VATTNPO: VISUAL ATTENTION-GUIDED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reinforcement learning with verifiable rewards commonly assigns one sequence advantage to an entire response. This outcome-level signal distinguishes overall response quality but leaves changes in visual-evidence use across generated tokens implicit. We introduce **VAttnPO** (Visual Attention-Guided Policy Optimization), which uses answer-conditioned visual attention to refine token-level credit while preserving the verifier-derived trajectory-mean advantage. A frozen old policy constructs a visual reference by teacher forcing the known correct answer and produces a visual distribution for each rollout token. VAttnPO measures their agreement in visual-attention space using Jensen–Shannon divergence, rather than reading guidance from teacher output-token distributions or introducing an attention-matching objective. Trajectory centering, relative-position weighting, and credit balancing turn this agreement into a zero-mean, position-aware correction to the GRPO sequence advantage. Token intervention, visual-index permutation, and trajectory analysis provide diagnostic evidence linking this signal to model behavior, visual correspondence, and generation position. On Qwen3-VL-8B-Instruct, VAttnPO raises six-task image AVE from 70.97 with GRPO to 73.50 and scores higher on every reported structured, video, and MME metric. Gains over GRPO also hold on Qwen2.5-VL-7B-Instruct and InternVL3-8B-Instruct.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) uses automatic feedback, such as answer correctness and format compliance, to improve reasoning. Methods such as GRPO estimate policy advantages from relative rewards among multiple responses to the same prompt, avoid training a separate value network, and have expanded from text reasoning to vision-language models (Shao et al., 2024; Guo et al., 2025; Huang et al., 2025; Shen et al., 2025; Meng et al., 2025; Chen et al., 2025). Yet the environment usually returns only one scalar reward per response, and the resulting sequence advantage is applied to the entire generated trajectory. Every token in a response therefore receives the same advantage coefficient, leaving differences in credit within the trajectory implicit.

This granularity gap is especially important in multimodal reasoning: a correct answer does not imply consistent use of relevant visual regions, and equally rewarded responses may follow different visual reasoning paths. Existing fine-grained approaches rely on preference or process supervision (Yu et al., 2024; Wang et al., 2025a), or read dense guidance from teacher- and privileged-context output-token distributions (Agarwal et al., 2024; Yang et al., 2026b; Pan et al., 2026). A vision-language model also produces attention over visual inputs at every generated token, providing an internal readout distinct from output behavior. We treat visual attention as a proxy to be evaluated through behavioral diagnostics and controlled training comparisons, rather than as a causal attribution or certificate of correctness (Liu et al., 2026; Zhao et al., 2026a; Lu et al., 2026; Yang et al., 2026a). This leads to our central question: *Can token-varying internal visual statistics refine credit within a trajectory while preserving the verifier-derived trajectory-mean advantage?*

Attention magnitude itself has no intrinsic task direction, so directly rewarding larger attention does not address this question. We instead use the visual distribution formed by the frozen old policy $\pi_{\theta_{\text{old}}}$ under the known correct answer as a reference and measure each generated token's agreement with it. Token intervention tests whether attention rank tracks behavioral sensitivity; visual-index permutation tests whether correct–incorrect separation depends on patch identity rather than the ref-

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.

erence's marginal shape; and trajectory analysis tests whether this separation varies with generation position. These diagnostics provide evidence for a behaviorally relevant token proxy, reference-patch correspondence, and position-dependent variation. The position-dependent alignment profile further motivates us to adopt a smooth, bounded relative-position sigmoid and evaluate this design through controlled ablations.

Motivated by this evidence, we introduce **VAttnPO**. Within each update, the frozen old policy both samples responses and produces the visual reference by teacher forcing the correct answer, placing reference and rollout attention in the same parameter state; the answer supplies conditioning context without process labels or region annotations. VAttnPO converts token-wise visual similarity into trajectory-centered credit, modulates it with a smooth bounded function of relative position, and balances it into a zero-sum correction. Adding this correction to the GRPO sequence advantage allows token coefficients to vary while preserving the verifier-derived trajectory-mean advantage.

To summarize, our contributions are as follows:

- **Diagnostic Evidence.** We identify behavioral, spatial, and positional structure in internal visual attention that motivates token-level credit differentiation.
- **Visual Attention-Guided Credit.** VAttnPO derives a token-level correction from answer-conditioned visual attention through similarity centering, relative-position weighting, and sign-preserving balancing, varying credit within a response while preserving the trajectory-mean advantage.
- **Empirical Results.** On Qwen3-VL-8B-Instruct, VAttnPO improves AVE over six image tasks by 2.53 points over GRPO and records higher scores on every reported metric for structured understanding, video understanding, and MME.

2 Related Work

2.1 Multimodal RLVR and Credit Assignment

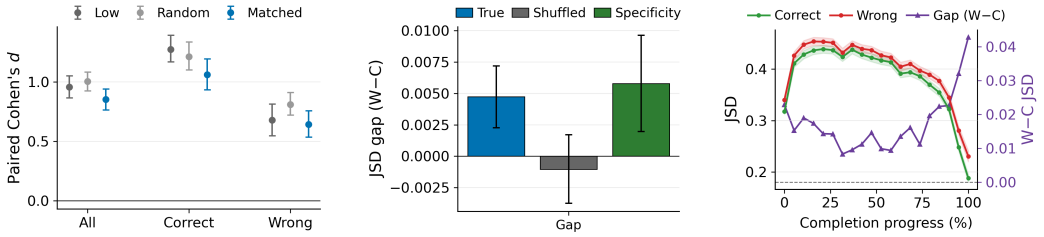
GRPO and DeepSeek-R1 establish group-relative RLVR (Shao et al., 2024; Guo et al., 2025); Vision-R1, VLM-R1, MM-Eureka, R1-V, and Video-R1 extend it to image and video reasoning (Huang et al., 2025; Shen et al., 2025; Meng et al., 2025; Chen et al., 2025; Feng et al., 2025), while SAPO further modifies policy optimization (Gao et al., 2025). These methods retain response-level verifiable outcomes as task supervision, and standard GRPO shares one sequence advantage across a completion. Preference learning and process supervision offer finer feedback but generally require preference data, step labels, a process verifier, or an additional reward model (Yu et al., 2024; Sun et al., 2024; Li et al., 2023a; Lightman et al., 2024; Wang et al., 2024b; 2025a). VAttnPO adds none of these sources and leaves verifier rewards unchanged; it only redistributes the resulting sequence advantage within each response.

2.2 On-policy Distillation and Self-Distilled RL

GKD and OPD provide token-level targets from teacher output distributions on student trajectories (Agarwal et al., 2024; Lu & Lab, 2025), while OPSD uses the same model as teacher and student under different contexts (Zhao et al., 2026b). RLSD reweights verifier-derived advantages using token-likelihood relations between original and correct-answer contexts (Yang et al., 2026b), whereas RLCSD contrasts policy distributions under correct and incorrect hints (Pan et al., 2026). All read reference behavior from output-token distributions. VAttnPO instead reads an answer-conditioned visual-attention distribution and constructs a zero-sum correction to verifier-derived advantages, without an output-distribution or attention-matching objective; verifiable rewards remain the source of task supervision.

2.3 Attention-Derived Supervision and Visual Credit

Attention transfer and distillation match teacher attention maps for knowledge transfer (Zagoruyko & Komodakis, 2016; Zhou et al., 2025), whereas diagnostic and intervention work studies links between visual attention, grounding, hallucination, and behavior (Liu et al., 2026; Zhao et al., 2026a;



(a) Token intervention (b) GT vs. shuffled reference (c) Alignment profiles

Figure 1: Diagnostic evidence from visual attention. (a) Paired Cohen's d for High tokens against Low, Random, and Position-matched controls. (b) Wrong-minus-correct JSD gaps under the correct-answer and patch-shuffled references. (c) Mean JSD for correct and incorrect responses and their wrong-minus-correct gap over completion progress. The final point denotes the last valid marker. Error bars in (a) and (b), and shaded bands around the correct and incorrect curves in (c), show confidence intervals obtained by bootstrap resampling at the example level.

Lu et al., 2026; Yang et al., 2026a). These studies show that attention is neither a causal explanation nor a certificate of correctness, yet it is not behaviorally arbitrary. VAttnPO neither increases attention at prescribed locations nor matches attention maps: the same frozen old policy produces the answer-conditioned reference and rollout attention, and the derived statistics are used only to construct the token-level GRPO correction.

3 Pilot Study

We examine whether visual attention tracks behavioral sensitivity, whether correctness-related separation depends on visual-patch correspondence, and how this separation varies over completion progress. The answer-conditioned reference P_{gt} pools the frozen model's attention over correct-answer queries and normalizes it over visual patches (Equation (2)); lower JSD indicates that a rollout token's visual distribution is closer to this reference. We conduct three targeted pilot studies to examine these questions. All studies use layer 20 of Qwen3-VL-8B-Instruct, the same fixed 200 examples, and $G = 8$ responses per input. The analysis includes 1,429 valid trajectories, comprising 771 correct and 658 incorrect responses. Correct–incorrect statistics are aggregated within input before inputs are weighted equally, with input-cluster bootstrap intervals; Appendix E provides additional protocol details.

