GRADIENT ELIGIBILITY IS NOT LEARNING UTILITY: HIDDEN CURRICULA IN VISION-LANGUAGE RL

Anonymous authors
Paper under double-blind review

ABSTRACT

General-purpose vision–language reinforcement learning often trains a single policy across heterogeneous perception and reasoning tasks using group-relative advantages with binary correctness feedback. Only rollout groups containing both correct and incorrect answers yield nonzero relative advantages through this outcome channel, making each query's contribution policy-dependent. Dynamic sampling amplifies this selection by discarding zero-variance groups and refilling the batch, so queries that remain mixed accumulate more updates, creating a hidden curriculum. We ask whether this persistence signals useful learning opportunities or merely reflects continued outcome variability. Before RL, we predict which queries will remain mixed using initial solvability and *response demand*, measured as the thinking length elicited by a fixed probe VLM. Across Qwen2.5-VL and Qwen3-VL, higher demand predicts longer persistence after accounting for initial solvability and query composition. Under dynamic sampling, the highest-demand decile consequently passes the filter 2.7–$4.0\times$ as often as the lowest-demand decile. Persistence, however, does not imply learning utility. Across matched three-seed interventions at two policy states, high-demand pools pass the filter more often and consume more generated response tokens. Starting from the base policy, both arms improve, but low-demand training gains 2.17 percentage points more at pass@1; all six paired pass@1 comparisons favor low-demand training. A past-only inverse-acceptance controller reduces exposure drift by 67.2% at near-matched cost, with higher observed pass@8 and pass@16. Thus, controlling which queries receive updates is distinct from deciding which updates are useful.

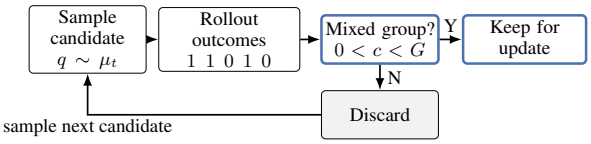
1 INTRODUCTION

Outcome-based reinforcement learning with group-relative advantages has become a standard recipe for eliciting reasoning in language and vision–language models (Shao et al., 2024; Liu et al., 2025; Zheng et al., 2025a). With a binary correctness signal, only rollout groups containing both correct and incorrect answers provide a relative outcome signal. Each query's contribution is therefore already policy-dependent. Dynamic sampling amplifies this selection by discarding zero-variance groups and continuing to sample until the optimizer batch is refilled (Yu et al., 2025).

This selection is especially consequential in general-purpose vision–language RL, where chart understanding, spatial reasoning, mathematics, and science share one policy and update budget despite differing in visual dependence, reasoning horizon, response format, and initial difficulty (Shen et al., 2025; Zhan et al., 2025; Wang et al., 2025b; V Team et al., 2026; Kimi Team, 2026). A task-agnostic mixed-group filter can therefore admit query groups from some task families more often than others. The resulting exposure imbalance can contribute to uneven task progress alongside objective or gradient conflict (Ma et al., 2025; Sarch et al., 2026).

We call a mixed-outcome group *gradient-eligible* through the outcome channel. Repeated filter-and-refill sampling turns this local signal into a hidden curriculum, which we formalize as a policy-dependent change of measure from the candidate mixture to the accepted query-group distribution (Figure 1(a)). Operationally, this mechanism treats continued gradient availability as a reason for further allocation. *Do queries remain eligible because they still offer useful learning opportunities, or simply because their outcomes remain variable?*

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.



(a) Dynamic sampling repeatedly keeps mixed groups

Starting policy	Base	Step 40
D_1 acceptance	.224	.691
D_4 acceptance	**.597**	**.816**
D_4/D_1 token cost	**1.68×**	**2.12×**
$D_1 - D_4$ pass@1	**+.0217**	+.0124

(b) More eligible, not more useful

Figure 1: **Higher filter acceptance does not imply greater learning gains. (a)** All-pass and all-fail groups are discarded and replaced; queries that remain mixed therefore enter more updates. **(b)** With 1,280 accepted groups per arm, D_4 passes the filter more often and uses more generated tokens, but mean held-out pass@1 favors D_1 from both policies. Values are three-seed means.

We address this question in two steps. First, we show that eligibility persistence is partly predictable before training from two simple measurements. Initial solvability measures how reliably the target model answers a query before training, while *response demand*, the median thinking-token length elicited by a fixed probe VLM, records how much deliberation the probe produces for it. Across Qwen2.5-VL and Qwen3-VL, higher demand predicts longer persistence after accounting for initial solvability and query composition. Under dynamic sampling, the highest-demand decile consequently passes the filter 2.7–4.0× as often as the lowest-demand decile.

Second, we test whether allocating more updates to these persistent queries produces greater held-out gains. In matched three-seed interventions starting from the Qwen3-VL base and step-40 policies, each low- and high-demand arm receives the same number of accepted groups (Figure 1(b)). Starting from the base policy, both arms improve substantially, but low-demand training gains 2.17 percentage points more at pass@1. Across both starting policies, all six seed-paired pass@1 contrasts favor low-demand training, while high-demand training consumes 1.68–2.12× as many generated response tokens. Persistent eligibility therefore does not by itself justify further allocation.

This allocation–utility mismatch motivates controlling the realized update distribution. We introduce a past-only inverse-acceptance sampler that proposes more candidates from strata that pass the filter less often while leaving the mixed-group criterion unchanged. Without task labels or future information, it removes 67.2% of demand-dependent exposure drift at 1.02× the generated-token cost of standard dynamic sampling, with higher observed pass@8 and pass@16. Controlling which queries receive updates is therefore distinct from deciding which updates are useful.

Our contributions are:

- **A two-stage account of query selection.** We separate policy-dependent eligibility inherent to the binary outcome channel of group-relative RL from its amplification by filter-and-refill, and show how persistent eligibility becomes cumulative update allocation.
- **Predictable persistence and an allocation–utility mismatch.** Pre-RL response demand predicts eligibility persistence beyond initial solvability, and dynamic sampling converts this persistence into exposure imbalance. Controlled interventions show that high-demand queries are accepted more often and consume more generated response tokens without providing greater observed held-out benefit.
- **Control of realized allocation.** Past-only acceptance statistics enable a lightweight inverse-acceptance controller that reduces exposure drift at near-matched cost.

2 Related Work

Query-level dynamics in multi-task VLM RL. General-purpose VLM RL trains shared policies across heterogeneous perception and reasoning tasks (Shen et al., 2025; Zhan et al., 2025; V Team et al., 2026; Kimi Team, 2026; Choi et al., 2026). Uneven benchmark progress has motivated task-aware rewards and verifiers (Xiaomi LLM-Core Team, 2025; Ma et al., 2025; Sarch et al., 2026), perception-aware or role-specific objectives (Wang et al., 2025b; Miao et al., 2026), task-level normalization and length shaping (Hu et al., 2026), and mixture optimization (Liang et al., 2025). These methods mainly intervene at the task or domain level. We study a preceding query-level selection channel: within the same task, queries differ in initial solvability, visual dependence,

response demand, and how long they remain eligible. Filtering can therefore turn policy–query dynamics into task-level exposure imbalance before task gradients are compared.

Curriculum and data allocation in RL. Classical curricula order examples by difficulty or learner competence; adaptive methods use learning progress, future learning potential, or domain staging (Bengio et al., 2009; Platanios et al., 2019; Graves et al., 2017; Jiang et al., 2021; Wang et al., 2025a). Recent RL selectors target the policy's current learning frontier. LILO prioritizes high reward variance, while online difficulty filtering and related predictors select intermediate-difficulty prompts (Foster et al., 2025; Bae et al., 2026; Qu et al., 2025; Gao et al., 2025). GRESO uses past reward dynamics to skip likely zero-variance prompts before rollout, whereas DAPO filters zero-variance groups after generation and refills the batch (Zheng et al., 2025b; Yu et al., 2025). Our focus is temporal and downstream: even at matched current solvability, queries can remain eligible for different durations. We study how refill converts this persistence into exposure and whether that exposure yields held-out value. Inverse acceptance realizes a chosen accepted mixture rather than defining a new priority, so it can be combined with these selectors.

Policy state and the value of learning signals. Methods that recover signal from zero-variance groups (Le et al., 2026) or introduce dense stepwise rewards (Zhang et al., 2025) expand the set of trainable queries. We ask a downstream question: once a query supplies a learning signal, does its persistence make it valuable for the current policy? Controlled chess experiments show that pretraining changes subsequent RL (Shen et al., 2026), while TailSFT removes already-fit sequences to improve coverage and later RL performance (Malladi et al., 2026). Within RL, pass@1 can diverge from high-k coverage (Yue et al., 2025), motivating analyses of overtraining and marginal-coverage-based update gating (Yuan et al., 2026). We use pre-RL target solvability and a response-demand probe to predict eligibility and generation cost, then test whether persistent signals yield greater held-out benefit under a fixed update budget.

3 Gradient Eligibility and Training Allocation

We distinguish three questions: which queries can supply an outcome-based learning signal, how that eligibility accumulates into training exposure, and whether the resulting allocation improves held-out performance.

3.1 From Mixed Outcomes to Accepted Exposure

For query q, let $p_t(q)$ denote the probability that one rollout receives a correct binary outcome under policy π_t. With $G \geq 2$ conditionally independent rollouts, the probability of observing both a success and a failure is

$$A_G(p) = 1 - p^G - (1-p)^G. \tag{1}$$

A_G is zero at $p = 0, 1$ and maximized at $p = 1/2$. We call it an *eligibility kernel*: the probability that a group supplies nonzero relative advantages through the binary outcome channel.

Let $\mu_t(q)$ be the candidate-query distribution and $\bar{A}_t = \mathbb{E}_{q\sim\mu_t}[A_G(p_t(q))]$ its mean eligibility. Suppose the policy and candidate distribution are fixed within a batch. Candidates are sampled independently, non-mixed groups are discarded, and sampling continues until the accepted batch is full. Provided $\bar{A}_t > 0$, the accepted-query distribution is

$$\mu_t^{\text{eff}}(q) = \frac{\mu_t(q) A_G(p_t(q))}{\bar{A}_t}. \tag{2}$$

Without refill, B_t candidate groups yield $B_t \mu_t(q) A_G(p_t(q))$ mixed groups from query q in expectation. Refill instead targets B_t accepted groups, scaling this count by $1/\bar{A}_t$ while preserving the relative distribution among mixed groups. Thus eligibility determines who can contribute through the outcome channel, whereas refill determines how strongly persistent contributors are represented. Cumulative accepted exposure is therefore $\sum_t B_t \mu_t^{\text{eff}}(q)$.

The symmetry $A_G(p) = A_G(1-p)$ also matters: low eligibility can reflect either reliable success or reliable failure. Mixed occupancy alone does not distinguish which boundary a query approaches.

