

Break the Loop with Good Attention: Visual-Attention-Based Probes in GRPO Rewards to Mitigate Repetition Collapse

Isabella Beltran Shapland¹, Hardy Chen¹, Cihang Xie¹, Yuyin Zhou¹

¹ University of California, Santa Cruz

isabshapland@gmail.com, hchen403@ucsc.edu, cixie@ucsc.edu, yzhou284@ucsc.edu

Abstract

Reasoning vision-language models (VLMs) trained with GRPO tend to fail on hard or out-of-distribution inputs by collapsing into degenerate repetition, looping indefinitely without committing to an answer. We show this collapse can be mitigated by roughly a third while raising accuracy with a visual-attention probe that we find detects and combats repetition collapse. We identify a key signal: the vision-text attention probe predicts correctness through commitment, where successful rollouts maintain grounded visual attention while failed rollouts drift into textual loops. Correctness is predictable via attention because committing to an answer correlates with correctness. While repetition collapse is known to stem from attention-drift, existing remedies target surface symptoms via length or repetition penalties. Standard penalties instead reduce in-domain accuracy by up to 20.9 points or otherwise require complex implementation. To tackle the problem, we introduce **FT-GRPO**, which injects this signal directly into the GRPO advantage as a magnitude-neutral, provably order-preserving within-tie perturbation. Our method raises held-out in-domain accuracy (+3.4 points) and out-of-distribution accuracy simultaneously, concentrated in collapse-prone MathVision (+6.3 points), VLAA-Thinking Math (+8.4 points) and MCQ (+4.4 points).

1 Introduction

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has become a practical framework for post-training reasoning models because it learns from groups of sampled responses without a separate value model. Extending this framework to vision-language models (VLMs), however, exposes a recurrent failure on difficult and out-of-distribution inputs: instead of using the image to reach and state an answer, the model repeats fragments of its own reasoning until it exhausts the

generation budget. Prior work documents related symptoms as repetitive decoding loops (Pipis et al., 2025; Xu et al., 2022), low-entropy generation, and length inflation under reinforcement learning (Cui et al., 2025; Li et al., 2026b; Wang et al., 2026). We use *repetition collapse* to denote their joint manifestation in our setting: self-repetition that prevents a VLM from completing an answer within the generation budget. This failure also coincides with attention drifting away from the visual prompt and toward the model’s generated text (Huang et al., 2026; Chu et al., 2025).

Existing remedies largely target the surface form of collapse. Length and repetition penalties discourage long or duplicated outputs, but can also suppress the extended reasoning needed for a correct answer; adaptive schemes add gates and thresholds to manage this brevity-accuracy trade-off (Liu et al., 2025; Yu et al., 2025). Visual-attention signals offer a more mechanism-aligned alternative, yet they are imperfect: a model can attend to an image and still interpret it incorrectly. Adding such a signal directly to the task reward can therefore change the preference between a correct and an incorrect response. What is missing is both a signal for the model’s failure to remain visually grounded and a constrained way to use that signal without allowing it to override task correctness.

We first ask whether visual-attention patterns contain a response-level signal associated with correctness and collapse. Using Qwen3-VL-8B-Thinking, we collect eight responses for each of 6,000 VLAA-Thinking samples (Chen et al., 2025) spanning counting, multiple choice, mathematics, and localization. We represent each response with 1,152 features: for each of the model’s 36 layers and 32 attention heads, we compute the fraction of attention assigned from generated response tokens to image tokens and average it over the response. A logistic-regression probe inspired by Lookback Lens (Chuang et al., 2024) maps this vector to a

scalar in $[0, 1]$ estimating how likely the response content is to be correct.

On 9,600 held-out responses, the probe predicts correctness with an AUROC of 0.796, compared with 0.478 after shuffling the training labels; within the 515 prompts containing both correct and incorrect responses, it ranks the correct response higher in 74.9% of pairwise comparisons. Yet the verifier frequently provides no such distinction: 95.4% of these 1,200 eight-response groups contain a reward tie, and 34.4% are fully tied (*i.e.*, answers are all correct or wrong), giving every response zero vanilla-GRPO advantage. This combination—an informative but imperfect continuous signal and a coarse verifier that often declares responses equivalent—motivates using the probe only within reward ties, never across distinct reward levels. Its relevance to repetition collapse is supported by its negative Pearson correlations with duplicated 4-gram sequences ($r = -0.47$), response length ($r = -0.45$), and truncation ($r = -0.28$) on 1,000 vanilla-policy responses (all $p < 0.001$).

We therefore introduce **FT-GRPO**. The method scores each rollout with the frozen visual-attention probe, but uses the score only to rank responses tied under the task reward. Within each tied cluster, it applies a zero-sum perturbation bounded so that responses cannot cross distinct task-reward levels; for an all-tied group, it supplies a scaled signal where vanilla GRPO supplies none. Thus, the probe provides a fine-grained preference only where the outcome verifier is indifferent, while task correctness remains the primary ordering. This design directly accommodates an informative but fallible internal signal, unlike an additive reward that can trade task performance against the probe score.