Token Intervention. We first test whether attention rank tracks behavioral sensitivity. Using the top 20% of visual patches under P_{gt} as the reference region, we rank tokens by the preceding query's normalized attention to that region and form equal-size high, low, random, and position-matched sets. We then mask the selected tokens' complete self-attention key–value columns. Over all responses, high-attention tokens cause larger drops in the original answer log-probability than all controls, with paired Cohen's $d = 0.85$–1.01 and intervals excluding zero; high and position-matched sets have similar mean positions. Figure 1(a) additionally reports the correct and incorrect response subsets, and its legend abbreviates Position-matched as Matched. Attention rank therefore distinguishes tokens with different intervention effects even after matching average position.

GT Versus Shuffled Reference. We compare rollout-token JSD to the correct-answer reference and to a visual-index-permuted version that preserves its attention values and entropy but breaks visual correspondence. Incorrect trajectories are farther from the correct-answer reference on average, whereas the shuffled-reference gap is near zero; the interval for the difference between these gaps excludes zero. In Figure 1(b), True denotes the wrong-minus-correct gap under the correct-answer reference, Shuffled denotes the gap under the permuted reference, and Specificity is the True gap minus the Shuffled gap. Correct–incorrect separation therefore depends on reference-patch correspondence rather than only the marginal distribution of reference values.

Alignment Over Completion Progress. Correct responses have lower mean JSD than incorrect responses at every displayed position. The wrong-minus-correct gap varies non-monotonically and

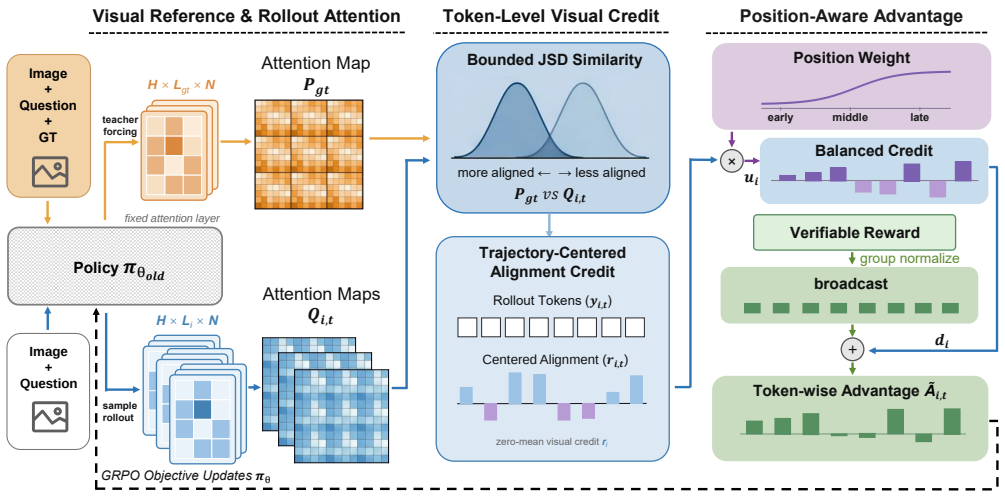


Figure 2: VAttnPO overview. Within each update, the frozen old policy extracts visual distributions by teacher forcing the correct answer and by processing sampled responses. Their similarity is centered within each trajectory, weighted by relative position, and balanced into a token correction of the sequence advantage.

is largest at the final valid token, showing position-dependent alignment without implying a general late-reasoning mechanism (Figure 1(c)).

These diagnostics motivate using answer-conditioned visual attention to differentiate tokens and examining its variation over generation progress. The position-dependent alignment profile motivates us to adopt a smooth, bounded relative-position sigmoid and evaluate this design through controlled ablations in Section 5.3. Centering and credit balancing preserve the trajectory mean of the verifier-derived advantage.

4 METHOD

VAttnPO adds a token-level visual correction to the GRPO sequence advantage while preserving its trajectory mean. The verifiable reward and its group-relative advantage evaluate the complete response, while the internal visual signal redistributes that advantage within the response.

Constructing this correction requires explicit constraints: the answer-conditioned reference supplies task direction, trajectory centering removes response-wide offset, bounded position weights modulate local credit, and credit balancing restores zero mean while retaining nonzero credit signs. The three stages construct reference and rollout attention, compute centered credit, and form position-aware token advantages.

Concretely, P_{gt} and $Q_{i,t}$ produce visual similarity $s_{i,t}$; trajectory centering yields $r_{i,t}$; position modulation forms $u_{i,t}$; and credit balancing produces $d_{i,t}$. The final coefficient $\widetilde{A}_{i,t} = \hat{A}_i + \alpha d_{i,t}$ retains $\hat{A}_i$ as the verifier-derived response-level evaluation while $d_{i,t}$ controls its local allocation within the response.

4.1 PRELIMINARIES

Given multimodal input x, the frozen old policy $\pi_{\theta_{\mathrm{old}}}$ samples G responses $y_i = (y_{i,1}, \ldots, y_{i,L_i})$, where L_i is the valid completion length. The old policy remains fixed within an update, placing sampling probabilities, response attention, and the visual reference obtained under the correct an-

swer in the same parameter state. The policy π_{ref} used for KL regularization is a separate frozen model with a different role.

The verifiable reward R_i gives the group-relative advantage

$$\hat{A}_i = \frac{R_i - \text{mean}(R_{1:G})}{\text{std}(R_{1:G}) + \epsilon}. \tag{1}$$

Standard GRPO broadcasts $\hat{A}_i$ to every valid token, so token-level differences arise only from likelihood ratios, clipping states, and score functions rather than explicit credit. VAttnPO leaves R_i and the group-normalized sequence advantage $\hat{A}_i$ unchanged. It replaces the broadcast $\hat{A}_i$ with token-varying $\widetilde{A}_{i,t}$ having the same trajectory mean. Figure 2 summarizes the path from visual attention to these coefficients.

4.2 Stage 1: Visual Reference and Rollout Attention

For each input, we run two forward paths through the same $\pi_{\theta_{\text{old}}}$. The reference branch teacher-forces the known correct answer y^{gt}, while the response branch reads attention from responses sampled by that old policy. Both paths share model parameters, input visual tokenization, and the target attention layer. Therefore, their comparison focuses on the distribution of visual attention generated under the same strategy state across two text conditions, rather than on representation drift across models or updates.

At the fixed layer, we retain only post-softmax attention from textual queries to the N visual-patch keys. Let $\mathbf{a}^{\text{gt}}_{m,h,:} \in \mathbb{R}^N_+$ and $\mathbf{a}^{(i)}_{t,h,:} \in \mathbb{R}^N_+$ denote visual attention at correct-answer token m and rollout token t for head h. The reference branch pools over L_{gt} valid answer queries and H heads, whereas the rollout branch retains one distribution for every valid generation position. Normalizing the pooled weights over visual patches gives

$$P_{\text{gt}} = \text{Norm}\Big(\tfrac{1}{L_{\text{gt}}H}\textstyle\sum_{m=1}^{L_{\text{gt}}}\sum_{h=1}^{H}\mathbf{a}^{\text{gt}}_{m,h,:}\Big), \qquad Q_{i,t} = \text{Norm}\Big(\tfrac{1}{H}\textstyle\sum_{h=1}^{H}\mathbf{a}^{(i)}_{t,h,:}\Big). \tag{2}$$

All responses to an input share P_{gt} and are therefore compared with the same visual reference. The reference is obtained from the old policy under the known correct answer; the answer supplies conditioning context but no region supervision. After generation, VAttnPO assigns $Q_{i,t}$ to token t at the same index, enabling relative position normalization by the final valid length without altering sampling. Appendix C outlines attention aggregation and the query-index distinction from the intervention in Section 3. Appendix I.1 examines the layer-wise profile of correctness-related visual-attention separation across three backbones, while Appendix I.2 illustrates spatial attention patterns across layers.

4.3 Stage 2: Token-Level Visual Credit

Both $Q_{i,t}$ and P_{gt} are probability distributions over the same visual-patch support. We use Jensen–Shannon divergence (JSD) because it is symmetric, always finite, and bounded above by the sample-independent constant $\ln 2$. These properties provide a stable mapping from distance to $[0, 1]$ without requiring identical nonzero support. The resulting bounded similarity is

$$s_{i,t} = 1 - \frac{\text{JSD}(Q_{i,t} \| P_{\text{gt}})}{\ln 2} \in [0, 1]. \tag{3}$$

Absolute similarity can vary with task content, the size of the visual support, and the concentration of the reference distribution. Treating $s_{i,t}$ directly as an added reward would carry the similarity offset shared by a response into local credit and could duplicate sequence-level differences. We therefore subtract the mean over valid generated tokens:

$$r_{i,t} = s_{i,t} - \frac{1}{L_i}\sum_{u=1}^{L_i} s_{i,u}, \qquad \sum_{t=1}^{L_i} r_{i,t} = 0. \tag{4}$$

The sign of $r_{i,t}$ indicates whether token t is more or less aligned than its trajectory average. Centering removes the response-wide offset and yields zero-sum local credit. It preserves every within-trajectory similarity difference because $r_{i,t} - r_{i,u} = s_{i,t} - s_{i,u}$, but does not rescale by variance, so different responses retain their own credit dispersion.