In our implementations, the filter uses binary correctness and retains a group exactly when $0 < c < G$, where c counts correct rollouts. The optimization reward additionally includes a format term. Equation 2 therefore characterizes accepted query-group counts through the outcome gate, not gradient magnitude after reward weighting, clipping, or loss aggregation.

3.2 Initial Solvability and Eligibility Persistence

Current eligibility is only a snapshot; cumulative exposure depends on how long it persists. To make this distinction explicit, write $p_t(q) = \sigma(\theta_t(q))$ and define a band of appreciable eligibility,

$$\mathcal{B}_\varepsilon = \{\theta : A_G(\sigma(\theta)) \geq \varepsilon\} = [\theta_-, \theta_+],$$

where $0 < \varepsilon < A_G(1/2)$. Consider a local model in which each accepted query-specific update moves the query toward the success boundary by a fixed increment $\alpha(q; \pi_0) > 0$. A query initialized inside the band then requires approximately

$$N_{\mathrm{acc}}(q) \approx \frac{\theta_+ - \theta_0(q; \pi_0)}{\alpha(q; \pi_0)} \tag{3}$$

accepted updates to cross that boundary.

At the same starting point, a smaller α keeps a query eligible for more updates. Pre-RL solvability $s_0(q)$ estimates the starting position but not the rate of leaving the band, so similarly solvable queries can accumulate different exposure. Residence counts accepted query-local updates; because an accepted update requires $1/A_G(p)$ candidate presentations in expectation, accepted-update residence and candidate-presentation cost need not coincide.

This motivates looking for pre-RL predictors of persistence beyond initial solvability. We define *response demand* $D(q)$ as the median thinking-token length elicited before RL by a fixed probe VLM distinct from the target policy. We test whether $D(q)$ adds predictive information about subsequent mixed occupancy after controlling for $s_0(q)$; we do not equate it with the update increment α. Probe configurations and the successful-rollout-only alternative appear in Appendix C.1.

3.3 From Accepted Exposure to Learning Utility

Eligibility determines access to updates; learning utility concerns their downstream effect. Let a denote a query-allocation rule and let π_B^a be the policy obtained from a common checkpoint $\pi_\star$ after training on B accepted query groups, with the remaining training protocol held fixed. For a held-out performance metric M, define finite-budget utility as

$$\mathcal{U}_B(a; \pi_\star) = \mathbb{E}[M(\pi_B^a) - M(\pi_\star)], \tag{4}$$

where the expectation is over training randomness. This definition evaluates the resulting shared policy, including cross-query transfer and interference, without assuming additive query utilities.

The kernel and local model describe access to updates, not downstream value. At two policy states, we therefore match current solvability, fix the accepted-group budget, and test whether a higher-demand, more persistent pool produces greater held-out gains (Section 5.1). Generation cost is reported separately because equal accepted-group counts do not imply equal token budgets.

3.4 Experimental Setting and Measurements

We study Qwen2.5-VL-3B-Instruct on a 2,000-query ViRL39K subset and Qwen3-VL-2B-Thinking on a 6,000-query subset, with Qwen3-VL-2B-Instruct as an additional target on the latter (Bai et al., 2025b;a; TIGER-Lab, 2025). All primary runs use $G = 5$; full training, split, reward, and evaluation configurations appear in Appendix A.

For a group with c correct rollouts, we distinguish all-fail ($c = 0$), mixed ($0 < c < G$), and all-pass ($c = G$). For each query, *mixed occupancy* is the fraction of its attempted presentations that are mixed; it measures persistence across observations rather than uninterrupted residence. *Accepted exposure* counts groups admitted to optimizer batches. We evaluate downstream utility with held-out pass@k and report response-token costs for generated candidates and accepted groups separately.

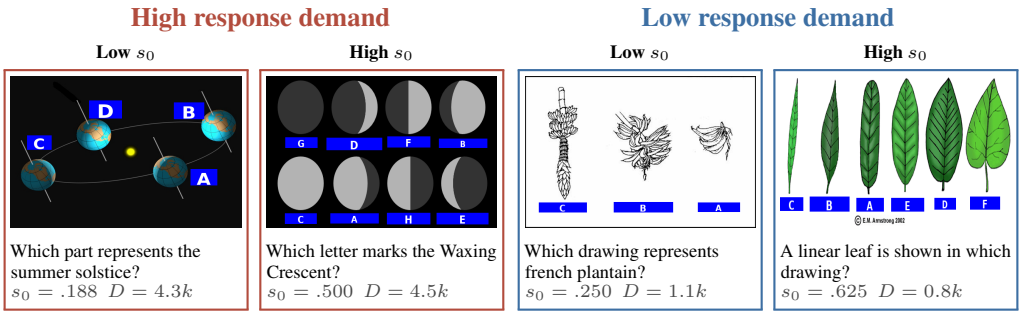


Figure 2: **One task category contains distinct query regimes.** AI2D examples span low/high pre-RL solvability and response demand. Thresholds are the Qwen2.5 medians ($s_0 = .375$; $D = 2{,}207$ probe thinking tokens).

4 Persistent Eligibility Creates a Hidden Curriculum

We test whether pre-RL query properties predict persistent eligibility and whether dynamic filtering converts that persistence into unequal training exposure. We profile queries before target RL and log all attempted rollout groups, including those discarded by the filter.

4.1 Task Labels Hide Distinct Query Regimes

Even within one benchmark and fine category, queries can have very different pre-RL profiles. Figure 2 illustrates this within AI2D Tables/Diagrams/Charts. Initial solvability measures how reliably the target already answers a query, whereas response demand records how much deliberation the probe produces for it.

Visually similar lookup questions occupy all four regimes. The summer-solstice and waxing-crescent questions elicit long probe responses but differ in target solvability; the shorter leaf-identification questions likewise include both lower- and higher-solvability examples. Thus neither a shared task label nor elicited response length determines where a query begins. The same separation appears within MMK12 Geometry (Appendix B.1).

4.2 Response Demand Predicts Persistent Eligibility

Figure 3 shows two related patterns. Within every broad pre-RL solvability band, mean mixed occupancy increases with response demand. The trajectories then separate over training: under Qwen2.5 standard dynamic sampling, occupancy in D_1 falls from .640 to .162, whereas D_4 remains at .526 in the final window.

An event-time analysis supports this temporal interpretation. A repeatedly observed all-pass state is reached by 86.2% of D_1 queries but only 38.6% of D_4 queries. Among those that reach it, median candidate exposure is 3 for D_1 and 14 for D_4. The ordering also holds without filtering and within the first 160 steps (Appendix C.2).

The association survives retrospective adjustment for initial solvability, query composition, and realized presentation count. A one-standard-deviation increase in log demand is associated with 9.4 percentage points more mixed occupancy for Qwen2.5 and 8.4 points more for Qwen3-Thinking. It remains positive without filtering (8.3 and 9.3 points), while Qwen3-Instruct shows the same direction at smaller magnitude. Adding demand to solvability and composition features improves repeated held-out prediction by $\Delta R^2 = .041$–$.081$.

The demand ordering also survives after parse-failed and capped rollouts are excluded (Appendix C.4). The incremental value of probe demand is target-dependent. It adds substantial information beyond target-self length for Qwen2.5 ($\Delta R^2 = .0725$–$.0870$), but little for Qwen3-Thinking once richer probe statistics are included ($+.0042$). For Qwen3-Instruct, target-self length mainly predicts token cost, while probe demand retains a modest increment for persistence. These results suggest that a separate probe is most useful when the target's own responses provide a weak or horizon-limited

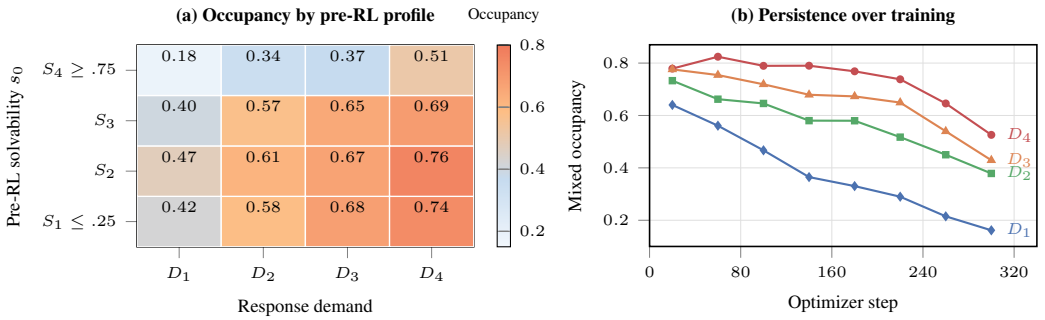


Figure 3: **Pre-RL profiles predict eligibility persistence.** Qwen2.5-VL-3B-Instruct under standard dynamic sampling, steps 1–320. **(a)** Within every pre-RL solvability band, mean mixed occupancy increases with response demand. **(b)** Across 40-step windows, occupancy falls from .640 to .162 for D_1 but remains .526 for D_4. Demand quartiles use all-rollout probe demand D_{all}; points mark window midpoints.

length signal. For targets that already exhibit stable long-form reasoning, target-self length may serve as a practical proxy. We use probe-based demand as a consistent cross-target measure, not because a separate probe is always necessary. Full predictor comparisons and group-size checks appear in Appendix C.

4.3 Filtering Turns Persistence into Exposure Drift

Dynamic sampling refills the batch until enough groups pass the mixed-outcome filter. Candidate shares are close to 25% in every demand quartile, but D_4 acceptance is 52.1%–71.6%, compared with 19.0%–35.8% for D_1 across the three targets (Table 1). Consequently, D_4 supplies 31.9%–37.8% of admitted optimizer groups, while D_1 supplies only 13.1%–15.6%. At decile resolution, the D_{10}/D_1 acceptance-rate ratio is $2.71\times$ for Qwen2.5 and 3.26–$4.00\times$ for Qwen3.

The induced allocation is predictable from past observations. On Qwen3 standard training, lagged acceptance estimates reduce next-window total-variation error from .173 to .043, with a .982 correlation between predicted and realized shares (Appendix E). Thus the change-of-measure model predicts how filtering reshapes allocation. Unlike hard-example mining, the filter rejects both consistently solved and consistently failed queries, favoring those that continue to straddle the outcome boundary.

4.4 The Same Task Imbalance Can Mean Different Things

We examine how this query-level curriculum appears at the task level. The *perception-associated* family contains Tables/Diagrams/Charts and Spatial Reasoning; the *reasoning-associated* family contains mathematics, geometry, science, and the remaining reasoning categories. Image corruption reduces accuracy more for the perception-associated family, providing behavioral support for this grouping (Table 5).

Which family remains mixed longer depends on the target policy. In a Qwen2.5 sequence-mean arm, reasoning queries remain mixed longer ($P - R = -.094$). The direction reverses under Qwen3-Thinking ($P - R = +.151$), and Qwen3-Instruct also shows higher perception occupancy in both halves of training ($+.079$ and $+.072$). These directions remain after balancing initial solvability, response demand, answer format, and data source. The Qwen2.5/Qwen3-Thinking reversal also appears on shared queries without filtering. Task-level imbalance is not a fixed property of the dataset.

