) Let $\mu = \max(1-\rho, \eta) < \kappa < 1$. By (A1), the first term is bounded by $\rho c \sum_{i=1}^{t} \mu^{t-i}\mu^i = \rho c\, t\, \mu^t$, and the second by $(1-\rho)^t c\, \eta \le c\, \mu^t$. Hence

$$|\text{EMA}_t - L_M| \le c\,(\rho t + 1)\,\mu^t \le c'\kappa^t, \qquad c' = c \sup_{t \ge 0} (\rho t + 1)(\mu/\kappa)^t < \infty, \tag{12}$$

which also holds for $t = 0$ because $|\text{EMA}_0 - L_M| = |L_1 - L_M| \le c\eta \le c'$. By (A2), $L_t - L_M = |L_t - L_M|$, and by the reverse triangle inequality $\big||L_t - \text{EMA}_{t-1}| - |L_t - L_M|\big| \le |\text{EMA}_{t-1} - L_M|$.

Summing over t,

$$\big||\hat{P}|_T - |P|_T\big| \le \sum_{t=1}^{T} |\mathrm{EMA}_{t-1} - L_M| \le \sum_{t=1}^{T} c'\kappa^{t-1} \le \frac{c'}{1-\kappa}. \tag{13}$$

□

Corollary 1. *Under the assumptions of Proposition 1, let $\Delta_t = |\hat{P}|_t - |\hat{P}|_{t-P}$ and $W_t = |P|_t - |P|_{t-P}$ be the growths over the last P steps, and $e_t = c'\kappa^{t-P}/(1-\kappa)$. Then for all $t > P$, $|\Delta_t - W_t| \le e_t$. Consequently, $W_t < \varepsilon - e_t$ implies $\Delta_t < \varepsilon$, and $\Delta_t < \varepsilon$ implies $W_t < \varepsilon + e_t$. For any margin $\xi > 0$, $e_t \le \xi$ for all $t \ge P + \log\big(c'/((1-\kappa)\xi)\big)/\log(1/\kappa)$, after which the plateau tests on $|\hat{P}|$ and on $|P|$ can disagree only when W_t lies within ξ of ε.*

Proof. By (A2) and the reverse triangle inequality as above, $|\Delta_t - W_t| = \big|\sum_{s=t-P+1}^{t} \big(|L_s - \mathrm{EMA}_{s-1}| - |L_s - L_M|\big)\big| \le \sum_{s=t-P+1}^{t} |\mathrm{EMA}_{s-1} - L_M| \le \sum_{u=t-P}^{t-1} c'\kappa^u \le \frac{c'\kappa^{t-P}}{1-\kappa} = e_t$. The two implications follow directly, and $e_t \le \xi$ is equivalent to the stated condition on t. □

Remark. Assumption (A2) is needed because $|\hat{P}|$ accumulates absolute residuals whereas Eq. (3) accumulates signed differences; without it the two sums can differ by an amount that grows with T. (A1) is a geometric-convergence assumption on the entropy trajectory. The proposition relates $|\hat{P}|$ to a prequential-style sum on the entropy stream; it makes no claim about the structural information of the test set.

D Definitions of the Diagnostic Metrics

For Figure 3, let $\mathcal{H}_i \in [0,1]$ be the normalized attention entropy of response i (Eq. 6) and $\hat{m}_i = m_i/\|m_i\|_1$ its normalized image attention map.

- **Smear rate**: fraction of responses with $v_i > 0.5$ and $\mathcal{H}_i > 0.9$, i.e., most of the attention on the image but spread almost uniformly over patches.
- **Collapse rate**: fraction of responses with $\max_j \hat{m}_{ij} > 0.5$, i.e., more than half of the image attention on a single patch.
- **Consensus entropy** (panel c): for each query, the unnormalized entropy $\log|\mathcal{J}| \cdot \mathcal{H}_{\mathrm{grp}}$ of the majority cluster, averaged separately over queries whose majority answer is correct and incorrect according to the ground-truth label; labels are used only for this analysis.

Rates are computed over the N votes of each query and averaged over the queries of a training step.

E Attention–Accuracy Correlation

We examine whether the attention signals carry information about correctness *before* any training. We sample $K = 8$ responses per prompt from the base InternVL3-2B at temperature 1.0 on 100 test prompts per dataset, and compute the visual share, the attention entropy, the output entropy and the correctness of every response. Table 6 reports accuracy by visual-share quintile and the Spearman correlation with correctness.

