GOLD-LP: DATA SELECTION FOR VISION-LANGUAGE REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards improves vision-language models, and which questions enter training affects what the model learns. Choosing them is itself costly: selectors that estimate each candidate's difficulty from sampled responses must generate over the whole pool before optimization begins, a fifth as many responses as the training run itself, paid whether or not the resulting ranking helps. Existing selectors fall into three categories: sampling responses; reusing an earlier training run's outcomes, which requires one to exist; and substituting the model's own probability, which reduces generation by an amount the published papers do not always make clear. However, these approaches leave a fundamental question unanswered: **can a useful subset be identified without generating a single response, and how much accuracy can such a subset deliver?** We introduce GOLD-LP: one forward pass per candidate scores the initial model's likelihood for the correct option letter under a fixed answer prefix, and a contiguous interval of the resulting ranking is retained, leaving group relative policy optimization unchanged. Training Qwen3-VL-8B-Instruct on 1,000 of 3,023 DOCCI-derived examples at a matched step budget, it reaches the highest mean accuracy among the evaluated fixed-size selections on both holdouts across three runs; the geometry gain over random selection is 1.35 points (unadjusted 95% CI $[0.41, 2.30]$). Which region is kept matters as much as how candidates are scored: with score and subset size held fixed, accuracy is higher at our interval than at the lowest-score tail on both holdouts, exceeding a ported mean-centred rule in distribution, while three fixed-size comparators adapting published selectors do not reach it. Selection-time generation can therefore be removed entirely without giving up downstream accuracy at a fixed subset size.

1 INTRODUCTION

Vision-language models have demonstrated a remarkable ability to integrate textual and visual information, and reinforcement learning with verifiable rewards (RLVR) improves that ability on tasks with a checkable answer, such as visual multiple-choice questions (Bai et al., 2025; Wu et al., 2026; Gao et al., 2026). For each question, the model generates responses, a verifier checks their correctness, and the model learns from the resulting rewards. In group-relative methods such as GRPO, learning depends on differences in reward among responses to the same question (Shao et al., 2024). Questions the model always answers correctly, and questions it never does, produce none of the differences that group-relative methods look for. Which questions enter training therefore affects what the model learns. This has motivated selecting questions according to the model's estimated success rate, rather than treating all candidates equally (Foster et al., 2025; Zhou et al., 2026). Obtaining that estimate means generating responses across the whole candidate pool before training starts. A selector must therefore justify both the subset it chooses and the workload required to choose it.

Existing approaches make that trade-off in different ways. Success-based selectors estimate difficulty or uncertainty from sampled responses, though the sampling is itself the expense. Reward-history selectors reuse outcomes from an earlier training run, and need one to exist. Online methods such as GRESO predict from reward dynamics which prompts will be uninformative and skip them before producing their rollouts (Zheng et al., 2025), changing the training procedure rather than

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.

the training set. A second line replaces outcomes with model likelihood: UFO-RL (Zhao et al., 2025) selects examples near the pool's mean answer-token confidence, and ZPD Detector (Yang et al., 2026) combines reference-answer likelihood with correctness-based calibration, though that calibration returns to sampled outcomes.

What these methods leave unquantified is the cost of the selection signal itself: the generation they front-load before optimization begins. Published work reports the accuracy a selected subset obtains, but rarely the responses spent to produce the ranking. A reader therefore cannot weigh one against the other. We take the offline case, which fixes the training subset once before optimization, and measure both under matched subset and training-step budgets.

We introduce GOLD-LP, a generation-free selector for visual multiple-choice RL. A single forward pass scores each candidate using the initial model's correct-answer token likelihood, without generating any responses. A success-rate probe instead needs sixteen samples of every candidate, which in our setting amounts to 21.5% of a subsequent training run's nominal response count. Our main rule, GOLDLP-MATCH, retains a contiguous interval of the ranking, excluding the lowest-scoring tail and the highest-scoring examples. The selected subset remains fixed, and training proceeds with standard GRPO. This separates two decisions often discussed together: how to score examples, and which scores to retain. A model's likelihood for an answer under a supplied prefix is not the same as its success after reasoning. We therefore assess the signal through the performance of models trained on the selected data, and compare the interval against both a lowest-score window and a ported mean-centred rule under the same measurement and subset size.

We train and evaluate on DOCCI-derived visual questions released by Wu et al. (2026) on two backbones: Qwen3-VL-8B-Instruct with three training runs per arm for the main comparisons, and Gemma-3-4B-IT with one run per arm as a provisional replication. On Qwen, we evaluate four outcome-dependent selectors against a size-matched random baseline: static InSight (Zhou et al., 2026), a ZPD-inspired rule (Yang et al., 2026), RIPE and UNRIPE, the last two ranking within the questions the probe never solved. None separates from random on either holdout. GOLD-LP obtains higher mean accuracy than the same baseline on both, gaining 1.35 percentage points on a separate 691-question geometry holdout of Geo170K images (Gao et al., 2023) and 2.54 in distribution; on geometry its mean is similar to that of full-pool training from one-third of the distinct examples, while full-pool training remains stronger in distribution. Gemma moves the same way in distribution at 3.44 points from a single run per arm, but its geometry ordering reverses, leaving cross-backbone transfer unconfirmed. On Qwen the results highlight what a generation-free signal buys: a gain on a held-out domain the model never trained on, at no generation cost.

Overall, GOLD-LP overcomes the generation cost of outcome-based selection by combining correct-answer token likelihood, a fixed rank interval, and an unchanged GRPO objective. This reframes offline selection for visual RLVR as a scoring problem rather than a sampling problem: the initial model's own probabilities already carry a usable selection signal, before a single response is generated. Our contributions are GOLD-LP itself (Section 3); the finding that its subset gains over size-matched random selection on a held-out geometry domain; a matched-budget measurement of four outcome-dependent controls, none of which separates from that baseline; and an isolation of the retained interval from the score, where accuracy is higher at our band than at the lowest-score tail, and in distribution than a ported mean-centred rule (Zhao et al., 2025) (Section 4).

2 SETUP AND THE CHALLENGE OF FRONT-LOADED GENERATION

We begin with the selection problem and the GRPO recipe that every arm shares, then with the cost of obtaining a selection signal at all, which together motivate the method of Section 3.

Selection as measurement and retention. Let $\mathcal{D} = \{(x_i, a_i^\star)\}_{i=1}^N$ be a labeled pool, where $x_i = (I_i, q_i)$ contains an image and a question with four answer options, and $a_i^\star \in \mathcal{A} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ is the correct option letter. Given an initial vision-language policy π_{θ_0} and a subset budget $B < N$, our goal is to construct a fixed subset for subsequent RL training. We separate this into a measurement rule, which scores each candidate, and a retention rule, which chooses an interval of the resulting ranking. The score is computed once, before optimization, and never becomes a reward, a loss term, or a sampling weight. The initial policy scores the pool without generating responses; the selected questions then support ordinary GRPO training with generated reasoning. The actor is initialized

