BETTER IN HINDSIGHT: WHY ORACLE FUSION GAINS ARE HARD TO LEARN AND WHERE EVIDENCE HELPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval systems often fuse a fast retriever's score with a stronger verifier's score, and the fusion weight that ranks the correct item first varies from query to query. In hindsight, choosing the best of 21 weights for each CIRR query adds 10.2–12.2 R@1 points over a tuned constant weight, whereas a ridge policy that predicts the weight from score statistics gains 0.32 points. We explain this gap with two exact identities. For R@1, the successful weights of a query form an interval that can be computed from its scores, so headroom is exactly the share of queries whose interval excludes the constant. Every policy's net gain is its rescued failures minus its harmed successes. Across 21 vision–language verifier configurations on CIRR and FashionIQ and five BEIR text datasets, learned policies gain through caution: they rarely harm solved queries and seldom rescue queries whose successful weights lie far from the constant. The bottleneck is knowing which candidate is the target. A label-derived hint lets the same procedure recover a median 89% of headroom, competitor-aware features stop helping under an equal tuning budget, and only decoders trained to identify the target beat the tuned baseline. Acting on this diagnosis, the winner-path duel asks a second verifier about only the candidates that some weight can rank first, 1.6–2.0 calls per query. Under protocols fixed in advance, it gains 2.17 and 1.16 R@1 points on new CIRR and FashionIQ queries, and 3.99 and 1.85 points (36% and 22% of headroom) when a larger verifier scores only those candidates. We recommend reporting reachable headroom, rescue and harm, and comparing adaptive policies under matched tuning budgets.

1 INTRODUCTION

Many retrieval systems rerank a fast retriever's shortlist with a slower but stronger verifier, such as a vision–language model (VLM) that reads the query together with each candidate. The final ranking usually fuses the two scores with a single weight tuned on held-out data (Bruch et al., 2022). The best weight, however, differs from query to query. Figure 1 shows a composed image retrieval (CIR) query, a reference image plus a requested change, for which the tuned weight ranks another candidate first while smaller weights rank the target first. Summed over queries, the opportunity is large. On CIRR, an oracle that picks the best of 21 weights for each query gains 10.2–12.2 R@1 points over the tuned constant (Figure 1, right).

Oracle gaps like this one are a common motivation for methods that gate, route or reweight a reranker per query (Yoon et al., 2025; Dabod et al., 2026; Genc et al., 2026). Algorithm selection measures the same kind of opportunity as the gap between the per-instance best solver and the single best solver (Rice, 1976; Kerschke et al., 2019). Here the opportunity is hard to realize. A ridge policy that predicts the value of each weight from score statistics gains $+0.32$ points, about 3% of the oracle gain. *Why is hindsight so much better than foresight, and what would close the gap?*

We first hold the retriever, the Top-20 shortlist, the verifier scores and the 21 candidate weights fixed, so any change in recall is due to the choice of weight. Two exact identities organize the study. First, for R@1 the weights that rank a target first form an interval that can be computed from the scores. Headroom is therefore exactly the fraction of queries whose interval excludes the constant (Section 3). Second, a policy's net gain equals the failures it *rescues* minus the successes it *harms* (Section 4), which separates never moving to the right weight from moving when it hurts. Across 21 verifier configurations from seven VLM families on CIRR and FashionIQ, we find:

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.

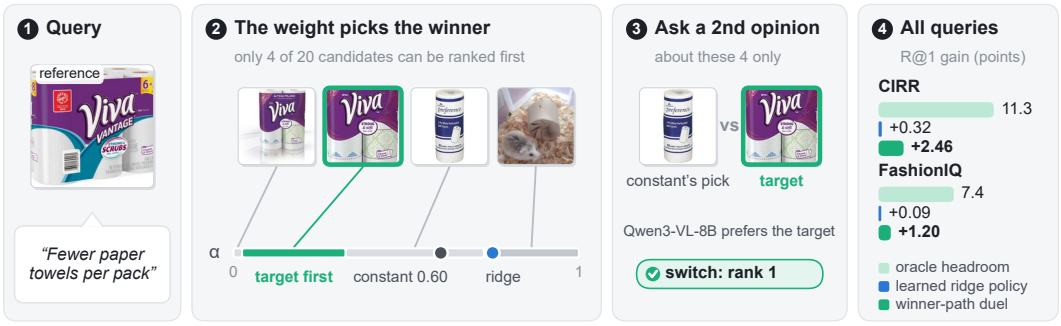

Figure 1: **Learning misses the successful weight; a targeted second opinion finds it.** A CIRR query (InternVL3-2B verifier; target outlined). Each fusion weight ranks one candidate first, and only four candidates win at some weight. Every weight in the green segment ranks the target first, but the tuned constant ($\alpha = 0.60$) and the ridge policy (0.60 or 0.75 across seeds) pick other winners. The winner-path duel asks a second verifier, here Qwen3-VL-8B, about these four candidates only and switches to the target. Right: over all development queries, oracle headroom and the gains of the ridge policy and the duel. Appendix A walks through this query, including other second opinions.