The same task-level direction can also arise for different reasons. Although both Qwen3 targets keep perception-associated queries mixed longer, Qwen3-Thinking reasoning queries more often remain all-fail. Under Qwen3-Instruct, reasoning queries instead move from mixed to all-pass more often. In the late Instruct window, the adjusted mixed-to-all-pass difference is $P - R = -.050$ (95% CI $[-.083, -.017]$), while the all-fail persistence difference is near zero. Low mixed occupancy can therefore reflect either consistent success or consistent failure. These transitions are observational

Table 1: **Filtering produces unequal optimizer exposure.** Acceptance is computed within each demand quartile; optimizer share is each quartile's fraction of all admitted groups and sums to 100% per target. Qwen2.5 uses steps 1–320 and Qwen3 uses steps 1–120.

Target	Measure	D_1	D_2	D_3	D_4
Qwen2.5-Instruct	Candidate pass rate	35.8%	55.1%	63.4%	71.6%
	Share of optimizer batch	15.6%	24.4%	28.1%	31.9%
Qwen3-Instruct	Candidate pass rate	19.0%	32.2%	41.9%	52.4%
	Share of optimizer batch	13.1%	21.9%	28.7%	36.3%
Qwen3-Thinking	Candidate pass rate	20.1%	28.5%	37.3%	52.1%
	Share of optimizer batch	14.5%	20.8%	26.8%	37.8%

rather than causal estimates of learning speed, and accepted task shares reveal neither task difficulty nor learning value. Full controls and transition estimates appear in Appendix C.5.

5 Gradient Eligibility Is Not Learning Utility

We test whether persistent eligibility predicts held-out value using matched low- and high-demand interventions at two policy states.

5.1 More Persistent Pools Do Not Yield Greater Gains

Matched interventions. The early intervention branches from the Qwen3-VL-2B-Thinking base policy, whereas the late intervention branches from its standard-filtering step-40 checkpoint. At each state, external Qwen3-VL-4B-Thinking demand defines low- and high-demand pools (D_1 and D_4). We independently construct 80-query pools matched on fine category, answer schema, and fresh current solvability.

The early pools are exactly pair-matched in measured solvability, with a shared mean of $.5445$ and a $7.25\times$ demand ratio. The late pools have mean solvability $.5508$ for D_1 and $.5375$ for D_4 ($n = 16$; caliper $\leq 3/16$), with a $7.12\times$ demand ratio. Across three seeds, every arm receives exactly 1,280 accepted groups and is evaluated on the same disjoint 400-query panel. This fixes the number of admitted training groups while allowing acceptance, exposure concentration, and generation cost to differ. Full construction and seed-level results appear in Appendix D.1.

Different residence, consistent performance ordering. The two interventions have sharply different eligibility dynamics (Table 2). Early D_1 rapidly thins: its mean acceptance is $.224$, it requires $2.67\times$ as many candidate groups as D_4 to fill the same update budget, and its exposure ESS contracts to 37.0 of 80 queries. Late D_1 remains healthier, with acceptance $.691$ and ESS 72.7.

Despite this difference, all six seed-paired pass@1 contrasts favor D_1. The early mean difference is $+2.17$ percentage points (pp), with a training-seed 95% t interval of $[+.57, +3.78]$ pp. The late difference is $+1.24$ pp, although its three-seed interval includes zero. Early pass@16 also favors D_1 by $+1.42$ pp ($[+.12, +2.71]$ pp); the other high-k contrasts are positive on average but less precisely estimated.

Higher persistence is also more expensive. Relative to D_1, D_4 uses $1.68\times$ the candidate tokens and $3.61\times$ the accepted tokens in the early intervention, and $2.12\times$ and $2.30\times$, respectively, in the late intervention. Nor does D_4 training show a special advantage on high-demand evaluation queries: the observed $D_1 - D_4$ pass@1 contrast on that quartile is $+2.67$ pp early and $+2.92$ pp late.

5.2 Interpreting the Early and Late Contrasts

Different endpoint pathways. The two interventions produce the same performance ordering through different endpoint components. At the early state, both allocations produce substantial learning gains: mean pass@1 rises from $.5192$ at the base policy to $.6392$ under D_1 ($+11.99$ pp) and $.6174$ under D_4 ($+9.82$ pp). Thus high-demand training is useful, but its greater persistence and $3.61\times$ accepted-token cost do not yield greater progress under the same accepted-update budget.

Table 2: **Higher eligibility and token cost do not yield greater held-out utility.** Early arms start from the base policy and late arms from step 40. Each trained arm receives 1,280 accepted groups; endpoints are three-seed means. All evaluations use 16 rollouts per query. Token ratios are D_4/D_1, performance differences are $D_1 - D_4$, and ESS is the final exposure effective sample size out of 80 training queries.

State	Training pool	Acceptance	Candidate tokens	Accepted tokens	ESS	pass@1	pass@8	pass@16
Early	Base policy	—	—	—	—	.5192	.8179	.8575
	Low demand (D_1)	.224	49.8M	14.1M	37.0	.6392	.8515	.8858
	High demand (D_4)	.597	83.7M	50.8M	70.9	.6174	.8427	.8717
	Cost ratio / metric Δ	—	1.68×	3.61×	—	+.0217	+.0088	+.0142
Late	Step-40 branch	—	—	—	—	.6270	.8301	.8675
	Low demand (D_1)	.691	27.5M	21.1M	72.7	.6332	.8421	.8717
	High demand (D_4)	.816	58.4M	48.6M	77.1	.6207	.8312	.8633
	Cost ratio / metric Δ	—	2.12×	2.30×	—	+.0124	+.0108	+.0083

Of the $+2.17$ pp $D_1 - D_4$ contrast, $+.86$ pp comes from parseable completion and $+1.31$ pp from correctness conditional on completion. The conditional-correctness contribution is positive in all three seeds, with a training-seed interval excluding zero.

In the late intervention, an exact accounting decomposition assigns $+1.25$ pp of the $+1.24$ pp pass@1 contrast to parseable completion, with approximately zero contribution from correctness conditional on completion. Relative to the step-40 branch, D_4 remains near its original response length, cap rate, parse failure, and pass@1, whereas D_1 shortens responses and improves completion reliability. The replicated ordering therefore does not depend on a single endpoint pathway.

Training eligibility is a separate phenomenon. The endpoint decompositions do not explain why D_4 passes the training filter more often. After parse-failed and capped training rollouts are removed, 92–97% of the observed $D_4 - D_1$ mixedness gap remains across the early and late streams (Appendix C.4). Thus the persistent training signal is primarily content-level outcome variation, even though the late held-out contrast is expressed through completion reliability.

Scope of the early–late comparison. The larger early contrast does not establish that utility changes with policy state. The two interventions use different matched pools, and the early-minus-late difference remains uncertain. Instead, the early experiment stress-tests the allocation under rapid signal depletion: forcing a shrinking D_1 pool to supply the same accepted-update count still does not reveal an advantage for the more persistent and more expensive D_4 pool. A refreshed or larger D_1 pool could reduce exposure concentration, but that counterfactual is not tested here.

The same low-versus-high-demand ordering appears at high k in a Qwen2.5 no-filter pilot (Appendix D.3) and in an exploratory MathVista testmini evaluation (Lu et al., 2024), where D_1 leads by 2.4 pp in accuracy (Appendix D.4). An observational trajectory audit finds that high-demand queries more often remain mixed without greater subsequent solve-rate progress (Appendix C.3).

6 Controlling Which Queries Reach the Optimizer

Dynamic sampling gives more optimizer slots to strata that pass the filter more often. We ask whether this realized allocation can be controlled without removing the filter. We control how a chosen accepted mixture is realized, not which mixture should be chosen.

6.1 Rebalancing the Accepted Mixture

Suppose demand stratum b is proposed with weight w_b and passes the filter with probability $\bar{A}_b > 0$. Its accepted mass is proportional to $w_b \bar{A}_b$. To obtain a desired accepted share ρ_b, we therefore use

$$\rho_b^{\mathrm{acc}} = \frac{w_b \bar{A}_b}{\sum_j w_j \bar{A}_j}, \qquad w_b \propto \frac{\rho_b}{\bar{A}_b}. \tag{5}$$

We use equal accepted shares across demand quartiles as a neutral control target. Drift is the total-variation distance from this target, and drift reduction is measured relative to standard dynamic sampling.

Table 3: **Past-only rebalancing reduces accepted-exposure drift at near-matched cost.** Accepted shares and generation cost cover Qwen3 steps 41–100; performance is evaluated at step 100. Cost is relative to standard dynamic sampling, and the in-sample oracle uses same-window acceptance rates.

Sampling	D_1	D_2	D_3	D_4	Drift removed	Gen. tokens	pass@1	pass@8	pass@16
Standard	12.8%	20.1%	27.7%	39.5%	—	$1.00\times$	.5953	.8083	.8401
Oracle	24.8%	25.0%	25.4%	24.8%	97.5%	$0.95\times$	.6279	.8409	.8723
EMA	20.0%	24.4%	26.3%	29.4%	67.2%	$1.02\times$	.6035	.8227	.8597

The *static oracle* derives fixed candidate-sampling weights from stratum acceptance rates observed in the completed standard dynamic-sampling run over the target window. A separate training arm applies these sampling weights, allowing us to test whether the retrospective calibration realizes the target allocation and to evaluate its endpoint. The deployable *past-only EMA* instead updates the sampling weights from earlier attempted and accepted groups, using neither task labels nor future information. Because the correction requires only positive, estimable stratum-level acceptance rates, it can be layered on other filters or candidate-priority rules.

6.2 From Oracle Rebalancing to Past-Only Control

The static oracle nearly equalizes the accepted shares, removing 97.5% of drift. In this run, it also uses only $0.95\times$ the generated tokens and coincides with positive pass@k differences, including $+.0326$ at pass@1. Because it uses acceptance rates from the scored window, this is an in-sample controllability reference: it shows that near-uniform exposure can be realized without additional generation cost, but does not identify a general performance effect. A separate Qwen2.5 static calibration likewise removes 79% of drift at $1.018\times$ token cost (Appendix E).

Using only past observations, EMA realizes much of this correction. It removes 67.2% of drift at $1.02\times$ the training-generation tokens and $1.01\times$ the wall time. Pass@1 remains similar ($+.0082$; query-paired 95% interval $[-.001, +.017]$), while pass@8 improves by .0145 ($[.003, .026]$) and pass@16 by .0196 ($[.005, .036]$). The largest pass@16 gain occurs on high-demand evaluation queries ($+.0634$), even though the corresponding training stratum receives less exposure. These single-run results reflect the joint mixture intervention, not an isolated D_4 effect.

