FIX THE REWARD OR FIX THE DATA? COSEC, A COUNTERFACTUAL SELF-EVOLVING CURRICULUM FOR GROUNDED VISION–LANGUAGE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards is widely used to post-train vision–language models (VLMs). Because an outcome-only reward scores final-answer correctness, it does not separate answers that rely on the image from answers that exploit prompt templates, option priors, or rendering artifacts. Counterfactual grounding rewards address this distinction by perturbing visual evidence and measuring the resulting change in support for an answer. When this diagnostic only shapes a reward on fixed data, however, it does not guide the next round's training examples, and repeated optimization can amplify response-length dependence. We introduce COSEC, a **Co**unterfactual **S**elf-**E**volving **C**urriculum in which a shared set of counterfactual interventions connects both loops of grounded VLM RL: reward shaping and the evolution of training data. A frozen scorer compares the same response across the original image, an answer-changing edit, and a label-preserving edit. These diagnostics shape the reward and prioritize replay, while pass-rate feedback controls task difficulty and allocation. A programmatic generator supplies fresh, constructively labeled geometric scenes each round, requiring neither human annotation nor real training images during post-training. Across five model scales in two Qwen-VL families on MME, MME-RealWorld, and VSR, COSEC improves seed-averaged endpoint Macro-3 accuracy by 2.37 percentage points over outcome-only static GRPO and training-average accuracy by 1.23 points, with gains at every scale. Visual probes connect learning on generated tasks to available image evidence. In Qwen2.5-VL-3B comparisons, data renewal also improves accuracy under two response-normalized rewards, showing that renewal complements reward normalization. These findings support joint design of visual rewards and adaptive training distributions, with counterfactual feedback guiding both policy optimization and data renewal.

1 INTRODUCTION

Reinforcement learning with verifiable rewards has become a standard tool for post-training vision–language models (VLMs): a group-relative method such as GRPO (Shao et al., 2024) optimizes the policy against a reward that checks final-answer correctness. On a multiple-choice visual question, a correct answer is nonetheless a coarse signal, because a model can be right by reading the image or by exploiting a prompt template, an option prior, or a rendering artifact (Niu et al., 2021; Agarwal et al., 2020). An outcome-only reward gives both the same value. Counterfactual grounding rewards provide a more specific signal: change the visual evidence and measure how strongly the model's response is supported under the original and edited views (Wang et al., 2026b; Yu et al., 2026; Huang et al., 2026a; Xiao et al., 2026). This approach complements groundedness feedback (Sun et al., 2024) and grounded-reasoning RL (Sarch et al., 2025) by using sensitivity to controlled visual interventions as an explicit training signal.

Turning this signal into sustained improvement raises three connected design questions. First, how should task-relevant visual sensitivity be distinguished from sensitivity to label-preserving changes? Comparing the same response under both edits provides a common scoring basis, while the

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.

aggregation and scaling of response scores determine their length dependence. Second, how should a sample's visual diagnostic influence what the model learns next? Used only to grade the current response, the diagnostic supplies no feedback for selecting the next round's examples. Third, how does a shaped reward behave under repeated optimization on fixed examples? In our Qwen2.5-VL-3B runs, a summed counterfactual reward on static data accompanies late accuracy loss and growing outputs, motivating joint study of reward design and data renewal.

The last two questions share a practical connection: the training distribution determines which examples the policy repeatedly optimizes. A counterfactual diagnostic can therefore serve two roles. In the inner loop, it distinguishes correct responses by their sensitivity to visual evidence. In the outer loop, it identifies examples to revisit, while measured pass rates indicate which task families need a change in difficulty or sampling weight. Regenerating examples lets this feedback change the next round's training distribution. This leads to the question in our title: should we fix the reward or fix the data? We study their combination, using a shared diagnostic to connect policy optimization with the selection and generation of future examples.

We introduce CoSEC, a **Co**unterfactual **S**elf-**E**volving **C**urriculum with three coordinated components. **(i) A dual-intervention grounding reward** uses a frozen scorer to compare the same correct response across the original image, an answer-changing semantic edit, and a label-preserving nuisance edit. The two likelihood contrasts share units and are combined with explicit scaling and weighting parameters. **(ii) A constructive data generator** computes answers from scene geometry and regenerates labeled examples each round; template–answer dependence, template-lookup, and question-only prediction audits gate the generated batches. **(iii) A feedback controller and replay archive** adjust task difficulty and allocation from pass rates, and prioritize replay using the visual diagnostics. Fresh generation and selected replay together form the next round's data. This implementation requires neither human annotation nor real training images during post-training.

Our evaluation connects this design to accuracy, visual evidence, and training behavior. The primary comparison uses 42 runs across five model scales and three out-of-distribution benchmarks, with the same fixed number of training updates and endpoint selection. A 127-run archive adds training trajectories, reward variants, and curriculum controls. The resulting evidence supports three contributions spanning method design, visual transfer, and training dynamics:

- **A shared counterfactual interface for reward and curriculum.** CoSEC carries the same visual diagnostics from reward shaping into replay selection, alongside pass-rate control of difficulty and task allocation, to connect policy updates with the generation and selection of constructively labeled training examples (§3).
- **Sustained transfer from generated visual tasks.** Mean endpoint Macro-3 improves by 2.37 percentage points over outcome-only static GRPO, with gains at all five scales and 22 of 22 CoSEC runs above their same-base baseline mean. Training-average accuracy improves at every scale, by 1.23 points on average; visual probes relate gains on generated tasks to task-relevant evidence under controlled image interventions (§5).
- **Evidence for combining reward design with data renewal.** In Qwen2.5-VL-3B comparisons, renewal improves endpoints by 3.21 and 2.14 points under word-count and scorer-token normalization, respectively. Together with the fixed-data and closed-loop trajectories, these results show how renewal can complement normalization, while the archive and curriculum analyses characterize the resulting training behavior (§6).

2 Related Work

Grounding rewards. GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and training systems such as `verl` (Sheng et al., 2025) support reward-based post-training. Visual grounding can be encouraged through groundedness feedback (Sun et al., 2024), grounded reasoning (Sarch et al., 2025), or counterfactual interventions (Wang et al., 2026b; Yu et al., 2026; Huang et al., 2026a; Xiao et al., 2026). These methods use different supervision and control conditions; Evidence-RL, for example, compares evidence removal with matched non-evidence regions. CoSEC combines constructively paired edits with a data controller, reusing the diagnostics across policy optimization and sample selection.

Curricula and self-evolution. Curriculum learning orders or regenerates examples according to learning progress (Bengio et al., 2009; Chen et al., 2026; Qi et al., 2025). Recent VLM approaches use

self-play and model-generated supervision (Thawakar et al., 2026; Sunil et al., 2026; He et al., 2026b; Wang et al., 2026a; Venkatraman et al., 2026); counterfactual generation also supplies contrastive training data (Huang et al., 2026b). Our generator instead derives answers from scene geometry, and the curriculum combines pass-rate feedback with counterfactual visual diagnostics. Table 25 and Appendix G provide the detailed positioning; the adapted comparison protocols appear in Appendix H.

Reward exploitation and length. Proxy optimization can diverge from the intended objective (Amodei et al., 2016; Skalse et al., 2022; Gao et al., 2023). Length bias in preference optimization motivates reward corrections (Singhal et al., 2024; Chen et al., 2024), while visual robustness methods control the use of degraded views (Liu et al., 2026). Potential-based shaping studies conditions for policy invariance (Ng et al., 1999). We examine a complementary question: how the length dependence and optimization behavior of a counterfactual reward change when visual diagnostics also guide the regeneration and selection of training examples.

3 CoSEC: Joint Grounding Reward and Curriculum Design

3.1 Dual-intervention diagnostics for visual grounding

Let $\pi_\theta(y \mid I, q)$ be a policy that answers question q about image I. We assume a generator G with verifiable labels and two controlled edits: a semantic edit I^{cf} changes the correct answer, and a nuisance edit I^{nu} preserves it. Each generated item also has a task family $f \in \{1, \dots, K\}$, a difficulty $d \in [0, 1]$, and a generation seed. These assumptions define the generator interface; our geometric implementation and its audits are specified in Appendix I.

For each correct policy response y, the frozen scorer π_s evaluates the *same token sequence* under the original, semantic-edit, and nuisance-edit views:

$$\ell_v = \log P_s(y \mid I^v, q), \qquad v \in \{o, cf, nu\}, \quad I^o = I. \tag{1}$$

Two frozen-scorer contrasts summarize the response's likelihood shifts:

$$g = \ell_o - \ell_{cf}, \qquad e = |\ell_o - \ell_{nu}|. \tag{2}$$

With q and the full response y fixed, g and e are frozen-scorer proxies for visual sensitivity. Positive g favors the original view; near-zero g indicates little semantic response contrast. The nuisance edit preserves the answer label, but need not preserve the validity of incidental descriptions in y. Thus e penalizes full-response likelihood displacement under that edit. Scoring is restricted to responses that pass the answer verifier.

3.2 Grounding rewards with cross-view consistency

With correctness $o \in \{0, 1\}$, the reward is

$$R_\lambda = o + \mathbf{1}[o = 1]\frac{\lambda}{S}(g - \beta e). \tag{3}$$

Here $\lambda \geq 0$ sets shaping strength, $\beta \geq 0$ weights nuisance displacement, and $S > 0$ is a *designer-chosen scaling scalar*. Incorrect responses receive zero. The two contrasts share units because the same scorer evaluates the same response. For $\beta \geq 1$, Proposition 1 gives

$$g - \beta e \leq \min\{\ell_o, \ell_{nu}\} - \ell_{cf}. \tag{4}$$

A positive bonus therefore requires both views to support the response over the counterfactual view. A fixed S controls magnitude while retaining the length dependence of summed log-probabilities (Section 6). Complete proofs are in Appendix P.

GRPO constructs group-relative advantages from these rewards. Policy optimization neither updates the scorer nor differentiates through the image edits. Data selection reuses the recorded diagnostics, providing a second feedback path without a separate learned critic.

3.3 Closed-loop control of difficulty and task allocation

Training alternates between policy optimization and dataset renewal:

$$\pi_{r+1} = \mathrm{GRPO}(\pi_r, D_r, R_\lambda), \qquad D_{r+1} = C(G, \mathcal{T}_r, A_r), \tag{5}$$

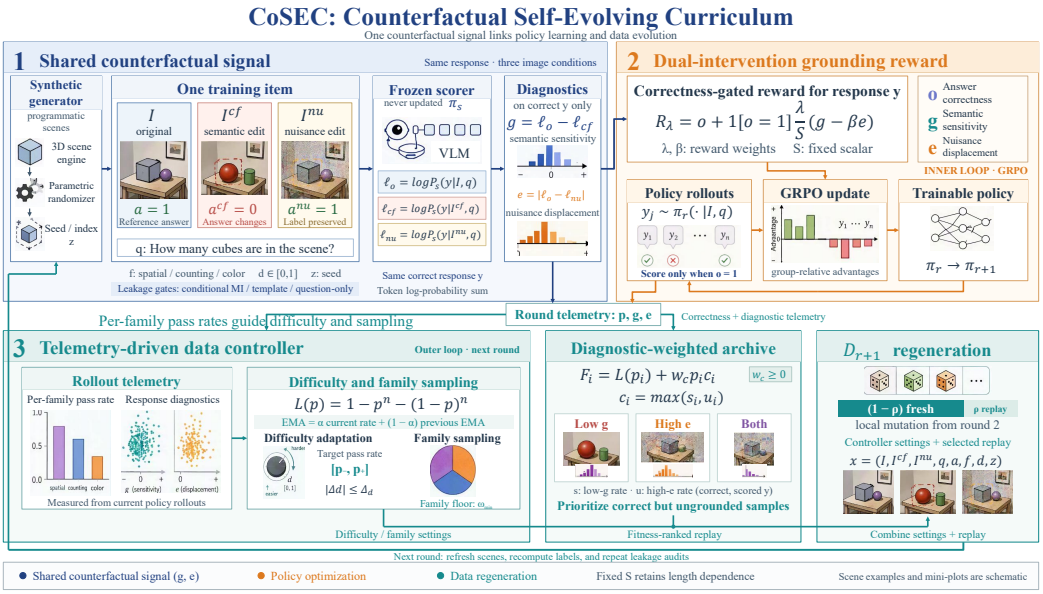


Figure 1: CoSEC reuses visual diagnostics for reward shaping and future data selection. Pass rates control difficulty and task allocation; a diagnostic-weighted archive supplies replay. The illustrated generator instantiates the framework with geometric scenes. Parameter definitions are in §3; experimental settings are in Appendix I.1.

where $\mathcal{T}_r$ contains pass rates and scored-response diagnostics, and A_r is the replay archive. We use GRPO for the inner update; the outer controller determines the next training distribution.

For a fixed item with pass probability p and n independent rollouts, the probability of observing both correct and incorrect responses in one group is

$$L(p) = 1 - p^n - (1-p)^n. \tag{6}$$

We evaluate L at observed pass rates as a learnability proxy. At the family level, $L(p_f)$ is a plug-in allocation heuristic, not the observed mixed-group frequency of heterogeneous items. For an all-correct group of $n \geq 2$ responses, Proposition 2 gives

$$R_j - \bar{R} = \frac{\lambda}{S}(Z_j - \bar{Z}), \qquad Z_j = g_j - \beta e_j, \tag{7}$$

where bars denote group means. For $\lambda > 0$, variation in Z_j thus preserves reward contrast after all responses become correct.

The controller smooths family pass rates using $\hat{p}_f^{(r)} = \alpha p_f^{(r)} + (1-\alpha)\hat{p}_f^{(r-1)}$. Within a target band $[p_-, p_+]$, difficulty remains unchanged. Outside it, with midpoint $p^\star$, the update is

$$d_f^{(r+1)} = \mathrm{clip}_{[0,1]}\left[d_f^{(r)} + \mathrm{clip}_{[-\Delta_d,\Delta_d]}\left(\eta_d(\hat{p}_f^{(r)} - p^\star)\right)\right], \tag{8}$$

Family weights use the current round's pass rates:

$$w_f = \omega_{\min} + (1 - K\omega_{\min})\frac{\max\{L(p_f^{(r)}), \epsilon\}}{\sum_{j=1}^{K}\max\{L(p_j^{(r)}), \epsilon\}}. \tag{9}$$

The designed scalars $\alpha \in (0,1]$, $\eta_d > 0$, and $\Delta_d > 0$ control smoothing, gain, and update size; $0 \leq \omega_{\min} < 1/K$ prevents family starvation, and $\epsilon > 0$ stabilizes normalization. Each regenerated batch must pass template–answer dependence, template-lookup, and question-only prediction audits before the batch is admitted to policy optimization.

Pass rates above the band increase difficulty; rates below it decrease difficulty. The deadband filters small fluctuations, and clipping limits changes between rounds. Family allocation favors intermediate pass rates while reserving a floor for every family. Figure 11 shows both controls over training.

3.4 DIAGNOSTIC-GUIDED REPLAY BEYOND ANSWER CORRECTNESS

Pass rates can mark an item as solved despite its diagnostic flags. The archive ranks sample i by

$$c_i = \max(s_i, u_i), \qquad F_i = L(p_i) + w_c p_i c_i, \tag{10}$$

where s_i and u_i are the fractions of correct, scored responses with $|g| < \tau_g$ and $e > \tau_e$, respectively. The positive thresholds select near-zero semantic contrast and high nuisance displacement; $w_c \geq 0$ sets their archive weight. The near-zero flag excludes strongly negative g, which the signed reward term still penalizes. Without scoring telemetry, selection uses $L(p_i)$.

For fixed $c > 0$ and $w_c > 0$, let p_c maximize $F(p, c) = L(p) + w_c pc$. Proposition 3 establishes

$$p_c > \tfrac{1}{2}, \qquad F(1, c) = w_c c > 0. \tag{11}$$

Diagnostic replay therefore favors higher-success flagged items and assigns them positive fitness even at saturation. The larger flag rate preserves sample-level distinctions hidden when pass rates are averaged over an entire task family.

The archive retains at most M items. From the second round, up to a fraction $\rho \in [0, 1)$ of each family's batch is sampled without replacement in proportion to $\max(F_i, \epsilon)$ and replayed with bounded difficulty mutation. Fresh generation fills the remaining slots. Thus (g, e) informs both the current reward and future sample selection. Appendix I.1 collects all settings and Appendix F gives the complete round algorithm.

4 EXPERIMENTAL SETUP

We evaluate Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B (Bai et al., 2025a). Post-training uses generated spatial, counting, and color questions. The main-text evaluation focuses on 5515 held-out examples from MME (Fu et al., 2025), MME-RealWorld (Zhang et al., 2025), and VSR (Liu et al., 2023). Macro-3 is the unweighted mean of question-level accuracy on these three frozen subsets. We additionally evaluate on BLINK (Fu et al., 2024); its complete results and the original four-benchmark aggregate are reported in Appendix O; Appendix M.2 specifies the evaluation subsets. The primary comparison uses GRPO, 210 updates, and the fixed final checkpoint, with four baseline runs per scale and four CoSEC runs except six at Qwen3-VL-2B. We report mean $\pm$ sample standard deviation across training runs, and compute percentage-point differences from unrounded seed means. Appendix I.1 documents scorer assignments, optimizer settings, generator parameters, and all hyperparameter values; Table 23 summarizes the shared training configuration.

The training-dynamics archive covers 127 completed runs across model bases, optimizers, and reward variants: 64 regenerated-data and 63 fixed-data runs, including 34 adapted concurrent-method runs. Main-text drawdown is peak-to-endpoint Macro-3 loss on the common evaluation grid. Run selection, training-reward diagnostics, and within-base GRPO comparisons are specified in Appendices D and M; method adaptations are in Appendix H. Excluding the logged reward-computation phase from both arms, the median CoSEC-to-baseline time ratio across five bases is $1.09\times$; full timings and per-base decompositions are in Appendix E.4.

5 RESULTS

5.1 ACCURACY ACROSS SCALES AND BENCHMARKS

CoSEC improves mean endpoint Macro-3 at all five scales, by 1.36–3.36 points (Figure 2). The average gain is 2.37 points, and all 22 individual CoSEC runs exceed their same-base baseline mean. Table 1 reports per-benchmark means and standard deviations: $13/15$ mean differences are non-negative, including all five MME comparisons. Additional BLINK results and the original four-benchmark aggregate are reported in Appendix O.

The advantage persists across the final training rounds: all five scales have positive seed-mean gaps at steps 150, 180, and 210 (Figure 3). Integrating Macro-3 over all eight checkpoints gives training-average gains of 0.52–2.51 points across scales, with an equal-scale mean of 1.23 (Figure 2b). Thus the endpoint gains are accompanied by higher accuracy over the training trajectory.

The advantage also persists across benchmark pairs. Omitting MME, MME-RealWorld, or VSR and averaging the remaining two benchmarks within each run gives positive seed-mean gains at all

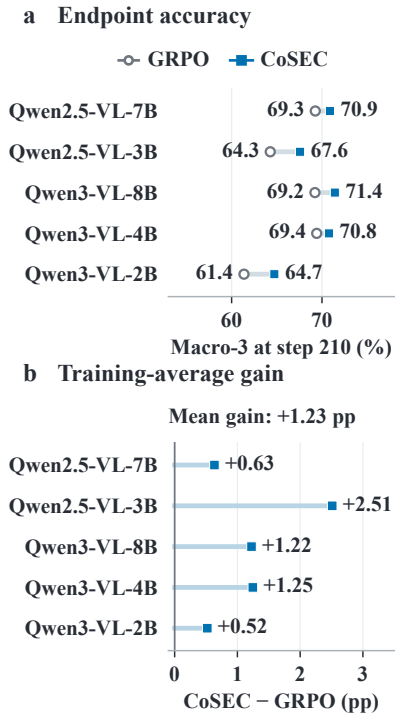


Figure 2: Cross-scale Macro-3. **a**, Step-210 seed means. **b**, Training-average CoSEC-minus-GRPO gain (trapezoidal area/210). Figure 3 and Table 1 report run counts and variation across independent training runs.