4.4 Stage 3: Position-Aware Advantage Construction

Position Modulation. We use a smooth, bounded position weight to modulate visual credit over completion progress without directly scaling $\hat{A}_i$. For $L_i \geq 2$, let $p_{i,t} = (t-1)/(L_i - 1)$ denote relative completion position. With $\kappa > 0$, location parameter μ, and $0 < g_{\min} \leq 1$, the endpoint-normalized sigmoid, position weight, and weighted credit are

$$\begin{aligned} S_\kappa(p) &= \frac{\sigma(\kappa(p-\mu)) - \sigma(-\kappa\mu)}{\sigma(\kappa(1-\mu)) - \sigma(-\kappa\mu)}, \\ g_{i,t}^{\text{pos}} &= g_{\min} + 2(1-g_{\min})S_\kappa(p_{i,t}), \\ u_{i,t} &= g_{i,t}^{\text{pos}} r_{i,t}. \end{aligned} \tag{5}$$

Here, μ locates the sigmoid transition, κ controls its steepness, and $g_{\min}$ sets the bounded weight range from $g_{\min}$ to $2 - g_{\min}$. When $\mu = 0.5$, the midpoint weight is one. The construction uses only valid length and relative position, placing different-length responses on the same smooth progress coordinate; for $L_i = 1$, we use unit weight. Position alone creates no independent advantage and changes only the magnitude of existing positive and negative visual credits.

Credit Balancing. Although $\sum_t r_{i,t} = 0$, position modulation generally gives $\sum_t u_{i,t} \neq 0$. Directly re-centering $\mathbf{u}_i$ would restore zero sum but could change a local credit's sign. Let $[z]_+ = \max(z, 0)$. We compute positive and negative weighted-credit totals separately and scale both sides to the smaller total:

$$P_i^+ = \sum_t [u_{i,t}]_+, \qquad N_i^- = \sum_t [-u_{i,t}]_+, \qquad C_i = \min(P_i^+, N_i^-). \tag{6}$$

$$d_{i,t} = C_i \left(\frac{[u_{i,t}]_+}{P_i^+} - \frac{[-u_{i,t}]_+}{N_i^-} \right), \qquad \sum_t d_{i,t} = 0. \tag{7}$$

C_i is the largest credit mass that both signs can retain, so the smaller side remains unchanged and only the larger side is contracted. In a nondegenerate trajectory, both scaling factors lie in $(0, 1]$; hence each $d_{i,t}$ retains the sign of its corresponding $u_{i,t}$ without increasing its magnitude, while the positive and negative totals cancel exactly. If either total does not exceed the numerical threshold τ, the entire $\mathbf{d}_i$ is set to zero.

Policy Update. The final token advantage uses correction strength α:

$$\widetilde{A}_{i,t} = \hat{A}_i + \alpha d_{i,t}, \qquad \sum_{t=1}^{L_i} \widetilde{A}_{i,t} = L_i \hat{A}_i. \tag{8}$$

Thus, $\alpha d_{i,t}$ changes token coefficients within a response, while $\sum_t d_{i,t} = 0$ preserves the trajectory mean of the verifier-derived sequence advantage. All attention, similarity, position-weight, and balance statistics are stop-gradient. During training, $\widetilde{A}_{i,t}$ directly replaces the broadcast $\hat{A}_i$ in the clipped GRPO objective, while likelihood ratios, clipping, and KL regularization against the separate frozen policy π_{ref} remain unchanged. Appendix C gives the full objective, degenerate branches, and Algorithm 1; Appendix D provides the corresponding properties and derivations.

5 Experiments

5.1 Experimental Setup

All primary experiments start from Qwen3-VL-8B-Instruct (Bai et al., 2025a) and train directly, without SFT warm-up, on a fixed set of 168,460 multiple-choice image–video examples from Video-R1-260k (Feng et al., 2025). Except in the training-step ablation, the training budget is 400 optimization steps. We post-train separate models with GRPO (Shao et al., 2024), SAPO (Gao et al., 2025), On-policy GKD (Agarwal et al., 2024), OPD (Lu & Lab, 2025), OPSD (Zhao et al., 2026b), RLSD (Yang et al., 2026b), RLCSD (Pan et al., 2026), and VAttnPO. Each method retains its original objective while sharing the training data, optimization budget, and evaluation protocol. GRPO and VAttnPO additionally share the verifiable reward, sampling, clipping, and KL regularization; their

Table 1: Image reasoning and fine-grained perception. AVE is the mean of the six tasks shown. Best results are shown in **bold**.

Model Type	Method	Fine-grained Perception			Reasoning			AVE
		VStar	HRBench 4K	MMVP	MMMU-Pro	MuirBench	BLINK	
Base	Base model	81.15	77.38	80.00	52.95	60.35	67.18	69.84
RLVR	GRPO	83.25	77.50	81.00	54.28	61.54	68.23	70.97
	SAPO	84.29	77.75	81.33	53.82	60.27	68.33	70.97
On-policy Distillation	GKD	79.06	79.00	76.33	53.87	62.62	63.97	69.14
	OPD	85.34	78.75	79.33	54.57	63.65	65.97	71.27
	OPSD	82.72	80.13	80.67	54.74	60.46	66.65	70.90
Self-distilled RL	RLSD	82.20	77.62	82.00	53.18	60.58	65.65	70.21
	RLCSD	83.77	77.50	79.67	54.10	61.85	69.02	70.99
Ours	**VAttnPO**	**87.43**	**80.62**	**82.33**	**55.95**	**65.62**	**69.07**	**73.50**

Table 2: Structured, document, and reliability results. HallusionBench reports aAcc only. Best results are shown in **bold**.

Model Type	Method	Structured & Document				Hallucination		CountBenchQA
		ChartQA	OCRBench	DocVQA	InfoVQA	POPE	HallusionBench	
Base	Base model	82.72	88.10	95.27	82.38	87.77	70.87	88.71
RLVR	GRPO	82.92	89.50	95.40	82.97	88.21	71.61	89.53
	SAPO	82.92	89.50	95.41	82.44	88.19	70.56	89.94
On-policy Distillation	GKD	81.36	88.70	94.86	82.11	88.44	71.50	85.42
	OPD	81.44	88.70	95.04	82.18	87.58	73.29	88.71
	OPSD	82.60	90.00	95.26	82.38	88.44	71.19	89.73
Self-distilled RL	RLSD	81.44	89.50	94.87	81.89	88.39	70.14	89.12
	RLCSD	83.20	89.40	95.24	82.85	87.92	72.03	89.12
Ours	**VAttnPO**	**83.84**	**90.10**	**95.59**	**83.64**	**88.67**	**73.82**	**91.17**

sole method difference is that VAttnPO replaces the broadcast sequence advantage with Equation (8). Appendices A and B provide the full data composition, reward, sampling, and method parameters.

Table 1 reports VStar (Wu & Xie, 2024), HRBench (Wang et al., 2025b), MMVP (Tong et al., 2024), MMMU-Pro (Yue et al., 2024), MuirBench (Wang et al., 2024a), and BLINK (Fu et al., 2024), whose mean is AVE. These tasks span general visual reasoning, multidisciplinary understanding, multi-image relations, fine-grained perception, and visual pattern discrimination, so AVE summarizes complementary image capabilities rather than mixing every metric into one scale. Table 2 separately reports ChartQA (Masry et al., 2022), OCRBench (Liu et al., 2024b), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), POPE (Li et al., 2023b), HallusionBench (Guan et al., 2024), and CountBenchQA (Beyer et al., 2024). Table 3 reports MVBench (Li et al., 2024), LongVideoBench (Wu et al., 2024), TempCompass (Liu et al., 2024a), Video-MME (Fu et al., 2025b), and MME (Fu et al., 2025a). ChartQA, all video metrics, MME, and the heterogeneously scaled metrics in Table 2 do not enter AVE.

5.2 Main Results

Image Reasoning and Fine-Frained Perception. VAttnPO raises AVE from 70.97 with GRPO to 73.50 ($+2.53$) and exceeds the strongest compared baseline, OPD at 71.27, by 2.23 points. All six image tasks improve over GRPO: the largest gains are 4.18 on VStar and 4.08 on MuirBench, while the other four tasks also remain positive. This distribution shows that the average gain spans both fine-grained perception and visual reasoning rather than being driven by one benchmark.

Structured, Document, and Reliability. VAttnPO obtains the highest score among compared methods on all seven structured, document, and reliability metrics in Table 2. Its largest margins over GRPO are 2.21 points on HallusionBench aAcc and 1.64 on CountBenchQA (Beyer et al., 2024), whereas gains on document tasks are smaller, including 0.19 on DocVQA. The magnitudes

Table 3: Video understanding and MME. MME-per and MME-rea denote MME-Perception and MME-Reasoning. Best results are shown in **bold**.

Model Type	Method	Video Understanding				MME	
		MVBench	LongVideoBench	TempCompass	Video-MME	MME-per	MME-rea
Base	Base model	70.72	62.2	69.10	67.7	1724.7	671.4
RLVR	GRPO	71.25	62.8	70.04	68.4	1736.1	673.9
	SAPO	71.58	62.9	69.97	68.4	1733.2	652.1
On-policy Distillation	GKD	68.53	57.3	70.04	64.4	1722.7	656.8
	OPD	70.97	62.1	71.38	66.6	1727.5	681.8
	OPSD	70.83	62.7	70.89	67.6	1732.4	657.9
Self-distilled RL	RLSD	70.08	61.0	68.77	66.4	1724.6	658.9
	RLCSD	70.60	62.6	70.46	68.0	1732.3	661.1
Ours	**VAttnPO**	**72.15**	**63.6**	**71.58**	**68.6**	**1755.0**	**683.9**

Table 4: Four controlled ablation groups. AVE denotes the mean over the six image tasks in Table 1.