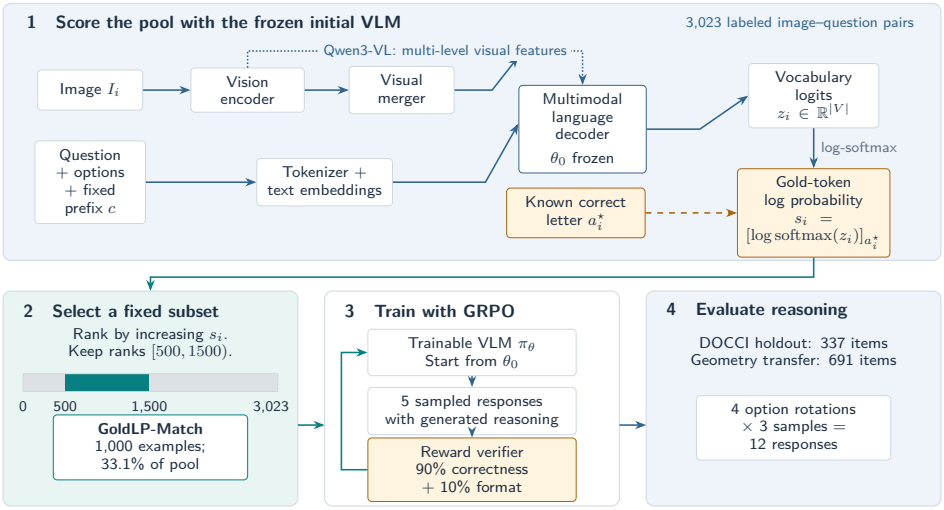


Figure 1: GOLD-LP selection and training. The initial VLM scores each image and question under a fixed answer prefix; the known label indexes the next-token log-probability vector and is never inserted into the input. A rank interval fixes the training subset for GRPO. The backbone shown is the existing Qwen3-VL, not an architectural modification.

from θ_0 and trained, while the scoring parameters stay fixed at θ_0. Section 4 gives the backbone and training configuration. Figure 1 illustrates the pipeline.

2.1 BALANCED TRAINING INPUTS

Correct-answer token likelihood depends on the option ordering and is not guaranteed to remain unchanged under permutation, and language models show selection bias tied to option position (Zheng et al., 2026). Before scoring or selection we therefore permute each training candidate's options so that the gold letter is uniform, and we recompute scores after balancing; Appendix G quantifies how far the ranking moves between orientations. Neither the target assignment nor the rotation uses model scores or rollout outcomes, and Appendix A gives the assignment procedure. Below, x_i and $a_i^\star$ refer to the balanced inputs and labels.

2.2 GROUP RELATIVE POLICY OPTIMIZATION

All static selection arms use the same GRPO training recipe. For a selected question x_i, the rollout policy $\pi_{\theta_{\text{old}}}$ generates $G = 5$ responses $y_{i,1}, \ldots, y_{i,G}$ under the normal reasoning prompt, without the scoring prefix defined in Section 3.1. Let $C(y, a^\star)$ indicate a correct extracted answer and $F(y)$ valid reasoning-tag and boxed-answer formatting. The recipe's outcome reward and group-relative advantage are

$$r_{i,g} = 0.9\, C(y_{i,g}, a_i^\star) + 0.1\, F(y_{i,g}), \tag{1}$$

$$\overline{r}_i = \frac{1}{G}\sum_{g=1}^{G} r_{i,g}, \qquad \widehat{\sigma}_i = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(r_{i,g} - \overline{r}_i)^2}, \qquad A_{i,g} = \frac{r_{i,g} - \overline{r}_i}{\widehat{\sigma}_i + 10^{-6}}. \tag{2}$$

The format verifier accepts either reasoning-tag spelling present in the prompt, requiring matching opening and closing tags. This shared modification applies to every arm; correctness scoring is unchanged. Appendix A gives the verification details.

2.3 THE COST OF A RANKING

Selecting a subset requires a way to rank candidates, and obtaining that ranking has a price of its own. Outcome-based selectors estimate each candidate's difficulty from sampled responses, so the

pool must be generated over before optimization begins, many times for every candidate. For our pool a sixteen-sample probe spends 48,368 responses on this. The cost is paid in full whether or not the resulting ranking helps, and it scales with the pool that is scored rather than with the subset that is finally kept.

Motivation. Existing selectors pay that cost in different ways: (i) success-based scoring estimates difficulty from sampled responses, where the sampling is itself the expense; (ii) reward-history scoring reuses outcomes from an earlier training run, and needs one to exist; and (iii) likelihood-based scoring replaces outcomes with the model's own probability, though some rules calibrate it against sampled outcomes anyway. The third reduces generation by an amount the published papers do not always make clear. What the three share is that each reports what its subset achieves without reporting what the ranking cost to obtain, so the two cannot be weighed against one another. This motivates the question the rest of this section answers: **can a useful subset be identified without generating a single response, and how much accuracy can such a subset deliver?** We argue that the key lies in scoring each candidate with the initial model's probability for the correct answer and in choosing which part of the ranking to keep; neither step generates a response.

3 GOLD-LP: GENERATION-FREE SELECTION WITH GRPO

To address this, we introduce GOLD-LP. It scores every candidate with the initial model's probability for the correct answer, retains a contiguous interval of the resulting ranking, and trains on that fixed subset with the objective left unchanged; no step of the three generates a response. Figure 1 illustrates the pipeline.

3.1 GENERATION-FREE CORRECT-ANSWER SCORING

Let $h(x_i)$ denote the model's multimodal chat rendering of the image, question, options, and user-side reasoning instruction. We append the same assistant prefix c to every candidate, stopping immediately before the answer letter. Specifically, c supplies a short image-inspection statement followed by an opening boxed-answer delimiter (Appendix A). The statement is supplied text, not a generated reasoning trace. The gold letter is withheld from the input and used only to index the output distribution.

For vocabulary V, let $z_i \in \mathbb{R}^{|V|}$ be the next-token logits at the end of $h(x_i) \,\|\, c$, and let $t(a)$ map an option letter to its token identifier. We define

$$\begin{aligned} s_i &\equiv \log \pi_{\theta_0}(t(a_i^\star) \mid h(x_i) \,\|\, c) \\ &= z_{i,t(a_i^\star)} - \log \sum_{v \in V} \exp(z_{i,v}). \end{aligned} \tag{3}$$

A larger score indicates greater probability assigned to the correct letter in this scoring context. We evaluate the model in inference mode, verify letter-token identities in the prefix context, and compute log-softmax in float32. Each candidate requires one forward pass and no decoding or backward pass. The score and selected example use the same option ordering.

The normalization in Equation 3 spans the full vocabulary rather than the option letters alone; Appendix A decomposes it. We use s_i directly, without option-only renormalization. It measures a conditional next-token likelihood, not the probability of solving the question after freely generated reasoning.

3.2 FIXED-INTERVAL SELECTION

Let σ be a permutation of the pool indices in increasing order of s_i, using image path as a deterministic tie-breaker. With zero-based ranks, offset L, and $0 \le L \le N - B$, retain

$$\mathcal{S}_{L,B} = \{(x_{\sigma(j)}, a^\star_{\sigma(j)}) : L \le j < L + B\}. \tag{4}$$

Our main rule, GOLDLP-MATCH, uses $L = 500$ and $B = 1{,}000$ for $N = 3{,}023$. The offset is not a free parameter fitted to accuracy. It was set so that the retained band would hold about as many initially unsuccessful questions as the probe-based arm it is compared against, 44.3% against 46.7%,

and the name records that match; Appendix A tabulates coverage against offset. Coverage enters only through this choice, never through the rule itself, which reads s_i alone. Matching it narrows what differs between the two arms without eliminating it: a subset chosen from one forward pass per candidate against one chosen from sixteen generations. The window removes the 500 lowest-scoring and 1,523 highest-scoring candidates. The retained interval lies approximately between the 16.5th and 49.6th percentiles; it is not centered at the median. The GOLDLP-HARD control instead uses $L = 0$ with the same score and subset budget. The two rules share ranks $[500, 1000)$ and exchange 500 examples, isolating the retention interval from the scoring procedure.

Both windows were specified before their respective training runs, following exploratory pilot analyses. Selection remains fixed throughout each run: as a procedural choice, we do not recompute scores or revise the interval during training, and retained questions receive no additional sampling weight, reward multiplier, or loss coefficient. The selected pool defines the empirical distribution

$$\widehat{P}_{L,B} = \frac{1}{B} \sum_{j=L}^{L+B-1} \delta_{(x_{\sigma(j)}, a^{\star}_{\sigma(j)})}, \tag{5}$$

with δ a point mass; training shuffles this fixed pool and revisits it as needed.