We evaluate FT-GRPO by training Qwen3-VL-8B-Thinking on a 5,400 samples from VLAA-Thinking (Chen et al., 2025). The method improves held-out in-domain accuracy from 69.3% to 72.7% (+3.4 points), with the largest gains on the collapse-prone mathematics (+8.4) and multiple-choice (+4.4) categories. Across seven out-of-distribution benchmarks, mean accuracy rises from 47.5% to 49.9%, including gains of +6.3 points on MathVision and +4.0 on MMMU-Pro. Over these benchmarks, mean truncation falls from 18.3% to 12.6% and duplicated 4-gram sequences from 15.3% to 10.6%, reductions of roughly one third. In contrast, a length-based tie-break improves out-of-distribution accuracy but costs 20.9 in-domain

points, while random and repetition-based tie-breaks do not reproduce the gains.

In summary, our contributions are three-fold: **(1)** we show that layer-head visual-attention patterns predict response correctness and are negatively associated with repetition and truncation; **(2)** we introduce an order-preserving GRPO tie-break that uses this imperfect signal without reversing preferences established by the task reward; and **(3)** we show that the resulting method improves in-domain and out-of-distribution accuracy while substantially reducing truncation and repetition, without the in-domain cost of a length penalty.

2 Preliminary

2.1 Vision-Language Autoregressive Policies

Let $x = (I, q)$ denote a prompt containing an image I and a question q . Given x , the current policy π_θ generates a response $y = (y_1, \dots, y_T)$. After applying the model’s chat template, we write the complete token sequence as $z = (z_1, \dots, z_N)$. Let $\mathcal{V}$ be the indices of the visual tokens between the model’s vision boundary markers, $\mathcal{R}$ the indices of the response tokens, and $\mathcal{C}(t)$ the causal context visible from response position t .

The language backbone contains L transformer layers and H query-attention heads. For layer ℓ and head h , let $\mathbf{q}_t^{\ell,h}$ and $\mathbf{k}_u^{\ell,h}$ denote the projected query and key vectors after the model’s query/key normalization and rotary positional embedding. Keys are repeated over query-head groups as required by grouped-query attention. The causal attention weight from response position t to token u is

$$a_{t,u}^{\ell,h} = \frac{\exp\left((\mathbf{q}_t^{\ell,h})^\top \mathbf{k}_u^{\ell,h} / \sqrt{d_h}\right)}{\sum_{v \in \mathcal{C}(t)} \exp\left((\mathbf{q}_t^{\ell,h})^\top \mathbf{k}_v^{\ell,h} / \sqrt{d_h}\right)}, \quad u \in \mathcal{C}(t), \quad (1)$$

where d_h is the head dimension. This notation follows the actual feature extractor: attention scores are computed in floating point, causally masked, and normalized over every allowed prompt and response token.

2.2 Group Relative Policy Optimization

For each prompt x , GRPO samples a group of G responses $\{y_i\}_{i=1}^G$ from π_θ and obtains one scalar reward r_i per response (Shao et al., 2024). In our main training arms,

$$r_i = s_i + \eta m_i, \quad (2)$$

where s_i is the task-verifier score, $m_i \in \{0, 1\}$ indicates the required response format, and $\eta = 0.1$ is a shared format bonus. The score s_i is binary for exact-match digit, multiple-choice, and mathematical-answer verification, and is continuous intersection-over-union for localization. Writing $\bar{r} = G^{-1} \sum_i r_i$ and using the unbiased within-group standard deviation σ_r , vanilla GRPO assigns

$$A_i^{(0)} = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}. \quad (3)$$

The same scalar advantage is applied to all valid response-token positions in y_i before the standard clipped policy update and KL regularization. When every reward in a group is equal, $A_i^{(0)} = 0$ for all i , so that prompt provides no policy-gradient signal.

3 Method

3.1 Attention Reveals Information Hidden by Outcome Rewards

Experiment setup. We use Qwen3-VL-8B-Thinking before RL training. The same frozen model generates the responses and provides the attention weights. Our data contain eight responses for each of 6,000 VLAA-Thinking questions from counting, multiple choice, mathematics, and localization. We sample across decoding temperatures (0.7, 1.0, and 1.3; one, five, and two of the eight rollouts respectively.) We split the data by question into 4,800 discovery questions and 1,200 held-out questions, corresponding to 38,400 and 9,600 responses. We label each response by content correctness while ignoring answer formatting. For localization, a response is correct when IoU ≥ 0.5 .

Following Lookback Lens (Chuang et al., 2024), we measure the attention from each response token to the image tokens. For layer ℓ and head h , the feature for response y is

$$\phi_{\ell,h}(x, y) = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \frac{\sum_{u \in \mathcal{V}} a_{t,u}^{\ell,h}}{\sum_{v \in \mathcal{C}(t)} a_{t,v}^{\ell,h}}. \quad (4)$$

This feature is the fraction of attention assigned to image tokens, averaged over all response tokens. The model has 36 layers and 32 heads, so each response gives $36 \times 32 = 1,152$ features.

We compare three representations. The first averages all 1,152 features into one score. The second averages the 32 heads in each layer, producing 36 features. The third keeps all 1,152 layer-head features. For the second and third representations, we

Attention representation	Held-out AUROC
Shuffled-label control	0.478
Raw mean attention	0.592
36 layer averages	0.674
1,152 layer-head features	0.796

Table 1: Held-out AUROC for predicting response correctness from alternative attention representations. The shuffled-label control reports the mean over ten fits with permuted discovery labels. Higher is better.

fit an ℓ_2 -regularized logistic regression

$$p(c = 1 | x, y) = \sigma(\mathbf{w}^\top \mathbf{u}(x, y) + b). \quad (5)$$

Here, $\mathbf{u}$ is either the 36-dimensional layer representation or the 1,152-dimensional layer-head representation. We set $C = 1$ and fit the models only on the discovery split. The training set contains 474 correct and 474 wrong responses from each task. We evaluate all three representations on every held-out response. As a random control, we repeat the 1,152-feature experiment ten times after shuffling the discovery labels.