(a) Core components and metric

Variant	Metric	Position	Balance	AVE
GRPO	None	No	No	70.97
Position-only	None	Yes	Yes	70.84
KL variant	KL	Yes	Yes	71.37
No position	JSD	No	Yes	72.35
VAttnPO	**JSD**	**Yes**	**Yes**	**73.50**

(b) Position design

Variant	Setting	AVE
GRPO	No visual credit	70.97
Position-only	Pure position	70.84
No position	Uniform	72.35
Position-shuffled	Shuffled sigmoid	72.19
VAttnPO	**Relative sigmoid**	**73.50**

(c) Reference and correspondence

Variant	Reference	Token credit	AVE
GRPO	None	N/A	70.97
Wrong ref.	Wrong GT	Aligned	68.92
Shuffled ref.	Shuffled GT	Aligned	70.46
Token-permuted	GT	Permuted	70.86
VAttnPO	**GT**	**Aligned**	**73.50**

(d) Credit balance

Variant	Guarantees zero sum	Preserves credit sign	AVE
GRPO	N/A	N/A	70.97
No balance	No	Yes	70.84
Mean recenter	Yes	No	71.35
VAttnPO	**Yes**	**Yes**	**73.50**

are therefore heterogeneous, but their direction extends the improvement beyond the six image tasks to visual reliability, counting, and document understanding.

Video Understanding and MME. VAttnPO's gains extend beyond the six image tasks. It improves all four video benchmarks over GRPO, by 0.90 on MVBench, 0.8 on LongVideoBench, and 1.54 on TempCompass, with a smaller margin of 0.2 on Video-MME. It also improves MME-per and MME-rea by 18.9 and 10.0 points, respectively (Table 3). Consequently, these performance improvements span short-form video, long-form video, temporal aspects, and other multimodal evaluations, with the magnitude of improvement varying across different benchmarks.

5.3 Ablation and Controlled Analysis

We examine VAttnPO through controls on its components and similarity metric, position design, reference and token correspondence, and credit balancing. Figure 3 additionally compares data compositions and training checkpoints. Every AVE uses the six tasks in Table 1.

Core components and metric. Position-only replaces visual credit with a centered, RMS-calibrated position curve; No position retains visual credit with uniform position weights; and the KL variant replaces JSD similarity with reference smoothing, forward KL, and an exponential map while retaining position modulation and credit balancing. Position-only reaches 70.84 AVE, close to GRPO at 70.97; No position reaches 72.35, and the full VAttnPO reaches 73.50. Visual credit therefore contributes beyond a positional prior, with position modulation providing a further gain. The

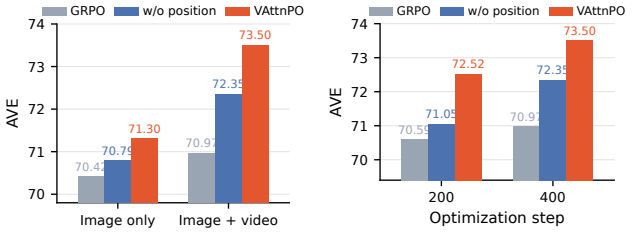
KL variant reaches 71.37; this comparison concerns only the construction defined in Appendix F and leaves policy KL regularization unchanged.

Position Design. Position-shuffled permutes the same position weights within each response, preserving their values while breaking their correspondence with tokens. Relative-position weighting raises AVE from 72.35 for No position to 73.50, whereas Position-shuffled reaches 72.19. Together with Position-only at 70.84, this pattern favors coupling visual credit to its original relative positions.

Reference And Token Correspondence. Wrong-GT reference replaces the teacher-forced correct option with a deterministic incorrect option; Shuffled reference permutes only the reference's visual-patch indices; and Token-permuted shuffles centered credits $r_{i,t}$ within each response. They reach 68.92, 70.46, and 70.86 AVE, respectively, compared with 73.50 for the complete VAttnPO, showing that reference content, visual correspondence, and token correspondence each affect the construction.

Credit Balancing. No balance directly uses the position-weighted credit $u_{i,t}$; Mean recenter restores zero sum by subtracting the trajectory mean; and the full rule rescales positive and negative credits separately to retain local signs. They reach 70.84, 71.35, and 73.50 AVE, respectively, supporting the combination of zero-sum restoration and local sign preservation.

Data Composition. VAttnPO improves over GRPO with both image-only and image–video training, with a larger gain on the mixed data: $+0.88$ and $+2.53$ AVE, respectively (Figure 3(a)). The variant without position weighting lies between GRPO and the full method in both settings.



(a) Data composition (b) Checkpoint comparison

Figure 3: (a) AVE under image-only and image–video training. (b) AVE at steps 200 and 400.

Checkpoint Comparison. VAttnPO leads GRPO and the variant without position weighting at both evaluated checkpoints, reaching 72.52 and 73.50 AVE at steps 200 and 400, respectively (Figure 3(b)). This comparison concerns two discrete training checkpoints rather than a continuous convergence trend.

5.4 Cross-Backbone Generalization

VAttnPO continues to demonstrate performance gains over GRPO across two additional backbone models. Specifically, VAttnPO improves the AVE by 0.91 on Qwen2.5-VL-7B (Bai et al., 2025b) and by 0.56 on InternVL3-8B (Zhu et al., 2025). The detail results for each benchmark are listed in Appendix H.

Table 5: Cross-backbone AVE. Δ denotes the VAttnPO gain over GRPO.

Backbone	Base	GRPO	VAttnPO	Δ
Qwen2.5-VL-7B-Instruct	59.44	61.64	**62.55**	**+0.91**
InternVL3-8B-Instruct	59.54	60.46	**61.03**	**+0.56**

6 Conclusion

We introduced VAttnPO, which converts answer-conditioned visual attention from a frozen old policy into a zero-mean, position-aware correction to the GRPO sequence advantage. On Qwen3-VL-8B-Instruct, VAttnPO raises six-task image AVE from 70.97 to 73.50, scores higher than GRPO on every reported structured, video, and MME metric, and retains positive gains on two additional backbones. Together, the diagnostics and controlled comparisons show that internal visual statistics can refine within-response credit while preserving the trajectory-mean advantage.

AI USE STATEMENT

We used generative AI tools to polish the writing, check mathematical derivations and formulas, and assist with writing and running parts of the research code, including submitting training jobs. The authors reviewed the edited text, verified the mathematical arguments, and inspected the AI-assisted code and resulting runs. We take full responsibility for the final manuscript and results.

REPRODUCIBILITY STATEMENT

We document the training-data filtering and composition in Appendix A, and report optimization settings, sampling parameters, random seeds, and the software and hardware environment in Appendix B. Section 4 and Appendix C specify the visual-reference construction, attention aggregation, numerical handling, and policy objective, with the update procedure summarized in Algorithm 1. Derivations of the credit-balancing and trajectory-mean preservation properties appear in Appendix D. Diagnostic protocols, statistical procedures, controlled variants, and benchmark evaluation are described in Appendices E–G. Appendices H and I provide additional cross-backbone results and layer-wise analyses.

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 *International Conference on Learning Representations*, 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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. GitHub repository, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 2024.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, Yuxuan Qiao, Mo Li, Tianhao Liang, Lin Zhu, Amit Agarwal, Xiaozhe Li, Shengyuan Ding, Jiazi Bu, Ziyu Liu, Zhangyang Qi, Yifei Li, Yuhang Zang, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Chaoyou Fu, Junbo Cui, Jixuan Chen, Enxin Song, Song Mao, Junming Lin, Xilin Wei, Jinsong Li, Zeyi Sun, Zhaowei Wang, Zicheng Zhang, Xiaoyi Dong, Junjun He, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models, 2026. URL https://arxiv.org/abs/2407.11691.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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.

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.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023a.

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*, 2023b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

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*, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), Dec 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, 2026.

Haolang Lu, Bolun Chu, WeiYe Fu, Guoshun Nan, Junning Liu, Minghui Pan, Qiankun Li, Yi Yu, Hua Wang, and Kun Wang. Reallocating attention across layers to reduce multimodal hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4157–4167, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

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

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.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591. IEEE, 2022.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation, 2026.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

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.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9426–9439, 2024b.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Chengxu Yang, Siqi Cai, Jingling Yuan, and Chuang Hu. Mitigating hallucination in multimodal large language models via cross-layer visual anchors. *Pattern Recognition*, pp. 114380, 2026a.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026b.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

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.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32582–32591, 2026a.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026b.

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18270–18280, 2025.

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.

A Training Data and Preprocessing

A.1 Data Sources and Composition

The main training data are cleaned multiple-choice samples selected from Video-R1-260k (Feng et al., 2025). All primary methods use the same frozen training manifest. The data-composition ablation changes only this manifest and leaves the other training settings unchanged.

Table A.1: Training-set composition.

Data type	Number of examples
Image	64,473
Video	103,987
Total	168,460

A.2 Filtering and Preprocessing

Starting from 263,071 examples in Video-R1-260k, we retain the 168,769 examples whose `problem_type` is multiple choice and whose `data_type` is image or video. The answer letter is extracted from the `<answer>...</answer>` field in the solution, and 10 examples without an extractable answer letter are removed. We then reserve the final 32 examples in the original data order for validation, leaving 168,727 candidate training examples. An independent video-availability scan reads the first and last frames of 26,896 unique video paths and identifies 15 invalid paths. After removing the 250 examples associated with these paths, we exclude 17 examples for which no valid incorrect answer can be constructed from the parsed options. The resulting fixed training manifest contains 168,460 examples: 64,473 image examples and 103,987 video examples. The 32 validation examples are not included in this training set.