3.3 OBJECTIVE ON THE SELECTED DISTRIBUTION

The same response-level advantage applies to every valid response token. We retain the clipped GRPO objective of Shao et al. (2024), with a sampled reference-policy penalty at coefficient $\beta = 0.01$ and a mask over padding and tokens beyond the response. Appendix B states the policy ratio, the objective, the clipping and the penalty exactly. This objective, reward and optimization recipe are identical across the static arms defined in Section 4; GOLD-LP changes only Equation 5.

4 EXPERIMENTS

Our evaluation is driven by three research questions. **RQ1 (Effectiveness):** does correct-answer likelihood select useful data at a fixed subset budget? **RQ2 (Placement):** does retaining an interval differ from selecting the lowest scores or from ranking only initially unsuccessful questions? **RQ3 (Cost):** how much generation does the selection procedure require?

4.1 EXPERIMENTAL SETUP

Datasets and splits. We use the 3,360 DOCCI-derived visual multiple-choice questions released by Wu et al. (2026): 3,023 training candidates and 337 held-out questions, with distinct images and no overlap between partitions. The main static selections contain 1,000 candidates. The released answer key is skewed, with 48.5% of gold letters B and 2.6% D, so we balance answer positions once on the full candidate pool before scoring (Section 2.1), keeping the orientation identical for questions shared across arms; balancing is exact to within one question on the pool, though not within selected subsets. The held-out set is left untouched, since rotation-averaged evaluation already makes its key uniform. For transfer we use a separate 691-question geometry subset drawn from Geo170K (Gao et al., 2023) via the VisualReasoning split of Wu et al. (2026), without additional training on these questions.

Comparators. We compare our two selections against eight others: six share the 1,000-example budget, while full-pool training and GRESO use every candidate. **Random** draws its subset without replacement and reuses it across runs. **Full pool** trains on the whole pool under balanced answer positions, with the original key reported for context; the untrained backbone measures the starting point. **Static InSight** (Zhou et al., 2026) applies its acquisition score once to the initial pool, without the online posterior updates. The **ZPD-inspired** control (Yang et al., 2026) ranks Rasch information from the negated score, omitting the published correctness-based calibration stage. **RIPE** and **UNRIPE** keep the highest- and lowest-scoring halves of the initially unsuccessful questions over a shared core, holding subset size and unsuccessful-question count fixed. The **ported UFO-RL band** (Zhao et al., 2025) applies the mean-centred retention rule of that method to our score, which places it at ranks $[1083, 2083)$; it is not a reimplementation of its scoring procedure. **GRESO**

(Zheng et al., 2025) filters before rollout using training history, on the full pool under original answer positions; a separate incomplete dynamic-sampling run is excluded from the matched-step comparison. Appendix A gives the implemented forms, the reduction behind the ported band, and the exact RIPE/UNRIPE construction. The term initially unsuccessful refers to the finite probe, not permanent unsolvability.

Training protocol. The main study trains Qwen3-VL-8B-Instruct on eight B200 GPUs. Each run uses 176 training steps, with 256 questions and five generated responses per rollout batch, prescribing 225,280 responses. The actor global batch size is 64, the learning rate is 10^{-6}, weight decay is 10^{-2}, and the reference-policy penalty coefficient is 0.01. We train in bfloat16 with both visual and language components unfrozen. All static arms share the reward and optimizer recipe in Section 3.3, and we evaluate the final prescribed checkpoint rather than selecting checkpoints by held-out accuracy.

The reported study comprises 28 Qwen training runs at the full step budget, corresponding to 6,307,840 prescribed responses. The eight core arms have three runs each. The ZPD-inspired, ported UFO-RL and GRESO comparisons each have one run and are exploratory. A further single-run, incompletely documented LIMR-inspired comparison is reported separately in Appendix I. Matching training steps holds the nominal response count fixed for static arms, but does not match epochs, distinct examples, or exact computation.

Rotation-averaged evaluation. For each held-out question, we generate three responses under each of four cyclic option rotations. Each item's correct answer therefore occupies every letter exactly once. Following MMBench's cyclic-permutation strategy (Liu et al., 2024), we control answer-position effects, but average correctness rather than require every permutation to be answered correctly. Our metric averages correctness over the four rotations and the three responses drawn under each, so a fixed-letter policy scores 25%, although the protocol does not make the model invariant to arbitrary option permutations; Appendix A states the estimator. We extract the boxed answer; missing or unparseable answers count as incorrect. Accuracy is separate from the training format reward.

Uncertainty and comparison scope. We report means and sample standard deviations across training runs. Pairwise contrasts use 20,000 bootstrap replicates, jointly resampling paired item identities and each arm's runs, with unadjusted 95% percentile intervals. We retain the stored Holm-adjusted tail probabilities over families of 13 in-distribution and 10 geometry comparisons against random, assembled after results became available rather than pre-registered. With three runs these estimates remain approximate, and single-run comparisons are not used to establish replicated improvements. Appendix A states the family membership, the bootstrap seed, and which run-to-run variation the reported spread does and does not capture.

4.2 Main results (RQ1: Effectiveness)

In-distribution accuracy. The score measures a next-token likelihood rather than success after reasoning, so its usefulness for selection is settled here, by the performance of models trained on the data it selects. GoldLP-Match reaches 49.91% on DOCCI, 2.54 points above size-matched random selection and 3.88 above static InSight (Table 1). Match has the highest mean among the evaluated 1,000-example selections, but its comparison with random has an unadjusted 95% interval of $[-0.21, 5.33]$ points and Holm-adjusted $p = 0.6354$. We therefore report a numerical advantage without claiming a resolved in-distribution improvement. Balanced full-pool training reaches 52.76%, exceeding Match by 2.85 points; the Match–full-pool interval is $[-5.15, -0.58]$ with Holm-adjusted $p = 0.0134$. The evidence does not support replacing full-pool training for maximum in-distribution accuracy at this budget.

Geometry transfer. On the separate geometry holdout, Match achieves 84.52%, 1.35 points above random (Table 1), with an unadjusted interval of $[0.41, 2.30]$, raw bootstrap $p = 0.0050$, and Holm-adjusted $p = 0.0500$. This is the strongest positive comparison among fixed-size selectors on geometry, but lies at the conventional adjusted threshold and depends on the stated post-hoc family. Full-pool training scores 84.01%; Match's $+0.51$-point difference has interval $[-0.79, 1.77]$ and adjusted $p = 1.000$. Similar means do not establish equivalence to full-pool training.

Table 1: Qwen rotation-averaged accuracy (%), as means ± sample standard deviations. Bold marks the best 1,000-example selection per holdout. Single-run rows are comparable to the three-run arms only at the shared seed (Section 4.3). Adapted selectors follow (Zhou et al., 2026; Yang et al., 2026; Zhao et al., 2025; Zheng et al., 2025).