Results. Table 1 reports held-out AUROC, where 0.5 indicates random ranking and 1 indicates perfect ranking.

Averaging attention hides useful correctness information. The shuffled-label control reaches an AUROC of 0.478, which is near random ranking. Raw mean attention reaches an AUROC of 0.592, while keeping one value per layer raises it to 0.674. Keeping every layer and head further raises the AUROC to 0.796. The result shows that correctness depends on which heads attend to the image, rather than only on the total amount of image attention. This finding motivates our layer-head probe. It does not yet show that the probe improves RL; we test that question in the training experiments.

3.2 Visual-Text Attention Probes for FT-GRPO

Our method uses a frozen attention probe to order responses only where the task reward cannot distinguish them. It has two parts: a response-level image faithfulness score and an order-preserving modification of the GRPO advantage.

Lookback features. Inspired by Lookback Lens, (Chuang et al., 2024) using correctness labels instead of hallucination labels, we re-forward the complete prompt-response pair (x, y_i) through the

frozen base VLM π_0 . For every layer and query head, we measure the fraction of attention allocated to visual tokens and average it over response positions:

$$\phi_{\ell,h}(x, y_i) = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \frac{\sum_{u \in \mathcal{V}} a_{t,u}^{\ell,h}}{\sum_{v \in \mathcal{C}(t)} a_{t,v}^{\ell,h}}. \quad (6)$$

The denominator is one after the causal softmax in Eq. (1), but we retain the ratio to make its meaning explicit. Flattening all (ℓ, h) pairs gives the feature vector $\Phi(x, y_i) \in \mathbf{R}^{LH}$; for our 8B backbone, $L = 36$, $H = 32$, and hence $LH = 1152$. The extractor uses neither bounding boxes nor ground-truth answers.

Probe training and inference. We fit an ℓ_2 -regularized logistic regression probe

$$f_i = f_\psi(x, y_i) = \frac{1}{1 + \exp[-(\mathbf{w}^\top \Phi(x, y_i) + b)]}, \quad (7)$$

where the target is format-agnostic content correctness: exact answer correctness for digit, multiple-choice, and mathematics, and IoU ≥ 0.5 for localization. Ground truth is used only to construct these training labels and is never an input to the probe. We split data by prompt, balance positive and negative examples within each verifier, and cap each verifier at the same number of examples per class. We retain the probe only when held-out AUROC, within-prompt score variation, and label-permutation checks pass. First, held-out AUROC > 0.55 , which means the probe predicts content correctness above chance. Second, the mean within-prompt standard deviation of probe scores must exceed 0.05. We sample across temperatures so that the probe assigns distinguishable scores to different rollouts of the same prompt. Third, refitting on shuffled correctness labels ten times yields a mean held-out AUROC below 0.55 (chance level), confirming the signal is not an artifact of the fitting procedure. During policy training, both π_0 and f_ψ remain frozen; each newly sampled policy response is re-forwarded through π_0 to obtain $f_i \in [0, 1]$.

Faith-tiebroken advantages. We first compute the vanilla advantages in Eq. (3). Consider a tied cluster C of $K \geq 2$ responses with the same $A_i^{(0)}$ inside a group that also contains another reward level. Let Δ_C be the smaller distance from this cluster to its nearest distinct lower or upper advantage, using the only existing side for an extreme

cluster. We rank the genuine probe scores in ascending order, with $\rho_i \in \{0, \dots, K-1\}$, and set

$$\tilde{A}_i = A_i^{(0)} + \alpha \Delta_C \left(\frac{\rho_i}{K-1} - \frac{1}{2} \right), \quad i \in C. \quad (8)$$

We use $\alpha = 0.5$. The offsets form a zero-sum arithmetic sequence and lie within $[-\alpha \Delta_C/2, \alpha \Delta_C/2]$. Thus non-tied responses retain their vanilla advantages exactly, the group mean is preserved, and for $0 < \alpha \leq 1$ no response can cross an adjacent task-reward band. If the available faith scores are fewer than two or are uniform, we apply no perturbation; an unavailable probe score is likewise excluded rather than treated as low faith.

An all-tied group needs separate handling because both σ_r and the vanilla advantage are zero. Over the genuine probe scores in such a group, we use

$$\tilde{A}_i = \beta \frac{f_i - \bar{f}}{\max(\sigma_f, \tau)}, \quad \tau = 0.01, \quad (9)$$

where σ_f is the unbiased within-group standard deviation. The scale β is the batch mean of $|A_j^{(0)}|$ over rollouts belonging to non-tied groups, matching the rescue to the current advantage scale. If the entire batch is all-tied, $\beta = 0$ and no signal is fabricated. This rescue remains zero-sum within the group, but uses image faithfulness where vanilla GRPO would discard the prompt altogether.