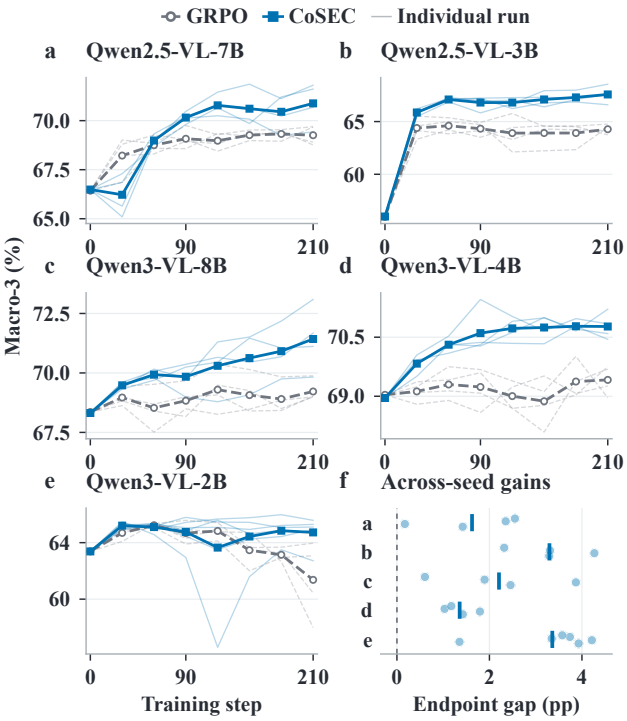


Figure 3: Macro-3 learning dynamics across 42 runs. **a–e**, Individual runs (thin) and seed means (thick) at eight checkpoints: four runs per arm and base, except six for CoSEC Qwen3-VL-2B. Vertical ranges differ. **f**, Each CoSEC endpoint minus its same-base baseline mean; vertical bars mark means, and a–e identify the corresponding base panels.

five scales. The equal-scale gains are 2.68, 2.52, and 1.90 points, respectively. Thus no single main benchmark is required for the positive cross-scale result.

Table 1: Step-210 accuracy (%). Entries are mean ± sample SD over n training runs; GRPO is the outcome-only static baseline. MME-RW denotes MME-RealWorld. Macro-3 averages MME, MME-RW, and VSR within each run. Bold marks the larger mean within each base.

Base model	Method	n	MME	MME-RW	VSR	Macro-3
Qwen2.5-VL-3B	GRPO	4	79.21 ± 0.81	36.36 ± 0.39	77.25 ± 0.83	64.27 ± 0.44
	CoSEC	4	**82.73** ± 0.84	**39.42** ± 1.63	**80.56** ± 2.51	**67.57** ± 0.80
Qwen2.5-VL-7B	GRPO	4	84.47 ± 0.85	41.49 ± 0.48	81.81 ± 0.74	69.26 ± 0.50
	CoSEC	4	**84.74** ± 2.50	**42.76** ± 0.38	**85.15** ± 1.42	**70.88** ± 1.09
Qwen3-VL-2B	GRPO	4	75.79 ± 1.81	**42.97** ± 0.51	65.34 ± 7.09	61.37 ± 2.70
	CoSEC	6	**77.84** ± 2.55	42.77 ± 0.69	**73.57** ± 1.11	**64.73** ± 1.03
Qwen3-VL-4B	GRPO	4	83.90 ± 0.52	41.39 ± 0.54	**82.96** ± 0.75	69.42 ± 0.37
	CoSEC	4	**85.30** ± 0.94	**44.24** ± 1.37	82.77 ± 0.39	**70.77** ± 0.33
Qwen3-VL-8B	GRPO	4	84.11 ± 0.49	41.35 ± 1.02	82.20 ± 0.41	69.22 ± 0.44
	CoSEC	4	**85.57** ± 1.47	**44.68** ± 1.93	**84.02** ± 0.70	**71.43** ± 1.35

5.2 Adapted methods and curriculum controls

CoSEC combines competitive endpoint accuracy with low drawdown across scales (Table 2). We compare adapted EvoLMM, iReasoner, and VisPlay methods with two curriculum controls: visual-token self-evolution (`SelfEvolve-VT`) and a fixed task schedule. The fixed schedule trains color, counting, and spatial tasks in three consecutive 70-step stages using outcome-only rewards. Initialization, update count, and evaluation are matched, with task, data, and reward adaptations specified in Appendix H. CoSEC reaches a 70.77 endpoint mean on Qwen3-VL-4B and a maximum

per-base mean drawdown of 1.05 points across five scales, compared with 3.46–14.74 for the other recipes.

Table 2: Adapted methods and curriculum controls on Qwen3-VL-4B. Accuracy is mean ± sample SD at step 210; Macro-3 averages MME, MME-RW, and VSR. The last column is the maximum of five per-base mean drawdowns for the same Macro-3, using each recipe's original stability cohort (Table 43). Supervision and adaptations are specified in Appendix H. Bold marks the highest mean or lowest drawdown.

Recipe	n	MME	MME-RW	VSR	Macro-3	Max. mean drawdown ↓
GRPO	4	83.90 ± 0.52	41.39 ± 0.54	82.96 ± 0.75	69.42 ± 0.37	3.93
iReasoner	4	$\mathbf{85.58} \pm 0.30$	43.32 ± 1.63	82.16 ± 0.35	70.35 ± 0.57	3.46
Fixed schedule	4	83.89 ± 0.48	41.46 ± 0.53	$\mathbf{83.00} \pm 0.30$	69.45 ± 0.27	5.19
VisPlay	4	85.08 ± 0.20	42.86 ± 0.73	82.82 ± 0.52	70.25 ± 0.36	14.74
EvoLMM	4	84.55 ± 0.48	43.88 ± 0.90	82.08 ± 0.41	70.17 ± 0.39	11.76
SelfEvolve-VT	4	84.48 ± 0.65	43.91 ± 0.69	82.59 ± 0.34	70.33 ± 0.16	5.37
CoSEC	4	85.30 ± 0.94	$\mathbf{44.24} \pm 1.37$	82.77 ± 0.39	$\mathbf{70.77} \pm 0.33$	**1.05**

5.3 Visual-evidence probes

Before RL, the frozen Qwen2.5-VL-3B scorer separates visual questions from text-answerable controls even when both responses are correct: summed-contrast AUROC is 0.894–0.955 across Qwen2.5-VL-3B/Qwen2.5-VL-7B generators and two controls (Figure 4; Table 37).

Accuracy gains track task-relevant visual evidence (Table 3). Intact-view gains span 12.5–30.4 points across scales; label-preserving views improve by 14.2–23.8 points, while blank and shuffled-image probes remain near their initial levels. Each probe contains 120 items. These results connect learning on the generated task to the presence of useful image evidence; Table 17 gives the protocol and start/end values.

Table 3: Visual probes: mean within-run accuracy change from step 0 to 210 (percentage points) for CoSEC on a fixed cohort. Each view contains the same 120 questions; intact and nuisance views preserve task evidence, while blank and shuffled views remove or replace it. n_C is the number of CoSEC runs; Table 17 adds the corresponding baseline comparison ($n_B = 4$ per base).

Base	n_C	Intact	Nuisance	Blank	Shuffled
Qwen2.5-VL-7B	4	+30.4	+23.8	−1.5	−4.4
Qwen2.5-VL-3B	3	+23.6	+21.4	−0.6	−0.8
Qwen3-VL-8B	4	+12.5	+18.8	−0.8	+4.2
Qwen3-VL-4B	3	+15.6	+18.1	−2.8	−0.8
Qwen3-VL-2B	5	+14.5	+14.2	+1.2	−0.5

6 Training Dynamics and Reward–Data Design

We study how visual reward design and data renewal work together. Normalized-reward comparisons test their combination, while training and curriculum trajectories trace the resulting behavior.

6.1 Reward normalization and data renewal

Data renewal improves accuracy with both normalized visual rewards tested here (Table 4). Word-count and scorer-token normalization reach endpoints of 67.41 and 67.90 with regenerated data, exceeding their fixed-data counterparts by 3.21 and 2.14 points. These results show that data renewal can complement response-length normalization.

An offline sweep applies g/T^p to the same 2880 responses from 360 questions, where T counts scorer tokens. Raising p from 0 to 1 reduces question-centered correctness correlation from 0.293 to 0.072 over 190 mixed-outcome groups. Normalization therefore changes the diagnostic's association with correctness, beyond rescaling its magnitude (Table 38).

Under the default summed reward, the representative Qwen2.5-VL-3B closed-loop run ends at 67.57 with median output length near 397 characters; the two fixed-data runs end at 57.28 and 31.42 as lengths grow to 648 and 2076 (Figure 5). Because Eq. 1 sums token log-probabilities, a fixed scalar S controls magnitude without removing length dependence. Across four bases,

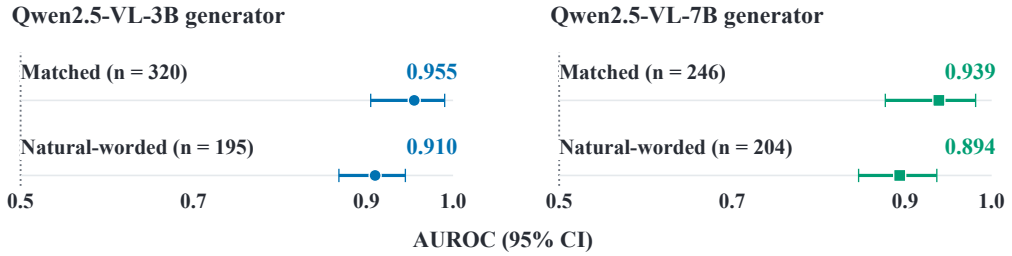


Figure 4: Visual–text discrimination with both answers correct. A frozen Qwen2.5-VL-3B scorer supplies summed-contrast AUROC; bars show 95% hierarchical paired-bootstrap intervals. Generators share image pools, and natural-worded controls use synthetic images. Table 37 gives sample counts and the paired-bootstrap resampling procedure.

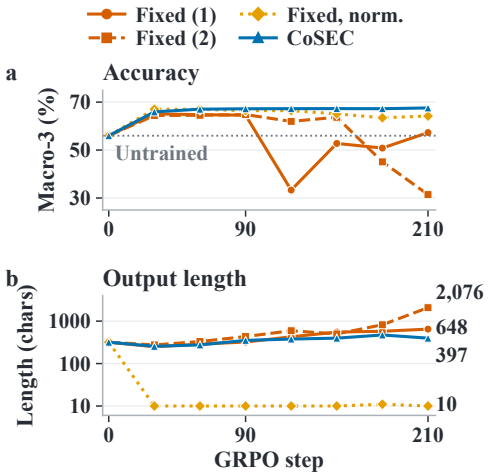


Figure 5: Qwen2.5-VL-3B training trajectories: Macro-3 (**a**) and median output characters (**b**, full evaluation pool, log scale) for four runs. "norm." denotes response-word normalization. Original four-benchmark curves: Figure 18.

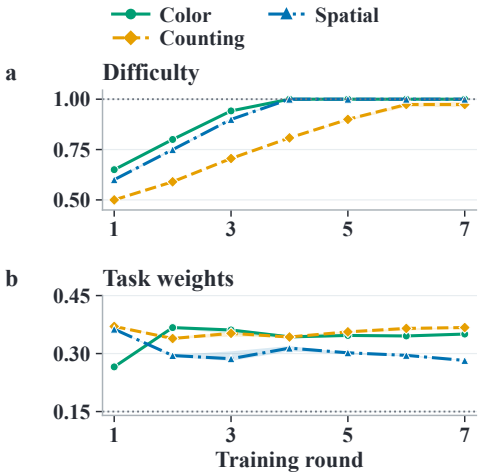


Figure 6: Qwen2.5-VL-3B curriculum dynamics: difficulty (**a**) and task weights (**b**). Curves and bands show medians and interquartile ranges across three complete seven-round histories. Dotted lines mark the difficulty ceiling (1.0) and the sampling floor for each task family (0.15).

Table 4: Normalized rewards with fixed and regenerated data on Qwen2.5-VL-3B, with a frozen Qwen2.5-VL-7B scorer and 210 GRPO updates. One run per cell. Values are three-benchmark endpoint Macro-3 (%); gaps use unrounded scores. Word-count normalization divides the summed bonus by whitespace-token count; scorer-token mean averages log-probabilities over model tokens. Full divisor variants and length diagnostics are in Table 13.

Normalization	Fixed	Regenerated	Δ (points)
Response word count	64.20	67.41	+3.21
Scorer-token mean	65.76	67.90	+2.14

grounding-score/response-length Spearman correlations rise from near zero to 0.22–0.60 during training (Figure 15, left; Table 29).

The stable Qwen2.5-VL-3B trajectory also maintains answer-marker emission: 99.8% of endpoint responses contain a marker, compared with 46.1% for the severe fixed-data run (Table 19). The latter falls from 99.7% at step 90 as repetition and budget exhaustion increase. Appendix M.2 defines marker emission and accuracy matching.

6.2 Curriculum dynamics

Further curriculum and configuration-aligned analyses appear in Appendices D and M.

Adaptive task allocation. Curriculum trajectories show how the evolving distribution is controlled (Figure 6). In three complete Qwen2.5-VL-3B histories, median color and spatial difficulty reach

the ceiling by round 4, while counting approaches it more gradually. Task-family weights continue changing after these difficulty plateaus. The controller therefore adjusts both within-family difficulty and between-family allocation; full five-scale trajectories are in Figure 11.

What the archive retains. Replay archives retain different task mixtures across policy bases (Table 5). Across 19 complete seven-round histories, all archives reach their 4000-item capacity by round 4. Counting is the largest stored category in all 14 runs at the four larger bases; spatial tasks dominate all five Qwen3-VL-2B runs. Final mean stored difficulty spans 0.66–0.73 across bases. These distributions describe the candidates available to fitness-guided replay, complementing the task-allocation weights in Figure 6.

Table 5: Final replay-archive content across n complete seven-round CoSEC histories. Entries are mean $\pm$ sample SD across runs. Task columns give percentages of the 4000 stored examples; $\bar{d}$ is their mean difficulty.

Base model	n	Color (%)	Counting (%)	Spatial (%)	$\bar{d}$
Qwen2.5-VL-3B	3	32.15 ± 1.40	41.42 ± 0.97	26.43 ± 0.74	0.698 ± 0.006
Qwen2.5-VL-7B	4	31.94 ± 1.56	37.83 ± 0.59	30.23 ± 1.05	0.725 ± 0.006
Qwen3-VL-2B	5	30.90 ± 2.85	28.69 ± 1.19	40.41 ± 1.82	0.656 ± 0.004
Qwen3-VL-4B	3	29.58 ± 0.51	39.04 ± 1.40	31.38 ± 0.97	0.681 ± 0.003
Qwen3-VL-8B	4	31.69 ± 0.60	38.09 ± 1.98	30.23 ± 2.31	0.729 ± 0.008

6.3 Component ablations and robustness

Table 6 compares design choices across scales. Removing the counterfactual signal changes both shaping and curriculum telemetry; setting $w_c = 0$ retains the reward and removes the archive diagnostic term. Uniform parent selection tests fitness weighting while retaining pass-rate feedback. The endpoint differences vary by base.

Table 6: Full-method mean minus ablated control in endpoint Macro-3 (points). Controls are seed means in the first row and single runs otherwise. Gaps use unrounded scores; Table 16 gives the full control configurations for each component comparison.

Component removed	Qwen2.5-VL 7B	Qwen2.5-VL 3B	Qwen3-VL 8B	Qwen3-VL 4B	Qwen3-VL 2B
CF reward and data renewal	+1.62	+3.30	+2.21	+1.36	+3.36
Regeneration	−0.13	+10.29	+1.01	+0.34	+3.98
Counterfactual signal	+0.40	+6.69	+1.14	−0.45	+1.67
Fitness-guided selection	+0.45	−1.41	+1.05	−0.67	+0.97
Diagnostic replay term w_c	+0.37	−0.31	+0.97	+0.20	+1.48

Reward strength. The full Qwen2.5-VL-7B configuration retains its advantage across a fourfold shaping-strength range. With diagnostic replay enabled ($w_c = 1$), $\lambda = 0.05$, 0.10, and 0.20 yield endpoint Macro-3 accuracies of 70.41, 70.88, and 71.51, compared with the outcome-only baseline mean of 69.26. The default strength and baseline summarize four runs each; the alternative strengths each use one run. Full-suite strength and scorer comparisons appear in Table 39 and Figure 10.

Extended training. Extending training to 420 steps and 14 rounds preserves the advantage in a single-run comparison from the same Qwen2.5-VL-3B initialization. CoSEC reaches 63.29 versus 61.00 for outcome-only GRPO on the four-benchmark Macro-4, with a median response length of 450 characters (Appendix D.3).

7 Conclusion

CoSEC uses shared counterfactual diagnostics to connect grounding rewards with the selection of future training examples. Combined with pass-rate control, this feedback drives a curriculum of constructively labeled scenes without human annotation or real training images during post-training. Across five model scales in two families, CoSEC improves endpoint and training-average Macro-3 accuracy over outcome-only static GRPO on out-of-distribution benchmarks. Visual probes link learning on generated tasks to available image evidence, while normalized-reward comparisons show that data renewal complements reward design. These results support joint design of the reward and the evolving training distribution as a practical route to grounded VLM reinforcement learning.

REFERENCES

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9687–9695. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00971. URL https://doi.org/10.1109/CVPR42600.2020.00971.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled reward mitigates hacking in RLHF. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7935–7952. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24bn.html.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for LLM reasoning. In *Conference on Language Modeling (COLM)*, 2026. URL https://www.microsoft.com/en-us/research/publication/self-evolving-curriculum-for-llm-reasoning/.

Lang Feng, Weihao Tan, Zhiyi Lyu, Longtao Zheng, Haiyang Xu, Ming Yan, Fei Huang, and Bo An. Towards efficient online tuning of VLM agents via counterfactual soft reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16884–16903. PMLR, 2025. URL https://proceedings.mlr.press/v267/feng25n.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer Nature Switzerland, 2024. doi:

10.1007/978-3-031-73337-6_9. URL https://link.springer.com/chapter/10.1007/978-3-031-73337-6_9.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Ruozhen He, Nisarg A. Shah, Qihua Dong, Zilin Xiao, Jaywon Koo, and Vicente Ordonez. Beyond referring expressions: Scenario comprehension visual grounding, 2026a. URL https://arxiv.org/abs/2604.02323.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. VisPlay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26274–26284, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/He_VisPlay_Self-Evolving_Vision-Language_Models_CVPR_2026_paper.html.

Haojie Huang, Xinlei Yu, Chengming Xu, Zhangquan Chen, Cheng Yang, Qingdong He, Yu Yang, Jiangning Zhang, and Xiaobin Hu. Evidence-RL: Towards evidence-intensive visual reasoning, 2026a. URL https://arxiv.org/abs/2608.08021.

Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting MLLMs' video understanding via counterfactual video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8153–8163, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Huang_Taming_Hallucinations_Boosting_MLLMs_Video_Understanding_via_Counterfactual_Video_Generation_CVPRF_2026_paper.html.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Rui Liu, Dian Yu, Haolin Liu, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, and Leoweiliang. Reinforcing multimodal reasoning against visual degradation, 2026. URL https://arxiv.org/abs/2605.09262.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287, 1999. URL https://people.eecs.berkeley.edu/~russell/papers/icml99-shaping.pdf.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705. IEEE, 2021. doi: 10.1109/CVPR46437.2021.01251. URL https://doi.org/10.1109/CVPR46437.2021.01251.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, pp. 79791–79821, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c66e1fcc9691aae706250638f36f681b-Abstract-Conference.html.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167625–167661, 2025. doi: 10.52202/085713-5049. URL https://doi.org/10.52202/085713-5049.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, pp. 1279–1297. Association for Computing Machinery, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471. Curran Associates, Inc., 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Meghana Sunil, Manikandarajan Venmathimaran, and Muthu Subash Kavitha. IREASONER: Trajectory-aware intrinsic reasoning supervision for self-evolving large multimodal models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29361–29372. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1468. URL https://aclanthology.org/2026.findings-acl.1468/.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman M. Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. EvoLMM: Self-evolving large multimodal models with continuous rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026. URL https://arxiv.org/abs/2511.16672v4.