7 Conclusion and Outlook

Binary outcome feedback makes GRPO-style learning selective even before explicit filtering: only mixed rollout groups provide relative advantages through the outcome channel. Dynamic sampling turns this local eligibility into an allocation rule by replacing zero-variance groups, allowing queries that remain mixed to absorb more optimizer slots. Across Qwen2.5 and Qwen3, pre-RL solvability and response demand predict which queries persist and the exposure and generation cost they accumulate. Yet two matched three-seed interventions show that more persistent pools do not provide greater held-out gains under equal accepted-update budgets. A past-only inverse-acceptance sampler can rebalance the resulting exposure without changing the eligibility criterion.

The central distinction is therefore among gradient availability, optimizer allocation, and learning utility. A nominal candidate mixture does not fully describe the effective training distribution: accepted exposure, generation cost, and held-out outcomes must be measured separately. Response demand is useful as a pre-RL prior over persistence and cost, not as a score of training utility.

Limitations and future work. Our evidence spans three Qwen2.5 and Qwen3 target-policy settings, including Instruct and Thinking policies, but all use Qwen-family VLMs, controlled ViRL39K mixtures, and primarily binary feedback. The utility interventions use one matched pool pair per policy state, so their seed intervals do not capture pool-construction uncertainty. We do not measure how selection-induced imbalance interacts with task-gradient conflict; separating them would require holding accepted exposure fixed. Finally, inverse acceptance realizes a specified allocation but does not determine its optimal target. Future controllers could select queries by expected held-out improvement per unit compute using demand, online progress, reliability, and generation cost.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, and analysis-code development. The authors reviewed AI-assisted text, verified reported claims against experiment artifacts, and checked generated or modified analysis code against reconstructed trainer metrics where applicable. The authors take responsibility for the final content, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Appendix A documents the model, data, reward, optimization, and evaluation configurations. Appendices D.1 and E provide seed-level intervention results and controller details, while Appendix A.3 records split, probe, and evaluation provenance. Together, these sections specify the training, matching, allocation, and evaluation procedures used for the reported results.

REFERENCES

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, Jeongyeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 700–719, 2026. doi: 10.18653/v1/2026.eacl-long.30.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

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

Eunbi Choi, Kibong Choi, Sehyun Chun, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Ahra Jo, Hyunjik Jo, Yeonsik Jo, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Changhun Lee, Haeju Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Kwangrok Ryoo, Minju Seo, Sejong Yang, Heuiyeen Yeen, Hwan Chang, Stanley Jungkyu Choi, Yejin Choi, Kyubeen Han, Joonwon Jang, Kijeong Jeon, Geunyeong Jeong, Gerrard Jeongwon Jo, Jiyeon Jung, Daeseong Kim, Dohoon Kim, Dohyun Kim, Hyunseo Kim, Minu Kim, Myoungshin Kim, Youchul Kim, Byungoh Ko, Christopher Lee, Edward Hwayoung Lee, Honglak Lee, Jiyoung Lee, Sangeun Lee, Seungwon Lim, Woohyung Lim, Jueun Mun, Jaewoo Park, Jimin Park, Jinho Park, Yongmin Park, Wooseok Seo, Yongwoo Song, Sihyuk Yi, Kyungjae Yoo, and Sangyeon Yoon. EXAONE 4.5 technical report. 2026. URL https://arxiv.org/abs/2604.08644.

Thomas Foster, Anya Sims, Mattie Fellows, Johannes Forkel, and Jakob Foerster. LILO: Learning to reason at the frontier of learnability. In *Advances in Neural Information Processing Systems*, 2025.

Zhaolin Gao, Joongwon Kim, Wen Sun, Thorsten Joachims, Sid Wang, Richard Yuanzhe Pang, and Liang Tan. Prompt curriculum learning for efficient LLM post-training. *arXiv preprint arXiv:2510.01135*, 2025. URL https://arxiv.org/abs/2510.01135.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, 2017.

Wenbo Hu, Xin Chen, Yan Gao-Tian, Yihe Deng, Nanyun Peng, and Kai-Wei Chang. OpenVL-ThinkerV2: A generalist multimodal reasoning model for multi-domain visual tasks. *arXiv preprint arXiv:2604.08539*, 2026.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *ICML*, 2021.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

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.

Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. MoDoMoDo: Multi-domain data mixtures for multimodal LLM reinforcement learning. *arXiv preprint arXiv:2505.24871*, 2025. URL https://arxiv.org/abs/2505.24871.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One RL to see them all: Visual triple unified reinforcement learning. *arXiv preprint arXiv:2505.18129*, 2025.

Sadhika Malladi, Samy Jelassi, Dylan Foster, Jordan T. Ash, and Akshay Krishnamurthy. TailSFT: Filtered fine-tuning improves post-training performance. *arXiv preprint arXiv:2608.25756*, 2026. URL https://arxiv.org/abs/2608.25756.

Ziqi Miao, Haonan Jia, Lijun Li, Chen Qian, Yuan Xiong, Wenting Yan, and Jing Shao. Seeing with you: Perception-reasoning coevolution for multimodal reasoning. *arXiv preprint arXiv:2603.28618*, 2026.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. Competence-based curriculum learning for neural machine translation. In *NAACL-HLT*, 2019.

Yun Qu, Qi Wang, Yixiu Mao, Vincent Tao Hu, Bjorn Ommer, and Xiangyang Ji. Can prompt difficulty be online predicted for accelerating RL finetuning of reasoning models? *arXiv preprint arXiv:2507.04632*, 2025.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open RL recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026. URL https://arxiv.org/abs/2604.04917.

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.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable R1-style large vision–language model. *arXiv preprint arXiv:2504.07615*, 2025. URL https://arxiv.org/abs/2504.07615.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training. *arXiv preprint arXiv:2607.16097*, 2026. URL https://arxiv.org/abs/2607.16097.

TIGER-Lab. ViRL39K: A curated collection of 38,870 verifiable vision–language QA pairs. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/TIGER-Lab/ViRL39K.

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

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yangyi Chen, Zihan Li, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-Cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. *arXiv preprint arXiv:2512.13607*, 2025a.

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. *arXiv preprint arXiv:2507.06448*, 2025b.

Xiaomi LLM-Core Team. MiMo-VL technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in RLVR via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026. URL https://arxiv.org/abs/2606.15455.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.13837.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-R1: Evolving human-free alignment in large vision–language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025. URL https://arxiv.org/abs/2503.18013.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2503.12937.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a. URL https://arxiv.org/abs/2507.18071.

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. In *Advances in Neural Information Processing Systems*, 2025b.

Table 4: **Primary training configurations.** "Steps" reports the trajectory range used by the main analysis, not a claim that all runs share a common optimizer-step or token budget.

Target	Queries	Horizon	G	Steps	Aggregation
Qwen2.5-VL-3B-Instruct	2,000	2,048	5	320	token mean
Qwen3-VL-2B-Thinking	6,000	16,384	5	120	sequence mean
Qwen3-VL-2B-Instruct	6,000	16,384	5	120	sequence mean

A Experimental Setup and Measurement Provenance

A.1 Targets, Training, and Controlled Splits

The primary trajectories use Qwen2.5-VL-3B-Instruct on 2,000 ViRL39K queries and Qwen3-VL-2B-Thinking on 6,000 queries. The Qwen3-VL-2B-Instruct bridge run uses the same 6,000 queries and training configuration as Qwen3-Thinking. Table 4 summarizes the settings that differ across targets.

The Qwen2.5 split was constructed from the high and low extremes of the original visual-evidence-sensitivity profile. Its two 1,000-query halves were matched on design-time solvability, fine category, answer schema, and MCQ status. Each half contains 173 perception-associated and 827 reasoning-associated queries. The Qwen3 split contains two 3,000-query design halves with the same category and answer-schema composition; each contains 796 perception-associated and 2,204 reasoning-associated queries. These controls make within-target contrasts interpretable, but the two target splits are not exchangeable: their perception shares are .173 and .265, respectively.

All primary arms use a global accepted-group batch size of 64, no KL penalty, and AdamW with `bf16` arithmetic, learning rate 10^{-6}, $\beta = (0.9, 0.999)$, weight decay .01, no warmup, and asymmetric clipping $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}}) = (.2, .3)$. The filter key is binary accuracy and retains groups with mean accuracy strictly between .01 and .99, which for $G = 5$ is exactly $0 < c < G$; it refills until 64 accepted groups are obtained. Rollouts use temperature 1.0 and top-$p = .95$; Qwen3 additionally uses top-$k = 20$.

A.2 Reward, Filtering, and Evaluation

The rule-based verifier extracts the final answer and normalizes multiple-choice labels against their option values. Training advantages use the format-augmented score 0.9 accuracy $+ 0.1$ format, but the filter key, every eligibility state, and every filtering statistic in this paper use binary accuracy: all-fail has $c = 0$, all-pass has $c = G$, and mixed has $0 < c < G$. Within each rollout group, scalar rewards are centered and divided by the group sample standard deviation plus $\epsilon = 10^{-6}$; no additional cross-group advantage normalization is applied.

For a query with c correct answers among n rollouts, we estimate

$$\widehat{\text{pass@}k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}},$$

and average this quantity across queries. End-policy pass@k therefore uses binary correctness rather than the format-inclusive training reward. Accepted exposure denotes query groups admitted through the outcome gate, rather than their eventual gradient magnitude. Parse failures and cap hits are reported separately in the reliability decomposition.

The primary observational trajectories, Qwen2.5 pilot, and controller arms are single training runs. The early and late D_1/D_4 interventions each have three paired training seeds. We use these interventions for the finite-budget eligibility–utility contrast and treat the remaining trajectories as mechanism, prediction, or supporting evidence at their stated scope.

A.3 Data, Panel, and Probe Provenance

Response demand is the median thinking-token count across all rollouts from a fixed probe VLM distinct from the target policy. Qwen2.5 analyses use Qwen3-VL-2B-Thinking with $n = 3$; Qwen3

analyses and utility-pool construction use Qwen3-VL-4B-Thinking with $n = 8$. Unless otherwise stated, primary observational analyses use all-rollout demand. Both probes use temperature 1.0, top-$p = .95$, top-$k = 20$, and a 16,384-token response horizon; demand is therefore defined relative to this fixed decoding contract.

Initial solvability is always measured with the target policy. The Qwen2.5 and Qwen3-Thinking trajectory regressions and both utility-pool matching designs use fresh $n = 16$ estimates; Qwen3-Instruct uses $n = 8$. The original Qwen3-Thinking $n = 3$ score is retained only for full-split construction summaries. The Qwen2.5 and Qwen3 training splits share 399 exact query identities; these form the paired target comparison in Appendix C.5.

Evaluation panels are query-disjoint from their corresponding training sets. The Qwen2.5 400-query panel contains 454 image references, ten of which also appear in the 2,000-query training set (2.2%); all compared Qwen2.5 arms share this condition. The early and late Qwen3 utility endpoints use the same 400 queries, 16 rollouts per query, and one common evaluation seed, so paired query intervals condition on that evaluation draw. The Qwen3 controller and Qwen2.5 matched pilot use 32 rollouts per query. We report training-seed variation separately from query-bootstrap uncertainty whenever multiple training seeds exist.