Selection	Rows	Runs	DOCCI	Geometry
Untrained Qwen	–	–	37.78	–
Full pool (balanced)	3,023	3	52.76 ± 0.54	84.01 ± 1.04
Full pool (original key)	3,023	3	49.33 ± 0.59	83.30 ± 0.64
Random	1,000	3	47.37 ± 1.11	83.17 ± 0.21
Static InSight	1,000	3	46.03 ± 0.75	83.00 ± 0.51
RIPE	1,000	3	48.53 ± 0.46	84.00 ± 0.24
UNRIPE	1,000	3	48.41 ± 1.06	82.34 ± 0.79
GoldLP-Hard (Ours)	1,000	3	48.83 ± 0.06	83.68 ± 0.97
GoldLP-Match (Ours)	1,000	3	$\mathbf{49.91 \pm 0.43}$	$\mathbf{84.52 \pm 0.63}$
ZPD-inspired	1,000	1	44.86	83.60
UFO-RL band (ported)	1,000	1	46.71	83.41
GRESO (online; original key)	3,023	1	49.65	82.68

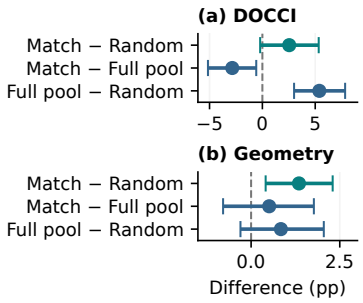


Figure 2: Paired accuracy contrasts in percentage points. Points are mean differences; bars are unadjusted 95% bootstrap intervals. The full-pool advantage over random largely disappears on geometry.

What outcome-based selection buys. None of the four outcome-dependent controls (static InSight, ZPD-inspired, RIPE, and UNRIPE) improves on random after the reported multiple-comparison adjustment on either holdout; their in-distribution mean differences from random are -1.34, -2.51, $+1.16$, and $+1.04$ points. The supplementary LIMR-inspired comparison likewise shows no adjusted positive difference (Appendix I), though its acquisition procedure is unverified and it does not support conclusions about the published method. This is specific to these implementations and budgets, and does not show outcome-based selection to be generally ineffective; nor does a significant comparison against random for one method and a nonsignificant one for another establish a difference between those methods. Figure 2 reports the Match and full-pool comparisons rather than every control. Two arms sit outside it. GRESO reaches 49.65% in distribution from the full pool in a single run, within the across-seed deviation of Match, but 82.68% on geometry; training on three times as many rows under a different answer orientation makes it a system comparison rather than a selector contrast. The ported UFO-RL band is the one arm without contrast statistics, and Section 4.3 compares it with Match at their shared seed.

Second-backbone check. The curated Gemma-3-4B-IT comparison has one training run per arm and remains provisional. Within its common evaluation cohort, Match scores 39.12% on DOCCI against 35.68% for random. However, geometry accuracy is 39.04% for Match and 40.55% for random. Thus the observed in-distribution ordering does not extend to geometry in this run. Cohorts are kept separate, and Table 3 in Appendix C reports all available values. These measurements do not establish replicated cross-backbone gains, and the geometry reversal leaves the transfer benefit unconfirmed.

4.3 Ablations and analyses (RQ2: Placement)

Which interval should be retained? Match and Hard use identical scores, selection cost, and subset size, sharing 500 questions. Shifting the retained interval away from the lowest scores yields mean increases of 1.08 and 0.84 points over Hard (Table 1). Appendix F plots the individual runs. A third window tests whether that gain continues above Match. At the seed all three share, the ported UFO-RL band reaches 46.71% on DOCCI against 49.53% for Match and 48.76% for Hard, and 83.41% on geometry against 83.84% and 82.57% (Figure 3b,c). Neither holdout therefore rises with the offset beyond Match: moving off the lowest scores helps, continuing past Match does not. Score, subset size and training budget are identical across the three, so the contrast isolates placement alone, and our rule finds the better position on it: higher than either alternative in distribution. The region that trains best differs from the mean-centred band in what it holds as much as in where it sits: 44.3% of its questions went unsolved by the initial probe, against 17.7% of theirs (Figure 3a). The in-distribution drop of 2.82 points is roughly three times the 0.99-point deviation expected for a difference of two single runs (Appendix A), while the 0.43-point geometry difference is smaller than

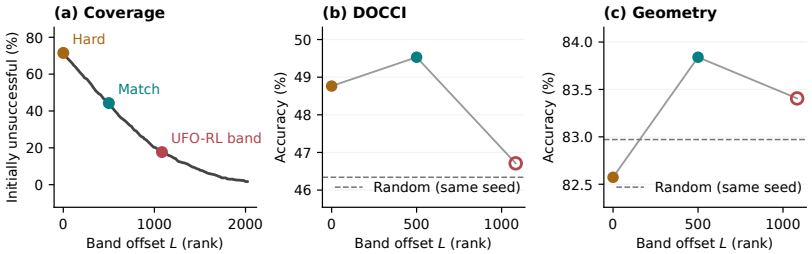


Figure 3: Where the retained band sits on the ranking. (a) Share of retained questions the initial probe never solved, against band offset L, with the three trained bands marked. (b, c) Rotation-averaged accuracy at those offsets, all at the shared seed; the open marker is the single-run ported band. Lines join measured points and do not interpolate untrained offsets.

Table 2: Offline selection-signal requirements. Scoring forwards and generated responses are distinct units; the scoring column counts standalone passes beyond those generation already performs. The ported band entry is the cost of the arm we ran, not of the published method. Adapted selectors follow (Zhou et al., 2026; Yang et al., 2026; Zhao et al., 2025; Zheng et al., 2025).

Selection	Scoring forwards	Generations
Full pool / Random	0	0
GoldLP-Match / Hard (Ours)	3,023	0
UFO-RL band (ported)	3,023	0
Static InSight	0	48,368
RIPE / UNRIPE	3,023	48,368
ZPD-inspired	3,023	48,368
GRESO (no offline probe)	0	0

the corresponding 0.98-point figure and we do not read it as a difference. These comparisons bound the useful interval between the extremes without identifying its optimum, and do not separately attribute the gain to removal of the lowest tail versus inclusion of the replacement questions. The ported band follows from the published rule rather than from any choice of ours.

Where the differences appear. Partitioning the holdout by initial probe outcome locates where the differences sit (Appendices D and E). GOLDLP-HARD exceeds Match by 5.3 points on the 119 initially unsuccessful questions but falls 4.6 points below it on the other 218; because the unsuccessful stratum is 35.3% of the holdout, Match stays ahead overall. Static InSight retains no initially unsuccessful training questions at this budget. This decomposition is descriptive and does not establish coverage as the cause.

Does likelihood identify better initially unsuccessful questions? RIPE and UNRIPE differ by 0.12 points in distribution despite their deliberately different unsuccessful-question halves, and by 1.66 points on geometry (Table 1). Both subsets contain 46.7% initially unsuccessful questions; their similar in-distribution means do not establish that coverage causes the observed outcome. These results do not support a universal claim that question identity is irrelevant. A separate predictive pilot finds no advantage either (Appendix G).

4.4 SELECTION OVERHEAD AT A FIXED TRAINING BUDGET (RQ3: COST)

Table 2 reports selection-signal requirements separately from training. The 16-attempt pool probe requires 48,368 responses, equal to 21.5% of the nominal 225,280-response training budget, whereas the generation-free rules score each candidate in a single forward pass and generate nothing. Static InSight alone pays that probe without also computing correct-answer scores.

Every completed static arm retains the same prescribed training budget, so fewer distinct training questions do not imply proportionately less training computation; Appendix A gives the per-question presentation counts. We report operation counts rather than training-hour comparisons here and make no end-to-end speedup claim. The supported efficiency result is the removal of front-loaded

generation from the selection stage relative to the evaluated probe-based controls, alongside the measured downstream accuracy.

5 Related Work

Which questions enter RLVR training is an established problem; obtaining the ranking is the hard part. Existing selectors fall into three categories by what they spend: sampled responses, an earlier training run, or the model's own probability.