For comparison, our additive baseline places the same probe output directly in the scalar reward,

$$r_i^{\text{comp}} = m_i [\lambda s_i + (1 - \lambda) f_i] + \eta m_i, \quad (10)$$

and then applies vanilla GRPO. Unlike Eq. (8), this baseline does not guarantee preservation of the task-reward ordering; it tests whether the bounded tiebreak mechanism is preferable to ordinary reward shaping.

4 Experiment

4.1 Implementation

Training. We use Qwen3-VL-8B-Thinking (Yang et al., 2025) as the base policy and train on a stratified 5,400-example subset of VLAA-Thinking (Chen et al., 2025) covering visual counting, multiple choice, mathematics, and localization. Every training variant uses full-parameter GRPO in Verl (Sheng et al., 2025) for 250 steps on two H200 GPUs with seed 42, group size $G = 8$, PPO mini-batch size 16,

learning rate 10^{-6} , KL coefficient 0.04, and a 4,096-token response budget. All variants share the same data order and optimization configuration; only the tie-breaking signal changes. Additional implementation and hyperparameter details are provided in Appendix A.

Baselines. Table 2 compares FT-GRPO against four matched GRPO baselines. Vanilla GRPO uses only the task-verifier and format rewards, with no auxiliary ranking within reward ties. The other three baselines retain the same magnitude-neutral, order-preserving within-tie perturbation as FT-GRPO, but replace the visual-attention score with a control signal: the length tie-break ranks tied rollouts by negative response-token count, the random tie-break uses seeded uniform scores, and the repetition tie-break ranks by negative seq-rep-4. All four use the same training data and optimization configuration as FT-GRPO; Appendix A gives the repetition metric and remaining implementation details.

Evaluation. We select a single comparison checkpoint using only the vanilla-GRPO validation trajectory: S^* is the first step at which the three-point moving average over the four in-domain task categories ceases to improve by more than 1% relative. This rule selects $S^* = 150$, which is then fixed for every variant (Figure 2). We use greedy decoding with a 4,096-token response budget and evaluate on the held-out VLAA-Thinking set and eight out-of-distribution benchmarks. We group these VQA benchmarks into **math** (VLAA-Heldout, MathVista (Lu et al., 2024), MATH-Vision (Wang et al., 2024), and MMMU-Pro (Yue et al., 2025)), **general** (MMStar (Chen et al., 2024), BLINK (Fu et al., 2024), and RealWorldQA (xAI, 2024)), and **medical** (VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021)). Scoring is identical across variants: we report task accuracy together with truncation rate and duplicated four-gram rate as collapse measures. Answer extraction, task-specific scoring, and metric definitions are detailed in Appendix A.

4.2 Main Results

FT-GRPO improves accuracy in and out of domain. At S^* , FT-GRPO raises mean in-domain accuracy from 69.3% to 72.7%, a gain of +3.4 points (95% CI [1.6, 5.4]; paired bootstrap over evaluation prompts; McNemar $p < 0.001$). The gain concentrates in the collapse-prone math and

MCQ subsets (+8.4 and +4.4 points), while the saturated digit subset falls by 2.4 points. Across the seven collapse benchmarks, mean out-of-distribution accuracy rises from 47.5% to 49.9%, including gains of +6.3 points on MathVision and +4.0 points on MMMU-Pro; performance is nearly unchanged on low-collapse SLAKE (Table 2).

Generic tie-break signals do not reproduce the joint gain. The length tie-break raises mean out-of-distribution accuracy from 47.5% to 51.4%, but lowers in-domain accuracy from 69.3% to 48.4% (−20.9 points), with the largest loss on math, where long reasoning is often necessary. Random and repetition tie-breaks stay near vanilla both in domain (70.3% and 68.1% versus 69.3%) and out of domain (47.8% and 47.0% versus 47.5%). Thus, the improvement does not come from perturbing tied advantages alone: FT-GRPO is the only main setting that improves both axes without sacrificing in-domain reasoning quality (Table 2).

FT-GRPO reduces collapse without cutting off useful reasoning. Across the seven collapse benchmarks, FT-GRPO reduces mean seq-rep-4 from 15.3% to 10.6% and mean truncation from 18.3% to 12.6%, reductions of roughly one third (Figure 1). The length tie-break suppresses collapse further (9.0% repetition and 10.9% truncation), but its large in-domain accuracy loss shows that brevity alone removes useful reasoning. Random tie-breaking leaves both rates essentially unchanged (15.3% and 18.0%), while repetition tie-breaking is slightly worse than vanilla (15.9% and 19.2%). The visual-attention signal therefore provides a better accuracy–collapse trade-off than generic anti-collapse controls.