A.4 Role and Replication of Each Training Arm

The training arms serve different evidential roles; a single arm denotes one training trajectory, not one evaluation rollout.

Persistence and prediction. The Qwen2.5 and Qwen3-Thinking standard and no-filter trajectories are single runs used to analyze persistence, exposure drift, saturation, and task projection. The single Qwen3-Instruct bridge trajectory tests whether the Qwen3 task ordering is specific to Thinking mode and supplies target-self length controls.

Utility interventions. The early and late Qwen3-Thinking D1/D4 interventions branch from the base and step-40 policies, respectively, and each use three training seeds per arm. Together they provide the primary finite-budget utility contrast across two policy states; the early intervention additionally stress-tests rapid signal depletion. The Qwen2.5 no-filter pilot and staging arms provide single-run cross-setting evidence about utility and presentation order.

Allocation control. The Qwen3 standard, static-oracle, and past-only EMA controller arms are single runs used to test prediction and realization of the accepted distribution. They are not treated as independent replications of the utility intervention.

B Qualitative Examples and Task Grounding

B.1 Qualitative-Map Selection and Full Prompts

The examples in Figure 2 and Figure 4 follow the same deterministic selection protocol. The medians of the full Qwen2.5 split define the axes: $s_0 = .375$ from 16 fresh target rollouts per query and $D = 2{,}207$ thinking tokens from the probe. The main-text figure fixes AI2D Tables/Diagrams/Charts; the appendix figure fixes MMK12 Geometry. Candidates must have exactly one locally available image. Within each quadrant, we select the query nearest its medoid in $(s_0, \log D)$ space. Training outcomes are not used, and no examples are manually substituted.

Full reasoning prompts. Figure 4 uses compact, complete summaries so that all four cards remain legible. The longer original prompts are reproduced below, with only the image placeholder removed and mathematical typography normalized.

1. **High demand, low initial solvability.** In the figure, in $\triangle ABC$, $\angle C = 90°$ and $\angle CAB = 60°$. Follow these steps to construct: (1) with A and B as centers, and a radius greater than $\frac{1}{2}AB$, draw arcs that intersect at points P and Q; (2) draw line PQ, intersecting AB at point D and BC at point E, and connect AE. If $CE = 4$, then find AE.

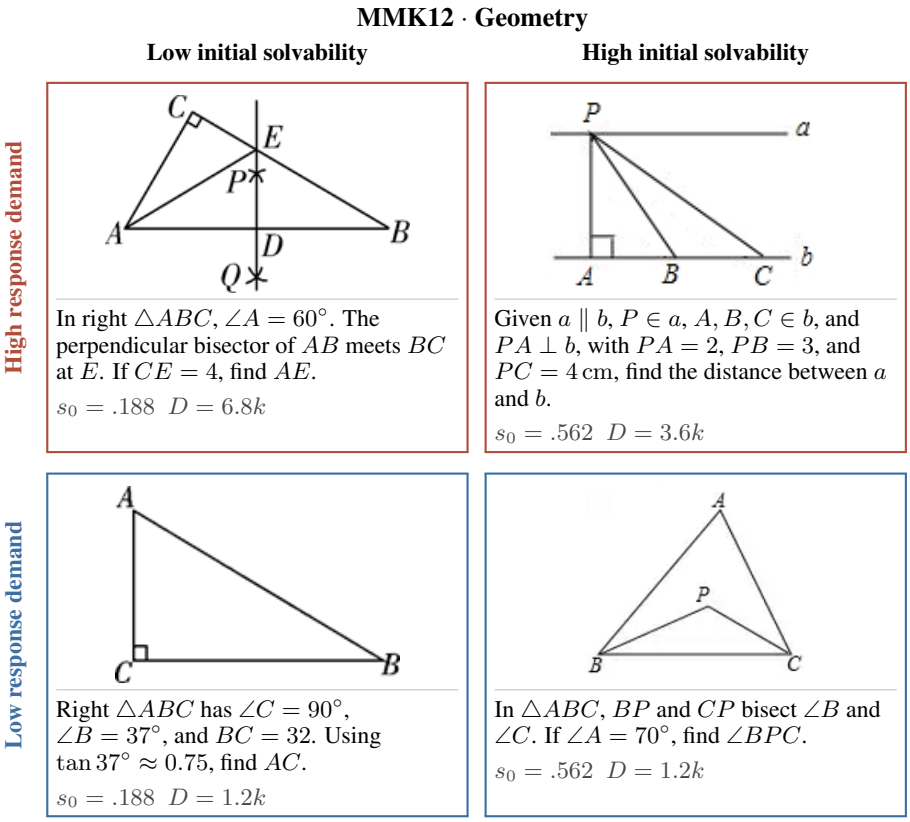


Figure 4: **Reasoning queries also occupy distinct pre-RL regimes.** MMK12 Geometry examples span low/high target solvability and response demand under the same medians and nearest-medoid selection protocol as Figure 2. Full original prompts are reproduced in Appendix B.1.

2. **High demand, high initial solvability.** As shown in the figure, $a \parallel b$, point P is on line a, and points A, B, and C are on line b. $PA \perp AC$, and $PA = 2\,\text{cm}$, $PB = 3\,\text{cm}$, and $PC = 4\,\text{cm}$. The distance between lines a and b is __ cm.
3. **Low demand, low initial solvability.** In the right triangle ABC, $\angle C = 90°$, $\angle B = 37°$, and $BC = 32$. Find AC. (Reference data: $\sin 37° \approx 0.60$, $\cos 37° \approx 0.80$, and $\tan 37° \approx 0.75$.)
4. **Low demand, high initial solvability.** In the figure, in $\triangle ABC$, BP and CP bisect $\angle ABC$ and $\angle ACB$, respectively, and $\angle A = 70°$. Then $\angle P =$ _____°.

B.2 Visual-Evidence Sensitivity and Task Grounding

Section 4.4 uses a deliberately coarse grouping of perception- and reasoning-associated categories. We validate that grouping with visual-evidence sensitivity (VES), rather than defining it by category names alone. Following the clean-versus-corrupted visual comparison of Wang et al. (2025b), we define an accuracy-based VES as the drop from clean to corrupted-image accuracy; adjusted analyses residualize this drop against an independently measured clean accuracy. The Qwen3-VL-2B pilot uses PAPO-style random patch corruption, whereas the Qwen2.5 profile replaces the image with one from another query in the same fine category. VES therefore measures reliance on the correct image, while response demand measures the deliberation elicited in the response.

The Qwen3 entries in both panels of Table 5 come from the same 250-query pilot. Panel (a) uses the 248 queries assigned to the two task families. Panel (b) retains the five categories with at least 15 Qwen3 examples and compares their adjusted means with the Qwen2.5 2,000-query profile.

In panel (a), the family-level VES difference is $+.2391$ (bootstrap 95% interval $[+.1368, +.3323]$) and grows to $+.2589$ after clean-accuracy adjustment. In panel (b), the rank correlation across the

Table 5: **The task grouping captures visual dependence across probes. (a)** Raw VES is clean-minus-corrupted accuracy in the Qwen3 patch-noise pilot. **(b)** Residualized category-level VES compares Qwen3 patch noise with the Qwen2.5 within-category image-mismatch profile.

(a) Task-family contrast

Group	n	Clean	Perturbed	VES
Perception	71	.5023	.1596	**.3427**
Reasoning	177	.5254	.4218	.1036

(b) Category-level transfer

Category	Qwen3	Qwen2.5	Rank
Tables/Diagrams/Charts	$+.1982$	$+.0981$	1 / 2
Spatial Reasoning	$+.1454$	$+.1110$	2 / 1
Non-geometric Math	$-.0234$	$-.0083$	3 / 3
Geometric Reasoning	$-.1508$	$-.0414$	4 / 4
Science	$-.1661$	$-.0487$	5 / 5

five sufficiently represented categories is $+.900$ and the value correlation is $+.975$. Because the Qwen3 profile is a small pilot with three rollouts per condition, these results support category-level grounding rather than a reliable query-level Qwen3 VES measure.

The Qwen2.5 split was originally stratified using an earlier VES profile, but fresh target profiling reveals a solvability difference of $-.0678$ between its high- and low-VES halves (95% interval $[-.0907, -.0449]$). Adding this fresh solvability estimate reduces the high-VES coefficient for all-pass occupancy from $+.1115$ $[+.062, +.161]$ to $+.0352$ $[-.012, +.082]$. Re-matching on the fresh VES profile likewise gives an all-pass contrast of $-.006$ $[-.050, +.037]$. We therefore do not use the original VES split as evidence that visual dependence predicts persistence; the main persistence result instead comes from response demand after fresh solvability control.

The two measurements are also empirically distinct before RL. On a separate 600-query profile with repeated corruptions, adding VES to a model of log response demand changes held-out R^2 only from .3290 to .3292 (partial $R^2 = .0003$). We therefore use VES to ground visual dependence and the task projection, not to rank queries by persistence or learning value.

C Robustness of Persistence and Exposure

C.1 When Does Probe-Derived Demand Add Information?

Measurements and baselines. Response demand is relative to a named probe and decoding contract. Because the target profiling used to estimate s_0 also produces responses, target-self length is a natural low-cost baseline. We therefore ask when a probe-derived thinking-length profile adds information beyond the target's own length and other pre-RL measurements, rather than assuming that it is always necessary or superior.

The Qwen2.5 analysis uses Qwen3-VL-2B-Thinking with three rollouts per query; the Qwen3 analyses use Qwen3-VL-4B-Thinking with eight. Response demand D_{all} is the median thinking-token length over all probe rollouts.

Primary persistence specification. The occupancy coefficients reported in Section 4.2 use query-level mixed occupancy as the outcome, exact s_0 fixed effects, and controls for fine category, answer schema, and presentation count, with HC3 intervals. Qwen2.5 uses steps 1–320; the Qwen3 standard and no-filter regressions use the healthy windows through steps 100 and 160, respectively. For the reported predictive increments, repeated 5×5-fold cross-validation compares a baseline containing s_0, source, category, answer schema, MCQ status, and presentation count with the same model plus log response demand. Because presentation count is realized during training, these controlled occupancy analyses are retrospective. Table 6 separately evaluates predictor sets composed only of measurements available before RL.

Predictor-ladder comparisons. Target-self length is the median full-response token count from target-policy profiling: eight rollouts per query for Qwen2.5-Instruct and Qwen3-Instruct, and three for Qwen3-Thinking. The Qwen2.5 and Qwen3-Thinking comparisons average six repeated five-fold splits; the Qwen3-Instruct analysis averages eight. Nested comparisons within a target use identical queries and folds; absolute R^2 values should not be compared across targets because their features and profiling contracts differ.

Table 6: **Response demand is predictive, but its incremental value is target-dependent.** Values are repeated held-out R^2 for query-level mixed occupancy. The rich profile contains s_0 and the probe's accuracy, reward variance, rollout disagreement, and truncation rate. The final row isolates the gain from adding demand after target-self length.