Outcome-based selection and allocation. One line scores candidates from sampled outcomes. Sampling for learnability keeps the questions the model solves about half the time (Foster et al., 2025), needing rollouts for every candidate before training begins. InSight ranks by a mutual-information objective over Beta beliefs refreshed from training rewards, so the cost recurs (Zhou et al., 2026). DEPO selects offline on difficulty, diversity and influence, then keeps filtering during optimization (Tang et al., 2025). Others allocate rather than select: dynamic sampling discards prompt groups whose rollouts are all correct or all incorrect (Yu et al., 2025), and GRESO skips prompts it predicts uninformative from reward dynamics (Zheng et al., 2025), which needs a training history to exist. Each pays for its ranking in generated responses. We evaluate InSight's rule as a static offline selector, so conclusions about it do not transfer to the online algorithm.

Likelihood signals and interval selection. A second line uses model likelihood, which reduces generation but not always to zero. Dataset Cartography uses the true-label probability and its variability across training epochs, which requires those epochs (Swayamdipta et al., 2020). ZPD Detector adds correctness-based calibration, so sampled outcomes re-enter (Yang et al., 2026). UFO-RL is closest to our aim of ranking from probability rather than sampled success: it scores average answer-token log probability and keeps examples near the pool's mean confidence (Zhao et al., 2025). GOLD-LP reads a single token instead, and keeps a band offset below the median rather than centred on it. Which part of a ranking to keep is itself established: Moderate Coreset shows the useful interval shifts with the setting and proposes examples near the median (Xia et al., 2023), so excluding extremes is not novel alone. What we claim is the location, which Section 4 tests against both alternatives.

Learning from difficult prompts and visual curricula. A third direction changes the learning signal rather than the training set. RL-ZVP shapes the advantages of groups whose rollouts share a reward, so groups conventionally filtered out still contribute gradient (Le et al., 2026). DIVA-GRPO generates harder or easier variants of the questions themselves (Gao et al., 2026). Both alter how examples are used rather than which are available, and GOLD-LP leaves the objective and the question content unchanged. A complementary direction stages visual post-training by capability (Wu et al., 2026); we use its DOCCI-derived questions to study retention within a stage.

6 Conclusion

We identified a cost that data selection for reinforcement learning with verifiable rewards leaves unaccounted: producing the ranking at all requires generating over the whole candidate pool before optimization begins, whether or not the ranking helps. To remove it we proposed GOLD-LP, which scores each candidate with the initial model's likelihood for the correct option letter in one forward pass, retains a contiguous interval of that ranking, and leaves the GRPO objective and the question content unchanged. No step generates a response, so what this removes is generation from selection rather than the prescribed training budget. Across three training runs the selected subset reaches the highest mean accuracy among the evaluated fixed-size selections on both holdouts, and with the score and subset size held fixed accuracy is higher at our interval than at the lowest-score window. Full-pool training remains stronger in distribution, the in-distribution advantage over random is not resolved, and a single-run second-backbone replication does not reproduce the geometry benefit (Appendices C and H). Within the studied setting, these results put the choice of retention interval on the same footing as the choice of scoring rule.

AI USE STATEMENT

Generative AI tools assisted with literature search, research-idea discussion, manuscript drafting and editing, LaTeX preparation, figure preparation, and analysis of experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The selection equations, scoring prefix, training configuration, and evaluation protocol are specified in Section 3, Section 4, and Appendix A. Accuracy is aggregated from individual evaluation questions and generated responses. We specify baseline adaptations, the coverage of timing measurements, and the sources of variation represented by repeated training runs.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Thomas Foster, Anya Sims, Johannes Forkel, Mattie Fellows, and Jakob Foerster. Learning to reason at the frontier of learnability. *arXiv preprint arXiv:2502.12272*, 2025. URL https://arxiv.org/abs/2502.12272.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. DIVA-GRPO: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.01106.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. URL https://arxiv.org/abs/2312.11370.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21880.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. LIMR: Less is more for RL scaling. *arXiv preprint arXiv:2502.11886*, 2025. URL https://arxiv.org/abs/2502.11886.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2307.06281.

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.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Xinyu Tang, Zhenduo Zhang, Yurou Liu, Wayne Xin Zhao, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Towards high data efficiency in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2509.01321*, 2025. URL https://arxiv.org/abs/2509.01321.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. *arXiv preprint arXiv:2605.20177*, 2026. URL https://arxiv.org/abs/2605.20177.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7D5EECbOaf9.

Bo Yang, Yunkui Chen, Lanfei Feng, Yu Zhang, and Shijian Li. ZPD detector: Data selection via capability–difficulty alignment for large language models. *arXiv preprint arXiv:2601.10986*, 2026. URL https://arxiv.org/abs/2601.10986.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, 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, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Zhao, Kai Xiong, Xiao Ding, Li Du, Yangou Ouyang, Zhouhao Sun, Jiannan Guan, Wenbin Zhang, Bin Liu, Dong Hu, Bing Qin, and Ting Liu. UFO-RL: Uncertainty-focused optimization for efficient reinforcement learning data selection. *arXiv preprint arXiv:2505.12457*, 2025. URL https://arxiv.org/abs/2505.12457.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for LLM reasoning via selective rollouts. *arXiv preprint arXiv:2506.02177*, 2025. URL https://arxiv.org/abs/2506.02177.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. Mitigating selection bias in large language models via permutation-aware GRPO. *arXiv preprint arXiv:2603.21016*, 2026. URL https://arxiv.org/abs/2603.21016.

Xinyu Zhou, Boyu Zhu, Haotian Zhang, Huiming Wang, and Zhijiang Guo. Efficient RLVR training via weighted mutual information data selection. *arXiv preprint arXiv:2603.01907*, 2026. URL https://arxiv.org/abs/2603.01907.

A IMPLEMENTATION AND REPORTING DETAILS

Scoring prefix. The exact assistant text appended after the model's chat template is:

```
<think>I have examined the image.</think>\boxed{
```

There is one literal backslash before `boxed`. The prefix is supplied, not generated, and contains no correct answer. Its wording is a measurement choice; it does not establish that the resulting probability isolates visual perception. The user-side suffix requests reasoning tags followed by a boxed answer. Its instruction names `<think>` while its example uses `<thinking>`; our format verifier accepts either matched pair. This modification is shared across training arms. The score computation checks option-token identity in the prefix context and uses a model-specific scoring checkpoint.

Decomposition of the score. The normalization in Equation 3 spans the full vocabulary. To clarify its meaning, define the total option-token mass m_i and the conditional probability among option letters $\widetilde{p}_i$:

$$m_i = \sum_{a \in \mathcal{A}} \pi_{\theta_0}(t(a) \mid h(x_i) \,\|\, c), \qquad \widetilde{p}_i = \frac{\pi_{\theta_0}(t(a_i^\star) \mid h(x_i) \,\|\, c)}{m_i}. \tag{6}$$

Thus $s_i = \log m_i + \log \widetilde{p}_i$: the score reflects both relative answer preference and the probability of emitting an option token.

Rotation validation and cached evaluation. Balancing orders candidates by a stable hash of their original question text and assigns target letters A–D round-robin, so letter counts differ by at most one; each candidate then receives the cyclic rotation that moves its correct option to its assigned letter. With M items the rotation-averaged estimator is

$$\widehat{\mathrm{Acc}} = \frac{1}{12M} \sum_{i=1}^{M} \sum_{k=0}^{3} \sum_{r=1}^{3} \mathbf{1}[\hat{a}_{ikr} = a_{ik}^\star]. \tag{7}$$

Non-identity rotations require exactly four consecutive option lines; malformed blocks raise an error in the rotation operator. The identity rotation returns the original prompt unchanged, allowing reuse of compatible cached responses. Evaluation records per-rotation accuracies as well as their aggregate. An original-orientation majority-letter frequency describes only the unrotated labels and is not the baseline for rotation-averaged evaluation; the latter gives a policy that always emits one fixed valid letter exactly 25% accuracy. The training-pool builder separately warns and retains a row if its option block cannot be parsed, so error handling in that caller should not be confused with the evaluation operator's validation.