4.3 Ablation

4.3.1 Additive Reward.

We also test an additive variant (Eq. 10; Table 3) which is common in Faith-as-Reward GRPO (Li et al., 2026a), (Gong et al., 2026), (Höth et al., 2026). This construction places the probe score f_i directly in the scalar reward, mixed with the task score s_i at weight $\lambda = 0.7$, before applying vanilla GRPO. The additive reward contributes to every rollout rather than only to tied ones, and it carries no ordering guarantee. Putting the faith signal directly into the reward can rank a high-probe incorrect response above a correct one once the faith weight is large enough to matter (for binary

Method	VLAA-Heldout				Math			General			Medical		Overall
	Digit	MCQ	Math	IoU	MathVista	MathVision	MMMU-Pro	MMStar	BLINK	RealWorldQA	VQA-RAD	SLAKE	
Zero-shot (no RL)	94.0	84.0	43.2	24.8	32.8	19.8	40.5	66.1	50.0	54.2	51.1	50.6	50.9
<i>Baseline</i>													
Vanilla GRPO	96.8	87.2	58.8	34.4	33.7	25.4	44.3	68.5	52.9	54.4	53.0	53.4	55.2
Length tie-break	82.4	65.6	9.2	36.4	35.9	37.7	49.0	70.5	52.7	59.7	54.2	52.4	50.5
Random tie-break	96.8	89.6	60.8	34.0	35.2	25.8	44.7	67.0	53.2	55.4	53.2	52.3	55.7
Repetition tie-break	95.2	86.8	57.6	32.8	34.6	24.0	42.2	67.7	51.9	55.4	53.3	52.2	54.5
FT-GRPO (Ours)	94.4	91.6	67.2	37.6	35.5	31.7	48.3	68.6	53.3	57.9	54.0	52.9	57.8

Table 2: Accuracy (%) on held-out benchmarks. VLAA-Heldout reports task-level in-domain accuracy; all other datasets are grouped by out-of-distribution domain. Overall is the unweighted mean of the 12 reported results in each row. Best results in each column are bold.

Method	VLAA-Heldout				Math			General			Medical		Overall
	Overall	Digit	MCQ	Math	IoU	MathVista	MathVision	MMMU-Pro	MMStar	BLINK	RealWorldQA	VQA-RAD	SLAKE
Vanilla GRPO	69.3	96.8	87.2	58.8	34.4	33.7	25.4	44.3	68.5	52.9	54.4	53.0	53.4
Faith $\alpha=0.25$	70.0	87.6	86.4	66.8	39.2	35.7	31.6	48.5	68.7	52.6	58.4	55.3	51.8
Faith $\alpha=0.50$ (main)	72.7	94.4	91.6	67.2	37.6	35.5	31.7	48.3	68.6	53.3	57.9	54.0	52.9
Faith $\alpha=0.75$	72.1	96.4	88.8	65.6	37.6	36.1	33.9	48.0	69.1	52.8	59.3	53.9	52.5
Faith $\alpha=1.00$	71.4	96.4	88.4	66.0	34.8	36.4	33.8	48.3	68.9	52.9	58.3	54.7	53.1
Additive $0.7c+0.3f$	70.9	94.4	91.2	63.2	34.8	35.4	36.1	47.7	69.7	53.1	62.1	56.1	52.5

Table 3: Accuracy (%) for faith-strength and additive-reward ablations on the same held-out benchmarks as Table 2.

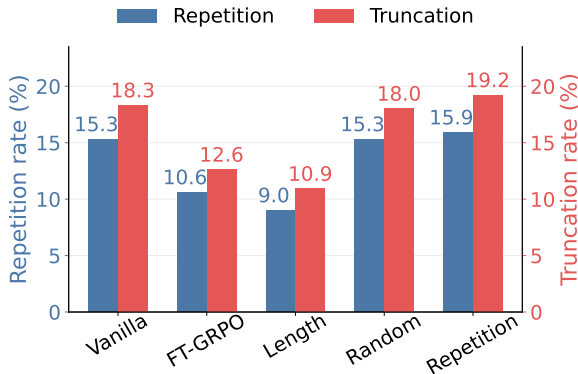


Figure 1: Repetition and truncation at S^* for the five trained settings in the primary comparison. Each bar is the unweighted mean of the corresponding rate across the seven collapse benchmarks (VQA-RAD–RealWorldQA), excluding SLAKE; lower is better. The two vertical axes use the same scale. Faith-strength and additive variants are reserved for the ablation analysis.

task rewards, when $\lambda < 0.5$). Even with a high correctness weight like $\lambda = 0.7$ faith can push incorrect responses over correct ones on continuous rewards like IoU. Empirically, the additive arm generalizes well and sometimes exceeds the tie-break out of domain, for example on MathVision, where vanilla reaches 25.4%, the tie-break 31.7%, additive 36.1%. In domain it shows a smaller accuracy gain than the tie-break on the math and IoU verifiers. On the 21 in-domain prompts where the

tie-break and additive arms disagreed, the tie-break was correct on 16 (76%). All of these incorrect responses have high scores from the probe. The reverse only happens 5 times. Injecting the vision-attention signal additively may be preferable for highest out-of-distribution accuracy and overall collapse reduction, but is not suitable for tasks that prohibit confident hallucination.

4.3.2 Alpha Arms

We test $\alpha \in 0.25, 0.5, 0.75, 1.0$ (Eq.8; Table 3), where $\alpha = 0$ reduces to baseline GRPO with only the all-tied rescue (Eq.9). The accuracy boost and collapse reduction are not sensitive to α : all four arms beat vanilla out of domain. The value of α mainly trades in-domain accuracy against brevity. Too small ($\alpha = 0.25$) matches vanilla in domain (70.0% vs. 69.3%) while giving the same OOD gain as $\alpha = 0.5$. Larger α cuts collapse harder but costs a little in-domain accuracy: $\alpha = 0.75$ reduces truncation and repetition the most, as much as the length penalty, but without the length penalty’s in-domain damage at a slight in-domain cost (72.1% vs. 72.7%). $\alpha = 1.0$ offers no further benefit over $\alpha = 0.75$. We use $\alpha = 0.5$, which gives the best in-domain accuracy while retaining the full OOD gain.