Predictors	Qwen2.5-Instruct		Qwen3-Thinking
	No filter	Standard	Standard
Compact profiles			
Initial solvability s_0	.1103	.1343	.0128
Response demand	.1865	.2356	.1147
s_0 + response demand	.2381	.2972	.1250
Stringent incremental test			
Rich pre-RL profile	.1240	.1518	.2002
+ target-self length	.1766	.2329	.2475
+ response demand	**.2491**	**.3199**	**.2517**
Δ from response demand	**+.0725**	**+.0870**	**+.0042**

Target-dependent predictive value. For Qwen2.5, adding response demand after the full target-self-length baseline raises R^2 by .0725 without filtering and .0870 under standard filtering. The target has a 2,048-token response horizon, whereas the probe can express longer deliberation. The gain may therefore partly reflect compression of the target-self length signal; this analysis establishes the probe's predictive value but does not isolate why it adds information.

For Qwen3-Thinking, the target's own long responses and the other probe statistics already capture most of the available signal. Adding response demand to the full specification raises R^2 by only $+.0042$, and demand alone (.1147) predicts less than the rich pre-RL profile (.2002). Probe-derived thinking length is therefore largely redundant once this richer baseline is available.

Qwen3-Instruct separates persistence from immediate cost. Beyond initial solvability, target-self length adds only $+.0009$ held-out R^2 for mixed occupancy but $+.6527$ for generated-token cost. Qwen3-VL-4B probe demand adds $+.0086$ for occupancy and still adds $+.0077$ after self length is included; their rank correlation is .515. Target-self length is therefore the better direct cost estimate, whereas the probe retains modest information about which queries remain mixed.

Late-stream persistence and generation cost. In Qwen3-Instruct, the demand-conditioned curriculum strengthens after step 60. The D_4:D_1 accepted-exposure ratio rises from 2.38 over steps 1–60 to 3.23 over steps 61–120; in the final 20 steps, D_4 supplies .399 of accepted groups and D_1 only .099. Meanwhile, candidate acceptance falls from .395 to .325, candidate groups per optimizer step rise from 162 to 197, and generated tokens per accepted group rise from 29.5K to 44.4K ($+50.7\%$). Parse failure, cap hits, and format success remain stable, so this is an efficiency and response-length drift rather than failure-mode collapse. Only about 12% of the accepted-response length increase is explained by the shift toward higher-demand quartiles; the remainder occurs within quartiles. These are training-stream diagnostics, not held-out utility measurements.

Probe robustness. Some Qwen2.5 sampling arms were originally constructed with D_{succ}, the median over correct probe rollouts, rather than D_{all}. The two log-demand definitions correlate at .985, assign 91.5% of queries to the same quartile, and change held-out R^2 by at most .003. We use D_{all} in Table 6 and retain the original bins only when evaluating an arm constructed from them.

Demand is not probe-invariant, but some ordering transfers. On the 399 queries shared by the Qwen2.5 and Qwen3-Thinking splits, one GLM-4.1V-9B profile adds positive held-out signal for both targets, although the Qwen-family probe is stronger. We therefore use probe-derived demand as a common pre-RL analysis coordinate, not as an intrinsic complexity score or a feature that must be collected for every target.

Interpretation of the intervention. The D1/D4 interventions compare pools selected by probe-derived demand. Because those pools can also differ in target-self response length, they estimate the finite-budget utility of the resulting allocations; they do not identify a causal effect of response demand or response length itself.

C.2 Demand Predicts Later Stable Success

Mixed occupancy averages repeated states but does not directly record when a query leaves the eligible region. We therefore define an event time on the Qwen2.5 trajectories. A query reaches *stable all-pass* at the first candidate exposure for which at least three of the next four observed groups are all-pass. Queries that never satisfy this criterion are right-censored; stable all-fail is tracked separately rather than counted as continued learning. Measuring time in query-level candidate exposures permits comparison despite the greater refill volume under filtering.

Table 7: **High-demand queries reach stable all-pass later and less often.** "Median exposure" is conditional on reaching the event. Demand quartiles use D_{all}.

	Standard filtering			No filter	
Quartile	Stable pass	Median exposure	Stable fail	Stable pass	Median exposure
D_1	.862	3	.044	.700	1
D_2	.700	6	.056	.456	2
D_3	.554	9	.116	.314	3
D_4	.386	14	.170	.134	4

The standard and no-filter traces contain medians of 23 and 10 candidate exposures per query, respectively. Event rates should therefore be compared across demand quartiles within an arm, not interpreted as a causal filtering effect between arms.

The ordering survives without filtering and within the first 160 steps, where degeneration is negligible. Across these specifications, a one-unit increase in log demand has a Cox log-hazard coefficient between approximately $-.40$ and $-.57$: higher demand is associated with a lower rate of reaching stable all-pass. Adding demand after target-self length raises event-time concordance from .7212 to .7644 without filtering and from .6960 to .7321 under standard filtering. These results support the temporal interpretation of persistence without treating queries absorbed at all-fail as useful or still learning.

C.3 Persistence Need Not Mean Progress

We analyze 5,948 transitions from 3,199 queries that are mixed at their current Qwen3 observation. For each transition, we record whether the next observed group remains mixed and the change in empirical solve probability. We residualize both outcomes within current solve probability, training window, answer schema, and pre-RL correct-count strata, then aggregate repeated observations with empirical-Bayes shrinkage.

The signs of residual persistence and progress define four descriptive regions: productive persistence (both positive), stalled persistence (persistence only), rapid acquisition (progress only), and terminal/low-yield (neither). These are observed trajectory regions, not intrinsic query types. From D_1 to D_4, stalled persistence rises from 14.9% to 31.2% and rapid acquisition falls from 63.3% to 32.7%; productive persistence changes little (15.8% to 17.2%).

An exact-matched audit additionally controls category, prior accepted exposure, and current truncation. Relative to D_1, D_4 has higher next-observation mixedness ($+.156$, clustered 95% interval $[+.129, +.190]$) and more subsequent included groups ($+.0625$, $[+.0474, +.0808]$), but lower solve-rate progress ($-.0649$, $[-.0779, -.0496]$). This audit is observational because exposure and re-observation remain policy-dependent; the matched interventions in Appendix D.1 supply the allocation contrast.

C.4 Failure Sources and Group-Size Robustness

We test whether high-demand groups remain mixed merely because long outputs are truncated or fail to parse, and whether the demand ordering depends on the rollout group size used by the eligibility rule.

For each Qwen3 candidate group, we remove every parse-failed or cap-hit rollout and recompute whether the remaining outcomes contain both a correct and a parsed, uncapped incorrect answer.

This *parsed-clean mixedness* is a sensitivity analysis, not the filter used in training. Both rates use all complete candidate groups as the denominator; a cleaned group with fewer than two remaining rollouts is counted as non-mixed. Counts are pooled within query for Qwen3-Instruct and within matched pair for the utility pulses. For both pulses, the three training seeds are summed within pair before paired cluster bootstrapping.

Table 8: **Demand-dependent mixedness persists after reliability failures are removed.** Values are candidate-group mixed-rate contrasts.

Sample	Original $D_4 - D_1$	Parsed-clean $D_4 - D_1$	Gap retained
Q3-Instruct, steps 1–60	$+.3267$	$+.3167$	97.0%
Q3-Thinking, early pulse	$+.3910$	$+.3577$	91.5%
Q3-Thinking, late pulse	$+.1360$	$+.1317$	96.8%

The training signal is primarily content-level. Cluster-bootstrap 95% intervals for the original and parsed-clean contrasts are, respectively, $[+.2970, +.3576]$ and $[+.2860, +.3464]$ for Qwen3-Instruct, $[+.3258, +.4509]$ and $[+.2929, +.4179]$ for the early pulse, and $[+.0791, +.1925]$ and $[+.0756, +.1885]$ for the late pulse. In Qwen3-Instruct, groups whose failures are all parsed, uncapped, and incorrect still yield a $D_4 - D_1$ gap of $+.2702$ $[+.2402, +.2991]$. Reliability involvement among D_4 mixed groups falls from 15.8% in steps 1–10 to 7.2% in steps 41–60, while D_4 mixed occupancy remains high (.632 to .540).

Late-intervention check. In the late intervention, training-time parse-failure and cap-hit rates are 0.047% and 0.061% for D_1, versus 0.774% and 0.901% for D_4; parsed-clean mixedness retains 96.8% of the eligibility gap. Thus the training-time selection difference is not primarily generated by completion failures. Appendix D.2 separately decomposes the held-out endpoint contrast. The Qwen2.5 trace lacks the same rollout-level parse/cap contract, so this failure-source analysis is scoped to Qwen3.

The original contrasts in Table 8 (.3910 early; .1360 late) are candidate-group mixed-rate contrasts. Table 2 instead reports realized accepted groups divided by all candidates, giving rounded differences of .373 and .125. The quantities differ because the final refill attempt at each step can produce surplus mixed groups. The trainer concatenates accepted groups in generation order and admits the first 64; any surplus remains in the candidate diagnostic but not in accepted exposure.

Rollout group-size boundary check. For each observed $G = 5$ group with c correct outcomes, we compute the exact probability that a uniformly sampled subset of size g is mixed. This preserves the realized rollouts while varying only the local eligibility calculation.

Table 9: **Exact group-size subsampling.** Entries are expected mixed rates; parentheses give the resulting D4/D1 acceptance ratio.

Panel	$g = 2$	$g = 3$	$g = 4$	$g = 5$
Qwen2.5 no-filter	.311 (1.84)	.466 (1.84)	.559 (1.81)	.621 (1.77)
Qwen2.5 standard	.281 (2.09)	.422 (2.09)	.507 (2.05)	.565 (2.01)
Qwen3 no-filter	.177 (2.40)	.266 (2.40)	.321 (2.38)	.359 (2.36)
Qwen3 standard	.169 (2.47)	.254 (2.47)	.307 (2.45)	.345 (2.43)

Eligibility increases monotonically with g, as predicted by the kernel, while the demand ordering changes little. This fixed-rollout calculation does not simulate how a genuinely different group size would alter refill stopping, optimizer updates, or the policy trajectory; it is a boundary check on the local selection model.

C.5 Task Projection and Boundary Dynamics

We define perception as Tables/Diagrams/Charts plus Spatial Reasoning and reasoning as mathematics, geometry, science, and the remaining reasoning categories. Table 10 reports mixed occupancy under standard filtering. The adjusted column standardizes to common support over target solvability,

Table 10: **Perception–reasoning eligibility is target-dependent.** Entries are perception minus reasoning mixed occupancy.

Target and window	P mixed	R mixed	Raw / adjusted P−R
Qwen2.5-Instruct, sequence-mean filter, 1–320	.1227	.2163	$-.0936/-.1650$
Qwen3-Instruct, filter, 1–60	.4614	.3828	$+.0786/+.0301$
Qwen3-Instruct, filter, 61–120	.3830	.3109	$+.0721/+.0600$
Qwen3-Thinking, filter, 1–120	.4545	.3039	$+.1506/+.0862$

Qwen3-VL-4B probe demand, answer schema, MCQ status, and source. Because source overlap is incomplete, these are common-support descriptions rather than causal task effects.