Shravan Venkatraman, Ritesh Thawkar, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Salman Khan, and Fahad Shahbaz Khan. Paying more attention to visual tokens in self-evolving large multimodal models. In *Computer Vision – ECCV 2026*, volume 17035 of *Lecture Notes in Computer Science*, pp. 41–59. Springer, 2026. doi: 10.1007/978-3-032-37464-6_3. URL https://link.springer.com/chapter/10.1007/978-3-032-37464-6_3.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-evolution via multi-agent self-play. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 106716–106744, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ae4893b2b9a4a7623e271c4e3443f17c-Paper-Conference.pdf.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 80502–80530, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/82846e19e6d42ebfd4ace4361def29ae-Abstract-Conference.html.

Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, Tianyang Wang, and Hao Xu. Staying VIGILant: Mitigating visual laziness via counterfactual visual alignment in MLLMs. In *Computer Vision – ECCV 2026*, volume 17014 of *Lecture Notes in Computer Science*, pp. 287–305. Springer, 2026. doi: 10.1007/978-3-032-37232-1_16. URL https://link.springer.com/chapter/10.1007/978-3-032-37232-1_16.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual policy optimization for multimodal reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65634.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, pp. 89655–89701, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

Metric conventions for supplementary results. Unless explicitly labeled Macro-3, aggregate accuracy, drawdown, and statistical tests in the following appendix sections use the original four-benchmark Macro-4. Per-benchmark scores, visual-probe results, reward diagnostics, and recorded timing retain their stated definitions.

A STEP-210 SUMMARY OF THE MAIN ARMS

Each row reports step-210 accuracy averaged over the same n independent runs for all four benchmarks. MME-RW denotes MME-RealWorld; Macro-4 is the unweighted mean of the four accuracies, and Δ_{base} is its gain over the untrained base. Bold rows identify the default CoSEC configuration. Machine-readable per-run and intermediate-step results accompany the paper.

Table 7: Qwen2.5-VL-7B, step-210 accuracy (%); untrained Macro-4 63.55.

Arm	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ_{base}
CoSEC (self-scorer)	1	56.87	87.41	46.17	85.68	69.03	+5.48
CoSEC (calibrated)	1	60.03	84.67	44.03	86.74	68.87	+5.32
control (w_c=0,λ=0.05)	1	60.15	87.36	41.58	85.84	68.74	+5.19
static+grounding	1	60.15	86.86	43.20	82.98	68.30	+4.75
CoSEC (lennorm)	1	59.90	87.03	42.11	83.96	68.25	+4.70
CoSEC (DAPO)	1	61.66	85.17	41.06	84.45	68.09	+4.54
CoSEC (roundmean)	1	61.41	84.46	41.22	84.62	67.93	+4.38
CoSEC	**4**	**58.92**	**84.74**	**42.76**	**85.15**	**67.89**	**+4.34**
ScenGround	1	60.15	84.54	42.26	84.37	67.83	+4.28
CoSEC (scoremean)	1	58.13	82.94	44.50	85.43	67.75	+4.20
SelfEvolve-VT	1	57.38	86.14	43.77	83.63	67.73	+4.18
no-counterfactual	1	59.27	86.77	40.23	84.45	67.68	+4.13
control (w_c=0,λ=0.20)	1	59.02	84.12	43.04	84.45	67.66	+4.11
control (w_c=0,λ=0.10)	1	58.51	84.67	41.27	85.60	67.51	+3.96
random-mutation	1	58.13	83.74	42.05	85.52	67.36	+3.81
EvoLMM	1	55.23	85.26	42.89	83.63	66.75	+3.20
static baseline	4	57.82	84.47	41.49	81.81	66.40	+2.85
static baseline(DAPO)	1	56.87	86.14	40.80	80.77	66.15	+2.60
iReasoner	1	56.87	83.91	41.38	81.91	66.02	+2.47
VisPlay	1	56.49	71.10	42.11	44.19	53.47	−10.08

Table 8: Qwen2.5-VL-3B, step-210 accuracy (%); untrained Macro-4 51.10.

Arm	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ_{base}
random-mutation	1	54.73	84.04	39.86	83.06	65.42	+14.32
CoSEC (lennorm)	1	57.38	81.89	37.78	82.57	64.90	+13.80
CoSEC (scoremean)	1	55.86	84.79	37.31	81.59	64.89	+13.79
control (w_c=0,λ=0.10)	1	55.11	83.07	38.98	81.59	64.68	+13.58
CoSEC (roundmean)	1	53.22	84.79	38.93	81.42	64.59	+13.49
CoSEC	**4**	**52.96**	**82.73**	**39.42**	**80.56**	**63.92**	**+12.82**
EvoLMM	1	58.39	77.59	38.41	81.10	63.87	+12.77
static+grounding(scoremean)	1	54.73	77.38	39.71	80.20	63.00	+11.90
static+grounding(lennorm)	1	57.38	72.16	39.76	80.69	62.50	+11.40
static baseline	4	55.80	79.21	36.36	77.25	62.16	+11.06
ScenGround	1	55.86	79.06	36.27	77.25	62.11	+11.01
SelfEvolve-VT	1	52.08	72.79	41.22	78.56	61.16	+10.06
CoSEC (self-scorer)	1	50.44	82.81	36.69	74.06	61.00	+9.90
static+grounding(roundmean)	1	55.99	77.63	36.11	72.50	60.56	+9.46
iReasoner	1	52.46	80.37	33.87	72.26	59.74	+8.64
no-counterfactual	1	52.33	81.09	34.71	66.86	58.75	+7.65
VisPlay	1	52.33	70.13	39.03	69.97	57.87	+6.77
static+grounding (draw 1)	1	53.72	58.72	36.37	76.76	56.39	+5.29
static+grounding (draw 2)	1	52.84	48.90	36.43	8.92	36.77	−14.33

Table 9: Qwen3-VL-8B, step-210 accuracy (%); untrained Macro-4 66.72.

Arm	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ_{base}
iReasoner	3	62.51	86.06	44.78	83.72	69.26	+2.54
CoSEC (scoremean)	1	61.16	86.44	45.49	82.57	68.91	+2.19
SelfEvolve-VT	1	62.55	84.67	44.45	83.96	68.91	+2.19
CoSEC	**4**	**61.07**	**85.57**	**44.68**	**84.02**	**68.84**	**+2.12**
static+grounding	1	63.30	83.53	44.50	83.22	68.64	+1.92
no-counterfactual	1	63.18	85.43	42.21	83.22	68.51	+1.79
CoSEC (DAPO)	4	63.59	84.38	41.43	84.27	68.42	+1.70
random-mutation	1	62.30	84.46	42.31	84.37	68.36	+1.64
CoSEC (calibrated)	1	60.15	84.63	44.71	83.80	68.32	+1.60
VisPlay	1	61.29	85.05	43.62	83.06	68.25	+1.53
static baseline(DAPO)	1	62.42	83.78	41.64	84.78	68.15	+1.43
CoSEC (self-scorer)	1	62.55	84.04	43.04	82.73	68.09	+1.37
EvoLMM	1	61.54	85.21	43.41	82.08	68.06	+1.34
control (w_c=0,λ=0.10)	1	60.78	83.82	43.51	84.04	68.04	+1.32
CoSEC (calibrated,quantile)	1	59.39	83.57	44.66	84.21	67.96	+1.24
static baseline	4	62.33	84.11	41.35	82.20	67.50	+0.78
ScenGround	1	61.41	84.16	39.86	83.96	67.35	+0.63

Table 10: Qwen3-VL-4B, step-210 accuracy (%); untrained Macro-4 66.69.

Arm	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ_{base}
random-mutation	1	63.68	84.88	45.65	83.80	69.50	+2.81
no-counterfactual	1	63.81	86.14	45.13	82.41	69.37	+2.68
CoSEC (calibrated,quantile)	1	61.54	85.43	46.22	82.41	68.90	+2.21
CoSEC (lennorm)	1	61.92	85.30	45.02	82.65	68.72	+2.03
CoSEC (scoremean)	1	62.42	85.38	44.19	82.82	68.70	+2.01
CoSEC	**4**	**62.39**	**85.30**	**44.24**	**82.77**	**68.68**	**+1.99**
SelfEvolve-VT	4	62.32	84.48	43.91	82.59	68.33	+1.64
CoSEC (DAPO)	1	62.80	83.36	44.87	83.06	68.52	+1.83
control (w_c=0,λ=0.10)	1	61.79	84.08	44.09	83.55	68.38	+1.69
static+grounding	1	62.04	84.37	44.35	82.57	68.33	+1.64
CoSEC (calibrated)	1	60.40	85.26	45.23	82.08	68.24	+1.55
EvoLMM	4	62.45	84.55	43.88	82.08	68.24	+1.55
VisPlay	4	61.38	85.08	42.86	82.82	68.03	+1.34
ScenGround	4	62.45	83.89	41.46	83.00	67.70	+1.01
static baseline	4	63.18	83.90	41.39	82.96	67.86	+1.17
static baseline(DAPO)	1	62.80	83.99	41.79	82.65	67.81	+1.12
iReasoner	4	58.89	85.58	43.32	82.16	67.49	+0.80

Table 11: Qwen3-VL-2B, step-210 accuracy (%); untrained Macro-4 59.77. Counterfactual-reward arms use a frozen Qwen3-VL-2B scorer initialized from the same base as the policy (§6.3).

Arm	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ_{base}
CoSEC (scoremean)	1	54.85	77.76	41.43	73.73	61.94	+2.17
CoSEC	**6**	**52.14**	**77.84**	**42.77**	**73.57**	**61.58**	**+1.81**
static baseline(DAPO)	9	54.15	77.30	41.93	72.16	61.39	+1.62
random-mutation	1	52.21	78.39	43.15	69.72	60.87	+1.10
CoSEC (DAPO)	9	55.74	74.38	41.99	70.47	60.64	+0.87
no-counterfactual	1	52.33	75.23	42.16	71.77	60.37	+0.60
control (w_c=0,λ=0.10)	1	50.06	77.72	42.47	69.56	59.95	+0.18
VisPlay	1	49.81	77.97	41.64	68.17	59.40	−0.37
static+grounding	1	53.72	77.93	42.21	62.11	58.99	−0.78
static baseline	4	51.83	75.79	42.97	65.34	58.98	−0.79
iReasoner	1	47.16	76.16	44.24	65.06	58.16	−1.61
ScenGround	1	53.22	75.11	41.12	62.77	58.05	−1.72
SelfEvolve-VT	1	51.58	78.22	42.42	59.08	57.83	−1.94
EvoLMM	1	47.92	78.22	41.79	39.28	51.80	−7.97

Concurrent-method comparison. The primary concurrent comparison uses Qwen3-VL-4B, where three concurrent arms each have four independent draws. Within-arm seed spread is 0.60–1.36 points and the evaluation jitter floor is 0.10. The expanded 8B iReasoner comparison contains three runs; per-arm counts are given in Tables 7–11.

B BEST-STEP CRITERION (SECONDARY)

Selecting the best of the eight evaluation steps gives seed-mean gains of $+1.59/+1.89/+1.36/+0.85/+0.17$, averaging $+1.17$ (5/5 positive). Best-step selection increases fixed-data accuracy by up to $+3.11$ points, at the 2B static baseline, compared with at most $+0.68$ for CoSEC; mean uplifts are $+0.87$ and $+0.43$, respectively. Both endpoint and best-step criteria favor CoSEC across scales. The primary analysis uses step 210, fixed before training, to evaluate sustained performance without selecting among intermediate checkpoints.

Comparing CoSEC's best seed against the weakest static-baseline draw gives an endpoint gap of $+3.48$, whereas the seed-mean comparison gives $+1.60$. Averaging over seeds removes this extremal selection; for example, the 2B VSR gap changes from $+19.31$ to $+8.22$. The main text consistently reports seed means.

C CASE STUDIES AND MECHANISM DETAILS

Case study 1: a worked counterfactual triple. Consider a counting item with three red squares and two blue squares. The question is "How many red squares are there?", with answer $a{=}3$ computed from the scene graph. The semantic edit I^{cf} repaints one red square blue, changing the answer to 2. The nuisance edit I^{nu} jitters positions and rerenders the background while preserving the red count at 3. Scoring the same response "3" under all three views measures whether its support decreases after the answer-changing edit and remains stable after the label-preserving edit. These contrasts define $g = \ell_o - \ell_{cf}$ and $e = |\ell_o - \ell_{nu}|$ in Eq. 2. Scene geometry supplies the labels and edits for each regenerated round, making the diagnostic available to both reward shaping and sample selection for the following round.

Case study 2: fixed-data training dynamics. The two fixed-data Qwen2.5-VL-3B runs illustrate the length channel in §6.1 (Figure 18, Table 28). Both track the outcome-only baseline through the first three rounds, at Macro-4 $\approx$62 around steps 60–90. Median output length then grows from 319 characters to 648 and 2076, while endpoint accuracy falls to 56.39 and 36.77; the latter is below the untrained base's 51.10. In the collapsed checkpoint, short responses score 59.4% versus 21.2% for long responses, compared with 63.3% before collapse. The outcome-only fixed-data control ends at 62.26 with length $1.04\times$ its initial value; the same grounding reward under regeneration ends at 63.73 with length $1.24\times$. These contrasts associate late collapse with repeated optimization of the grounding reward on a fixed training set.

Case study 3: detecting template shortcuts. A generated round exceeded the color-template/answer conditional mutual-information threshold: 0.033 against 0.020. The round 1 gate stopped the run before that dataset entered subsequent training; it was excluded from completed-run analyses. This example illustrates how the gate prevents detected template shortcuts from propagating through later rounds.

Case study 4: adaptation after difficulty saturation. Difficulty and family allocation adapt at different levels. One control run reached $d = 1.0$ in all three families within seven rounds. After the difficulty bound is reached, regeneration can still supply new scenes, and archive replay with local mutation can change sample composition. Data renewal therefore continues even when the scalar difficulty no longer increases; Figure 11 shows this distinction across the primary model scales.

Diagnostic thresholds and replay selection. The default criterion applies the same absolute threshold to a two-sided near-zero band on g and a one-sided upper tail on nuisance displacement. These rules can have different hit rates. On the main arm, groups with an exactly zero counterfactual channel account for 72.4%; the per-sample diagnostic rates have Spearman correlation -0.06.

Removing the diagnostic replay term changes the parent-sampling distribution by total variation 0.002–0.011. A paired criterion increases this shift to 0.06–0.35 and brings the diagnostic hit rates into a comparable range. Absolute thresholds applied to summed log-probabilities depend on score scale, including response length. These diagnostics characterize which samples the archive favors and help interpret the w_c ablation.

D MECHANISM ANALYSIS AND CONTROLLED INTERVENTIONS

Divisor variants and extended training examine the length-channel account of §6. Unadjusted archive tests and component definitions are reported here with the original Macro-4 metric.

D.1 REWARD LEVELS UNDER DATA REGENERATION

Reward trajectories provide a complementary view of the Macro-4 archive comparison in Table 42. Regeneration changes the examples used to optimize the reward after each round (Figure 7). Table 12 reports absolute training-reward levels. For the accompanying reward–accuracy diagnostic, divergence is a reward increase of at least 0.02 with a Macro-4 decrease of at least 2 points from the best evaluation step. Its interpretation depends on the training distribution and reward scale.

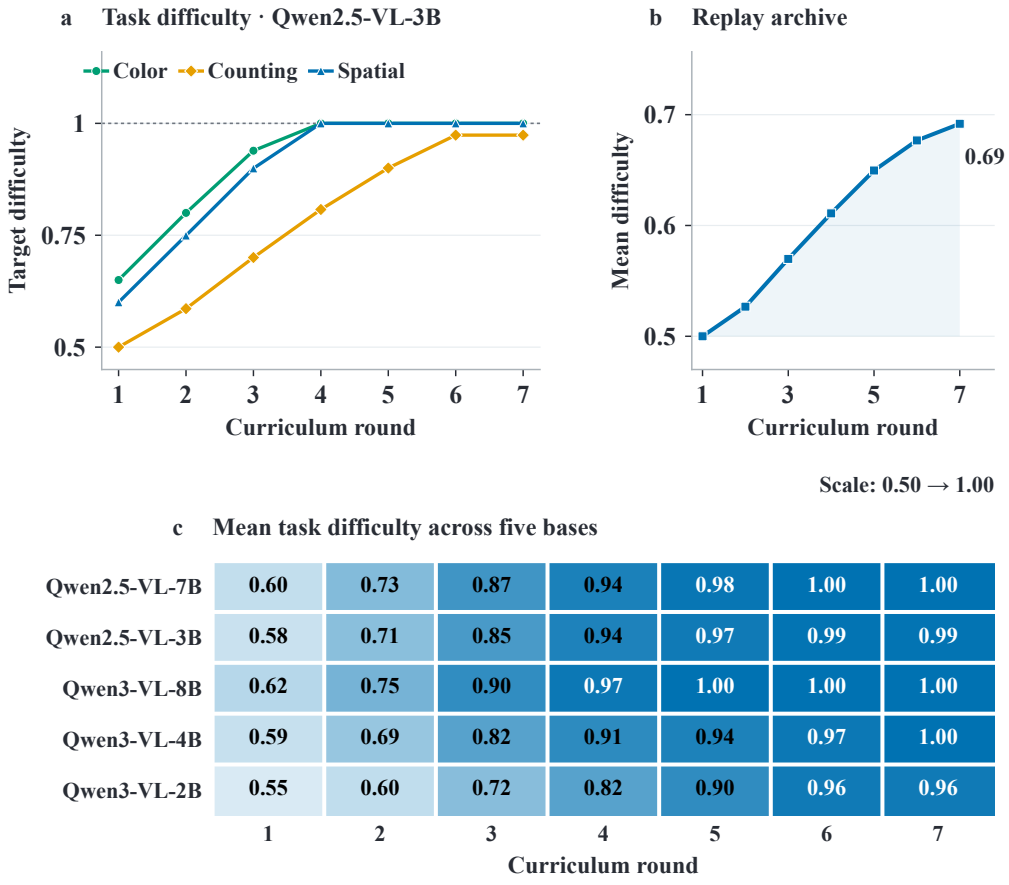


Figure 7: Difficulty adaptation across training rounds. **a**, Task-specific target difficulty for a representative Qwen2.5-VL-3B closed-loop run. **b**, The same run's replay-archive mean difficulty rises from 0.50 to 0.69. **c**, Mean target difficulty across the three task families for one representative run per base; values approach the upper bound across all five bases.

Table 12: Recorded training-reward levels across archived configurations. The endpoint threshold is 0.90; dataset distributions and reward definitions differ across runs. Counts summarize these traces rather than a common-scale measure of reward exploitation.

Group	n	reward @30	reward @210	mean curve max	**endpoint** > 0.90	$\Delta(210-30)$
closed-loop regeneration	64	0.861	0.753	0.918	**0/64**	-0.108
fixed data	29	0.545	0.852	0.944	9/29	$+0.307$
concurrent reproductions	34	0.737	0.898	0.943	16/34	$+0.160$

Training reward averages both the current policy and its data distribution: $\bar{R}_t = \mathrm{E}_{x\sim D_t, y\sim\pi_t}[R(x,y)]$. Regeneration changes D_t, and the archive also contains different reward constructions. The 0.90 threshold therefore describes the recorded reward scale, rather than a common measure of reward exploitation. Table 12 retains these trajectory summaries; OOD accuracy and drawdown provide the shared evaluation metric. The representative same-formula trajectories in Figure 8 locate reward growth alongside accuracy loss on fixed data.