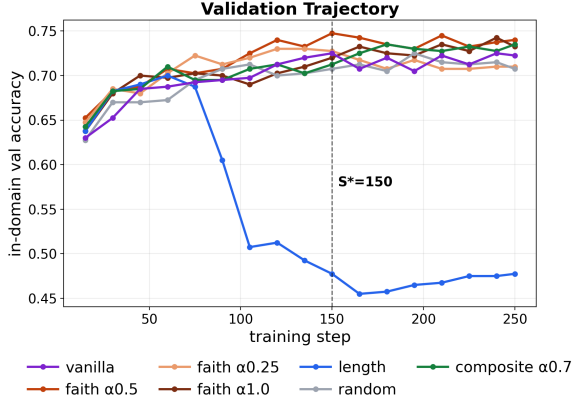


Figure 2: Val Trajectory Over 250 Steps

5 Case Study

We score policy responses with the frozen probe and inspect what it rewards.

5.0.1 Ex. 1

Ex. 1 is a physics MCQ question. We test Vanilla GRPO, Additive, and faith tiebreak arms. In this response, the vanilla-GRPO-trained model misreads the graph and commits 894 token wrong answer.

Prompt: What trend from graphs (c)/(d) as temperature increases? (GT: B)

"Vanilla GRPO (answers D) (faith 0.380):

... ϵ_1 decreases ... as temperature increases, Γ decreases. Wait, but option B says ... the linewidth Γ_1 increases. But if Γ_1 decreases, then B is wrong. Wait, maybe I misread ... Wait, let's check the graphs again ... both decrease. So option D.

To the probe, misreading and looping are attention signals indicative of an incorrect response. The same question answered by tiebreak is shorter (229 tokens) and reads the chart correctly, then commits.

Faith tie-break (answers B) (faith 0.750):

... as temperature increases, the ϵ_1 data points ... decrease ... so anisotropy ϵ_1 decreases. ... graph (d): ... as temperature increases, Γ increases ... So the trend is ϵ_1 decreases and Γ increases. That matches option B.

When the Additive reward arm (Eq. 10) responds to the same question, it shows this reward structure lacks the guarantee and can rank a confident

misread highly. This answer is the shortest of all, however. (105 tokens)

Additive arm (answers D) (faith 0.753):

... graph (c) shows ϵ_1 (anisotropy) vs temperature; as temperature increases, the data points decrease. Graph (d) shows Γ_1 (linewidth) vs temperature; as temperature increases, the data points also decrease. So both ϵ_1 and Γ_1 decrease with temperature. That matches option D.

Placing the probe score directly in the reward teaches the additive arm to commit more concisely than tie-break (229 vs 105 tokens, faith 0.75 vs 0.753), but with no guarantee that this commitment respects correctness. More examples scored by the probe can be found in Appendix B.

6 Related Works

Attention Probes Our probe builds on Lookback Lens (Chuang et al., 2024), which shows that the ratio of a model’s attention to its context versus its own generation is a strong linear predictor of contextual hallucination. Lookback Lens, and a broader line of work that reads internal attention to detect hallucinated outputs in language and vision-language models, uses attention as a post-hoc diagnostic trained on human hallucination labels. We instead fit the attention-faithfulness feature on verifier correctness in a visual context and repurpose it as a training signal inside reinforcement learning, using only labels the verifier already produces.

Internal signals as rewards in VLM/LLM RL. Several recent methods inject a probe or attention-derived signal into GRPO. In text, ProFIL (Parekh, 2026) trains a frozen probe on hidden states and verifier labels and zeroes the advantage of roll-outs it flags as post-commitment “theater,” shortening chain-of-thought. In vision, Saliency-R1 (Gong et al., 2026) rewards overlap between token saliency and human-annotated boxes, and Reinforced Attention Learning (Li et al., 2026a) optimizes a model’s attention distribution directly, treating “where to attend” as the objective. PAPO (Wang et al., 2026) adds an implicit-perception KL term rewarding divergence between outputs on the real and a masked image, but this invites “KL hacking,” which it must suppress with a double-entropy loss and further hyperparameters. Our approach differs on three axes: it is annotation-free, it never

optimizes outside correctness and unlike ProFIL, which shortens text-only reasoning when an answer is already implicit, we target repetition collapse in VLMs while preserving a learning signal on every rollout.

Relative rewards and tie handling in GRPO. GRPO derives its advantage from within-group reward differences, so groups with tied or compressed rewards yield weak gradients, motivating rank-based schemes. RankGRPO (Zhu et al., 2026) scores responses with an external preference model and heuristic length targets, while Ranking is Reward (Niu et al., 2026) breaks ties lexicographically by correctness, then length, then a reward model’s preference. We share their aim of extracting signal from tied responses but differ in signal, scope, and guarantee: our tie-break ranks by a visual-attention faithfulness score rather than a length term (which we show hurts in-domain) or a learned preference model, and it perturbs advantages only within a correctness tie, so it never overturns the task reward’s ordering. When a group is fully tied and vanilla GRPO produces no gradient, the case AVSPO (He et al., 2026) rescues with virtual samples and DAPO (Yu et al., 2025) handles by discarding the group; our all-tied rescue (Eq. 9) instead recovers a signal from image faithfulness.