Image and training settings. Images are converted to RGB and resized when necessary to lie between 262,144 and 4,194,304 pixels before model processing. Qwen uses fully sharded training, actor update/experience microbatches of 8/32 per device, maximum gradient norm 1.0, and no warmup. Rollout sampling uses temperature 1.0 and top-p 1.0, with prompt and response limits of 4,096 and 2,048 tokens; evaluation uses temperature 0.6, top-p 0.95 and a 2,048-token limit. Rotation k moves option text at position j to $(j + k) \bmod 4$ and remaps the gold letter with it. The balanced pool holds 756, 756, 756 and 755 gold answers at positions A–D. All selections are evaluated at the fixed training-step budget, without validation-based checkpoint selection. Gemma disables padding-free packing and uses microbatches of one for compatibility and memory constraints. Its results are separated by hardware platform.

Choice of retention offset. The offset L of GOLDLP-MATCH was fixed so that the retained band would sit at coverage comparable to the probe-based arm, whose subset holds 46.7% initially unsuccessful questions. Coverage falls monotonically with the offset, and $L = 500$ is the nearest match on a 250-rank grid:

Offset L	0	250	**500**	750	1,000	1,250	1,500	2,000
Unsuccessful (%)	71.5	57.6	**44.3**	31.6	20.3	13.2	7.9	1.7

A finer sweep places the exact coverage match near $L = 450$ at 46.5%; $L = 500$ was taken as the round value and fixed before the runs. These counts are reported after selection and are not consulted by `build_goldlp_arms.py`, which ranks on s_i alone.

Interpretation of seeds. The main Qwen runs vary data order and reuse a fixed subset per arm. Historical rollout-generation seed settings could not be verified for the reported checkpoints; current launcher defaults do not establish those historical settings. Consequently, the reported spread describes the available runs, without quantifying variability under independently varied generation seeds or independently redrawn subsets. Across-seed dispersion is reported as a sample standard deviation with the $n - 1$ denominator, per arm in Table 1 and pooled over the within-arm variances of the eight three-run arms, which gives 0.70 in distribution and 0.69 on geometry. Section 4.3 compares two single runs rather than two arm means, so the figure it quotes there is $\sqrt{2}$ times the pooled one, 0.99 and 0.98. The single-run arm contributes no variance estimate of its own; the deviation is carried over from the three-run arms.

Step budget and data reuse. At 176 batches of 256 questions, each static arm receives 45,056 question presentations. This corresponds nominally to 45.1 presentations per retained example for a 1,000-example subset and 14.9 for the full pool, with actual per-example counts determined by shuffling and batch construction. Therefore, using 33.1% of the candidate pool does not mean using 33.1% of the training computation. Preemption can also cause repeated work: one full-pool run recorded 253,440 generated responses despite the prescribed 225,280-response budget. We distinguish these infrastructure costs from the method's nominal rollout requirement. Measured wall-clock time supports the matched-budget design rather than any speedup: the completed static arms take 3.95–4.30 hours per run irrespective of subset size, of which 1.02–1.17 hours are rollout, while the online-filtering baseline takes at least 5.16 hours with 2.21 hours of rollout, because it must generate before it can filter. Runs that resumed after preemption report lower bounds. We draw no end-to-end speedup claim from these times.

Statistical reporting. The bootstrap pairs evaluation-item identities between arms and independently resamples each arm's training runs, using 20,000 replicates and random seed 20260921. Intervals are unadjusted percentile intervals. We retain the comparison families in the supplied statistical reports rather than redefining them to match only the main table. On DOCCI, the 13 contrasts against random include balanced and original-key full-pool training, Match, Hard, RIPE, UNRIPE, static InSight, ZPD-inspired, LIMR-inspired, GRESO, a shorter RIPE training run, and the untrained Qwen and Gemma references. On geometry, the ten contrasts contain the trained arms above except the shorter RIPE run; untrained references are absent. Separate families compare against the balanced full pool. The families were assembled after observing results, so the adjusted geometry value at 0.0500 should not be portrayed as robust to family choice. Single-run intervals omit training-run variability. Failure to reject a difference is not an equivalence test.

Baseline reproducibility. Static InSight uses $\alpha_i = 1 + k_i$, $\beta_i = 1 + 16 - k_i$, where k_i is the initial number of successes, and ranks

$$\bar{\phi}_i(1 - \bar{\phi}_i) \exp[-3(\bar{\phi}_i - 0.3)^2]\, I(R_{1:5}; \Phi_i), \qquad \bar{\phi}_i = \frac{\alpha_i}{\alpha_i + \beta_i}.$$

The posterior smoothing gives initially unsuccessful items nonzero scores; exclusion at a given subset budget is a ranking outcome, not a rule that assigns them zero value. Our ZPD-inspired control uses $b_i = 4(d_i - d_{\min})/(d_{\max} - d_{\min}) - 2$ for $d_i = -s_i$, fits θ using $\mathbf{1}[\hat{p}_i \geq 0.5]$ for the initial probe pass rate $\hat{p}_i$, and ranks $p_i(1-p_i)$ for $p_i = \mathrm{sigmoid}(\theta - b_i)$. All fitting uses the candidate pool. This omits the published error-aware calibration stage. The ported UFO-RL band keeps the candidates maximising $1 - (c_i - \mu)^2$ for confidence $c_i = \exp(\mathrm{Conf}(x_i))$ and pool mean μ, that is, those nearest the mean. Our answers are a single option letter, so their token-averaged confidence reduces to the score of Section 3.1; since $\exp$ is monotone the rule again selects a contiguous interval, ranks $[1083, 2083)$ here. We retain 1,000 rather than their 10% so that subset size matches the other arms. RIPE and UNRIPE split the 935 questions with zero correct responses in 16 attempts, taking the highest- and lowest-scoring 467 respectively, which leaves the median question in neither, over an identical 533-question core selected by static InSight score. The single-run LIMR-inspired result is descriptive; the comparison does not establish faithful reproduction of the published selector or its end-to-end cost.

B POLICY-UPDATE SPECIFICATION

For the sampled token at position t of response g, the policy ratio is

$$\rho_{i,g,t}(\theta) = \frac{\pi_\theta(y_{i,g,t} \mid h(x_i), y_{i,g,<t})}{\pi_{\theta_{\text{old}}}(y_{i,g,t} \mid h(x_i), y_{i,g,<t})}. \tag{8}$$

We retain the recipe's clipped GRPO objective (Shao et al., 2024); writing its tokenwise policy term as ψ and the sampled reference-policy penalty as $\widehat{k}$, the maximized minibatch surrogate is

$$\widehat{J}(\theta) = \frac{\sum_{i,g,t} M_{i,g,t}[\psi(\rho_{i,g,t}(\theta), A_{i,g}) - \beta\, \widehat{k}_{i,g,t}(\theta)]}{\sum_{i,g,t} M_{i,g,t}}, \qquad \beta = 0.01, \tag{9}$$

where M masks padding and tokens beyond the response. The objective retains asymmetric ratio clipping and the negative-advantage dual clip of the shared training recipe. Define

$$b(\rho, A) = \min\{\rho A, \text{clip}(\rho, 0.8, 1.3)A\}, \tag{10}$$

$$\psi(\rho, A) = \begin{cases} b(\rho, A), & A \geq 0, \\ \max\{b(\rho, A), 3A\}, & A < 0. \end{cases} \tag{11}$$

For numerical stability, the implementation clips the log policy ratio to $[-20, 20]$ before exponentiating the unclipped branch. The group standard deviation uses the sample normalization $G-1$, which follows the implementation rather than the population form of the original description (Shao et al., 2024). Advantages and old-policy probabilities are held fixed during each actor update, and policy and reference penalties are averaged over valid response tokens, rather than averaging per-response means.