Qwen3-Instruct follows Qwen3-Thinking rather than Qwen2.5-Instruct in both halves of training, rejecting a simple generation-mode explanation and showing that the ordering is not an early transient. Composition explains much of the early Qwen3-Instruct magnitude, while the adjusted late gap remains positive. In both Qwen3 policies, perception also receives more accepted groups per query after composition control; reasoning can nevertheless consume more generated tokens because its responses are longer. Query-group allocation and token compute are distinct.

Four controls narrow the reversal claim. First, Qwen3 remains positive without filtering, so filtering exploits rather than creates the ordering. Second, solvability and composition adjustment does not change the Qwen3 sign. Third, on the same 399 query identities under sequence-mean, no-filter training, Qwen2.5 has P−R $= -.2605$ (query-cluster 95% interval $[-.309, -.214]$), whereas Qwen3-Thinking has $+.1851$ $[+.080, +.289]$. Finally, moving Qwen2.5 from token to sequence mean makes its gap more negative rather than moving it toward Qwen3. The result establishes non-invariance of the task projection; it does not identify model generation, initial capability, response horizon, or training state as the cause.

Observed boundary transitions. Boundary audits distinguish repeated extreme outcomes from exits out of the mixed state. The Qwen3-Thinking standard-filter audit over steps 1–100 reports all-fail-to-all-fail rates of .431 for perception and .638 for reasoning. The perception-minus-reasoning contrast is $-.174$ after standardization over initial solvability and response-demand strata (query-cluster bootstrap 95% interval $[-.298, -.079]$).

The Qwen3-Instruct audit concatenates the original steps 1–60 with the successful resumed run at steps 61–120; an abandoned trace from a checkpoint-save failure is excluded. Mixed-to-all-pass exits remain more common for reasoning in both halves. The adjusted P−R contrast is $-.062$ (95% CI $[-.114, -.011]$) over steps 1–60 and $-.050$ ($[-.083, -.017]$) over steps 61–120. The late-by-perception interaction is not significant, providing no evidence that this boundary ordering changes after step 60. By contrast, late all-fail-to-all-fail rates are nearly identical for perception and reasoning (.579 and .577); their adjusted difference is $+.012$ ($[-.065, +.090]$). Windowed transition claims rely mainly on step 41 onward because few queries are re-observed earlier.

These audits describe different boundary patterns accompanying the same higher-perception-occupancy ordering. Their observation schedules, training windows, and adjustment specifications differ, so we do not interpret the cross-policy contrasts as acquisition-rate estimates. In particular, repeated all-fail observations need not represent permanent absorption.

C.6 Filtering, Aggregation, and Training Dynamics

The main analysis asks which queries filtering admits; it does not assume that filtering improves endpoint performance. A Qwen2.5 factorial comparison holds the 400-query, 32-rollout panel fixed while varying zero-variance filtering and token- versus sequence-mean loss aggregation. The primary Qwen2.5 observational trajectory elsewhere in the paper uses token-mean aggregation; the sequence-mean arms here are controlled factor arms. Table 11 reports matched endpoints at steps 160 and 320. Coverage equals pass@32 on this panel.

At step 320, under sequence-mean aggregation, filtering changes pass@1 by $+.0077$ (query-paired 95% interval $[-.015, +.030]$) but coverage by $-.0525$ ($[-.088, -.018]$). Under token mean, it changes pass@1 by $-.0284$ ($[-.041, -.016]$) and coverage by $-.0250$ ($[-.048, -.003]$); this arm

Table 11: **Filtering does not provide a late endpoint gain at matched steps in the Qwen2.5 factorial.** All entries use the same held-out panel.

Aggregation	Filter	Step	pass@1	pass@8	Coverage	Parse fail	Mean length
Token mean	Off	160	.4550	.8054	.8975	.0047	461
		320	.4666	.7782	.8750	.0688	403
Token mean	On	160	.4579	.7975	.8950	.0067	462
		320	.4383	.7548	.8500	.2015	528
Sequence mean	Off	160	.4752	.7699	.8675	.0016	345
		320	.5111	.7155	.7975	.0014	215
Sequence mean	On	160	.4935	.7414	.8350	.0014	274
		320	.5188	.6950	.7450	.0006	178

also develops substantial parse failure. These comparisons are not an argument that no filtering is universally preferable: an optimizer step, an accepted group, and a generated-token budget are different resources. They show that the extra groups admitted by refill cannot be counted as a performance gain without an endpoint and a stated budget.

A possible filtering–aggregation interaction. The token-mean arms are nearly indistinguishable at step 160: filtering changes pass@1 from $.4550$ to $.4579$ and parse failure from $.0047$ to $.0067$. The larger difference appears only later. One plausible explanation is a feedback between the two operations. Filtering repeatedly admits queries with persistent outcome variation, while token-mean aggregation gives a response influence roughly proportional to its number of valid tokens. Queries that are both persistent and token-heavy can therefore receive greater selection frequency and greater within-batch weight. This account is consistent with the late rise in response length and parse failure in the token-mean filtered arm, but it is not identified by this factorial: the coverage loss is not concentrated in high-demand or reasoning queries, and no arm holds response length fixed. We therefore treat the interaction as a mechanism hypothesis rather than an explanation established by the endpoint contrast.

D Utility Interventions and Supporting Checks

D.1 Matched Interventions at Two Policy States

Design. The early intervention branches from the Qwen3-VL-2B-Thinking base policy; the late intervention branches from its standard-filtering step-40 checkpoint. Each independently constructs 80 low-demand (D_1) and 80 high-demand (D_4) queries, pair-matched on fine category, answer schema, and current solvability from 16 fresh rollouts. The early arm means are both $.5445$, with a $7.25\times$ response-demand ratio. Late paired solvability differs by at most $3/16$; its arm means are $.5508$ and $.5375$, with a $7.12\times$ demand ratio.

Each run receives 20 optimizer steps, or exactly 1,280 accepted groups, under the same filtering and training configuration. We train three seeds per arm and evaluate every endpoint with 16 rollouts on the same disjoint 400-query panel. The estimand is the held-out effect of allocating the same accepted update budget to the two matched pools from a common policy state. Each intervention arm initializes a fresh optimizer and scheduler from its common starting policy.

For per-query accepted-group counts e_q, exposure effective sample size is $\mathrm{ESS} = (\sum_q e_q)^2 / \sum_q e_q^2$. It equals 80 when accepted exposure is uniform across all queries in an arm and decreases as exposure concentrates on fewer queries.

Seed uncertainty. In the late intervention, all 12 seed-paired pass@k contrasts favor D_1. The pass@1 differences are $+.0009$, $+.0220$, and $+.0144$, with a mean of $+.0124$. A query bootstrap conditional on these trained policies gives $[+.0045, +.0204]$, whereas the 95% t interval across only three training seeds is $[-.0141, +.0390]$. The observed direction is consistent, but its magnitude remains imprecisely estimated across training seeds.

Generation cost. To obtain the same 1,280 accepted groups, D_1 uses 1,853 candidate groups with acceptance $.691$, whereas D_4 uses 1,568 with acceptance $.816$. Nevertheless, D_4 consumes 58.4M candidate and 48.6M accepted response tokens, compared with 27.5M and 21.1M for D_1. Higher acceptance therefore makes the group budget easier to fill but does not make it cheaper in tokens.

Table 12: **Seed-level endpoints of the late intervention.** All policies are evaluated on the same 400-query panel.

Arm	pass@1	pass@4	pass@8	pass@16
Step-40 branch policy	.6270	.7817	.8301	.8675
D_1, seed 1	.6322	.7948	.8446	.8725
D_4, seed 1	.6313	.7920	.8382	.8675
D_1, seed 2	.6366	.7941	.8398	.8725
D_4, seed 2	.6145	.7781	.8270	.8625
D_1, seed 3	.6308	.7941	.8418	.8700
D_4, seed 3	.6164	.7789	.8286	.8600

Table 13: **Seed-level endpoints of the early intervention.** All policies are evaluated on the same 400-query panel with 16 rollouts per query.

Arm	pass@1	pass@4	pass@8	pass@16
Base policy	.5192	.7486	.8179	.8575
D_1, seed 1	.6425	.8021	.8508	.8800
D_4, seed 1	.6134	.7824	.8366	.8700
D_1, seed 2	.6384	.8027	.8575	.8950
D_4, seed 2	.6192	.7908	.8452	.8750
D_1, seed 3	.6366	.7967	.8462	.8825
D_4, seed 3	.6197	.7945	.8464	.8700

Early seed-level endpoints. Relative to the common base policy, the mean gains are $+.1199/+.0519/+.0336/+.0283$ for D_1 and $+.0982/+.0406/+.0248/+.0142$ for D_4 at pass@1/4/8/16. Thus both allocations are useful at the early state, while D_1 produces the larger gain at every reported k under the same accepted-update budget.

The early pass@1 differences are $+.0291$, $+.0192$, and $+.0169$, with a mean of $+.0217$ and a training-seed 95% t interval $[+.0057, +.0378]$. Every per-seed query-bootstrap interval also excludes zero. Early pass@16 is positive in all three seeds (mean $+.0142$; $[+.0012, +.0271]$ across seeds). The pass@4 and pass@8 means are positive, but their training-seed intervals include zero.

Table 14: **Early low-demand eligibility depletes in every seed.** Slash-separated entries report D_1/D_4 within each seed. Token counts are in millions; ESS is out of 80 training queries.

Seed	Acceptance	Candidate groups	Candidate tokens	Accepted tokens	ESS
1	.228/.590	5,624/2,168	48.7/84.1	13.9/50.2	36.8/69.9
2	.209/.613	6,120/2,088	53.6/82.7	14.1/51.4	36.5/72.0
3	.237/.588	5,408/2,176	47.2/84.4	14.1/50.8	37.6/70.7

On average, early D_1 requires $2.67\times$ as many candidate groups to fill the same accepted-update budget. Its shorter responses still keep its candidate-token cost lower: D_4 uses $1.68\times$ the candidate tokens and $3.61\times$ the accepted tokens. These residence and cost patterns are stable across all three training seeds.

D.2 What Drives the Endpoint Difference

The logged format-augmented score is 1.0 for a correct parsed answer, 0.1 for a parsed but incorrect answer, and 0.0 for an unparseable answer; the binary accuracy field used for evaluation is one only in the first case. Hence endpoint accuracy factors as FC, where F is parseable completion and C is correctness conditional on parsing. We use the exact symmetric decomposition

$$\Delta(FC) = \frac{\Delta F(C_1 + C_4)}{2} + \frac{\Delta C(F_1 + F_4)}{2}.$$

Late intervention. Across the three late seeds, the mean $D_1 - D_4$ pass@1 gap is $+.0124$; $+.0125$ comes from parseable completion and approximately zero from conditional correctness. The seed-

level completion contributions are $-.0011$, $+.0243$, and $+.0142$, while the corresponding conditional-correctness contributions are $+.0020$, $-.0023$, and $+.0002$. Mean conditional correctness is .6869 at the step-40 branch, .6810 after D_1 training, and .6811 after D_4 training.