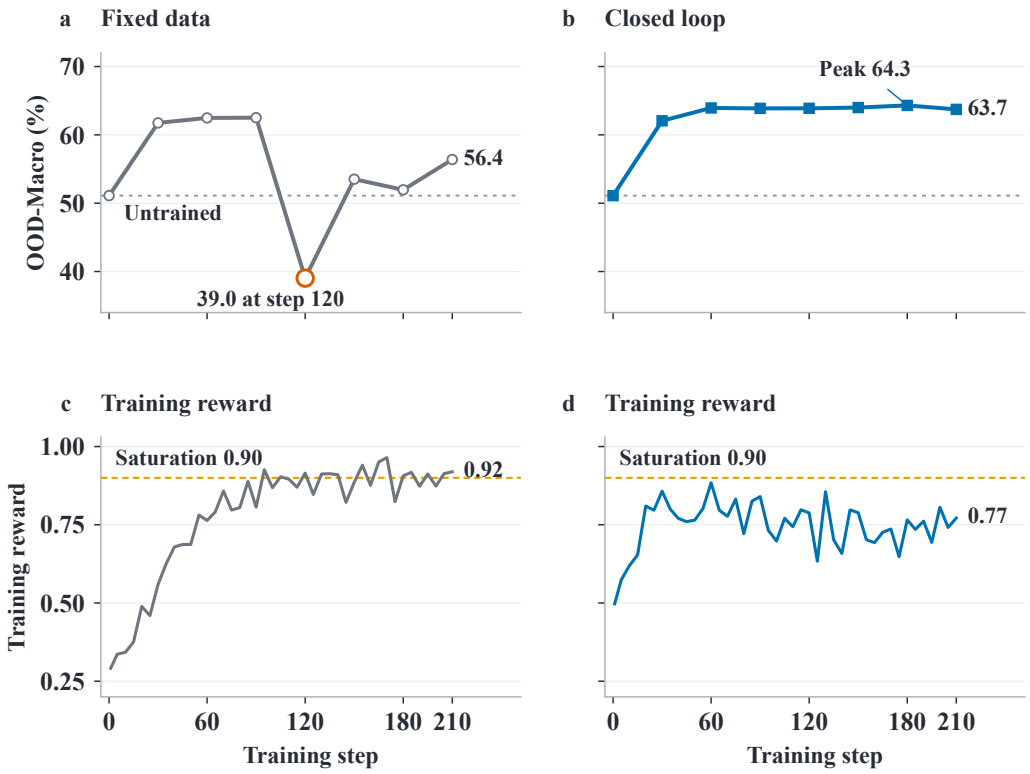


Figure 8: Accuracy and reward under two data regimes with the same counterfactual-grounding reward (Qwen2.5-VL-3B). **a,b**, Macro-4 trajectories; dotted lines mark the untrained base. Fixed-data accuracy falls to 39.0 at step 120 and ends at 56.4; closed-loop accuracy ends at 63.7, close to its 64.3 peak. **c,d**, Corresponding training rewards, with the 0.90 reference; the final-step rewards are 0.92 for fixed data and 0.77 for the closed loop.

D.2 Normalization and data-renewal comparisons

Table 13 compares the fixed scaling scalar with response-dependent and history-dependent divisors, on fixed and regenerated data. The contrasts test how normalization changes the observed length and accuracy behavior.

Table 13: Divisor intervention (Qwen2.5-VL-3B; length ratios are token-based here, so the two runs at $2.03\times$ / $6.51\times$ in §6.1's character metric read as $2.13\times$ / $5.15\times$). Response-dependent divisors produce shorter outputs; the history-based divisor retains length growth on fixed data. These are individual-run comparisons of normalization and data regime.

Arm	Divisor S	Data	len. ratio	drawdown	step-210 OOD
`static+grounding`, draw 1	64 (const.)	fixed	$2.13\times$	6.12	56.39
`static+grounding`, draw 2	64 (const.)	fixed	$5.15\times$	**25.57**	**36.77**
`static+grounding`(lennorm)	$\lvert y\rvert$	fixed	$\mathbf{0.04}\times$	1.90	**62.50**
`static+grounding`(scoremean)	per-token mean	fixed	$0.12\times$	1.03	63.00
`static+grounding`(roundmean)	history EMA	fixed	$2.44\times$	2.07	60.56
CoSEC	64 (const.)	**regenerated**	$1.36\times$	0.58	63.73
CoSEC(roundmean)	history EMA	**regenerated**	$1.14\times$	**0.00**	64.59
CoSEC(lennorm)	$\lvert y\rvert$	**regenerated**	$0.15\times$	0.20	**64.90**
CoSEC(scoremean)	per-token mean	**regenerated**	$0.27\times$	0.01	64.89

Replacing $S = 64$ with the response's whitespace-token count $|y|$ yields shorter outputs ($319 \to 10$ characters) and endpoint accuracy 62.50 on fixed data. Per-token mean log-probabilities give 63.00. The corresponding regenerated-data runs reach 64.90 and 64.89, respectively. These observations support compatibility between normalization and renewal in the tested 3B configurations; shorter output is not itself a task failure. Normalization also changes the effective shaping strength, so these contrasts jointly alter scale and length dependence.

The `roundmean` implementation uses the preceding responses' length EMA as its divisor (EMA rate 0.01), falling back to 64 before any history exists. Despite the historical arm name, the divisor is updated across responses rather than held fixed for a whole round. On fixed data it gives length amplification $2.44\times$, drawdown 2.07, and endpoint accuracy 60.56; under regeneration the corresponding values are $1.14\times$, 0.00, and 64.59. This comparison is consistent with the data-renewal effect observed under the fixed divisor.

For a fixed positive numerator, division by response length favors shorter outputs. The numerator also evolves during training. Table 13 reports the resulting accuracy and length behavior directly; the primary multi-run comparison uses a fixed scalar, while normalized rewards provide separate single-run controls for the divisor comparison.

D.3 Stability over an extended training horizon

We extend one run per arm from the same Qwen2.5-VL-3B initialization to 420 steps / 14 rounds, doubling the primary budget. The CoSEC curriculum seed is 3407; fresh scenes and archive replay supply each renewed training batch.

Both arms sustain accuracy over the extended horizon. CoSEC ends at 63.29 against 61.00 ($+2.29$ points), with median output length 450 characters ($1.41\times$ its initial value). Table 14 reports peak checkpoints as a secondary diagnostic. This single-run comparison extends the observed advantage to the tested training budget.

Table 14: Long-horizon extension (Qwen2.5-VL-3B, CoSEC curriculum seed 3407, 420 steps / 14 rounds). Both arms stay far above the untrained base (51.1) and the 36.77 fixed-data Macro-4 endpoint in Table 28; CoSEC holds a $+2.58$-point peak and $+2.29$-point endpoint lead while its median output length remains bounded through the extended training budget.

Arm	untrained	peak (step)	endpoint (420)	median len (420)
CoSEC (closed loop)	51.11	**65.15** (120)	**63.29**	450 ($1.41\times$)
`static baseline` (outcome)	51.08	62.57 (330)	61.00	338 ($1.06\times$)
CoSEC − baseline		**+2.58**	**+2.29**	

D.4 TABLES REFERENCED FROM THE MAIN TEXT

The following tables report the original Macro-4 archive tests and component contrasts for the experimental designs discussed in §6.

Table 15: Unadjusted descriptive tests over the 127 archived runs that carry both a training-reward curve and per-step OOD evaluation. Fisher's exact test for incidences, 200k-draw permutation for drawdown means; configuration mixtures differ.

Comparison	closed loop	fixed data	test	p
reward–accuracy divergence rate	**2/64**	12/63	Fisher	**0.00450**
—KL-anchored (GRPO) subset	**0/51**	11/51	Fisher	**0.00053**
endpoint below untrained base	**3/64**	11/63	Fisher	**0.02520**
peak→endpoint drawdown, mean (pts)	**0.60**	1.78	permutation 200k	**0.0033**
—vs. static-data group only	**0.60**	2.13	permutation 200k	**0.0031**
—vs. concurrent-method group only	**0.60**	1.48	permutation 200k	**0.0135**

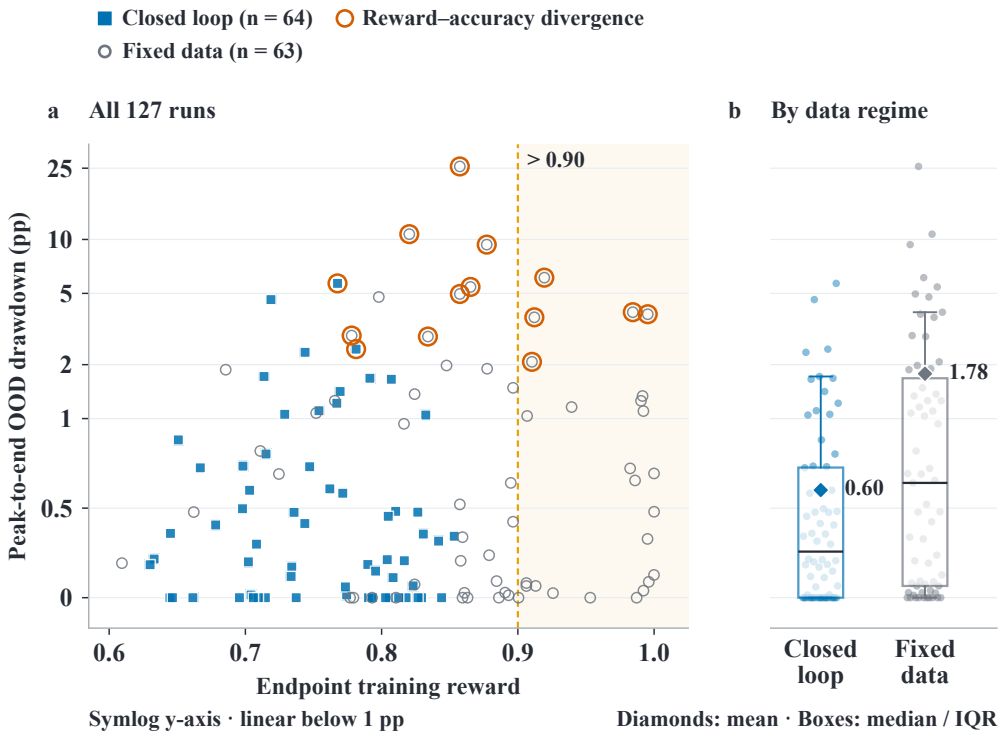


Figure 9: Recorded reward levels and accuracy drawdown across 127 runs. **a**, Each point shows one run's endpoint reward and peak-to-endpoint Macro-4 drawdown. Warm rings mark reward–accuracy divergences ($2/64$ closed-loop, $12/63$ fixed-data); reward exceeds 0.90 in $0/64$ and $25/63$ runs, respectively. **b**, The same drawdowns grouped by data regime, with median/IQR boxes, $1.5\times$ IQR whiskers, and mean diamonds (0.60 vs. 1.78 points). Both vertical axes use a symlog scale with a linear region below 1 point of drawdown.

Sensitivity to pre-protocol pilots (Table 42). Including the two 2026-08 fixed-data pilots gives divergence counts of $2/64$ against $14/65$ ($p = 0.00232$), preserving the direction of the primary 127-run comparison between the two data regimes.

Table 16: Ablation matrix: CoSEC − control at step 210 (points), with each control's position in the reward × data 2 × 2. The first row compares seed means; other controls are single runs. The 3B curriculum cell is parenthesized because that scale's fixed-data arm collapses (§6.1) and so is not usable for reward-vs-curriculum attribution.

Component removed	Control arm (2 × 2 cell)	7B	3B	8B	4B	2B	pos.
both interventions	`static baseline` (fixed, outcome)	+1.49	+1.76	+1.34	+0.82	+2.60	**5/5**
closed-loop regeneration only	`static+grounding` (fixed, grounding)	−0.41	(+7.53)	+0.20	+0.34	+2.59	4/5
counterfactual signal only	`no-counterfactual` (loop, outcome)	+0.21	+5.17	+0.33	−0.69	+1.21	4/5
fitness-guided selection	`random-mutation`	+0.53	−1.50	+0.48	−0.83	+0.71	3/5
diagnostic replay term w_c	`control` (w_c=0)	+0.38	−0.77	+0.80	+0.30	+1.63	4/5

D.5 Ablation matrix: per-scale reading

The `no-counterfactual` control uses outcome-only rewards and does not launch the scorer. It therefore removes the counterfactual diagnostic from both reward shaping and curriculum telemetry, while retaining the generator, pass-rate controller, archive, and data audits. This intervention assesses the shared diagnostic as a whole.

Regeneration and the shared signal. On fixed data, adding grounding changes endpoint accuracy by $+1.90/+1.14/+0.47$ points at 7B/8B/4B and $-5.77/+0.01$ at 3B/2B. Relative to the regenerated-data outcome-only control, the complete method changes accuracy by $+0.21/+5.17/+0.33/-0.69/+1.21$ at 7B/3B/8B/4B/2B. The combined design is the cross-scale result: it improves on outcome-only static training at every scale. Fitness-guided selection and diagnostic replay variants show how sample prioritization affects each base (Table 16); these controls each use one run.

Seeds, reward strength, and scorer configuration. The primary comparison includes at least four runs per scale, with $21/22$ CoSEC runs above their same-base baseline mean and cross-seed SD 0.65–1.08 points. In the 7B $w_c = 0$ control, sweeping $\lambda \in \{0.05, 0.10, 0.20\}$ gives endpoints $68.74/67.51/67.66$, covering a fourfold strength range under the same controller setting.

Using each policy's own frozen base as scorer gives endpoints 69.03, 61.00, and 68.09 at 7B, 3B, and 8B. The primary 2B experiment already uses this arrangement. These configurations allow a shared policy/scorer initialization while keeping scorer parameters fixed throughout policy optimization. The swaps change both scorer identity and capacity.

E Visual evidence, reward validation, and training diagnostics

The reward targets two properties of visual support: sensitivity to answer-changing edits and stability under answer-preserving edits. The following probes evaluate these properties, followed by reward-signal validation, endpoint inference, compute accounting, and response-level diagnostics.

E.1 Visual-evidence and perturbation probes

The two channels evaluate complementary responses to visual change: g measures sensitivity to an answer-changing edit, while e measures displacement under an answer-preserving edit (Eq. 2). Four probes hold the question fixed: `spectrum` uses intact images across a difficulty grid; `sham` applies label-preserving uniform scaling and translation; `blank` substitutes a blank canvas; and `shuffle` substitutes another item's image. Each contains 120 items and uses the same eight checkpoints as the primary evaluation.

Training improves intact accuracy by $+12.5$ to $+30.4$ points and nuisance-perturbed accuracy by $+14.2$ to $+23.8$ points at all $5/5$ bases (Table 17). Over 210 steps, blank-image accuracy changes by -2.8 to $+1.2$ and shuffled-image accuracy by -4.4 to $+4.2$, remaining near the initial 23–32% range. Gains are therefore concentrated in the probes that preserve task evidence.

The 23–32% range under `blank` and `shuffle` is already present before training, with little subsequent improvement. At step 210, `sham` accuracy is close to `spectrum`: 82.1 versus 86.2 at 7B and 84.6 versus 84.2 at 8B. Together, these observations support learning that uses visual evidence

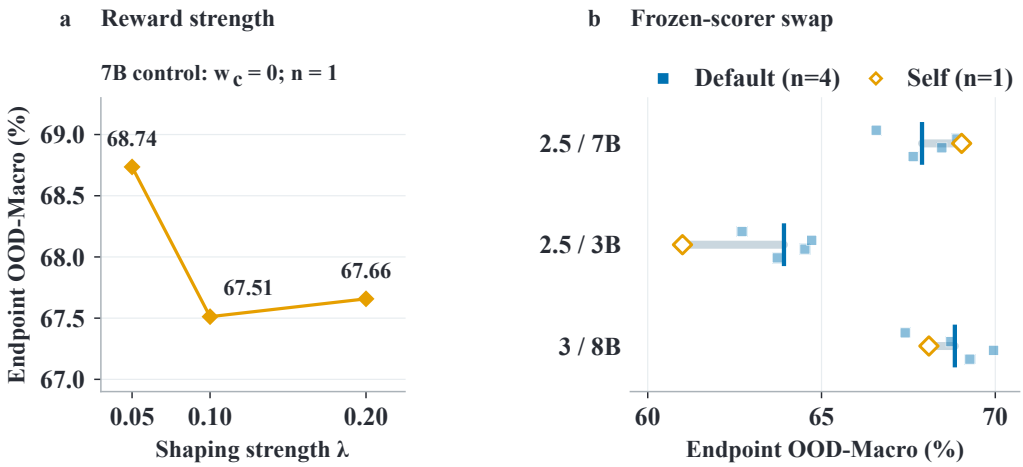


Figure 10: Reward-strength and scorer-configuration checks. **a**, Endpoint accuracy across a fourfold shaping-strength range in the 7B $w_c = 0$ control, with one run per setting. **b**, Individual primary runs under the default frozen scorer (four per scale; vertical bars mark means) and one self-scorer run per scale. The self-scorer shares the policy's initialization and remains frozen.

Table 17: Visual probe accuracy (%), step 0 → 210, on fixed cohorts: outcome-only GRPO (Baseline) and CoSEC. Every included run is evaluated at both steps on all four views of the same 120 questions. n_B and n_C denote baseline and CoSEC run counts, respectively. Values are run means; the main-text changes average within-run differences over these same cohorts.

Base	Arm	n	Intact	Nuisance	Blank	Shuffled
Qwen2.5-VL-7B	Baseline	4	55.8 → 85.4	58.3 → 81.9	30.0 → 29.4	32.5 → 29.8
	CoSEC	4	55.8 → 86.2	58.3 → 82.1	30.0 → 28.5	32.5 → 28.1
Qwen2.5-VL-3B	Baseline	4	59.2 → 83.3	60.0 → 81.5	32.5 → 29.0	28.3 → 27.3
	CoSEC	3	59.2 → 82.8	60.0 → 81.4	32.5 → 31.9	28.3 → 27.5
Qwen3-VL-8B	Baseline	4	71.7 → 80.6	65.8 → 78.5	24.2 → 23.8	25.8 → 29.0
	CoSEC	4	71.7 → 84.2	65.8 → 84.6	24.2 → 23.3	25.8 → 30.0
Qwen3-VL-4B	Baseline	4	75.0 → 82.7	67.5 → 79.6	30.0 → 31.0	27.5 → 25.8
	CoSEC	3	75.0 → 90.6	67.5 → 85.6	30.0 → 27.2	27.5 → 26.7
Qwen3-VL-2B	Baseline	4	60.8 → 76.5	60.8 → 76.7	26.7 → 27.7	26.7 → 25.4
	CoSEC	5	60.8 → 75.3	60.8 → 75.0	26.7 → 27.8	26.7 → 26.2

and tolerates the tested label-preserving perturbations. The corresponding fixed-data arms show the same qualitative probe pattern.

The dataset gates test complementary properties before training. Six generated datasets × three question families satisfy the template-normalized conditional mutual-information limit of 0.020 and question-only accuracy-lift limit of 0.100, across 36/36 slices. Per-slice majority-answer accuracy is at chance (§3.3; Appendix I). These construction checks complement the policy-level visual probes.

E.2 Quantitative validation of the semantic contrast

Before RL, a frozen Qwen2.5-VL-3B scorer evaluated Qwen2.5-VL-7B responses using the summed semantic contrast g in Eq. 2. Stage 1a reviewed visual-versus-text separation with this statistic post hoc. Stage 1c then fixed the statistic before evaluation on 360 held-out original/counterfactual image pairs, with eight responses per question (2880 total). Offline validation scores all responses; the correctness gate applies during RL.

Table 18: Quantitative validation of g. In Stage 1a, n counts visual/text question pairs; both controls share the same visual cohort. In Stage 1c, n counts questions. Mixed groups contain correct and incorrect responses; correct-response subsets contain at least two correct responses. Brackets give 95% confidence intervals.