For the sampled token at context $u_{i,g,t} = (h(x_i), y_{i,g,<t})$, the reference penalty is

$$\begin{aligned} d_{i,g,t} &= \text{clip}\left(\log \frac{\pi_{\text{ref}}(y_{i,g,t} \mid u_{i,g,t})}{\pi_\theta(y_{i,g,t} \mid u_{i,g,t})}, -20, 20\right), \\ \widehat{k}_{i,g,t} &= \text{clip}(\exp(d_{i,g,t}) - d_{i,g,t} - 1, -10, 10)\,. \end{aligned} \tag{12}$$

This sampled penalty is added to the actor loss with coefficient 0.01; it is not subtracted from the verifier reward before group normalization. It is an implementation of reference-policy regularization, not an exact vocabulary-summed KL divergence on each observed token. The actor minimizes $-\widehat{J}$ using AdamW. Neither the correct-answer score nor the selection rank appears in this loss.

Zero group advantage. One property of the design bounds what these measurements establish. When every response in a group receives the same total reward, Equation 2 gives zero reward advantage. This implies neither that the reference-policy term has zero gradient nor that a question unsolved by the initial probe stays uninformative during training, and GOLD-LP does not alter group advantages or add a gradient-recovery mechanism.

C SECOND-BACKBONE MEASUREMENTS

Cohort I corresponds to the common evaluation run used for the provisional Gemma selector comparisons; Cohort II is a separate run containing the full-pool reference. The random and InSight measurements differ across these cohorts by 2.52 and 1.31 points, respectively. Their causes have not been resolved, so neither cohorts nor duplicate evaluations are treated as independent training replications. The complete comparison requires matched re-evaluation. In particular, the 42.28% full-pool result from Cohort II is not a controlled comparator for Match's 39.12% in Cohort I. Each trained arm has one seed, and the geometry comparison is available only for Match and random in Cohort I.

Item-paired contrasts within Cohort I. The Cohort I evaluations share 337 questions, so item-level contrasts are available under the same 20,000-replicate procedure used for Table 1. Against random, Match gains 3.44 points with an unadjusted 95% interval of $[0.05, 6.80]$ and raw bootstrap $p = 0.0478$; the difference does not survive the Holm adjustment applied within this backbone's

Table 3: Provisional Gemma-3-4B-IT rotation-averaged accuracy (%). Each trained arm has one run. Cohort I contains the mutually comparable evaluations used for selector comparisons. Cohort II is a separate evaluation cohort and is not pooled with Cohort I.

Evaluation cohort	Selection	DOCCI	Geometry
I	Untrained	28.04	–
I	GoldLP-Match (Ours)	39.12	39.04
I	GoldLP-Hard (Ours)	36.97	–
I	RIPE	35.34	–
I	UNRIPE	37.24	–
I	Random	35.68	40.55
I	Static InSight	33.83	–
II	Full pool	42.28	–
II	Random	38.20	–
II	Static InSight	35.14	–

family of six comparisons against random. Every arm here has one training run, so these intervals resample items and not runs, and they capture none of the run-to-run variation that Table 1 reports for Qwen. They are stated so that the provisional Gemma numbers carry the same uncertainty information as the rest of the paper, not to upgrade them. We do not report contrasts against the Cohort II full-pool arm, for the reason given above.

How far the initial probe transfers across backbones. Whether the retained region is a property of the questions or of the scoring model bears on the geometry reversal. We compare the 16-attempt probe for Qwen3-VL-8B-Instruct with the same probe for Qwen2.5-VL-7B-Instruct over the same 3,023 rows in the same released orientation, so the two differ only in the backbone. The never-solved sets hold 890 rows (29.4%) and 543 rows (18.0%) with 320 in common, a Jaccard overlap of 0.2875 against 0.1256 expected from the marginals alone, or $2.29\times$ chance; the two pass-rate vectors agree at Spearman $\rho = 0.458$. Initial difficulty is therefore shared across backbones well above chance and far below identity. This is what a score defined on the initial model's own probabilities predicts, and it is consistent with a selection that transfers to a different backbone only in part. Two runs of one model set the scale for that overlap. This probe is a separate 16-attempt run from the one behind Appendix G; over the same rows and orientation the two never-solved sets are 890 and 1,068 with 767 in common, neither contained in the other, so they disagree on 424 of the 3,023 rows at a Jaccard of 0.6440. The released probe draws its 16 responses at temperature 1.0; the earlier run's decoding settings are not stored with its output, so the gap is reported as run-to-run variation rather than attributed to a setting. Cross-backbone agreement at 0.2875 is well below that, so the difference between backbones is not explained by probe variability alone. Coverage figures elsewhere in this paper come from a single balanced probe and carry the same variability, which is why Appendix H treats a zero-success result as an observation rather than a property.

D ACCURACY BY INITIAL PROBE OUTCOME

The in-distribution holdout contains 119 questions the original 16-attempt probe never solved and 218 it solved at least once. Table 4 partitions accuracy by this initial outcome and reports training *coverage*: the fraction of each training set that the balanced-pool probe never solved. These probes use different option orientations, so coverage and held-out strata are defined separately. All trained rows use the same three runs as Table 1; their overall accuracies reproduce that table.

The slices locate three differences hidden by aggregate accuracy. First, GOLDLP-HARD exceeds Match by 5.3 points on initially unsuccessful questions but falls 4.6 points below it on the other questions. Because the unsuccessful stratum accounts for 35.3% of the holdout, the latter difference outweighs the former, leaving Match ahead overall. Second, static InSight retains no initially unsuccessful training questions at this budget; its deficit relative to Match is 10.7 points on the unsuccessful holdout stratum, compared with 3.9 points overall. Third, balanced full-pool training is numerically close to Match on the unsuccessful stratum (27.05% versus 27.82%), but scores higher on the other questions (66.79% versus 61.96%), accounting for its higher aggregate accuracy.

These are descriptive decompositions, not isolated effects of coverage. Arms differ in question identity and, for the full-pool reference, subset size. Even the nearly coverage-matched Match and

Table 4: Rotation-averaged accuracy (%) by initial held-out probe outcome, averaged over the final three training runs. Unsuccessful denotes zero successes in the original 16-attempt probe; other denotes at least one success. Coverage uses the balanced training-pool probe. The untrained reference is a single checkpoint.

Arm	Runs	Coverage	Overall	Unsuccessful (119)	Other (218)
Static InSight	3	0.0%	46.03	17.13	61.80
Random	3	30.6%	47.37	19.70	62.47
Balanced full pool	3	30.9%	52.76	27.05	66.79
GoldLP-Match	3	44.3%	49.91	27.82	61.96
RIPE	3	46.7%	48.53	21.34	63.38
UNRIPE	3	46.7%	48.41	20.12	63.85
GoldLP-Hard	3	71.5%	48.83	33.12	57.40
Untrained backbone	–	–	37.78	10.36	52.75

Table 5: Training accuracy on the candidate pool by initial probe outcome, in ten-step bins, from one full-pool run under the original answer key. Unsuccessful denotes zero successes in the 16-attempt probe, always solved denotes sixteen, and partial the remainder.

Steps	Unsuccessful	Partial	Always solved
1–10	0.053	0.483	0.934
11–20	0.066	0.532	0.948
21–30	0.096	0.563	0.965
31–40	0.126	0.628	0.969
41–50	0.158	0.655	0.976
51–60	0.198	0.698	0.974
61–70	0.229	0.734	0.984
71–80	0.275	0.752	0.975
81–90	0.295	0.787	0.987
91–100	0.337	0.798	0.980
101–110	0.364	0.817	0.987
Change	$+0.311$	$+0.334$	$+0.053$

RIPE selections differ on both strata. The table therefore identifies where the accuracy differences occur without establishing coverage as their cause.