B Reproducibility and Training Configuration

B.1 Optimization and Sampling

GRPO and VAttnPO share the training data, learning rate, random seed, 400-step budget, sampling, and optimization settings. VAttnPO replaces the broadcast sequence advantage with the token advantage constructed in Equation (8). Each input has $G = 8$ sampled responses. VAttnPO constructs visual attention from all 32 query heads at zero-based layer 20 and uses the dataset answer as the reference. We use JSD similarity with $\alpha = 1$, $\mu = 0.5$, $\kappa = 8$, and $g_{\min} = 0.05$, followed by position weighting and credit balancing for both image and video responses. This layer is the 21st of 36 decoder blocks; grouped-query attention expands its eight key–value heads to 32 query heads.

B.2 Verifiable Reward

The sequence reward for multiple-choice training combines accuracy, format, and length terms:

$$R = 0.9R_{\text{acc}} + 0.1R_{\text{fmt}} + 0.1R_{\text{over}}.$$

The accuracy reward extracts an answer from `<answer>...</answer>`, the final `\boxed{...}` expression, or the full response, in that order. If both prediction and ground truth resolve to option letters A–J, they are compared case-insensitively; otherwise, the scorer uses mathematical equivalence. A match gives $R_{\text{acc}} = 1$ and a mismatch gives 0. The format reward checks whether the response ends with the case-sensitive form `<answer>\boxed{...}</answer>`; a match gives $R_{\text{fmt}} = 1$ and otherwise 0. Let L denote the completion length in tokens. The length penalty starts at $L_{\text{start}} = 3072$ and reaches -1 at the maximum completion length of 4096:

$$R_{\text{over}}(L) = \min\left(-\frac{L - 3072}{1024}, 0\right) = \begin{cases} 0, & L \le 3072, \\ -\dfrac{L - 3072}{1024}, & 3072 < L \le 4096. \end{cases}$$

Table B.1: Shared training configuration for GRPO and VAttnPO.

Item	Setting
Backbone	Qwen3-VL-8B-Instruct
SFT warm-up	None
Training steps	400
Loss type	GRPO
Learning rate	2×10^{-6}
Optimizer / schedule	AdamW Torch Fused / cosine
Per-device batch / processes / accumulation	1 / 8 / 32
Nominal global batch size	256
Response group size G	8
Temperature / top-p / top-k	1.0 / 0.85 / unrestricted
Maximum sequence / completion length	8192 / 4096
Dynamic sampling / resampling attempts	Enabled / 3
Overlong filtering	Enabled
Clipping ε / ratio range	0.20 / $[0.80, 1.20]$
Policy KL coefficient β_{KL}	0.04
Training / data seed	42 / 42
Parameter updates	Full fine-tuning
Warm-up ratio	0.05
Precision / distributed strategy	bfloat16 / ZeRO-2

B.3 Software and Hardware

Training uses ms-swift 4.4.1 with its default visual preprocessing and video-sampling settings, running on eight NVIDIA L20X GPUs with bfloat16 precision and DeepSpeed ZeRO-2. Except for the explicit checkpoint comparison, the main results use the prespecified 400-step endpoint rather than selecting checkpoints by test performance.

C VAttnPO Implementation Details

C.1 Visual Reference from the Correct Answer

For each input, the frozen old policy enters an assistant turn after the original system and multimodal user prompt and teacher-forces the dataset answer. All 168,460 training targets strictly use `<answer>X</answer>`, where $X \in \{\mathrm{A}, \ldots, \mathrm{J}\}$ is the correct option letter. They contain no explanation, chain of thought, or `\boxed{}`, and the reference branch appends neither EOS nor `<|im_end|>`. Valid ground-truth queries are defined by `labels != -100` within the assistant response. The tokens pooled into P_{gt} therefore include the opening tag, correct option letter, and closing tag, but exclude the assistant header and padding. Their attention to visual keys is averaged over the L_{gt} queries and all H heads and then normalized over the N visual positions. This branch is stop-gradient; the correct answer supplies conditioning context but no region annotation.

C.2 Sampled Response Attention

The main method uses the attention query at the same index as $y_{i,t}$ and constructs retrospective credit after the complete response has been generated. The behavioral intervention in Section 3 instead uses the query immediately before the predicted target token. These indexing choices correspond to different analyses and should not be interchanged. At the target layer, attention softmax is first computed over all causal keys, after which only key columns whose input token IDs denote image or video patches are selected. Text, system/user/assistant, visual-boundary, and padding tokens are excluded. Selected patches preserve their original order in the model input sequence. Video-frame patches are flattened in the processor's spatiotemporal order. We do not normalize individual frames separately: all frames of a video, and all visual patches in multi-image or multi-video inputs, form

one shared support $\mathcal{V}$. The selected post-softmax weights are renormalized in L1:

$$Q_{i,t,p} = \frac{a_{i,t,p}}{\sum_{p' \in \mathcal{V}} a_{i,t,p'} + 10^{-8}}, \qquad p \in \mathcal{V}.$$

The same visual mask, patch order, and normalization are used for P_{gt}. Both distributions therefore describe spatial attention allocation over one common visual support rather than the total attention weight assigned to visual inputs. Main training uses zero-based layer 20 and averages all 32 query heads after grouped-query expansion, without single-head, top-k-head, or other head selection.

C.3 Masks, Length, and Numerical Boundaries

Let L_i denote the number of valid completion tokens in a response. The terminal token in the response contributes to L_i and follows the same-index attention-credit rule as the other valid tokens. Relative position is defined only over these valid tokens. For $L_i \geq 2$, we use $p_{i,t} = (t-1)/(L_i-1)$; when $L_i = 1$, the position weight is one. Model computation and the query/key projections use BF16, and the attention-logit matrix multiplication follows the model precision. Logits are explicitly cast to FP32 before softmax; head averaging, patch L1 normalization, JSD, position weighting, balancing, and final credits are then computed in FP32. Both the L1 denominator and logarithms use an epsilon of 10^{-8}. With natural logarithms, let $M = (P+Q)/2$ and

$$\mathrm{JSD}(Q\|P) = \frac{1}{2}\,\mathrm{KL}(Q\|M) + \frac{1}{2}\,\mathrm{KL}(P\|M).$$

The raw JSD lies in $[0, \ln 2]$, and the implementation clips $1 - \mathrm{JSD}\,/\ln 2$ to $[0,1]$. The KL metric variant's $\epsilon_{\mathrm{KL}} = 10^{-4}$ is not used by the main JSD construction. All visual statistics are stop-gradient. Credit balancing uses $\tau = 10^{-8}$; if $P_i^+ \leq \tau$ or $N_i^- \leq \tau$, the entire $\mathbf{d}_i$ is set to zero. The fail-closed zero-sum check uses tolerance $10^{-6}\max(\sum_t |d_{i,t}|, 1)$.

C.4 Full Clipped Optimization Objective

Define the likelihood ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid x, y_{i,<t})}.$$

The frozen old policy $\pi_{\theta_{\mathrm{old}}}$ is used both to sample responses and to extract visual signals; a separate frozen π_{ref} is used only for KL regularization. VAttnPO inserts $\widetilde{A}_{i,t}$ into the clipped GRPO objective, with symmetric clipping $\varepsilon = 0.20$, corresponding to the ratio range $[0.80, 1.20]$:

$$J(\theta) = \mathbb{E}\left[\frac{1}{L_i}\sum_t \Big\{ \min\Big(\rho_{i,t}(\theta)\widetilde{A}_{i,t}, \mathrm{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\widetilde{A}_{i,t}\Big) - \beta_{\mathrm{KL}} D_{\mathrm{KL},i,t}\Big\}\right].$$

The position weight acts only on visual credit and never directly scales $\hat{A}_i$; gradients propagate only through this policy objective.

Algorithm 1 summarizes the VAttnPO update.

D Properties and Derivations

D.1 Endpoint-Normalized Position Weight

For $\kappa > 0$ and finite μ, the sigmoid is strictly increasing and the denominator in Equation (5) is positive. Thus S_κ increases strictly from $S_\kappa(0) = 0$ to $S_\kappa(1) = 1$. With $0 < g_{\min} \leq 1$, the position weight is positive and nondecreasing on $[0,1]$, with range $[g_{\min}, 2 - g_{\min}]$; it is strictly increasing when $g_{\min} < 1$. For $\mu = 0.5$, sigmoid symmetry gives $S_\kappa(0.5) = 0.5$ and a midpoint weight of one. For $L_i = 1$, we use unit weight directly.