Evaluation	n	Statistic	Estimate [95% CI]
Stage 1a: summed-score review			
Matched text controls	360	AUROC	0.892
Natural-text controls	360	AUROC	0.841
Stage 1c: confirmatory cohort			
Mixed-outcome groups	190	Pearson r	0.293 [0.233, 0.354]
Token-length adjustment	190	Partial r	0.282
Correct–incorrect within prompt	190	Mean Δg	2.155 [1.653, 2.682]
Correct-response subsets	312	$\mathrm{Var}(g) > 10^{-6}$	98.72% (308/312)

Stage 1c satisfies the prespecified correlation, variance, and length-correlation gates. The scorer contrast remains associated with correctness after within-question centering and token-length adjustment, while retaining variation among correct responses (Table 18). The within-correct variance measures availability of a nonconstant signal for group-relative optimization. Correlation pools group-centered observations; the correct–incorrect contrast averages question-level differences. Confidence intervals resample all 360 question clusters 5000 times, carrying their responses together; mixed-group statistics use the eligible groups in each resample. Section 6.1 examines how length dependence develops during training.

E.3 Exact inference for endpoint gains

The base-blocked null permutes CoSEC/baseline labels within each base and averages the five endpoint accuracy gaps. Across 5,042,100,000 assignments, 2592 statistics are at least the observed $+1.6032$ and 2788 have absolute value at least as large. Thus the one-sided p is 5.14×10^{-7} and the original four-benchmark two-sided p is 5.53×10^{-7}. Dropping 2B leaves 24,010,000 assignments and two-sided $p = 1.88 \times 10^{-5}$.

Leave-one-base-out mean gaps are $+1.63$, $+1.56$, $+1.67$, $+1.80$, and $+1.35$ when dropping 7B, 3B, 8B, 4B, and 2B, respectively; all remaining $4/4$ scales retain positive gains. At step 0, the arms start from identical weights and the mean gap is $+0.0012$, with per-base values from -0.09 to $+0.07$. This initial arm-to-arm variation is below 0.1 points, smaller than the endpoint effects.

The cross-scale mean gap develops over training: $+0.00$, $+0.12$, $+0.56$, $+0.81$, $+0.72$, $+1.24$, $+1.28$, and $+1.60$ from steps 0 through 210. Each of the final three evaluations is positive at $5/5$ bases. The fixed-endpoint gain therefore reflects a sustained late-training advantage.

E.4 Training compute and scorer cost

Table 34 reports recorded training-step cost over 210 updates in the primary runs with complete timing traces. CoSEC uses approximately 1.00–1.08× the baseline's recorded policy-token count. Summed step time is 3.07–4.44× the baseline value. The logged reward phase, including frozen-scorer calls, accounts for 60.3–77.9% of CoSEC's recorded step time. Rollout generation denotes policy response sampling. The timing scope comprises logged training steps; evaluation, initialization, and round-boundary dataset refresh lie outside these timers. For the reward-excluded comparison, we subtract each run's logged reward time from its step time in both arms, average within each base and arm, and divide the CoSEC mean by the baseline mean. The resulting ratios are 1.11, 0.87, 1.09, 1.07, and 1.59 for 7B, 3B, 8B, 4B, and 2B, respectively. Their cross-base median is 1.09 and equal-base mean is 1.15. This decomposition isolates the remaining logged phases; Table 34 reports total and reward times separately.

E.5 RESPONSE-LEVEL COLLAPSE DIAGNOSTICS

§6.1 characterizes the fixed-data collapse by length and accuracy. Its content is visible directly in the responses. Take one counting item (`visual_counting_0038`, "How many red triangles are in the image?", answer 8) under the collapsing `static+grounding` draw and under the closed loop at the same step:

> **Fixed data + reward, step** 90 **(before collapse),** 46 **tokens.** "There are 9 red triangles in the image. I counted each distinct red triangle shape, which includes those that are fully visible and those that are partially obscured. … FINAL: 9"
>
> **Fixed data + reward, step** 210 **(after collapse),** 477 **tokens /** 2780 **characters.** "There are 8 red triangles in the image. They are distributed throughout the image in various positions and sizes. … [2330 characters] … The image does not show any other orange shapes that could be mistaken for diamonds. The image does not show any other orange shapes that could be mistaken for diamonds. The image does"
>
> **Closed loop, step** 210**,** 10 **tokens.** "There are 8 red triangles in the image. FINAL: 8"

The collapsed response opens with the correct count, then repeats two boilerplate sentences four times each across 2780 characters and runs out of budget without ever emitting the answer marker. The closed-loop response answers the same item in 10 tokens and marks it. On a second matched item one sentence repeats 21 times. Across matched items the medians are 47 tokens before collapse, 472 after, and 10 under the loop.

Answer completion during collapse. We measure answer-marker emission across the full 6308-row out-of-distribution evaluation grid to distinguish answer completion from endpoint accuracy:

Table 19: Answer-marker emission rate on the 6308-row OOD grid (Qwen2.5-VL-3B). Each row is one representative run. Marker emission is reported separately from alias-aware answer accuracy. It remains stable for the closed loop and the outcome-only baseline, while declining in the severely degraded fixed-data run shown in this table.

Arm	step 90	step 210	Macro-4 @ 210
CoSEC (closed loop)	99.9%	**99.8%**	63.73
`static baseline` (outcome only)	96.4%	99.6%	62.26
`static+grounding`, draw 1	91.1%	90.5%	56.39
`static+grounding`, draw 2	99.7%	**46.1%**	36.77

CoSEC and the outcome-only baseline retain 99.6–99.8% answer emission at the endpoint (Table 19). The severe fixed-data draw falls to 46.1%, with over half its responses reaching the length cap without an answer marker. The response examples show how repetition can displace answer completion as output length grows.

The two fixed-data draws differ in how accuracy deteriorates. Draw 1 retains 90.5% answer emission but reaches 56.39 accuracy; draw 2 has 46.1% emission and 36.77 accuracy. The matched regenerated-data run sustains both answer completion and accuracy under the same grounding reward and training budget.

In-distribution and transfer performance. In-distribution validation accuracy changes from 75.8% → 70.0% in the collapsed run, while Macro-4 falls to 36.77, below the untrained base. This contrast motivates tracking held-out transfer together with training reward and in-distribution accuracy over the training trajectory.

F METHOD IMPLEMENTATION DETAILS

Archive replay and mutation. Replayed samples receive a small local difficulty mutation $d_i' = \text{clip}(d_i + \varepsilon, 0, 1)$ with $\varepsilon \sim U(-0.05, 0.05)$, and generation indices are reassigned when a mutation crosses a task-mode or answer-cardinality boundary to preserve balance. The archive has capacity 4000 and, from round two, supplies ≈30% of each family's samples as replays against 70% freshly generated training examples (§3.4).

Algorithm: one round of CoSEC

input: policy π_r, generator G, frozen scorer π_s, steps per round T
$D_r \leftarrow G(\text{difficulty } d_f, \text{ category weights } w_f, \text{ archive } A)$ §3.3
audit D_r: conditional MI / template lookup / question-only §3.3
for $t = 1 \dots T$:
  sample rollouts $\{y\} \sim \pi_r(\cdot \mid I, q)$, n per prompt
  $o \leftarrow$ answer match; if $o = 1$, score (g, e) with π_s Eq. 2
  $R_\lambda \leftarrow o + \mathbf{1}[o = 1]\,(\lambda/S)\,(g - \beta e)$; one GRPO step Eq. 3
telemetry $\mathcal{T}_r \leftarrow \{\text{pass rates } p_f, \text{ per-sample } (g, e)\}$
update d_f from $\hat{p}_f$ and w_f from $L(p_f)$; $F_i \leftarrow L(p_i) + w_c\, p_i\, c_i$ Eqs. 8–10
$A \leftarrow$ capacity-bounded archive by F; mutate replayed samples

A leakage gate firing. In the example described in Appendix C, the color-template/answer conditional mutual information is 0.033, exceeding the 0.020 threshold. The round 1 gate rejects that dataset before subsequent training; the stopped run is excluded from completed-run analyses.

Observed controller trajectories. Figure 11 shows how the curriculum adapts across all five policy scales, using complete seven-round histories. Color and spatial difficulty reach the upper bound earlier than counting in most scales, while task weights continue to evolve after difficulty saturation. These trajectories show the two controls working at different levels: difficulty tracks progress within a task family, and allocation redistributes training across families.

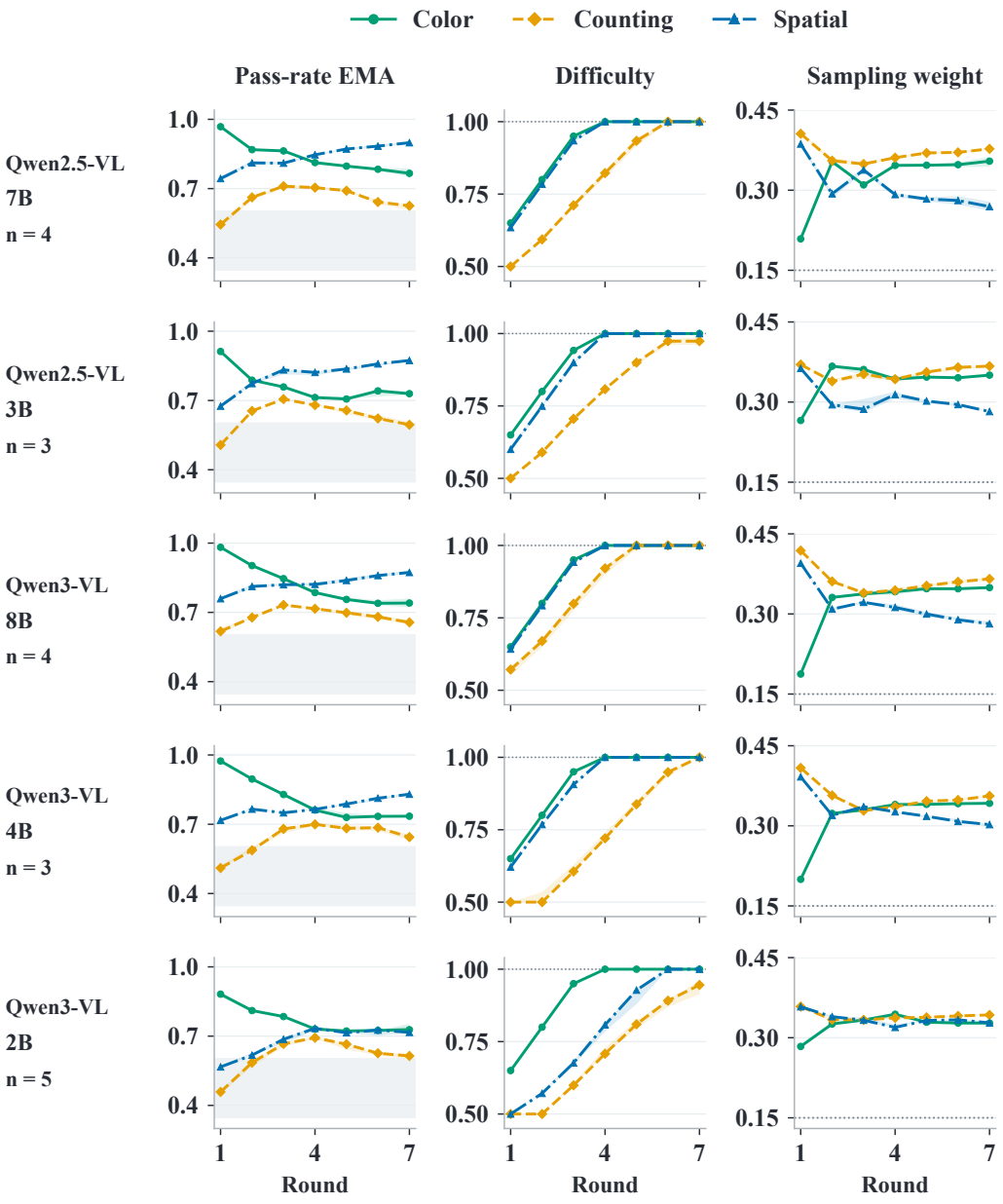


Figure 11: Curriculum adaptation across five policy scales. Columns show pass-rate EMA, difficulty, and task-family sampling weights after each round. Lines and bands denote medians and interquartile ranges of complete seven-round histories, with fixed $n = 4/3/4/3/5$ in row order. The gray band marks the configured pass-rate target $[0.35, 0.60]$; dotted lines mark the difficulty ceiling 1.0 and sampling floor 0.15. Difficulty and task allocation provide complementary adaptation over successive rounds of policy training.

G Extended related work

The following comparisons expand §2, distinguishing interventions on visual evidence, data evolution, and counterfactual credit assignment.

Counterfactual grounding rewards. Counterfactual grounding methods differ in the intervention, scoring signal, and training-data construction. PAPO (Wang et al., 2026b) adds an implicit perception loss inside GRPO: it masks image patches and measures KL divergence between the intact and masked answer distributions. This policy-derived signal encourages visual dependence on a fixed image set without a separate scorer. CFPO (Yu et al., 2026) replaces attention values to form a counterfactual internal state and regularizes the gap between factual and counterfactual policies on fixed data. Evidence-RL (Huang et al., 2026a) uses weak object-level proposals to neutralize an evidence region, then compares the support drop with matched non-evidence regions. Its Counterfactual Evidence Disentanglement diagnostic is combined with answer correctness in GRPO without question-specific evidence annotations. CoSEC builds on the use of controlled visual contrasts through constructively paired answer-changing and label-preserving edits, and reuses the diagnostics in data regeneration. VIGILant (Xiao et al., 2026) forms a blind counterfactual by zeroing cross-modal attention, then uses a DPO-style preference between grounded and blind answers on a static preference set.

Data evolution and self-play. Data evolution provides a complementary route. Duality-Forge (Huang et al., 2026b) synthesizes counterfactual videos for contrastive training under a binary reward, placing the intervention in data construction. CoSEC instead computes labels directly from scene geometry. Self-play VLMs (Thawakar et al., 2026; Sunil et al., 2026; He et al., 2026b; Wang et al., 2026a; Venkatraman et al., 2026) derive training signals from proposer/solver interactions, questioner/reasoner roles, or group consensus. We adapt EvoLMM, iReasoner, and VisPlay within our harness; Appendix H specifies the objectives and adaptations. Vision-Zero and ViSE provide broader multi-agent and self-improvement contexts. The distinction of CoSEC is to reuse a counterfactual grounding diagnostic in both policy rewards and selection of the next round's data.

A different sense of "counterfactual." CoSo (Feng et al., 2025) performs counterfactual soft RL for VLM agents by estimating each action token's contribution to return. Its counterfactual operation addresses token-level credit assignment. ROMA (Liu et al., 2026) studies multimodal reasoning under visual degradation and uses a teacher-forced KL anchor to control the influence of degraded views. These constructions address credit assignment and robustness, respectively; our diagnostic measures support for the same answer under controlled visual edits.

H Adapted methods and curriculum controls: protocols

Common setup. All adapted arms share the base checkpoint, 210-step budget, static MCQ images, and OOD evaluation grid used for `ScenGround` and `SelfEvolve-VT` (§5.2). Reward objectives and question generation are adapted below. The two intrinsic-reward methods score each response relative to the consensus of N responses for its prompt. Our batch reward manager reconstructs these groups using stable prompt identifiers. The constants below follow the released implementations.

Curriculum controls and archive labels. The fixed task schedule in Table 2 is archived as `ScenGround`. This local label differs from the scenario-grounding task introduced by He et al. (2026a). `SelfEvolve-VT` denotes the visual-token self-evolution control. Tables 7–11 retain these archive labels for run-level lookup.

EvoLMM (Thawakar et al., 2026). Label-free self-rewarding: for a prompt with $N{=}5$ sampled answers, each answer a_i receives $r_i = p_i^{0.7}\left(1 - 0.10 \cdot \mathrm{clip}((w_i - 6)/6,\, 0,\, 1)\right)$, where p_i is the fraction of the N normalized answers equal to a_i and w_i is the word count before the answer span. The reward never reads the ground-truth label—we log outcome correctness as telemetry only. On our fixed questions this reproduces the Solver reward; the Proposer half is supplied by our data.

iReasoner (Sunil et al., 2026). The adapted objective adds a trajectory-aware term to the consensus reward: $r_{\mathrm{sol}} = (1-\lambda)\, r_{\mathrm{ans}} + \lambda\, r_{\mathrm{step}}$, with $r_{\mathrm{ans}} = p_i^{0.7}(1 - 0.05\,\ell)$ (a mild penalty toward 64

words) and r_{step} the discounted cosine agreement between each reasoning-step embedding and the dominant-answer group's per-step mean. The schedule $\lambda(t)$ is 0 for the first 30% of steps, then ramps linearly to 0.7. Step embeddings come from the frozen scorer, as specified in adaptation 3 of the implementation details below.

VisPlay (He et al., 2026b). We adapt VisPlay's Image-Conditioned Questioner pipeline offline. A frozen base VLM emits a ⟨type, question, answer⟩ triple for each image and answers each question $m{=}9$ times to obtain a majority silver label and self-consistency score c. We retain $0.3 \le c \le 0.7$, following the released code (the paper's Algorithm 1 uses $0.25/0.75$), then train the Reasoner with GRPO and outcome rewards on these questions. Each arm uses its own frozen base as questioner, producing distinct self-proposed question sets at 8B and 4B.

Adaptation details. (1) EvoLMM and iReasoner contribute adapted reward objectives, while VisPlay contributes an offline question-generation procedure. Proposer/Questioner co-training is excluded, as in `SelfEvolve-VT`. (2) The 4–5-way letter answers discretize consensus mass p_i to values near $\sim\{0.2, \ldots, 1.0\}$, making $p_i^{0.7}$ coarser than with open-ended answers. (3) iReasoner uses label-free, policy-independent frozen-scorer embeddings instead of policy-side embeddings; VisPlay images are stored as on-disk paths.

Table 20: Adapted methods and curriculum controls at Qwen3-VL-4B, step-210 Macro-4 (%). Initialization, optimization steps, and evaluation are matched; objectives and supervision are specified in Appendix H. Parentheses give run counts. Gaps use unrounded seed means.

Qwen3-VL-4B arm	supervision signal	step-210 Macro-4 (n)	CoSEC − arm
`iReasoner`	label-free in-group consensus	67.49 (4)	+1.19
`ScenGround`	hand-designed category curriculum	67.70 (4)	+0.97
`VisPlay`	self-proposed silver labels	68.03 (4)	+0.64
`EvoLMM`	label-free in-group consensus	68.24 (4)	+0.44
`SelfEvolve-VT`	static visual-conditioned grounding	68.33 (4)	+0.25
CoSEC	**verifiable answer + counterfactual**	**68.68** (4)	—

Cross-scale stability. Cross-scale drawdown separates the adapted recipes more clearly than the nearby 4B endpoint means (Table 21). CoSEC's mean drawdown remains below one point at every base, reaching 0.74 at 2B; its maximum individual-run drawdown is 1.66. At 2B, the concurrent recipes have drawdowns of 1.9–9.4, and four of the five show divergence at some scale. This pattern extends the stability comparison beyond the outcome-only baseline under the adaptation protocols specified above.

Table 21: Peak-to-endpoint drawdown (points) by base for the concurrent recipes, CoSEC, and the pure-outcome static baseline, from the fixed 127-run Macro-4 stability archive summarized in Table 42; endpoint cohorts are listed separately in Appendix A. Each cell is the mean over available runs at that base (replicate coverage differs by arm); "worst" is the maximum of these five means; "div." counts the bases at which the arm diverged (reward up, Macro-4 down; §D.1). CoSEC is the only arm whose maximum per-base mean drawdown stays below one point, and has no observed divergences in these main-configuration runs.