The visual share correlates with correctness in a task-dependent way. The correlation is positive on DTD ($\rho = 0.16$, $p = 3.4 \times 10^{-5}$) and remains significant after controlling for output entropy (partial $\rho = 0.10$, $p = 0.008$). It vanishes on SEED and MathVerse and is negative on the reasoning-dominant MathVista ($\rho = -0.10$, $p = 0.008$). We attribute this to a difficulty confound: on reasoning benchmarks the figure mainly supplies the problem statement, so a high attention share concentrates on prompts the policy fails to solve, while easy prompts are answered with little visual attention. The attention entropy behaves consistently and correlates negatively with correctness on every dataset (-0.14 on DTD, -0.24 on SEED, -0.18 on MathVista, all $p < 10^{-5}$). Focused attention therefore predicts correct responses across task types, whereas the visual share alone is not a universally safe signal and must be paired with the focus penalty. This is consistent with group visual consensus being the strongest single component in Table 2 and with the gains of PrimeFocus concentrating on visually dependent benchmarks.

F Layer-wise Representation and Attention Analysis

The rewards use the final decoder layer and average over heads. To examine this choice, we probe the base InternVL3-2B before training on 20 test samples each from DTD (visually dependent) and MathVista (reasoning-dominant). We mean-pool image-token and text-token hidden states at every layer and measure cross-modal alignment with linear and RBF-kernel CKA, 10-nearest-neighbour Jaccard overlap, text-to-image retrieval MRR, and the effective rank of each modality. Figure 6 shows the full trajectories and Table 8 reports representative layers.

Cross-modal alignment of representations is a middle-layer phenomenon. On DTD, linear CKA peaks at layer 20 (0.649) and drops to 0.269 at the final layer, with k-NN overlap following the same trend, while RBF CKA stays near 0.97 at all depths; the decay coincides with a drop in the effective rank of image representations. Alignment exists only at the sample level: token-level linear CKA is about 0.005 on both datasets, and the mean per-sample cosine similarity at the final layer is only 0.23 (DTD) and 0.13 (MathVista). Alignment is also task-dependent: on MathVista, linear CKA never exceeds 0.43 and peaks at layer 3, while k-NN overlap stays high, suggesting that visual content is integrated early and then dominated by text-driven reasoning.

Figure 5 in the main text recomputes the answer-to-image attention of every layer with the same reward definitions and correlates it with correctness. On DTD, the visual-share signal is concentrated in the final layer, while all other layers stay below $|\rho| = 0.15$ with inconsistent signs; attention entropy correlates negatively in most layers. On MathVista the pattern reverses: the visual share correlates negatively with correctness at most depths and attention entropy positively. No layer offers a usable positive visual-share signal on MathVista, consistent with the difficulty confound in Appendix E. Thus, although representations align best in middle layers, the attention signal that predicts correctness on visually dependent tasks resides in the final layer, which motivates our choice. As a lightweight probe on 20 samples per dataset, this scan complements the larger analysis in Appendix E.

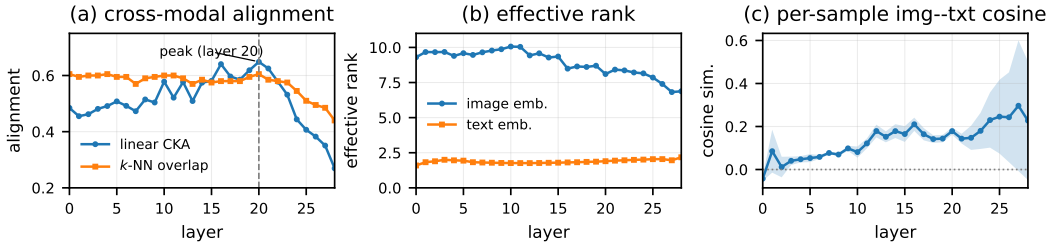

Figure 6: Layer-wise cross-modal alignment of the base InternVL3-2B on 20 test samples each from DTD (top) and MathVista (bottom), measured between mean-pooled image and text hidden states.

Dataset	Layer	Linear CKA	k-NN overlap	t2i MRR	Img. rank	Txt. rank
DTD	0 (embedding)	0.485	0.605	0.213	9.30	1.59
	7	0.473	0.570	0.179	9.65	1.80
	14	0.574	0.585	0.180	9.28	1.78
	20	**0.649**	**0.605**	0.180	8.10	1.90
	24	0.444	0.545	0.163	8.15	2.01
	27	0.351	0.485	0.329	6.82	1.96
	28 (final)	0.269	0.440	0.205	6.88	2.18
MathVista	0 (embedding)	0.102	0.720	0.184	7.35	1.23
	3	**0.429**	0.655	0.186	8.36	1.78
	14	0.242	0.620	0.180	8.48	1.58
	23	0.360	0.675	0.181	7.61	1.91
	28 (final)	0.180	0.685	0.181	6.71	1.74

Table 8: Layer-wise cross-modal alignment of the base InternVL3-2B at representative layers.