Algorithm 1 VAttnPO update

Require: Input x, ground-truth answer y_{gt}, rollout policy $\pi_{\theta_{\mathrm{old}}}$, group size G, and credit coefficient α

Ensure: Updated policy π_θ

```
Sample {y_i}_{i=1}^G ~ π_{θ_old} and compute {R_i, Â_i}_{i=1}^G        ▷ sequence-level signal
Teacher-force y_gt and construct P_gt with Equation (2)               ▷ shared reference
for i = 1 to G do
    for t = 1 to L_i do
        Extract Q_{i,t} and compute s_{i,t} with Equation (3)
    end for
    Compute r_{i,t} and u_{i,t} with Equations (4)–(5)
    Compute d_{i,t} with Equations (6)–(7)                            ▷ zero-mean correction
    Set Ã_{i,t} = Â_i + α d_{i,t} for all valid t
end for
Update π_θ with the clipped objective above
return π_θ
```

D.2 Centering and Credit Balancing

Centering removes the shared similarity offset and gives $\sum_t r_{i,t} = 0$, but leaves within-trajectory variance unchanged and does not standardize credit scales across responses. After position weighting, the residual satisfies $\sum_t u_{i,t} = L_i \operatorname{Cov}_t(g_i^{\mathrm{pos}}, r_i)$, where the covariance is over valid token positions. In a nondegenerate trajectory, Equation (7) rescales positive and negative credits by positive factors, preserving their signs. The total positive credit after balancing is

$$\sum_{t:u_{i,t}>0} d_{i,t} = \frac{C_i}{P_i^+} P_i^+ = C_i,$$

and the total negative credit is

$$\sum_{t:u_{i,t}<0} d_{i,t} = -\frac{C_i}{N_i^-} N_i^- = -C_i.$$

Their sum is zero, and entries with $u_{i,t} = 0$ remain zero. In either degenerate branch, $\mathbf{d}_i = \mathbf{0}$, which also has zero sum. This sign property concerns $d_{i,t}$ relative to $u_{i,t}$. Without position weighting, $u_{i,t} = r_{i,t}$; in exact arithmetic and the nondegenerate case, the two totals are equal, so balancing reduces to the identity $d_{i,t} = r_{i,t}$.

D.3 Preservation of the Trajectory-Mean Advantage

Because $\widetilde{A}_{i,t} = \hat{A}_i + \alpha d_{i,t}$ and $\sum_t d_{i,t} = 0$,

$$\frac{1}{L_i} \sum_{t=1}^{L_i} \widetilde{A}_{i,t} = \hat{A}_i.$$

This identity preserves the trajectory mean of advantage coefficients, not token gradients, their vector sum, or their norm. The final advantage can change sign when $\alpha d_{i,t}$ opposes a nonzero $\hat{A}_i$ and exceeds it in magnitude; when $\hat{A}_i = 0$, a nonzero correction can still produce positive and negative token advantages with zero mean.

D.4 Computational Complexity

Equation (3) computes JSD over an N-dimensional visual distribution for each valid rollout token. For T valid rollout tokens in a group, this credit computation requires $O(NT)$ operations.

E Diagnostic Protocols and Supplementary Analyses

E.1 Shared Sampling and Statistics

All three diagnostics use layer 20 of Qwen3-VL-8B-Instruct and the same image-only candidate pool. The candidate record contains 10,199 responses to 1,485 prompts from the image distribution of the training set; 732 prompts contain both correct and incorrect responses. We sample 200 of these prompts without replacement using seed 42 and obtain 1,429 valid trajectories, comprising 771 correct and 658 incorrect responses. For correct–incorrect comparisons, responses are aggregated within each prompt before prompts receive equal weight. The 95% intervals use 2,000 bootstrap resamples with the prompt as the cluster unit.

E.2 Analysis 1: Token Intervention

For each example, select the top 20% of visual patches under P_{gt} as S_{gt}. Let $\bar{a}^{\mathrm{probe}}_{i,t-1,p}$ be head-aggregated attention from the query before $y_{i,t}$ to visual patch p. Define

$$Q^{\mathrm{probe}}_{i,t,p} = \frac{\bar{a}^{\mathrm{probe}}_{i,t-1,p}}{\sum_{p' \in \mathrm{visual}} \bar{a}^{\mathrm{probe}}_{i,t-1,p'} + \epsilon}, \qquad w_{i,t} = \sum_{p \in S_{\mathrm{gt}}} Q^{\mathrm{probe}}_{i,t,p}.$$

The reasoning mask covers reasoning tokens before the committed answer span. For reasoning length L_r, every intervention set has size $m = \max(3, \mathrm{round}(0.15 L_r))$. High and low select the m tokens with the largest and smallest $w_{i,t}$, respectively. Random samples without replacement from the middle pool after removing high and low tokens. Position-matched greedily assigns each high token the nearest unused non-high token in relative position. The selected tokens' complete self-attention key–value columns are then masked for subsequent queries. The intervention effect is the log-probability drop of the model's original prediction:

$$\Delta\ell(S) = \log p_{\mathrm{full}}(y_i^{\mathrm{pred}}) - \log p_{\mathrm{mask}(S)}(y_i^{\mathrm{pred}}).$$

The paired effects are Cohen's $d = 0.957$ (95% CI $[0.867, 1.052]$) for High versus Low, 1.005 ($[0.925, 1.083]$) for High versus Random, and 0.852 ($[0.764, 0.942]$) for High versus Position-matched. The mean relative positions of the High and Position-matched sets are 0.666 and 0.654, respectively; the effect direction is positive in both the correct and incorrect response subsets. All effect sizes use paired differences within the same prompt and the prompt-cluster bootstrap above; position matching is deterministic and greedy.

E.3 Analysis 2: GT versus Shuffled Visual Reference

We compute each sampled token distribution's JSD from the GT reference P_{gt} and from a patch-index-shuffled version. The permutation preserves the multiset of reference values but breaks visual-patch identity. For an example e containing both correct and incorrect responses, the full-completion paired gap is

$$\delta_e = \mathrm{mean}_{i \in \mathcal{W}_e} \mathrm{mean}_t \, D_{i,t} - \mathrm{mean}_{i \in \mathcal{C}_e} \mathrm{mean}_t \, D_{i,t},$$

where $\mathcal{W}_e$ and $\mathcal{C}_e$ are its incorrect and correct responses. The comparison statistic is the difference between the GT-reference and shuffled-reference gaps. One patch permutation is generated and fixed for each (`source_row_idx`, `rollout_index`) pair. All token subsets of the same response share it, and it is not resampled per token or bootstrap replicate. This pilot uses the image-only candidate pool, so no video-frame boundary is involved.

E.4 Analysis 3: Correct–Incorrect JSD over Completion Progress

For $L_i > 1$, valid tokens in each response are assigned to 20 bins by relative generation position:

$$b_{i,t} = \left\lfloor 19 \frac{t-1}{L_i - 1} \right\rfloor.$$

The final bin contains only the terminal token of each response rather than an equal-width interval over its final segment. We first aggregate within each response and bin, then across correct or incorrect responses within each example, and finally weight examples equally.

F Ablation and Control Construction

F.1 Position and Balance Variants

Table F.1: Controlled variants and their sole change from VAttnPO.

Variant	Sole change
Position-only	Center and unit-RMS normalize each trajectory's position curve, then multiply by the fixed r-scale proxy 0.14
No position	Set $g^{\mathrm{pos}}_{i,t} = 1$, so $u_{i,t} = r_{i,t}$
Position-shuffled	Deterministically permute $g^{\mathrm{pos}}_{i,t}$ over valid tokens per response with seed 42
No balance	Set $d_{i,t} = u_{i,t}$
Mean recenter	Set $d_{i,t} = u_{i,t} - L_i^{-1} \sum_v u_{i,v}$

Table F.2: Per-benchmark results for position and credit-balance variants.

Variant	VStar	MMMU-Pro	MuirBench	BLINK	HRBench 4K	MMVP	AVE
GRPO	83.25	54.28	61.54	68.23	77.50	81.00	70.97
Position-only	83.34	54.18	62.23	67.85	78.12	79.33	70.84
No balance	83.77	54.41	62.04	67.75	77.38	79.67	70.84
Mean recenter	84.82	54.47	62.04	68.23	77.88	80.67	71.35
No position	85.34	55.55	63.35	68.70	79.50	81.67	72.35
Position-shuffled	84.95	55.48	63.25	68.39	79.38	81.67	72.19
VAttnPO	**87.43**	**55.95**	**65.62**	**69.07**	**80.62**	**82.33**	**73.50**

Similarity metric control. The KL variant uses the forward divergence $D_{\mathrm{KL}}(Q_{i,t} \| P^{\epsilon}_{\mathrm{gt}})$. We smooth the reference with the uniform distribution $U_j = 1/N$, using $\epsilon_{\mathrm{KL}} = 10^{-4}$, and map the divergence to similarity with a fixed temperature $\tau_{\mathrm{KL}} = 1$:

$$
\begin{aligned}
P^{\epsilon}_{\mathrm{gt}} &= (1 - \epsilon_{\mathrm{KL}}) P_{\mathrm{gt}} + \epsilon_{\mathrm{KL}} U, \\
D^{QP}_{i,t} &= D_{\mathrm{KL}}(Q_{i,t} \| P^{\epsilon}_{\mathrm{gt}}), \\
s^{\mathrm{KL}}_{i,t} &= \exp\!\left(-D^{QP}_{i,t} / \tau_{\mathrm{KL}} \right), \\
r^{\mathrm{KL}}_{i,t} &= s^{\mathrm{KL}}_{i,t} - \frac{1}{L_i} \sum_{v=1}^{L_i} s^{\mathrm{KL}}_{i,v}.
\end{aligned}
$$

Centering is performed over valid completion tokens. The same position weights then form $u^{\mathrm{KL}}_{i,t} = g^{\mathrm{pos}}_{i,t} r^{\mathrm{KL}}_{i,t}$, followed by the same credit-balancing rule to obtain $d^{\mathrm{KL}}_{i,t}$. This replaces the similarity construction in Stage 2; policy KL regularization remains unchanged. The variant scores 84.82, 54.16, 60.92, 68.12, 77.88, and 82.33 on VStar, MMMU-Pro, MuirBench, BLINK, HRBench 4K, and MMVP, respectively, for an AVE of 71.37. The exponential map changes the scale of within-trajectory similarity differences, so this comparison concerns the specified KL construction.