E Training dynamics by initial difficulty

Table 4 splits the *holdout* by initial probe outcome. Table 5 reports the training-time counterpart over the candidate pool, refining that table's *other* column into questions the probe solved sometimes and questions it always solved. The source is the rollout log of a single full-pool run under the original answer key: 140,800 records over the 110 steps recorded before the run was interrupted. It is one run, it is not the balanced pipeline behind Table 1, and it is descriptive.

The two moving strata gain comparable amounts, while the always-solved stratum has almost no headroom left. They also start moving at different times: measured against its own first-bin baseline, the partial stratum clears $+0.02$ at steps 11–20 and the unsuccessful stratum only at steps 21–30. Read alongside the coverage column of Table 4, this is consistent with the two ends of the ranking being weak for different reasons, a lowest-score window holding a large share of questions that begin near zero and move late, and a window near the pool mean holding a large share already close to ceiling. One run under a different answer key does not establish the mechanism behind the interval result in Section 4.3, and we do not read it as doing so. It does show that questions with zero successes in a finite probe acquire reward signal as training proceeds, which is the observation Appendix H relies on.

F Per-run selection ablations

Figure 4 plots the individual training runs behind the selection comparisons in Section 4.3. The arms left of the vertical rule are the three retention windows; those right of it are the initially unsuccessful split. Means and sample deviations for every arm appear in Table 1.

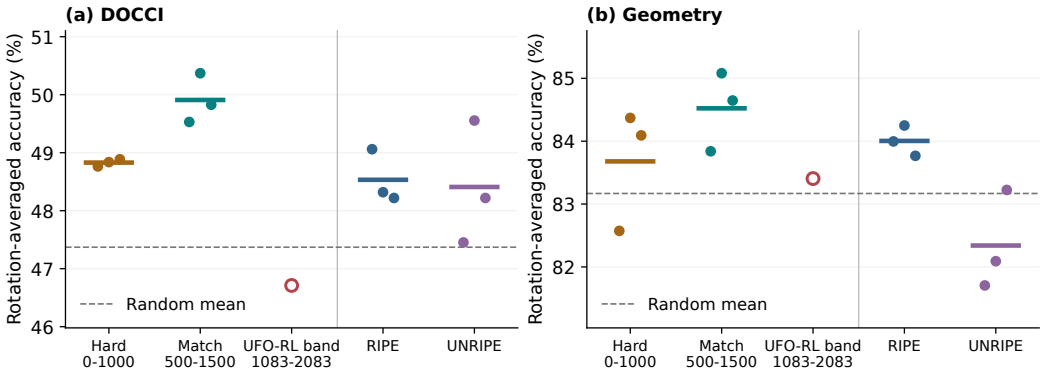


Figure 4: Selection ablations across individual Qwen runs. Dots are run accuracies, segments their means, and the dashed line the random-selection mean. Left of the vertical rule, arms are ordered by band position; right of it, RIPE and UNRIPE differ only in the unsuccessful-question half. The open marker is the single-run ported band, without a mean segment. Dots are offset horizontally and do not imply paired seeds; these plots are descriptive, not significance tests.

G PILOT ANALYSES: PREDICTION, SELECTION, AND ANSWER POSITION

These preliminary experiments use the original answer positions and precede the balanced-pool study. They informed the method's development and are not additional independent replications of its main evaluation.

Predicting subsequent solvability. Among 119 held-out questions with zero successes in the initial 16-attempt probe, we label a question as subsequently solved when a full-pool-trained model answers it correctly in at least two of three evaluation responses at the original option orientation. Initial correct-answer log probability predicts this label with AUC 0.702. The higher-scoring half has a subsequent success rate of 55.9%, compared with 23.3% for the lower-scoring half; the difference is 32.6 percentage points with a 95% bootstrap interval of $[15.8, 49.4]$. Using a second full-pool training run to define the outcome gives AUC 0.723 on the same questions. This checks sensitivity to the training outcome, not generalization to a fresh set of questions. The initial analysis verified a previously inspected result and should not be interpreted as a blind discovery.

Testing selection utility. In the original-orientation training pool, 1,068 examples had zero initial successes. Pilot RIPE and UNRIPE selections each contain 534 of these examples and the same 466 other examples, differing only in taking the higher- or lower-likelihood half. Under rotation-averaged evaluation, their accuracies are 48.34% and 48.32%; the paired difference is 0.02 points with an item-bootstrap interval of $[-2.42, 2.45]$. The interval does not establish equivalence. Each selection has one training run, so this interval omits training-run variability. These pilot subsets differ from the balanced-pool controls in the main study and are not pooled with them. Thus, predictive discrimination did not yield a resolved selection advantage in this pilot.

Sensitivity to answer positions. For the pilot full-pool model, original-orientation accuracy is 64.69%, compared with 49.58% after rotation averaging. The corresponding base-model values are 38.38% and 37.98%. This difference motivated balancing the training answer positions and using rotation-averaged evaluation in the main study. It describes sensitivity in this experimental setting, not an estimate of inflation on other benchmarks.

Answer orientation and subset composition. Balancing makes the candidate pool uniform in gold-letter counts, but selection can change its composition: Match's A–D shares are 25.4%, 24.1%, 21.4%, and 29.1%, versus 26.0%, 24.5%, 24.1%, and 25.4% for Hard. Rotation-averaged evaluation controls the evaluation key, not the training subset's composition. An exploratory comparison of 755 questions shared by the originally oriented and balanced initially unsuccessful strata gives a Spearman score correlation of 0.8558; 565 questions undergo a non-identity option rotation. The

latter count describes changed orientations, not necessarily changed ranks. This supports recomputing scores after balancing, but does not isolate answer position as the cause of the selection differences. Original-key full-pool training scores 49.33% on DOCCI, compared with 52.76% for balanced full-pool training; this contextual comparison also changes the training inputs and should not be interpreted as a pure evaluation effect.

Prediction after changing the scoring orientation. An auxiliary analysis labels initially unsuccessful training examples by whether their mean response accuracy exceeds 20% during training steps 91–110. Of 1,068 candidate examples, 1,052 have observations in this window and 445 meet the threshold. Scores at the original option positions yield AUC 0.6900; scores after answer-position balancing yield AUC 0.6639. This check uses training-pool observations and a different label from the held-out analysis above; the AUCs are not directly interchangeable.

H Scope and outstanding tests

Our study uses one DOCCI-derived selection pool, a 337-question in-distribution holdout, and one geometry transfer subset. Three Qwen training runs and single-run auxiliary controls limit statistical precision. Correct-letter scores may depend on answer position, prefix wording, tokenization, and answer-format probability. The current experiments do not isolate these effects or establish an optimal interval. Direct confidence-based comparisons, neighboring windows, coverage-matched random selection, additional backbones and training pools, and independent rollout seeds are needed. A zero-success initial probe is a finite-sample observation; it is not proof that an example will always produce zero reward variance during training. These limitations also preclude claiming universal transfer gains, full-pool equivalence, or a new mechanism for recovering gradients from zero-variance groups.

I Supplementary trajectory-based comparison

An available 1,000-question subset is attributed to reward trajectories from a preceding full-pool run and described as LIMR-inspired (Li et al., 2025). Its exact acquisition rule has not been verified, so we exclude it from the main baseline table and do not treat it as a reproduction of the published method. One training run obtains 47.13% in distribution and 83.83% on geometry, differences of -0.24 and $+0.66$ points from random. Its stored Holm-adjusted values are 1.00 and 0.78, respectively. It remains in the original statistical families; moving its presentation does not change the reported adjustments. The recorded signal requirement is a preceding training run, nominally 225,280 generated responses if those trajectories are unavailable. This is a descriptive accounting assumption, not a verified end-to-end cost measurement.