Repetition collapse and its remedies. Degenerate repetition was first characterized in neural text generation, where maximization decoding drives models into loops (Holtzman et al., 2020), quantified by n -gram metrics such as seq-rep- n (Welleck et al., 2020) and attributed to a self-reinforcement effect (Xu et al., 2022). In RL-trained reasoning models the same symptom re-emerges as training-induced entropy or diversity collapse (Cui et al., 2025). Existing remedies mostly treat the symptom rather than the cause: length penalties discount long generations (Zhang and Zuo, 2025; Li et al., 2026b; Xiang et al., 2025) and repetition penalties suppress verbatim loops (Ginart et al., 2025), but neither addresses the failure to commit, and both add gated machinery to avoid sacrificing in-domain reasoning. Entropy regularization and exploration control (Cui et al., 2025) preserve diversity but do so content-blindly, while advantage- and optimization-level fixes (Yu et al., 2025) stabilize training without any content signal. Our approach is complementary: it rewards grounded commitment, so collapse falls as a consequence of committing from visual evidence rather than as a

directly optimized objective.

7 Conclusion

Instead of suppressing the symptoms, we cut down repetition collapse by encouraging high visual attention and committing to what is seen. In our method, collapse falls by roughly a third and accuracy rises both in- and out-of-distribution. Additive faith signals promise denser learning but risk probe scores outweighing accuracy. Our within-tie framework is highly generalizable. While this work utilizes within-tie vision attention ratios to mitigate VLM repetition collapse, the same order-preserving perturbation wrapper can be adapted to isolate any high-utility internal state where outcome-only verifiers suffer from severe gradient plateaus.

Limitations

The current experiment tests on one seed, with one model. Previous experiments have held up results on smaller models, but those configurations are not directly comparable to this setup so they are not displayed. The probe is also model-family specific (Qwen-trained probe must be retrained to transfer to Llama family.) Probes cannot be transferred down model sizes within the same model family (8B→4B) but can be transferred within type (Thinking→Instruct).

References

- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. [Sft or rl? an early investigation into training rl-like reasoning large vision-language models](#). *Preprint*, arXiv:2504.11468.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#) *Preprint*, arXiv:2403.20330.
- Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. 2025. [Qwen look again: Guiding vision-language reasoning models to re-attention visual information](#). *Preprint*, arXiv:2505.23558.
- Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. [Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. [The entropy mechanism of reinforcement learning for reasoning language models](#). *Preprint*, arXiv:2505.22617.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [Blink: Multimodal large language models can see but not perceive](#). *arXiv preprint arXiv:2404.12390*.
- Antonio A. Ginart, Naveen Kodali, Jason Lee, Caiming Xiong, Silvio Savarese, and John R. Emmons. 2025. [Lz penalty: An information-theoretic repetition penalty for autoregressive language models](#). *Preprint*, arXiv:2504.20131.
- Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. 2026. [Saliency-rl: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward](#). *Preprint*, arXiv:2604.04500.
- Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. 2026. [Advantage collapse in group relative policy optimization: Diagnosis and mitigation](#). *Preprint*, arXiv:2605.21125.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text de-generation. In *International Conference on Learning Representations (ICLR)*.
- Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Zefeng He, Muxin Fu, Daizong Liu, Wei-Long Zheng, and Yu Cheng. 2026. [Persistent visual memory: Sustaining perception for deep generation in lvm](#). *Preprint*, arXiv:2605.00814.
- Max Henning Höth, Kristian Kersting, Björn Deiseroth, and Letitia Parcalabescu. 2026. [Atmanrl: Towards faithful reasoning via differentiable attention saliency](#). *Preprint*, arXiv:2604.16158.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. 2026a. [Reinforced attention learning](#). *Preprint*, arXiv:2602.04884.
- Zichao Li, Jie Lou, Fangchen Dong, Zhiyuan Fan, Mengjie Ren, Hongyu Lin, Xianpei Han, Debing Zhang, Le Sun, Yaojie Lu, and Xing Yu. 2026b. [Tackling length inflation without trade-offs: Group relative reward rescaling for reinforcement learning](#). *Preprint*, arXiv:2603.10535.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. [Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering](#). *Preprint*, arXiv:2102.09542.
- Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. 2025. [Learn to reason efficiently with adaptive length-based reward shaping](#). *Preprint*, arXiv:2505.15612.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). *Preprint*, arXiv:2310.02255.
- Wenzhe Niu, Wei He, Zongxia Xie, Jinpeng Ou, Huichuan Fan, Yuchen Ge, Yanru Sun, Ziyin Wang, Yizhao Sun, Chengshun Shi, Jiuchong Gao, Jinghua Hao, and Renqing He. 2026. [Ranking is reward](#).

Intra-group preference ranking for group relative policy optimization.

Swapnil Parekh. 2026. Drop the act: Probe-filtered rl for faithful chain-of-thought reasoning. *arXiv preprint arXiv:2605.11467*.

Charilaos Pipis, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. 2025. Wait, wait, wait... why do reasoning models loop? *Preprint*, arXiv:2512.12895.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Preprint*, arXiv:2402.14804.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiushi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2026. Perception-aware policy optimization for multimodal reasoning. *Preprint*, arXiv:2507.06448.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*.

xAI. 2024. Realworldqa: A real-world vision-language benchmark. <https://huggingface.co/datasets/xai-org/RealworldQA>. Accessed: 2026-07-26.