F.2 Reference and Token-Correspondence Controls

Shuffled reference applies a fixed permutation to the visual-patch indices, $P^{\pi}_{\mathrm{gt}}[j] = P_{\mathrm{gt}}[\pi(j)]$, preserving the multiset of probabilities, entropy, and concentration while disrupting spatial correspondence. Image and video patches are permuted separately, without crossing modalities; patches within each modality share one support, allowing permutations across images or video frames. Each example in the current training manifest contains either one image or one video. The permutation is determined per example with seed 42, shared by all responses to that input, and fixed throughout training. Token-permuted directly permutes the normalized centered similarity credit $r_{i,t} = s_{i,t} - \bar{s}_i$ from Equation (4) among valid completion positions, where $s_{i,t} = 1 - \mathrm{JSD}(Q_{i,t} \| P_{\mathrm{gt}}) / \ln 2$. Let π_i denote the permutation for response i; the position weights remain at their original positions:

$$
u^{\mathrm{perm}}_{i,t} = g^{\mathrm{pos}}_{i,t} r_{i,\pi_i(t)}, \qquad \mathbf{d}^{\mathrm{perm}}_i = \mathrm{Balance}(\mathbf{u}^{\mathrm{perm}}_i),
$$

where $\mathrm{Balance}$ follows Equations (6)–(7) and the same degenerate-case rule. With seed 42, the permutation is determined by the input and complete response-token sequence, so different responses generally receive different mappings while identical responses receive the same mapping. It is not resampled by training step, and padding positions are excluded. The operation preserves the multiset, mean, and RMS of $\mathbf{r}_i$, but need not preserve the multiset or RMS of $\mathbf{u}_i$ or $\mathbf{d}_i$ after position weighting and balancing; the balanced credit still has zero sum. The Wrong-GT reference parses valid options, orders them alphabetically, and selects the first option different from the ground truth. It teacher-forces `<answer>X</answer>` without `\boxed{}`, where X is the selected incorrect letter; its training manifest and all other settings match VAttnPO.

G Evaluation Protocol

All benchmark evaluations use VLMEvalKit (Duan et al., 2026) and follow its default protocol for each benchmark, including prompting, decoding, answer extraction, and scoring. We apply no benchmark-specific protocol overrides. AVE is the equally weighted mean of VStar, HRBench 4K, MMVP, MMMU-Pro, MuirBench, and BLINK. ChartQA, all video metrics, MME, and the heterogeneously scaled metrics in Table 2 are excluded.

H Cross-Backbone Results

The VAttnPO runs for Qwen2.5-VL-7B-Instruct and InternVL3-8B-Instruct both use zero-based layer 14. Apart from the backbone and attention layer, their settings follow Table B.1.

Table H.1: Cross-backbone results on the six tasks used to compute AVE.

Backbone	Method	VStar	MMMU-Pro	MuirBench	BLINK	HRBench 4K	MMVP	AVE
Qwen2.5-VL-7B-Instruct	Base	73.82	38.44	53.42	51.76	68.88	70.33	59.44
	GRPO	75.92	39.08	53.69	54.55	72.25	74.33	61.64
	VAttnPO	**76.39**	**40.02**	**55.58**	**54.71**	**73.62**	**75.00**	**62.55**
InternVL3-8B-Instruct	Base	68.59	26.36	55.65	56.60	70.37	79.67	59.54
	GRPO	71.20	27.59	56.42	56.81	70.75	80.00	60.46
	VAttnPO	**71.68**	**28.36**	**56.98**	**57.08**	**71.38**	**80.67**	**61.03**

I Layer-Wise Visual Attention Analysis

I.1 Layer-Wise Separation During Inference

We compare correctness-related visual-attention separation across network layers during inference. All three frozen models use the same fixed set of 200 diagnostic prompts as Section 3, with $G = 8$ sampled responses per prompt, temperature 1.0, top-p 1.0, and generation seed 271828. Each backbone contributes only prompts containing both valid correct and incorrect responses, yielding 198, 44, and 173 paired prompts, respectively; the eligible subsets can therefore differ across backbones.

We use the 0-based completion-progress bins defined in Appendix E.4. At each layer, visual attention is first aggregated across heads as in Equation (2), after which JSD is computed. For layer ℓ, prompt p, and progress bin b, the within-prompt wrong-minus-correct gap is

$$\delta_{p,b}^{(\ell)} = \mathrm{mean}_{i \in \mathcal{W}_p} \mathrm{mean}_{t \in b(i)} D_{i,t}^{(\ell)} - \mathrm{mean}_{i \in \mathcal{C}_p} \mathrm{mean}_{t \in b(i)} D_{i,t}^{(\ell)},$$

where $\mathcal{W}_p$ and $\mathcal{C}_p$ denote the incorrect and correct sampled responses. We summarize late-completion separation by

$$\Delta_\ell^{\mathrm{late}} = \frac{1}{5} \sum_{b=15}^{19} \mathrm{mean}_p \, \delta_{p,b}^{(\ell)}.$$

This statistic gives equal weight to the final five bins, including the terminal-token bin, and covers approximately 78.9–100% completion progress; it is not a pooled average over all tokens in that span. A positive value indicates that incorrect responses are farther from the answer-conditioned

Table I.1: Layer-wise wrong-minus-correct JSD gaps, averaged equally over the final five completion-progress bins. Positive gaps indicate lower reference alignment for incorrect responses; intervals are 95% prompt-cluster bootstrap. Bold marks Qwen3-VL layer 20, used in the main experiments, and does not denote statistical superiority.

(a) Qwen3-VL-8B-Instruct

Layer	Depth	$\Delta_\ell^{\text{late}}$	95% CI
L8	23%	4.4×10^{-3}	$[1.8 \times 10^{-3}, 7.1 \times 10^{-3}]$
L12	34%	3.6×10^{-3}	$[7.0 \times 10^{-4}, 6.6 \times 10^{-3}]$
L16	46%	1.0×10^{-2}	$[6.3 \times 10^{-3}, 1.4 \times 10^{-2}]$
L20	**57%**	$\mathbf{1.1 \times 10^{-2}}$	$\mathbf{[6.5 \times 10^{-3}, 1.5 \times 10^{-2}]}$
L24	69%	6.7×10^{-3}	$[3.2 \times 10^{-3}, 1.0 \times 10^{-2}]$
L28	80%	6.9×10^{-3}	$[3.2 \times 10^{-3}, 1.1 \times 10^{-2}]$
L32	91%	5.3×10^{-3}	$[2.4 \times 10^{-3}, 8.2 \times 10^{-3}]$

(b) Qwen2.5-VL-7B-Instruct

Layer	Depth	$\Delta_\ell^{\text{late}}$	95% CI
L0	0%	8.2×10^{-4}	$[9.5 \times 10^{-5}, 1.6 \times 10^{-3}]$
L4	15%	3.5×10^{-4}	$[-4.1 \times 10^{-3}, 4.5 \times 10^{-3}]$
L9	33%	1.1×10^{-3}	$[-4.1 \times 10^{-3}, 5.6 \times 10^{-3}]$
L14	52%	4.8×10^{-3}	$[-1.5 \times 10^{-3}, 1.1 \times 10^{-2}]$
L18	67%	3.5×10^{-3}	$[-8.2 \times 10^{-4}, 7.3 \times 10^{-3}]$
L22	81%	2.8×10^{-3}	$[-3.3 \times 10^{-3}, 8.9 \times 10^{-3}]$
L27	100%	2.8×10^{-3}	$[-2.3 \times 10^{-3}, 8.1 \times 10^{-3}]$

(c) InternVL3-8B-Instruct

Layer	Depth	$\Delta_\ell^{\text{late}}$	95% CI
L0	0%	6.9×10^{-4}	$[1.1 \times 10^{-4}, 1.3 \times 10^{-3}]$
L4	15%	1.3×10^{-3}	$[-1.1 \times 10^{-3}, 3.3 \times 10^{-3}]$
L9	33%	2.8×10^{-3}	$[-3.2 \times 10^{-4}, 5.5 \times 10^{-3}]$
L14	52%	4.3×10^{-3}	$[-2.0 \times 10^{-3}, 1.1 \times 10^{-2}]$
L18	67%	4.3×10^{-3}	$[1.6 \times 10^{-4}, 8.7 \times 10^{-3}]$
L22	81%	3.9×10^{-3}	$[-7.7 \times 10^{-4}, 8.6 \times 10^{-3}]$
L27	100%	2.3×10^{-3}	$[-1.6 \times 10^{-4}, 5.0 \times 10^{-3}]$

visual-attention reference. Confidence intervals use 2,000 prompt-cluster bootstrap resamples. For zero-based layer index ℓ in a model with L_{dec} decoder layers, normalized depth is $\ell/(L_{\text{dec}} - 1)$.

Among the scanned layers, the largest point estimates occur at intermediate normalized depths: 0.57 for Qwen3-VL-8B and 0.52 for both Qwen2.5-VL-7B-Instruct and InternVL3-8B-Instruct. In Qwen3-VL-8B, L20 is close to L16: their paired difference is 6.6×10^{-4}, with a 95% bootstrap interval of $[-2.5, 3.6] \times 10^{-3}$. The Qwen2.5-VL estimates are less precise with 44 paired prompts, while the displayed InternVL3 estimates for L14 and L18 both round to 4.3×10^{-3}. The observed profile therefore provides supporting context for using Qwen3-VL layer 20 rather than identifying a uniquely preferred layer.