Arm	7B	3B	8B	4B	2B	worst	div.
CoSEC	0.45	0.25	0.12	0.11	0.74	**0.74**	**0**
`ScenGround`	0.09	0.00	0.04	0.72	3.93	3.93	1
`SelfEvolve-VT`	0.07	1.48	0.00	0.01	4.77	4.77	0
`iReasoner`	0.69	1.26	0.00	0.61	3.82	3.82	1
`EvoLMM`	0.48	1.10	0.42	0.23	9.36	9.36	1
`VisPlay`	10.71	4.96	0.19	0.22	1.87	10.71	2
`static baseline`	0.00	0.07	0.34	1.16	3.11	3.11	1

Vision-Zero. Vision-Zero (Wang et al., 2026a) trains through a multi-agent "Who Is the Spy" game. We discuss this interaction-based approach in §2; the adapted experiments focus on objectives and question-generation procedures applicable to single-image MCQ training.

I Reproduction and Data Availability

Accuracy uses the common alias-aware matcher; per-step validation outputs and archived evaluation summaries supply the recorded results; the rank-correlation, length, and bucket numbers are computed from the same validation mirrors and the rollout dumps. A verification script checks every accuracy number in the paper, and a second checks the correlation, length, and bucket numbers, both against the machine-readable results committed with the artifacts. Run-directory-to-arm-label mapping and the full evaluation grid are provided with the code. We provide archived Stage-2 run logs, the 127-run subset that carries both a training-reward curve and per-step OOD evaluation, and the scripts that produce every statistical table in this paper (Fisher's exact test and 200k-draw permutation tests with the random seed fixed at 0, bit-reproducible).

I.1 Method parameters and scorer assignments

Table 22 maps the abstract scalars in §3 to the evaluated implementation. The generator uses three task families: spatial relations, counting, and color. Difficulty lies in $[0, 1]$. The smoothing and controller settings below parameterize the update equations; they are design choices rather than learned variables. The first round is freshly generated; later rounds combine fresh examples with archive replay from earlier curriculum rounds.

Table 22: Method hyperparameters and their roles. Optimizer and rollout settings are given separately in Table 23.

Symbol	Role	Experimental setting
λ	Grounding-reward strength	0.1; sweep $\{0.05, 0.1, 0.2\}$
S	Fixed reward-scaling scalar	64
β	Nuisance penalty in reward	1
K	Number of task families	3
α	Pass-rate EMA coefficient	0.5
$[p_-, p_+]$	Target pass-rate band	$[0.35, 0.60]$
$p^\star$	Target midpoint	0.475
η_d	Difficulty-controller gain	0.5
Δ_d	Maximum difficulty increment	0.15
$\omega_{\min}$	Per-family sampling floor	0.15
w_c	Diagnostic replay weight	1; control 0
M	Archive capacity	4000
ρ	Target replay fraction after round one	0.30 (subject to archive availability)
τ_g, τ_e	Diagnostic flag thresholds	0.5, 0.5 (summed-log-probability units)
ϵ	Numerical floor	10^{-6}
δ	Local mutation half-width	0.05; $\varepsilon \sim U(-\delta, \delta)$
N	Rows in each round's dataset	1200
R, T	Rounds and updates per round	7, 30

Reward and archive weights. All 22 primary CoSEC runs use $\beta = 1$, independently verified from 160 logged scores with nonzero nuisance error. The reward weight β controls nuisance displacement, while the archive weight w_c controls diagnostic-based sample selection.

Scorers. Qwen2.5-VL-3B policies use a frozen Qwen2.5-VL-7B scorer, and 7B policies use a frozen 3B scorer. All Qwen3 policies use Qwen3-VL-2B. Consequently, the default 2B experiment shares the scorer's initialization with the policy while keeping the scorer frozen. Self-scorer ablations use each policy's own frozen base. Outcome-only baselines do not invoke the scorer.

Generator and gating details. Replayed difficulty is $d'_i = \mathrm{clip}_{[0,1]}(d_i + \varepsilon)$ with $\varepsilon \sim U(-\delta, \delta)$. The offline dataset audit uses normalized template–answer mutual information at most 0.020 and

question-only accuracy lift at most 0.100. The online question-only lift limit defaults to 0.10 and is overridden to 0.15 in launch scripts, including Qwen2.5 and Qwen3 runs; the released mission files specify these overrides. The primary archive uses the original diagnostic rules $|g| < \tau_g$ and $e > \tau_e$. Calibrated diagnostic variants are separate controls. Difficulty uses the EMA and is unchanged within the target band; family allocation uses the current round's unsmoothed pass rates. The positive ϵ in Eq. 9 and replay sampling is a numerical guard, not a swept scientific parameter.

Training configuration. Table 23 specifies the primary GRPO protocol. GPU count, rollout memory fraction, and scorer assignment vary by scale as shown in Table 24. EvoLMM and iReasoner adaptations use five rollouts per prompt. The DAPO-style variant removes the KL loss and group-advantage standard-deviation normalization, and uses asymmetric clipping at $0.20/0.28$. Dynamic sampling, group filtering, and overlong-reward shaping are disabled in this variant.

Table 23: Primary GRPO training configuration. The framework is `verl` (`main_ppo`) with an FSDP actor and a vLLM rollout backend. The EvoLMM and iReasoner adaptations follow their own protocol at `rollout.n`= 5; all CoSEC and same-optimizer baseline arms use the common GRPO settings listed in this table.

Setting	Value
advantage estimator	GRPO
actor learning rate	1×10^{-6}
train / mini-batch size	64 / 32
rollouts per prompt (`rollout.n`)	8
max prompt / response length	2048 / 512 tokens
KL penalty	loss-side, `low_var_kl`, coefficient 0.01
KL added to reward	disabled
total steps (rounds $\times$ steps/round)	210 (7×30)
save / eval frequency	every 30 steps
nodes / tensor parallelism	1 / 1 (GPU count in Table 24)
rollout GPU-memory fraction	Scale-dependent (Table 24)
gradient checkpointing / padding-free	on / on

Table 24: Primary configuration by policy scale, from the supplied training-configuration cache. Memory fractions refer to the rollout engine; the frozen scorer shares the last GPU.

Policy	GPUs	Memory: CoSEC	Memory: baseline	Frozen scorer
Qwen2.5-VL-7B	4	0.45	0.60	Qwen2.5-VL-3B
Qwen2.5-VL-3B	4	0.45	0.60	Qwen2.5-VL-7B
Qwen3-VL-8B	8	0.45	0.50	Qwen3-VL-2B
Qwen3-VL-4B	4	0.40	0.45	Qwen3-VL-2B
Qwen3-VL-2B	4	0.45	0.45	Qwen3-VL-2B

Dataset construction and availability. A CPU-side procedural generator synthesizes training and validation data. PIL renders geometric primitives; templates generate spatial, counting, and color questions; scene geometry determines the answers. Each item includes its original image, an answer-changing semantic edit, and a label-preserving perturbation. This construction supplies labels directly without model-generated answers. The train and validation splits use disjoint seeds (2041 and 2042), and each regenerated round must pass the three dataset gates in §3.3. The released generator, seeds, and pinned sample pool reproduce the synthetic data without a GPU; training additionally requires the public base VLM weights and compute.

J Additional Tables and Figures

This section holds the detail floats deferred from the main text: the positioning table (§2), the per-benchmark decomposition (§5.1), the evaluation jitter floor (§4), the per-step collapse trajectory (§6.1), the per-step length–score correlation table (§6.1), the per-scope reward $\times$ data maxima, and

the per-scale length-containment table and figure. Accuracy aggregates in these archived displays use Macro-4 over the four original evaluation benchmarks.

Table 25: Positioning relative to the two closest design points (§2). CoSEC is the only one in which a single counterfactual diagnostic drives both the reward and the round-to-round data distribution.

Design point	Counterfactual signal	Drives reward	Drives next-round data
Counterfactual reward / eval (fixed data)	yes	reward *or* eval	no
Curriculum / data evolution	no (outcome-only)	no	preset / pass-rate
CoSEC (ours)	dual intervention	yes	yes (telemetry-driven)

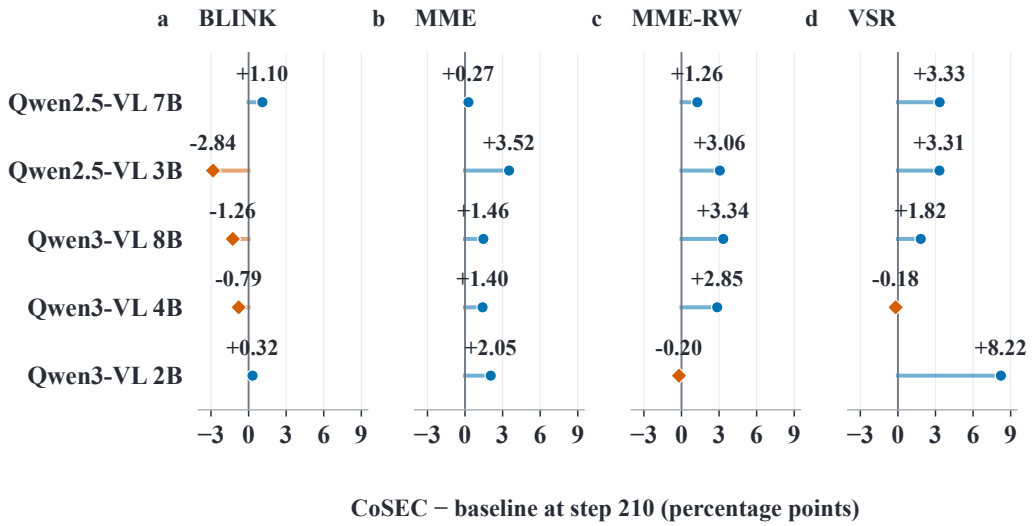


Figure 12: Per-benchmark gains (CoSEC minus the static baseline; seed means at step 210). Four panels share a percentage-point axis. Circles and warm diamonds indicate positive and negative differences, respectively. Gains are non-negative in $15/20$ comparisons and positive on MME at all five scales; seed variation is reported in §5.1.

Benchmark-level variation (§5.1). Three negative cells occur on BLINK. At 3B, the gap is -2.84: the untrained base scores 36.44, outcome-only GRPO 55.80, and CoSEC 52.96, so both trained arms improve by more than $+16$ points. At 8B the gap is -1.26 ($61.88 \rightarrow 62.33/61.07$), and at 4B it is -0.79 ($59.85 \rightarrow 63.18/62.39$). The corresponding CoSEC seed ranges are 5.04, 1.89, and 3.40 points. The other negative cells are 4B VSR (-0.18) and 2B MME-RealWorld (-0.20), with seed ranges of 0.82 and 1.82 points. Table 40 retains the full benchmark-level scores used to calculate each of these differences.

Short-output subsets. Among outputs of at most 600 characters, the mean gap is 3.67 points (Table 26). Accuracy is recomputed within each arm's subset; filtering can retain different question subsets for the two arms.

Table 26: CoSEC-minus-baseline accuracy gaps at step 210 (percentage points), within short outputs (≤ 600 characters) and across all outputs. Each gap compares seed means using the original four-benchmark Macro-4. Within the short-output subset, the advantage is larger at $5/5$ scales (mean $+3.67$ against $+1.60$ all-output). Filtering can retain different items by arm; this is not a matched-item or causal comparison.

Base	Short-output gap	All-output gap	short − all
Qwen2.5-VL-7B	+5.44	+1.49	+3.95
Qwen2.5-VL-3B	+4.26	+1.76	+2.50
Qwen3-VL-8B	+3.21	+1.34	+1.87
Qwen3-VL-4B	+2.53	+0.82	+1.71
Qwen3-VL-2B	+2.89	+2.60	+0.29
Mean	+3.67	+1.60	+2.07

Table 27: Evaluation jitter floor: spread of untrained (step-0) Macro-4 across arms that share identical initial weights.

Base	Untrained Macro-4	Spread across arms (pts)	n
Qwen2.5-VL-7B	63.55	0.27	29
Qwen2.5-VL-3B	51.10	0.16	25
Qwen3-VL-8B	66.72	0.05	28
Qwen3-VL-4B	66.69	0.10	32
Qwen3-VL-2B	59.77	0.02	41

Table 28: Fixed data + grounding reward collapses at 3B in two independent draws (Macro-4 by step); the per-step trajectory behind Figure 18. Draw 2 ends below the untrained base (51.10).

Draw	30	60	90	120	150	180	210	length 0→210
draw 1	61.8	62.5	62.5	**39.0**	53.5	52.0	**56.39**	319→648 (2.03×)
draw 2	61.6	62.3	62.3	59.9	61.2	**46.8**	**36.77**	319→2076 (6.51×)

Table 29: The length incentive is learned. Spearman(grounding score, response length) over correct-and-scored rollouts ($\approx$250–450 per step). Near zero at step 1, rising to $+0.22$–$+0.60$ by step 210. The last row is the optimizer control: the same reward under DAPO (7B) reproduces the mid-training rise ($+0.42$ at step 120).

Base	step 1	60	120	180	210
Qwen2.5-VL-7B	+0.02	+0.54	+0.60	+0.62	**+0.54**
Qwen2.5-VL-3B	+0.09	+0.26	+0.42	+0.64	**+0.60**
Qwen3-VL-8B	−0.08	+0.36	+0.23	+0.68	**+0.46**
Qwen3-VL-4B	+0.01	+0.23	+0.29	+0.15	**+0.22**
CoSEC (DAPO), 7B	+0.02	+0.19	**+0.42**	+0.40	+0.18

Table 30: Descriptive reward–data groups in the run archive. Base, optimizer, and normalization mixtures differ across cells. Among archived counterfactual-reward variants, the largest observed drawdowns are 25.57 points on fixed data and 5.69 under regeneration; the largest length ratios are 5.15× and 3.22×. Configuration-aligned comparisons are in Appendix M.

Scope	Cell	n	mean len. ratio	**max**	mean drawdown	**max drawdown**
all bases	fixed data + outcome-only	20	0.78	1.18	1.11	5.43
all bases	fixed data + counterfactual	9	1.67	**5.15**	4.40	**25.57**
all bases	closed loop + outcome-only	5	1.17	1.59	1.28	4.62
all bases	**closed loop + counterfactual**	59	1.16	**3.22**	**0.54**	**5.69**
3B+7B	fixed data + outcome-only	3	1.00	1.06	0.10	0.24
3B+7B	fixed data + counterfactual	6	1.89	**5.15**	6.12	**25.57**
3B+7B	closed loop + outcome-only	2	1.34	1.59	2.47	4.62
3B+7B	**closed loop + counterfactual**	24	1.45	**3.22**	**0.45**	**2.34**

Table 31: Length containment in representative draws (7B uses seed 123): median output length at step 210 divided by step 0; these are not seed means. CoSEC is no worse than fixed data at 5/5 scales; fixed data + reward runs away at 3B.

Base	`static+grounding` (fixed)	CoSEC (closed loop)	`no-counterfactual` (loop, outcome only)
Qwen2.5-VL-7B	1.43×	1.41×	1.08×
Qwen2.5-VL-3B	**2.03**× (draw 2: 6.51×)	**1.24**×	1.40×
Qwen3-VL-8B	1.23×	1.10×	1.09×
Qwen3-VL-4B	1.08×	**0.84**×	1.08×
Qwen3-VL-2B	1.44×	1.20×	1.05×

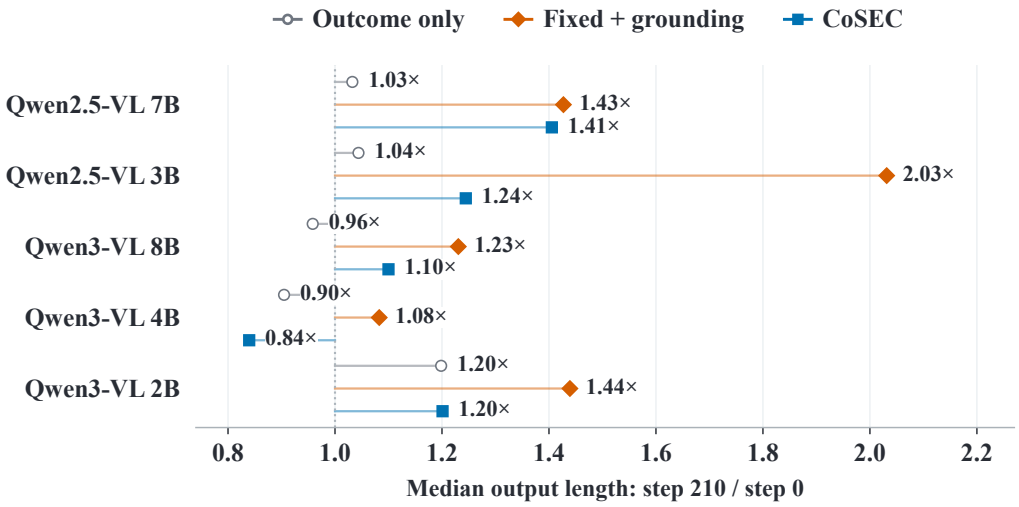


Figure 13: Length containment across five scales and three arms in representative runs (7B uses seed 123). Each point is the ratio of median output length at step 210 to step 0; the vertical reference is 1×. These are individual runs. Fixed-data arms with grounding exceed 1× (up to 2.03×); closed-loop ratios are no larger at any scale.

K BENCHMARK SENSITIVITY AND TRAINING DYNAMICS

Sensitivity to benchmark composition. The seed-mean advantage stays positive at all five model scales under every single-benchmark omission (Table 32). The mean across scales ranges from +1.04 to +2.37 points across the four omissions.

Table 32: Benchmark omission sensitivity: CoSEC minus outcome-only static GRPO at step 210 (points). Each omission re-averages the remaining three benchmarks using the primary run cohorts. The bottom row weights the five scales equally.

		Benchmark omitted			
Base	All four	BLINK	MME	MME-RW	VSR
Qwen2.5-VL-7B	+1.49	+1.62	+1.90	+1.57	+0.88
Qwen2.5-VL-3B	+1.76	+3.30	+1.18	+1.33	+1.25
Qwen3-VL-8B	+1.34	+2.21	+1.30	+0.67	+1.18
Qwen3-VL-4B	+0.82	+1.36	+0.63	+0.14	+1.16
Qwen3-VL-2B	+2.60	+3.36	+2.78	+3.53	+0.72
Mean	**+1.60**	**+2.37**	**+1.56**	**+1.45**	**+1.04**

Where the fixed-data trajectories change. Figure 14 resolves the 3B accuracy trajectories by benchmark. From step 90 to 210, fixed-data draw 2 falls from 78.40 to 8.92 on VSR and from 79.44 to 48.90 on MME. The corresponding CoSEC run retains 78.48 and 83.32 at the endpoint.