Mean response length, cap rate, and parse failure are 7,114, .077, and .070 for D_1, compared with 7,879, .100, and .089 for D_4. The branch values are 7,718, .095, and .087. Thus D_1 shortens responses and improves completion reliability, whereas D_4 remains near the branch; neither arm shows an observed conditional-correctness gain.

Early intervention. The base policy has mean response length 6,803 and parse failure .0522. The corresponding values are 6,921 and .058 after D_1 training, and 7,316 and .071 after D_4 training. Both arms therefore achieve their substantial base-relative pass@1 gains despite slightly higher parse failure.

Across the three early seeds, the mean $D_1 - D_4$ pass@1 gap is $+.0217$. Parseable completion contributes $+.0086$, with seed-level values $+.0134$, $+.0056$, and $+.0068$. Correctness conditional on completion contributes $+.0131$, with seed-level values $+.0157$, $+.0136$, and $+.0101$; its training-seed 95% t interval is $[+.0061, +.0201]$. Mean response length, cap rate, and parse failure are 6,921, .064, and .058 for D_1, compared with 7,316, .077, and .071 for D_4. The early contrast therefore contains a replicated conditional-correctness component and is not solely a completion effect.

Query-bootstrap intervals condition on the trained policies and do not represent training-seed uncertainty.

Evaluation strata. Across held-out demand quartiles from D_1 through D_4, the early mean $D_1 - D_4$ pass@1 contrasts are $+.0288$, $+.0067$, $+.0248$, and $+.0267$; the late contrasts are $+.0021$, $+.0090$, $+.0096$, and $+.0292$. High-demand training therefore shows no special advantage on high-demand evaluation queries, although these post hoc point estimates do not establish separate stratum-specific effects. The early contrast is larger on perception than reasoning ($+.0457$ versus $+.0150$), but the perception panel contains only 88 queries and this comparison is exploratory.

Policy-state comparison. The observed early pass@1 contrast is larger than the late contrast ($+.0217$ versus $+.0124$), but their difference is uncertain: the early-minus-late paired-query interval is $[-.0033, +.0221]$ conditional on the observed seeds. The interventions also use different query pools. We therefore interpret them as two-state replications of the eligibility–utility mismatch, not as a causal policy-state interaction.

D.3 Cross-Target Check and Staging

Qwen2.5 staging and mixture order. The Qwen2.5 pilot uses no filtering and token-mean aggregation. Its 211-query D_1 and D_4 pools are matched on fresh s_0, fine category, and MCQ status, while their median demands differ by $14.0\times$. Three additional arms allocate equal per-query exposure to both strata in opposite orders or interleave them 1:1.

Table 15: **Demand staging does not reveal a benefit from presenting the high-demand stratum first or last.** Values are single-run endpoints on the same 400-query panel with 32 rollouts per query.

Arm	pass@1	pass@8	pass@16	pass@32	D4-eval pass@8
D1 only	.4146	.8026	.8701	.9100	.7722
D4 only	.4056	.7831	.8466	.8900	.7242
Joint 1:1	.4082	.7958	.8623	.9025	.7561
D1 → D4	.4118	.7945	.8607	.8975	.7509
D4 → D1	.4020	.7894	.8562	.8975	.7491

None of the three mixed-stratum schedules is distinguishable from the joint arm on pass@1, pass@8, or pass@16. The direct D_1-only minus D_4-only contrast is $+.0090$ at pass@1 ($[-.0009, +.0188]$), $+.0195$ at pass@8 ($[+.0042, +.0352]$), and $+.0235$ at pass@16 ($[+.0065, +.0415]$). On the 100 queries with the lowest base-policy solvability, the pass@16 contrast is $+.0739$ ($[+.0129, +.1371]$). This single-seed experiment preserves the low-versus-high demand ordering but detects no benefit from presenting either stratum first. It does not identify an optimal mixture ratio.

D.4 Cross-Benchmark Evaluation on MathVista

As an exploratory benchmark-transfer check, we evaluate the base policy and the seed-1 early-intervention endpoints on all 1,000 MathVista testmini queries (Lu et al., 2024). Each policy generates one response per query using temperature 1.0, top-$p = .95$, top-$k = 20$, a 16,384-token response horizon, and the same evaluation seed. We find no exact matches in normalized question text or image SHA-256 hashes between the 160 early-intervention training queries and MathVista testmini. Because this check uses one training-seed pair, we do not use it for the primary seed-level inference.

Table 16: **Single-seed evaluation on MathVista testmini.** All policies are evaluated on the same 1,000 queries with one sampled response per query. Conditional accuracy is accuracy among parseable responses.

Policy	Accuracy	Parse fail	Cond. accuracy	Cap hit	Mean tokens
Base	68.7%	6.5%	73.5%	4.6%	3,515
Low demand (D_1), seed 1	**70.3%**	5.5%	74.4%	4.9%	3,924
High demand (D_4), seed 1	67.9%	6.5%	72.6%	5.5%	3,918

The $D_1 - D_4$ accuracy difference is $+2.4$ percentage points (query-paired bootstrap 95% interval $[+0.2, +4.6]$ pp; exact McNemar $p = .0416$), with 76 queries answered correctly only by D_1 and 52 only by D_4. The two endpoints have nearly identical mean response length (3,924 versus 3,918 tokens), so the contrast is not associated with greater evaluation-time generation by D_1. Its lower parse-failure rate contributes to the difference, while correctness conditional on parsing also favors D_1.

E Controller Implementation and Validation

Implementation. For demand quartile b, the sampler assigns a proposal multiplier proportional to $1/\max(\widehat{A}_b, .05)$ and caps any quartile's candidate share at $.45$. The acceptance estimate uses an EMA half-life parameter of 15 and a 1,024-row allocation window, approximately 4.3 optimizer steps, initialized from the common steps 1–40. Static and EMA arms use the same expanded-multiset, without-replacement sampler; fractional slots are assigned by seeded stochastic rounding. Multipliers are read when the next iterator is constructed, creating a one-window lag and preventing a window from affecting its own candidate sampling.

Past acceptance predicts the next accepted mixture. For each five-step window, we predict its accepted mixture by normalizing the product of its observed candidate shares and acceptance rates estimated from preceding windows. Under Qwen3 standard filtering, this predictor reduces mean total-variation error from $.1726$ for a candidate-share-only baseline to $.0431$, a 75.0% reduction. The paired window-bootstrap 95% interval for the absolute reduction is $[+.1161, +.1430]$, and predicted and realized bin-window shares correlate at $.982$.

The corresponding past-only check on the EMA arm gives error $.0448$ and correlation $.810$. An independent replay of the inverse-acceptance rule approximates the observed EMA candidate composition with mean error $.0257$. The archived replay uses five-step windows to approximate the original 4.3-step update cadence and is therefore a consistency check rather than a bit-for-bit reproduction. The result is insensitive to temporal and binning choices: ten- and twenty-step windows preserve the same ordering, while drift over two, four, and eight bins ranges only from $.0986$ to $.1013$ for Qwen2.5 and remains $.1678$ for Qwen3 over the full standard trajectories.

Realized allocation and cost. The static oracle derives fixed candidate-proposal probabilities from acceptance rates measured in the completed standard run over the scored window, then applies them in a separate training arm. Its calibration is therefore in-sample with respect to that standard-run window, making it a controllability reference rather than a deployable baseline. A controller frozen after step 40 removes only 57% of the oracle correction even though the rank ordering of bin acceptance is stable; the level of acceptance continues to fall. The past-only EMA tracks that changing level and reduces TV-to-uniform from $.1713$ to $.0562$, removing 67.2% of drift over post-branch optimizer

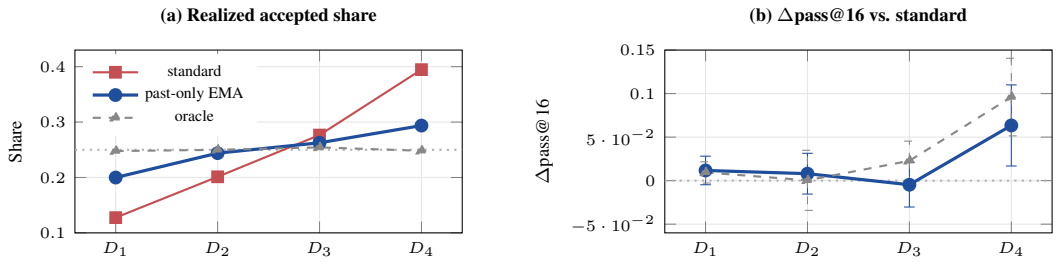


Figure 5: **Inverse-acceptance control flattens realized exposure. (a)** EMA reduces the standard D1–D4 skew using only past acceptance; the oracle is fitted in-sample. The dotted line is the uniform diagnostic target. **(b)** Repeated-sampling gains are largest in D4 even though the controller reduces its training exposure. Error bars are query-paired 95% intervals.

steps 41–100. It uses $1.02\times$ the generated response tokens and $1.01\times$ the wall time of standard filtering; the corresponding oracle ratios are $.95\times$ and $.94\times$. Table 3 reports the realized shares, costs, and aggregate evaluation results.

Where the coverage gain appears. Evaluation quartiles are formed by applying the same Qwen3-VL-4B probe to the held-out panel. EMA's pass@16 gain is largest in D_4, whose training share is reduced from .3947 to .2937: $+.0634$ with query-paired 95% interval $[+.020, +.113]$. The D_1 effect is $+.0117$ $[-.001, +.032]$. Regressing the per-query effect on continuous log demand gives $+.0207$ $[+.005, +.038]$ for EMA and $+.0339$ $[+.019, +.051]$ for the oracle. This is a consequence of the joint redistribution, not identification of one stratum's direct causal effect. All intervals are query-paired and condition on this single controller run; all policies use the same 400-query ordering and binary scorer.

Qwen2.5 replication. Static calibration under the Qwen2.5 token-mean setting reduces exposure imbalance from .1973 to .0421 (79%) while increasing generated tokens from 83.5M to 85.0M (1.8%). Relative to standard filtering, pass@1 changes by $-.0081$ $[-.024, +.008]$ and pass@8 by $+.0233$ $[+.0033, +.0433]$. These arms use the legacy D_{succ} partition that determined their training-time sampler; rescoring them with new bins would change the intervention's ruler after the fact.

Uniform accepted exposure is a zero-drift diagnostic, not a claim that it is the optimal curriculum. These experiments establish target realization: a chosen accepted distribution can be approached despite unequal acceptance, but they do not solve target design. A utility-aware system must still choose the desired distribution.