I.2 Qualitative Layer-Wise Attention Examples

The quantitative analysis summarizes correctness-related separation across layers; the following examples illustrate spatial attention patterns at individual layers. Figure I.1 presents nine benchmark examples as one continuous visualization. Each PDF shows the original input and head-mean attention heatmaps from layers 8, 12, 16, 20, 24, 28, and 32, with the question and reference answer printed above. Layers within an example use a shared q95 color scale, fixed opacity, and identical smoothing, allowing qualitative inspection of spatial migration across layers within that example; absolute color intensity should not be compared across examples. The figure illustrates qualitative changes in attention patterns across layers rather than establishing causal attribution or correctness.

Question: During the IQ test, you'll be presented with existing picture example, and four picture options. Your task is to identify the one picture that follows the same pattern or rule established by the previous pictures. Here are some strategies to help you determine the right choice:
1. Look for patterns related to quantity or numbers that progress from one picture to the next.
2. Check if there is a consistent way the images are rotated or flipped in sequence.
3. Identify a common feature that each successive picture shares with the previous one, while also paying attention to any variations that might indicate a regular progression or change.
Select from the following choices.
A. Picture A
B. Picture B
C. Picture C
D. Picture D
Answer with the option letter.
Reference answer: B



Question: What is the Fund No.?
Reference answer: 3293F



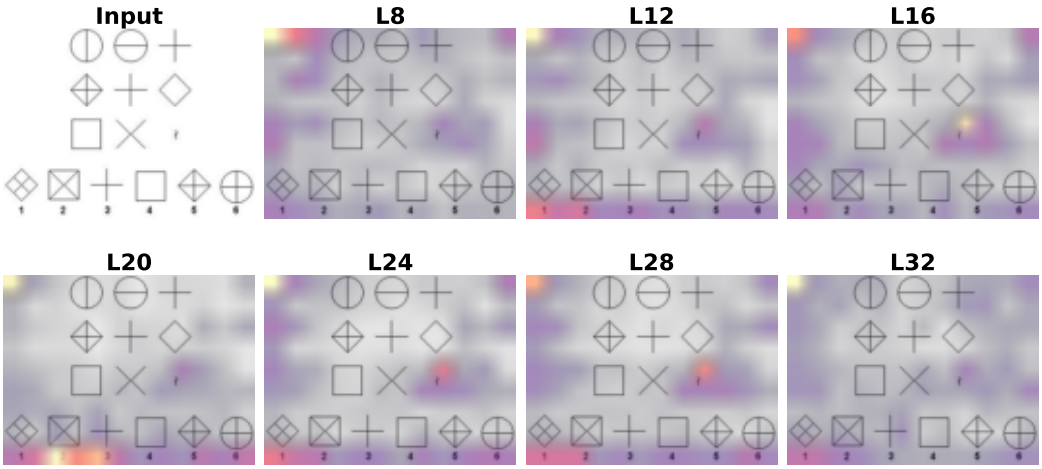
Question: How many matches were played in the 2015 ICC Cricket World Cup?
Reference answer: 49



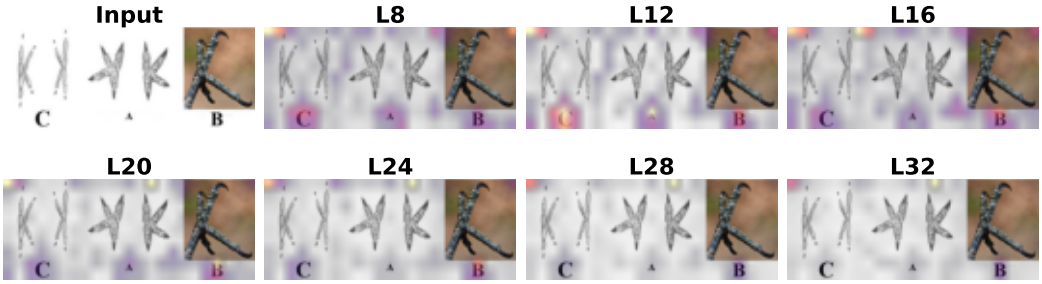
Question: Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.
Question: Subtract all small green cubes. Subtract all large cylinders. How many objects are left?
Reference answer: 1



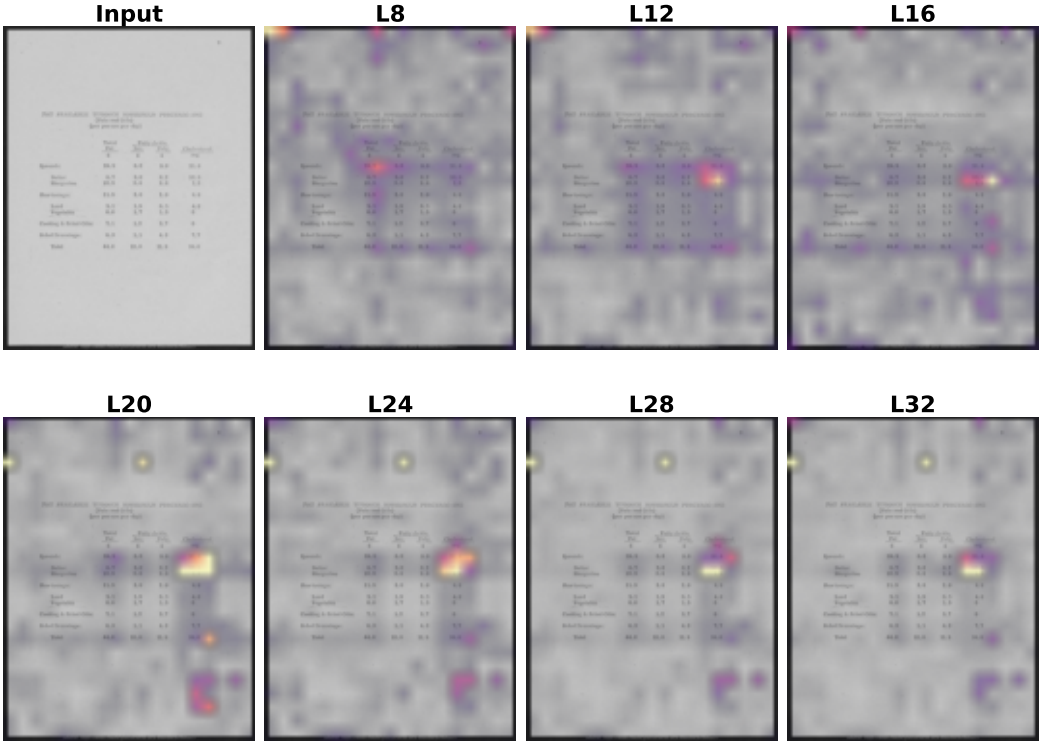
Question: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: What is the missing pattern in the picture?
Choices:
(A) 1
(B) 2
(C) 3
(D) 4
(E) 5
(F) 6
Reference answer: 2



Question: which label define downy woodpecker
A. c
B. none
C. b
D. a
Answer with the option letter.
Reference answer: C



Question: What is the total cholesterol in butter (mg)?
Reference answer: 20.6



Question: what is the total amount of this receipt? Answer this question using the text in the image directly.
Reference answer: 19.00



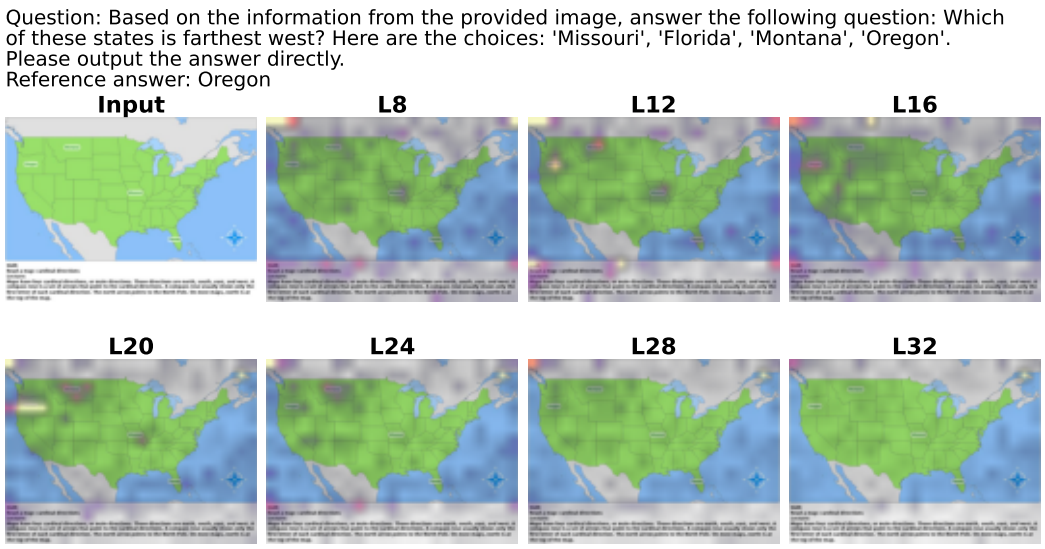


Figure I.1: Layer-wise visual attention across nine examples, ordered as BLINK, DocVQA, InfoVQA, two MathVista (Lu et al., 2023) examples, MMStar (Chen et al., 2024), two OCRBench examples, and OCRBench v2. Each example uses one shared scale across its displayed layers.