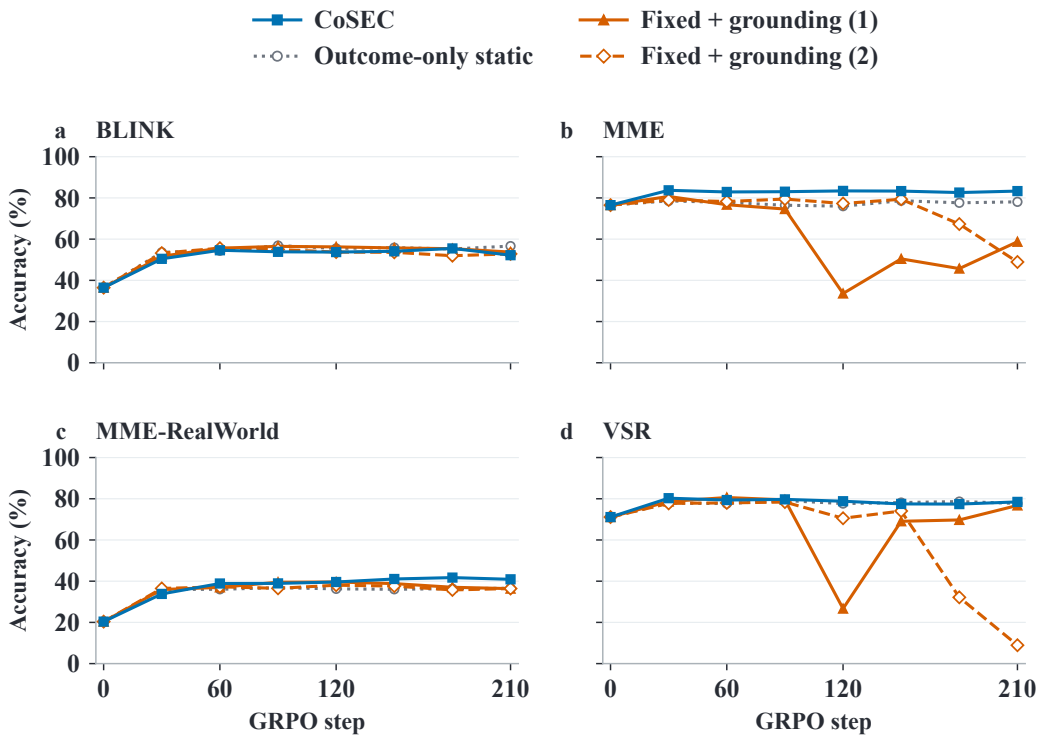


Figure 14: Qwen2.5-VL-3B accuracy by benchmark over all eight evaluation steps. Each line is one run; all panels share a 0–100% scale. The CoSEC and two fixed-data grounding runs also appear in Figure 18; the outcome-only static baseline provides an additional reference.

L RUN VARIATION AND RECORDED TRAINING COSTS

Variation across runs and throughout training. Table 33 reports endpoint variation across the primary runs and summarizes performance throughout training. We integrate each run's Macro-4 accuracy over steps $0, 30, \ldots, 210$ by the trapezoidal rule and divide by 210. We then average within each arm and subtract baseline from CoSEC. This training-average gap is positive at all five scales, with an equal-scale mean of $+0.79$ points.

Table 33: Run variation and training-average accuracy in the 42 primary runs. Endpoint values are mean $\pm$ sample standard deviation of step-210 Macro-4 accuracy (%); n_B/n_C gives baseline/CoSEC run counts. Δ_{end} and Δ_{train} are CoSEC minus baseline in percentage points.

Base	n_B/n_C	Baseline	CoSEC	Δ_{end}	Δ_{train}
Qwen2.5-VL-7B	4/4	66.40 ± 0.42	67.89 ± 1.02	$+1.49$	$+0.59$
Qwen2.5-VL-3B	4/4	62.16 ± 0.36	63.92 ± 0.91	$+1.76$	$+1.54$
Qwen3-VL-8B	4/4	67.50 ± 0.42	68.84 ± 1.08	$+1.34$	$+0.81$
Qwen3-VL-4B	4/4	67.86 ± 0.44	68.68 ± 0.65	$+0.82$	$+0.70$
Qwen3-VL-2B	4/6	58.98 ± 2.08	61.58 ± 0.84	$+2.60$	$+0.32$

Resource accounting. We first sum the logged costs over each run's 210 updates, then average over runs (Table 34). Step time covers the measured training step; rollout and reward times identify two components within it. Policy-token counts describe the training batches. These measurements complement the fixed-update accuracy comparison with the observed resource cost.

Table 34: Mean recorded cost in 39 primary runs with complete 210-update timing traces. n denotes runs; each run contributes its summed logged step, policy-rollout and reward times, and policy-token count. Baseline is outcome-only static GRPO. Reward time includes frozen-scorer calls; timing scope is defined in Appendix E.4.

Base	Arm	n	Step (h)	Rollout (h)	Reward (h)	Tokens (M)
Qwen2.5-VL-7B	Baseline	4	1.81	0.55	0.006	43.39
Qwen2.5-VL-7B	CoSEC	4	8.05	0.65	6.053	47.01
Qwen2.5-VL-3B	Baseline	4	1.81	0.70	0.006	39.24
Qwen2.5-VL-3B	CoSEC	3	7.15	0.60	5.568	41.21
Qwen3-VL-8B	Baseline	4	1.63	0.76	0.006	36.75
Qwen3-VL-8B	CoSEC	4	5.01	0.86	3.239	38.38
Qwen3-VL-4B	Baseline	4	1.53	0.66	0.006	37.10
Qwen3-VL-4B	CoSEC	3	4.74	0.75	3.109	38.01
Qwen3-VL-2B	Baseline	4	1.26	0.59	0.006	39.22
Qwen3-VL-2B	CoSEC	5	5.03	1.17	3.032	39.15

M CONFIGURATION CHECKS AND EVALUATION PROTOCOL

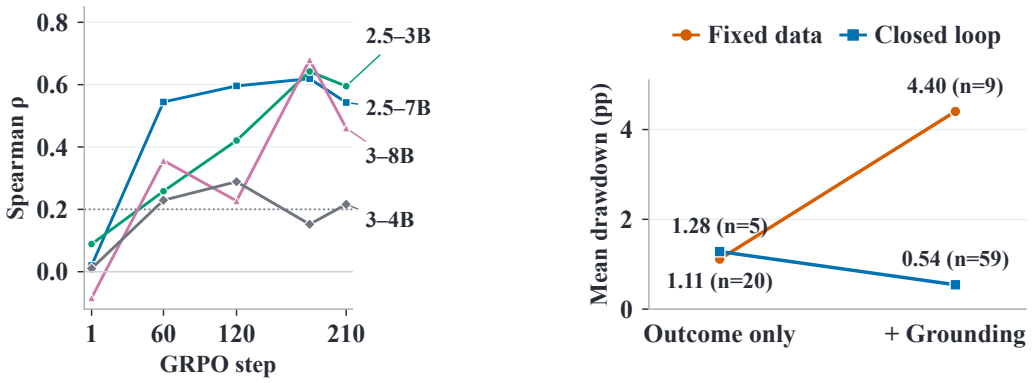


Figure 15: Length association and archived drawdown. Left: grounding-score/response-length Spearman correlations among correct-and-scored rollouts at four bases. Right: mean peak-to-endpoint drawdown as grounding is added under fixed or regenerated data; counts are shown. The pooled descriptive interaction is -4.03 points; base, optimizer, and normalization mixtures differ across cells. Appendix M resolves the base dependence.

M.1 CONFIGURATION-ALIGNED REWARD–DATA SUMMARIES

Within-base, optimizer- and controller-aligned comparison. We restrict the four-cell comparison to GRPO with the default reward construction ($\lambda = 0.1$, $S = 64$, $\beta = 1$) and the shared $0, 30, \ldots, 210$ evaluation grid. Both regenerated-data arms use $w_c = 0$, replacing the full-method arm with its matching controller ablation. We compute each run's peak-to-endpoint drawdown, average within each base and cell, and then weight the five bases equally. This descriptive interaction is concentrated at 3B: the equal-base mean is -3.35 points, while leaving out 3B gives $+0.85$ points. These archived executions do not form a seed-paired randomized factorial design.

Table 35: Within-base reward–data comparison under GRPO with aligned $w_c = 0$ regenerated-data controllers. Cells show mean peak-to-step-210 Macro-4 drawdown in points, with run counts in parentheses. Outcome uses the correctness reward; CF uses the default counterfactual reward. The interaction is $(\text{Regen}_{\text{CF}} - \text{Regen}_{\text{Outcome}}) - (\text{Fixed}_{\text{CF}} - \text{Fixed}_{\text{Outcome}})$. The final row weights bases equally. All 36 eligible executions, including both 3B fixed-CF runs, are retained.

	Fixed data		Regenerated data		
Base	Outcome	CF	Outcome	CF ($w_c = 0$)	Interaction
Qwen2.5-VL-7B	0.35 (4)	0.00 (1)	0.32 (1)	1.05 (1)	$+1.08$
Qwen2.5-VL-3B	0.33 (4)	15.84 (2)	4.62 (1)	0.00 (1)	-20.13
Qwen3-VL-8B	0.21 (4)	0.00 (1)	0.06 (1)	0.61 (1)	$+0.76$
Qwen3-VL-4B	0.34 (4)	0.03 (1)	0.00 (1)	0.73 (1)	$+1.04$
Qwen3-VL-2B	3.11 (4)	2.91 (1)	1.42 (1)	1.72 (1)	$+0.51$
Equal-base mean	0.87	3.76	1.28	0.82	-3.35

Influence diagnostics. Leaving out 7B, 3B, 8B, 4B, and 2B gives equal-base contrasts of -4.46, $+0.85$, -4.38, -4.45, and -4.31 points, respectively. Omitting only the more severe 3B fixed-CF execution changes the five-base contrast to -1.41; omitting only the other 3B fixed-CF execution gives -5.29. These are influence diagnostics, not exclusions from the main comparison. The regenerated outcome and regenerated CF cells each contain one execution per base, and the fixed-CF cell contains one per base except two at 3B; these contrasts are therefore reported descriptively, without a new significance test.

M.2 FROZEN EVALUATION SUBSETS AND ANSWER MATCHING

The full evaluation suite contains 6308 questions (Table 36), evaluated at steps $0, 30, \ldots, 210$. Main-text Macro-3 uses the 5515 MME, MME-RealWorld, and VSR questions; Macro-4 additionally includes the 793 BLINK questions. Benchmark names in this paper refer to these frozen subsets. Accuracy is the fraction of correct questions within each subset; Macro-4 weights the four subset accuracies equally. In particular, the MME column reports question-level accuracy rather than the official composite score for that benchmark.

Table 36: Frozen OOD evaluation grid. The converted manifests specify every retained question and image path. The evaluation pipeline uses these frozen manifests and one offline matcher.

Benchmark	Questions	Evaluated subset and conversion rule
BLINK	793	Validation rows with one image; all retained rows from six tasks, without a per-task cap.
MME	2374	Yes/no questions from the archived evaluation shards, retaining all 14 categories.
MME-RealWorld	1919	Frozen five-domain evaluation manifest; retain rows whose source images exist and decode.
VSR	1222	Zero-shot test split; retain its original label balance (593 false, 629 true).

BLINK includes Counting (120), IQ Test (150), Object Localization (122), Relative Depth (124), Relative Reflectance (134), and Spatial Relation (143). MME retains perception and reasoning categories; its 2374 questions reference 1187 image paths. MME-RealWorld includes Autonomous Driving (750), Diagram and Table (200), Monitoring (469), OCR with Complex Context (350), and Remote Sensing (150). Conversion caps the longest image side at 448 pixels for BLINK, MME, and VSR, and 1024 for MME-RealWorld.

The matcher normalizes case, whitespace, surrounding punctuation, number words one through five, and Boolean words. For a response with a `FINAL:` field, it accepts the last field's normalized full text or its restricted first answer token. For multiple-choice questions it accepts the ground-truth letter or the corresponding option text, including a letter-prefixed form. Without that field, it compares the normalized full response and any recognized letter-prefixed variants, without searching arbitrary embedded answers. MME and VSR use normalized yes/no equality. Answer-marker emission is reported separately from accuracy. Consequently, the completion diagnostic does not by itself partition errors into visual mistakes and formatting failures.

N VISUAL DIAGNOSTICS AND EVALUATION CONTROLS

N.1 VISUAL–TEXT SEPARATION AMONG CORRECT ANSWERS

Each synthetic image pair supports a visual question and two text-answerable controls: an answer-aligned control and a naturally worded control. We retain pairs for which both the visual and control responses have correct final answers. A frozen Qwen2.5-VL-3B scorer evaluates responses generated by either 3B or 7B, using the full-response summed contrast g from Eq. 2. Thus correctness is held at one while question dependence on the image varies. Table 37 reports all task families and both controls. Overall AUROC ranges from 0.894 to 0.955, showing that the contrast retains visual–text separation within correctly answered pairs.

Table 37: Visual questions versus text-answerable controls, conditional on both responses being correct. The signal is the summed semantic contrast g; n counts eligible image pairs out of 360 per generator/control condition (120 per task family). Confidence intervals use 5000 hierarchical paired bootstrap resamples of template pairs and image pairs. Controls and generators share image pools. Natural text denotes question wording on synthetic images.

Generator	Control	Task	n	AUROC	95% CI
3B	Matched	Overall	320	0.955	[0.905, 0.990]
3B	Matched	Color	116	0.999	[0.996, 1.000]
3B	Matched	Counting	87	0.828	[0.739, 0.903]
3B	Matched	Spatial	117	0.998	[0.995, 1.000]
3B	Natural text	Overall	195	0.910	[0.868, 0.945]
3B	Natural text	Color	70	0.962	[0.902, 0.992]
3B	Natural text	Counting	54	0.891	[0.777, 0.971]
3B	Natural text	Spatial	71	0.958	[0.910, 0.991]
7B	Matched	Overall	246	0.939	[0.877, 0.981]
7B	Matched	Color	94	0.981	[0.953, 0.998]
7B	Matched	Counting	79	0.799	[0.680, 0.899]
7B	Matched	Spatial	73	0.992	[0.975, 1.000]
7B	Natural text	Overall	204	0.894	[0.846, 0.937]
7B	Natural text	Color	72	0.910	[0.836, 0.957]
7B	Natural text	Counting	71	0.918	[0.827, 0.985]
7B	Natural text	Spatial	61	0.987	[0.958, 0.999]

N.2 RESPONSE-DEPENDENT NORMALIZATION CHANGES THE SIGNAL

We transform the same 2880 recorded responses to 360 held-out questions as g/T^p, where T is the scorer token count and $p \in \{0, 0.25, 0.5, 0.75, 1\}$. Within-question correlations use all 190 groups containing both correct and incorrect responses, centered by question. Increasing p reduces the within-question correctness correlation from 0.293 to 0.072 and shifts the global length correlation from -0.025 to -0.394 (Table 38). Response-dependent normalization therefore changes the signal's associations as well as its magnitude. This offline score analysis complements the trained reward variants in Table 4.

Table 38: Offline exponent sweep on a fixed set of responses. The first two columns correlate question-centered scores with question-centered correctness over mixed-outcome groups; the final column correlates scores with token length over all responses. All five transformations use the same recorded answers without regenerating policy responses.

p	Within-question r	Within-question ρ	Length r
0.00	0.293	0.270	-0.025
0.25	0.262	0.227	-0.138
0.50	0.207	0.166	-0.239
0.75	0.140	0.095	-0.324
1.00	0.072	0.034	-0.394

N.3 SHAPING STRENGTH WITH DIAGNOSTIC REPLAY

The full 7B configuration combines reward shaping with diagnostic-weighted replay ($w_c = 1$). Table 39 spans a fourfold range of shaping strengths: $\lambda = 0.05$, 0.10, and 0.20. Endpoint Macro-4 is 67.31, 67.89, and 68.20, respectively, compared with the outcome-only static baseline mean of 66.40. These settings yield gains of 0.91–1.80 points. The distinct $w_c = 0$ strength controls are shown in Figure 10.

Table 39: Qwen2.5-VL-7B strength sweep for the full CoSEC configuration at step 210. n counts runs; default strength and baseline use the primary cohorts. Values are four-benchmark accuracy (%), their unweighted Macro-4, and the gap from the baseline mean in points. All three CoSEC settings use diagnostic replay ($w_c = 1$).

λ	n	BLINK	MME	MME-RW	VSR	Macro-4	Δ
Baseline	4	57.82	84.47	41.49	81.81	66.40	+0.00
0.05	1	58.01	81.51	44.29	85.43	67.31	+0.91
0.10	4	58.92	84.74	42.76	85.15	67.89	+1.49
0.20	1	58.26	86.39	44.09	84.04	68.20	+1.80

O FULL EVALUATION SUITE AND ORIGINAL FOUR-BENCHMARK AGGREGATE

The main-text Macro-3 was defined during manuscript revision as the equal-weight mean of MME, MME-RealWorld, and VSR. This section preserves the full evaluation suite, including BLINK (Fu et al., 2024), and the original equal-weight four-benchmark aggregate, denoted Macro-4. The same training runs, checkpoints, and benchmark scores underlie both aggregates. Across the primary runs, the original Macro-4 endpoint gain is 1.60 points and its training-average gain is 0.79 points. Tables 40 and 41 retain all benchmark scores and cross-run standard deviations; Appendix A reports the additional experimental arms. Figure axes labeled OOD-Macro in this appendix denote Macro-4.

Table 40: Full-suite step-210 accuracy (%) across five base models. Entries are mean $\pm$ sample standard deviation across training runs; n is the number of runs. GRPO denotes the outcome-only static baseline. MME-RW denotes MME-RealWorld. Macro-4 averages all four benchmarks within each run before summarization. Bold marks the larger mean within each base.

Base model	Method	n	BLINK	MME	MME-RW	VSR	Macro-4
Qwen2.5-VL-7B	GRPO	4	57.82 ± 1.44	84.47 ± 0.85	41.49 ± 0.48	81.81 ± 0.74	66.40 ± 0.42
	CoSEC	4	$\mathbf{58.92} \pm 0.92$	$\mathbf{84.74} \pm 2.50$	$\mathbf{42.76} \pm 0.38$	$\mathbf{85.15} \pm 1.42$	$\mathbf{67.89} \pm 1.02$
Qwen2.5-VL-3B	GRPO	4	$\mathbf{55.80} \pm 0.60$	79.21 ± 0.81	36.36 ± 0.39	77.25 ± 0.83	62.16 ± 0.36
	CoSEC	4	52.96 ± 2.19	$\mathbf{82.73} \pm 0.84$	$\mathbf{39.42} \pm 1.63$	$\mathbf{80.56} \pm 2.51$	$\mathbf{63.92} \pm 0.91$
Qwen3-VL-8B	GRPO	4	$\mathbf{62.33} \pm 0.95$	84.11 ± 0.49	41.35 ± 1.02	82.20 ± 0.41	67.50 ± 0.42
	CoSEC	4	61.07 ± 0.88	$\mathbf{85.57} \pm 1.47$	$\mathbf{44.68} \pm 1.93$	$\mathbf{84.02} \pm 0.70$	$\mathbf{68.84} \pm 1.08$
Qwen3-VL-4B	GRPO	4	$\mathbf{63.18} \pm 0.99$	83.90 ± 0.52	41.39 ± 0.54	$\mathbf{82.96} \pm 0.75$	67.86 ± 0.44
	CoSEC	4	62.39 ± 1.68	$\mathbf{85.30} \pm 0.94$	$\mathbf{44.24} \pm 1.37$	82.77 ± 0.39	$\mathbf{68.68} \pm 0.65$
Qwen3-VL-2B	GRPO	4	51.83 ± 1.41	75.79 ± 1.81	$\mathbf{42.97} \pm 0.51$	65.34 ± 7.09	58.98 ± 2.08
	CoSEC	6	$\mathbf{52.14} \pm 1.10$	$\mathbf{77.84} \pm 2.55$	42.77 ± 0.69	$\mathbf{73.57} \pm 1.11$	$\mathbf{61.58} \pm 0.84$

Table 41: Full-suite adapted methods and curriculum controls on Qwen3-VL-4B at step 210. Accuracy (%) is mean ± sample standard deviation over n runs. Bold marks the highest mean accuracy or lowest drawdown. Macro-4 averages the four accuracies. The last column is the maximum of five per-base mean drawdowns (points), using the separate archive cohorts in Table 21. Supervision and adaptations are specified in Appendix H.