Violet Xiang, Chase Blagden, Rafael Rafailov, Nathan Lile, Sang Truong, Chelsea Finn, and Nick Haber. 2025. Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning. *Preprint*, arXiv:2506.05256.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.

Jixiao Zhang and Chunsheng Zuo. 2025. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *Preprint*, arXiv:2504.09696.

Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Vito Ostuni, Jundong Li, and Nathan Kallus. 2026. Rank-grpo: Training llm-based conversational recommender systems with reinforcement learning. *Preprint*, arXiv:2510.20150.

A Implementation and Metric Details

A.1 Training

We implement full-parameter GRPO with Verl on two H200 GPUs and use seed 42 for every training variant. Each variant is trained for 250 steps with a training batch size of 16 prompts, eight sampled responses per prompt, and a PPO mini-batch size of 16. We use a learning rate of 10^{-6} with a cosine schedule and 5% warmup, a KL coefficient of 0.04, and maximum prompt and response lengths of 4,096 tokens each within an 8,192-token model-length envelope. The required response format is `<think>...</think><answer>...</answer>`; the prompt includes a format reminder and an example. Data order, optimization settings, token budgets, and checkpoint cadence are shared across variants.

A.2 Evaluation and Answer Scoring

We decode greedily with temperature 0, top_p=1.0, a maximum prompt length of 4,096, and a maximum response length of 4,096. Scoring is strict and identical across variants. We identify closed `<answer>...</answer>` spans case-insensitively and score only the final closed span; a response without one receives zero credit. Digit answers use exact match, and multiple-choice answers use the first standalone option letter from A–E in the final answer span. Mathematical answers are scored by symbolic equivalence

with `math_verify`, including equivalent fractional, decimal, and boxed forms. Localization uses intersection-over-union, with $\text{IoU} \geq 0.5$ treated as correct when a binary decision is required. For short-answer and medical benchmarks, we use exact match for yes/no and integer targets, entity-name containment for single-token targets, and token-level F1 otherwise.

A.3 Collapse Metrics

We report two response-level measures of repetition collapse. A response is counted as truncated when it does not close its reasoning block with `</think>` before reaching the generation budget. For repetition, we use `seq-rep-4` (Welleck et al., 2020). Let $x = (x_1, \dots, x_T)$ be the whitespace-tokenized response and $\mathcal{G}_n(x)$ the multiset of its contiguous n -grams. We define

$$\text{REP-}n(x) = 1 - \frac{|\text{set}(\mathcal{G}_n(x))|}{|\mathcal{G}_n(x)|}, \quad (11)$$

and report $n = 4$. The score is zero when no four-gram repeats and approaches one as the response becomes increasingly repetitive; responses shorter than four tokens receive a score of zero.

B Probe Scores over Baseline GRPO Responses

The following responses illustrate edge cases for the probe (Ex. 2 – 4) and are made by the vanilla GRPO arm.

B.0.1 Ex. 2

A long math rollout the probe scores highly (988 tokens).

Prompt: Solve for $\cos^2 O$.

GT: $\cos^2 O = \frac{2809}{4356}$ (faith 0.835):

"...by the Law of Cosines, $19^2 = 11^2 + 9^2 - 2 \cdot 11 \cdot 9 \cos O$, so $361 = 202 - 198 \cos O$ and $\cos O = -\frac{159}{198} = -\frac{53}{66}$.

Wait, let me verify the Law of Cosines step again to make sure I didn't make a mistake: $361 = 121 + 81 - 198 \cos O$, so $361 - 202 = -198 \cos O$, giving $\cos O = -\frac{53}{66}$. Correct. Then $\cos^2 O = \left(\frac{53}{66}\right)^2 = \frac{2809}{4356}$."

The probe does not penalize this response for length because it finishes the question without repeatedly attending to its own previous text. Even when the model self-doubts, each step builds off of the last.

B.0.2 Ex. 3

An incorrect response that was scored highly, using 349 tokens.

Prompt: Find the area of the parallelogram.

GT: $C = 1520$ (Faith 0.839)

... the area of a parallelogram is base times height... the base is probably the 38 m side, and the height is the 28 m... area = base $\times$ height = 38×28 . 38 times 28. 30×28 is 840, 8×28 is 224, so $840 + 224 = 1064$. So the area is 1064... the answer should be A.

While the final answer was wrong in this case because the model incorrectly assumed the height of the parallelogram to be 28m, this response works through each step with clean logic, never falling into a loop. We re-rank within ties with faith so that within wrong responses, the policy can be encouraged towards those that perceive and respond over those that loop endlessly.

B.0.3 Ex. 4

This MCQ response was correct (it answered D after 2,738 tokens of looping) but the answer had a lower faith score of 0.357.

Prompt: What trend in time over successive iterations in figures A and B?

GT: $D = \text{Time decreases}$. (Faith 0.357)

... Wait, but the question is about the trend... Wait, maybe I misread... Wait, none of the options fit... Wait, maybe the question is wrong, but let's check again...

The continuous back-and-forth makes this response less preferable to other correct responses that are more concise. In practice, this response is only scored against other correct responses, and when there are none, it gets the highest reward.