Recipe	n	BLINK	MME	MME-RW	VSR	Macro-4	Max. mean drawdown ↓
Outcome-only							
GRPO	4	**63.18** ± 0.99	83.90 ± 0.52	41.39 ± 0.54	82.96 ± 0.75	67.86 ± 0.44	3.11
iReasoner	4	58.89 ± 0.72	**85.58** ± 0.30	43.32 ± 1.63	82.16 ± 0.35	67.49 ± 0.47	3.82
Fixed schedule	4	62.45 ± 1.49	83.89 ± 0.48	41.46 ± 0.53	**83.00** ± 0.30	67.70 ± 0.55	3.93
VisPlay	4	61.38 ± 0.77	85.08 ± 0.20	42.86 ± 0.73	82.82 ± 0.52	68.03 ± 0.33	10.71
EvoLMM	4	62.45 ± 0.64	84.55 ± 0.48	43.88 ± 0.90	82.08 ± 0.41	68.24 ± 0.26	9.36
SelfEvolve-VT	4	62.32 ± 0.62	84.48 ± 0.65	43.91 ± 0.69	82.59 ± 0.34	68.33 ± 0.25	4.77
CoSEC	4	62.39 ± 1.68	85.30 ± 0.94	**44.24** ± 1.37	82.77 ± 0.39	**68.68** ± 0.65	**0.74**

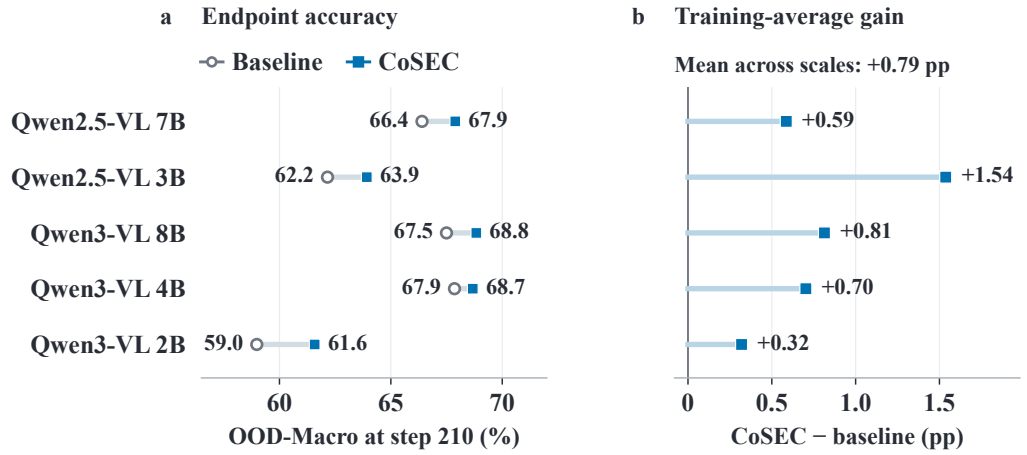


Figure 16: Original four-benchmark evaluation. Endpoint and training-average accuracy across five scales. Left: step-210 baseline and CoSEC seed means ($n = 4$ per arm, except $n = 6$ for CoSEC Qwen3-VL-2B). Right: the seed-mean training-average gain; each run uses trapezoidal integration over eight checkpoints from step 0 to 210, divided by 210.

Table 42: Original Macro-4 outcomes across 64 regenerated-data and 63 fixed-data runs. The archive covers multiple bases, optimizers, and reward variants with unequal run counts. Full archive summaries are in Table 15.

Measure	Regenerated	Fixed
Mean Macro-4 drawdown (points)	0.60	1.78
Endpoint below initial model	3/64	11/63

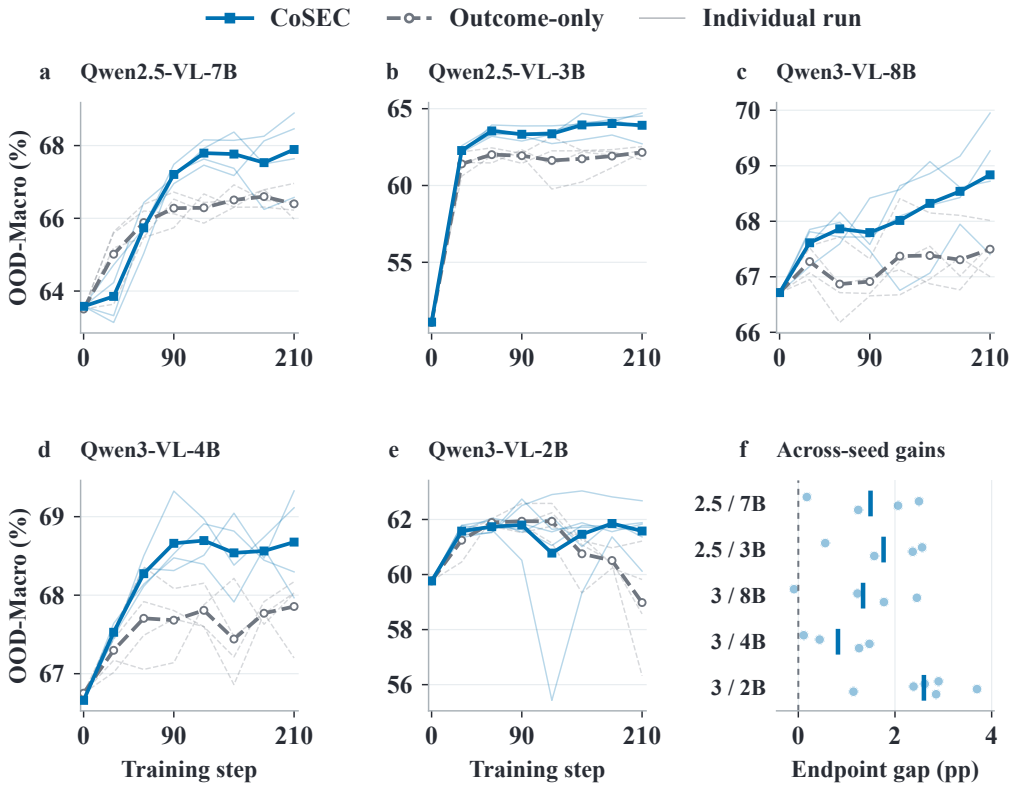


Figure 17: Original four-benchmark evaluation. Learning dynamics behind the cross-scale gains. **a–e**, All 42 runs at eight evaluation steps; thin lines show individual runs and thick lines show seed means. There are four runs per arm and scale, except six for CoSEC Qwen3-VL-2B. Panels use different vertical ranges. **f**, Each CoSEC endpoint minus its same-base baseline mean; vertical bars mark mean gaps. Plot prefixes 2.5 and 3 denote Qwen2.5-VL and Qwen3-VL, respectively; complete base names appear in Table 40.

Table 43: Macro-3 peak-to-endpoint drawdown (percentage points) underlying Table 2. Each cell is the mean over the original recipe/base stability cohort, with its run count in parentheses. These cohorts are separate from the endpoint-accuracy cohorts. The corresponding original Macro-4 summaries are in Table 21.

Recipe	Qwen2.5-VL 7B	Qwen2.5-VL 3B	Qwen3-VL 8B	Qwen3-VL 4B	Qwen3-VL 2B
Outcome-only GRPO	0.06 (1)	0.15 (1)	0.00 (1)	1.05 (1)	3.93 (4)
iReasoner	1.43 (1)	2.39 (1)	0.00 (1)	0.31 (4)	3.46 (1)
Fixed schedule	0.20 (1)	0.18 (1)	0.00 (1)	0.46 (1)	5.19 (1)
VisPlay	14.74 (1)	6.95 (1)	0.00 (1)	0.13 (4)	2.54 (1)
EvoLMM	0.00 (1)	1.59 (1)	0.71 (1)	0.25 (4)	11.76 (1)
SelfEvolve-VT	0.00 (1)	1.77 (1)	0.00 (1)	0.19 (1)	5.37 (1)
CoSEC	0.24 (4)	0.00 (3)	0.12 (3)	0.31 (2)	1.05 (3)

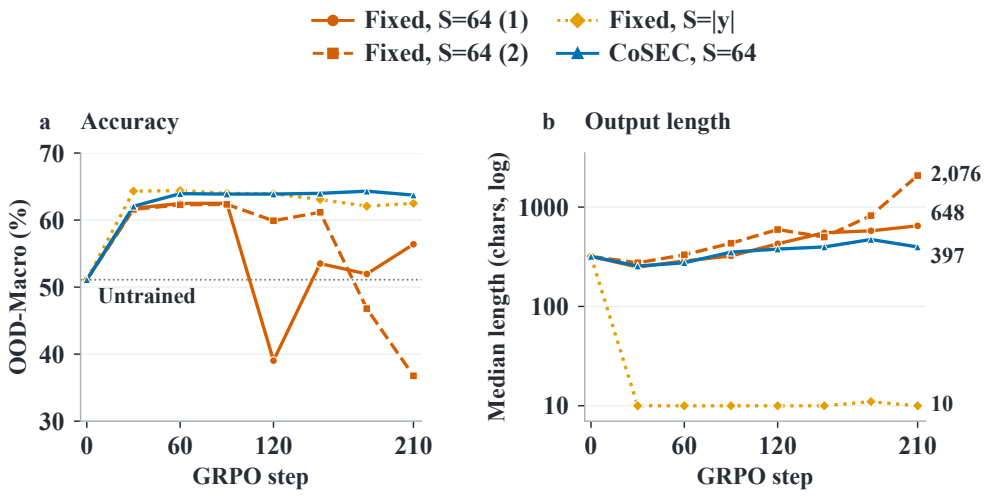


Figure 18: Original four-benchmark evaluation. Closed-loop regeneration sustains Qwen2.5-VL-3B accuracy under the same grounding reward. Left: Macro-4; right: median output characters on a logarithmic axis. Fixed-data training collapses in two draws; regeneration maintains accuracy with moderate length growth. A response-length divisor yields shorter responses. Curves are individual runs from Table 28.

P Analytical Properties of the Shared Diagnostic

We derive three properties of Eqs. 2, 3, and 10: a conservative cross-view likelihood margin, reward contrast within correct responses, and the success-rate preference induced by diagnostic replay. The policy and frozen scorer are fixed throughout each response-distribution calculation. The replay analysis holds the recorded diagnostic rate fixed and studies the continuous score over the full pass-probability range $p \in [0, 1]$.

P.1 A conservative cross-view likelihood margin

Proposition 1 (Cross-view margin) *Fix a question q and the same response y in all three views. Suppose the frozen-scorer log-likelihoods $\ell_o, \ell_{cf}, \ell_{nu}$ are finite. Set $g = \ell_o - \ell_{cf}$, $e = |\ell_o - \ell_{nu}|$, and $B = g - \beta e$. For $\beta \geq 1$,*

$$B \leq \min\{\ell_o, \ell_{nu}\} - \ell_{cf}. \tag{12}$$

For every real m, $B \geq m$ consequently implies

$$\frac{\min\{P_s(y \mid I, q), P_s(y \mid I^{nu}, q)\}}{P_s(y \mid I^{cf}, q)} \geq \exp(m). \tag{13}$$

The condition $\beta \geq 1$ is sharp for a bound valid for all finite log-likelihood triples.

Proof. Write $\Delta = \ell_o - \ell_{nu}$, $\Delta_+ = \max(\Delta, 0)$, and $H = \min\{\ell_o, \ell_{nu}\} - \ell_{cf}$. Since $\min\{\ell_o, \ell_{nu}\} = \ell_o - \Delta_+$, the exact slack is

$$H - B = \beta|\Delta| - \Delta_+ = \begin{cases} (\beta - 1)\Delta, & \Delta \geq 0, \\ -\beta\Delta, & \Delta < 0. \end{cases} \tag{14}$$

Both branches are nonnegative when $\beta \geq 1$, proving Eq. 12. Thus $B \geq m$ gives $\ell_o - \ell_{cf} \geq m$ and $\ell_{nu} - \ell_{cf} \geq m$. Exponentiation preserves these inequalities. Finite log-likelihoods ensure a positive denominator, establishing Eq. 13.

Equation 14 also characterizes equality completely. For $\beta = 1$, equality holds exactly when $\Delta \geq 0$, or $\ell_{nu} \leq \ell_o$. For $\beta > 1$, it holds exactly when $\Delta = 0$. If the nuisance view increases the response likelihood ($\Delta < 0$), the penalty remains conservative: its slack is $\beta(\ell_{nu} - \ell_o)$.

To show sharpness, choose any $0 \leq \beta < 1$ and $0 < \eta < 1 - \beta$, and take

$$\ell_o = -1, \qquad \ell_{nu} = -2, \qquad \ell_{cf} = -1 - \beta - \eta.$$

These are finite, valid log-probability values. They give $B = \eta > 0$ but $H = \beta + \eta - 1 < 0$. Hence even a positive bonus need not certify both view margins when $\beta < 1$. □

For a verified response with $\lambda > 0$ and $S > 0$, a positive shaping increment is equivalent to $B > 0$. Under $\beta \geq 1$, both original and nuisance views must therefore assign that response a higher likelihood than the counterfactual view. This certificate concerns the frozen scorer and the same full response; the generator supplies the answer-changing and answer-preserving intervention semantics in Section 3.1.

P.2 Reward contrast after correctness saturation

Proposition 2 (Within-correct reward contrast) *Fix an item (I, q) and a response distribution. Let $p = \Pr(o = 1) > 0$, $a = \lambda/S$ with $\lambda \geq 0$ and $S > 0$, and $Z = g - \beta e$ on correct responses. Assume $Z \mid o = 1$ has finite mean μ and variance σ^2. Then*

$$\mathrm{Var}(R_\lambda) = pa^2\sigma^2 + p(1 - p)(1 + a\mu)^2. \tag{15}$$

For any all-correct group of $n \geq 2$ responses with scores $Z_1, \ldots, Z_n$,

$$R_j - \bar{R} = a(Z_j - \bar{Z}). \tag{16}$$

If $\lambda > 0$, its centered reward vector is nonzero exactly when its visual scores are not all equal.

Proof. Incorrect responses have reward zero; correct responses have reward $1 + aZ$. Conditioning on correctness gives

$$\begin{aligned}\mathbb{E}[R_\lambda] &= p(1 + a\mu),\\ \mathbb{E}[R_\lambda^2] &= p\,\mathbb{E}[(1 + aZ)^2 \mid o = 1]\\ &= p\big[(1 + a\mu)^2 + a^2\sigma^2\big].\end{aligned} \tag{17}$$

Subtracting $\mathbb{E}[R_\lambda]^2$ proves Eq. 15. Equivalently, its two terms are the expected conditional variance and the variance of the conditional mean. At $p = 1$, the second term vanishes and the variance is $a^2\sigma^2$; at $a = 0$, it reduces to the outcome-only variance $p(1 - p)$. When $p = 0$, all rewards are zero, without requiring conditional moments on the empty correctness event.

For an all-correct group, $R_j = 1 + aZ_j$ and $\bar{R} = 1 + a\bar{Z}$, proving Eq. 16. With $a > 0$, $a(Z_j - \bar{Z}) = 0$ for every j if and only if every $Z_j = \bar{Z}$. This groupwise identity does not require independence between its responses. □

Consequences for group-relative advantages. Define the sample standard deviation by

$$s_Z = \left[\frac{1}{n-1}\sum_{j=1}^{n}(Z_j - \bar{Z})^2\right]^{1/2}.$$

Using the same convention for rewards gives $s_R = a s_Z$ for $a \geq 0$. For GRPO's centered, standardized reward-derived advantages (Shao et al., 2024), with an additive denominator stabilizer $\varepsilon_A > 0$ and $a > 0$, the response-level advantage is

$$A_j = \frac{R_j - \bar{R}}{s_R + \varepsilon_A} = \frac{a(Z_j - \bar{Z})}{a s_Z + \varepsilon_A} = \frac{Z_j - \bar{Z}}{s_Z + \varepsilon_A/a}. \tag{18}$$

The denominator is common and positive, so $A_j > A_k$ if and only if $Z_j > Z_k$. If $s_Z = 0$, all advantages are zero; the stabilizer makes this case well-defined. If $a = 0$, the original advantage expression likewise gives zero for an all-correct group. When $s_Z > 0$ and $\varepsilon_A = 0$, positive rescaling cancels exactly. For positive ε_A, rescaling preserves ordering but can change advantage magnitude through ε_A/a. Thus the diagnostic supplies a response preference after correctness saturation, whereas its scalar normalization governs magnitude.

P.3 How diagnostic replay changes sample priority

Proposition 3 (Diagnostic shift in replay priority) *Let the integer group size satisfy $n \geq 2$, fix $c \in [0, 1]$, and set $b = w_c c$ with $w_c \geq 0$. On $p \in [0, 1]$, the unrounded fitness*

$$F(p, c) = 1 - p^n - (1 - p)^n + bp \tag{19}$$

has a unique maximizer p_c. If $b = 0$, then $p_c = 1/2$. If $0 < b < n$, then $p_c \in (1/2, 1)$ uniquely solves the following first-order stationarity condition:

$$n\big[p_c^{n-1} - (1 - p_c)^{n-1}\big] = b. \tag{20}$$

If $b \geq n$, then $p_c = 1$. In the interior regime, p_c increases strictly with c when $w_c > 0$. At correctness saturation, $F(1, c) = w_c c$ while $L(1) = 0$.

Proof. Continuity on $[0, 1]$ guarantees existence. In the interior,

$$\begin{aligned}F_p(p, c) &= n\big[(1 - p)^{n-1} - p^{n-1}\big] + b,\\ F_{pp}(p, c) &= -n(n - 1)\big[(1 - p)^{n-2} + p^{n-2}\big] < 0.\end{aligned} \tag{21}$$

The second derivative is strictly negative also for $n = 2$, when both powers equal one. Consequently, F is strictly concave and F_p strictly decreases, ensuring a unique maximizer. At the boundaries and midpoint, the derivative values are

$$F_p(0, c) = n + b, \qquad F_p(1/2, c) = b, \qquad F_p(1, c) = b - n.$$

For $b = 0$, the unique zero occurs at $1/2$. For $0 < b < n$, continuity and strict decrease place the unique zero in $(1/2, 1)$; setting $F_p = 0$ gives Eq. 20. For $b > n$, the derivative is positive throughout $[0, 1]$, so the maximum occurs at one. At the boundary $b = n$, $F_p(1, c) = 0$ but $F_p(p, c) > 0$ for every $p < 1$, giving the same unique endpoint maximizer.

For $0 < b < n$, implicit differentiation of Eq. 20 yields

$$\frac{dp_c}{dc} = \frac{w_c}{n(n-1)\left[p_c^{n-2} + (1-p_c)^{n-2}\right]} > 0 \quad (w_c > 0). \tag{22}$$

The denominator is positive. The same calculation gives dp_c/db as its reciprocal and dp_c/dw_c as c times that reciprocal. For $b > n$, the maximizer remains at one; the interior derivative does not apply across the transition. As a direct check, $n = 2$ gives $p_c = \min\{1/2 + b/4, 1\}$. Finally, substitution yields $F(0, c) = 0$ and $F(1, c) = b$, proving the saturation statement. □

From the score to archive sampling. This is a fixed-c comparison of the continuous fitness profile; recorded pass rates evaluate that profile at their observed values. In particular, $c = 0$ or $w_c = 0$ recovers the symmetric learnability score, whereas $c > 0$ and $w_c > 0$ retain positive unrounded fitness at $p = 1$.

Let $\widetilde{F}_i$ be fitness stored after rounding to six decimal places, and set $v_i = \max(\widetilde{F}_i, \epsilon)$ with $\epsilon > 0$. Rounding to the nearest such value changes the ideal score by at most 5×10^{-7} and can merge close scores into ties. For a fixed remaining candidate pool $\mathcal{C}_t$ within one family, the conditional probability of the next draw is

$$q_i^{(t)} = \frac{v_i}{\sum_{k \in \mathcal{C}_t} v_k}, \qquad \frac{q_i^{(t)}}{q_j^{(t)}} = \frac{v_i}{v_j}, \quad i, j \in \mathcal{C}_t. \tag{23}$$

Thus each next draw preserves stored-score ordering, with strict ordering for unequal scores above the floor. Without replacement, the selected item is removed and the denominator is recomputed. Equation 23 describes this conditional draw, rather than a proportional formula for final inclusion probabilities. Archive capacity M and family quotas determine the eligible pools. The two diagnostic pathways therefore have complementary algebraic roles: distinguishing correct responses when $\lambda > 0$, and retaining priority for diagnosed items after correctness saturation when $w_c > 0$.