- **The opportunity is countable.** A query that no weight can fix lacks its target in the shortlist or has it outranked at every weight; the 21-value grid misses a feasible interval for only 0.1–0.2% of queries. The same pattern holds in text retrieval on five BEIR datasets (headroom 4.8–11.7 points, ridge gain -0.16 to +0.31).
- **Learned policies gain by caution.** The ridge policy harms far fewer solved queries than a permutation reference, but on CIRR it rescues only 0.7% of queries whose successful weights lie at least 0.4 from the constant (reference 14.5%).
- **The missing ingredient is information about the target.** A label-derived hint of a successful weight lets the same ridge procedure recover a median 89% of headroom. Encoding the target's competitors helps a fixed decoder by over a CIRR point, mostly by reducing harm, but equal-budget tuning shrinks this to +0.14. Only decoders that identify the target beat the tuned baseline.
- **Targeted evidence recovers it.** Our winner-path duel (WPD) asks a second verifier about only the candidates that some weight can rank first (1.6–2.0 calls per query) and switches only when a learned judge is confident. On development data it recovers 22% of CIRR and 15% of FashionIQ headroom, several times more than any fixed-information policy. Under protocols fixed in advance it gains +2.17 and +1.16 R@1 points on new queries, and +3.99 and +1.85 (36% and 22% of headroom) when a larger verifier scores only the path candidates.

Contributions. (1) An exact, tie-aware accounting of per-query fusion opportunity: feasible intervals, unreachable queries, and the identity linking headroom to rescue and harm. (2) Controlled comparisons across 21 verifier configurations and five text datasets showing that learned gains are caution and that only target identification beats equal-budget tuning. (3) WPD, a budgeted cascade that consults a second verifier only where the fusion decision is contested, confirmed on new populations and needing about a tenth of a larger verifier's calls.

2 Problem Setup

Retrieval pipeline. A CIR query consists of a reference image and a text describing a requested change. The goal is to retrieve the target image. We use CIRR (Liu et al., 2021), with natural images, and FashionIQ (Wu et al., 2021), with fashion products. After availability checks and the exclusion of pilot target groups they contain 3,380 and 4,539 validation queries. Frozen SigLIP-L embeddings (Zhai et al., 2023) retrieve a Top-20 shortlist for each query. A VLM verifier then scores each candidate. It receives the reference image, the requested change and the candidate, and its score is the log-probability gap between an affirmative and a negative answer at the last non-padding token. We use 11 verifier configurations on CIRR and 10 on FashionIQ (Appendix B).

Fusion actions. Candidate j of query q receives the fused score

$$s_{qj}(a) = (1-a)\bar{r}_{qj} + a\bar{v}_{qj}, \qquad a \in \mathcal{A} = \{0, .05, \ldots, 1\}, \tag{1}$$

where $\bar{r}$ and $\bar{v}$ are retriever and verifier scores min–max normalized within the query's shortlist. Ties keep the retriever order, and a verifier that gives every candidate the same score leaves the retriever order unchanged. $Y_q(a) \in \{0, 1\}$ indicates R@k success at weight a. A target outside the shortlist is a miss at every weight.

Policies. A policy chooses one weight per query. Except for the oracle, it uses only information available before the target is known.

- *Constant*: the single weight c that maximizes training recall. *Oracle*: a weight with $Y_q(a) = 1$ whenever one exists, chosen in hindsight; it is an upper bound only.
- *Ridge policy*: from 25 query-level summaries of retriever and verifier scores (e.g., top-two gaps, entropies, rank agreement), ridge regression (penalty 10) predicts each weight's gain $Y_q(a) - Y_q(0)$ and picks the largest.
- *Candidate policy*: from 158 per-candidate features, gradient boosting predicts the probability that each candidate is the target; the policy picks the weight with the highest resulting expected recall.

Evaluation. Policies are trained and tested on five folds, with queries that share a target image kept in the same fold, and five seeds. Tuning and standardization use training folds only. We report gains over the constant in R@k points with paired 95% intervals from 10,000 target-group bootstrap draws. When verifiers are pooled, their outcomes are averaged within each query before resampling. The intervals treat fitted policies as fixed; Appendix D reports refit and partition variability separately.

3 How Much Opportunity Is There?

Successful weights form an interval. Fix a query whose single target t is in the shortlist. Let $d_j^r = \bar{r}_t - \bar{r}_j$ and $d_j^v = \bar{v}_t - \bar{v}_j$ be the target's retriever and verifier margins over competitor j. The target outranks j at weight a exactly when

$$d_j^r + a(d_j^v - d_j^r) \ \triangleright_j \ 0, \tag{2}$$

with $>$ if j precedes t in the tie order and $\geq$ otherwise. Each competitor thus contributes one linear constraint on a. When the verifier favors the target more than the retriever does ($d_j^v > d_j^r$), the constraint is a lower bound. When it favors the target less, the constraint is an upper bound. When the two margins are equal, the constraint either always holds or never holds. The target ranks first exactly on the intersection of these constraints with $[0, 1]$. This set is an interval I_q; empty intervals, singletons and open endpoints all occur. We call the competitors that set its endpoints *binding*. Figure 2a shows intervals of real queries. For R@k with $k > 1$, or with several relevant targets, the success set can be a union of intervals. Appendix C gives proofs and a counterexample.

Headroom equals the rescuable mass. Let $G_q = \{a \in \mathcal{A} : Y_q(a) = 1\}$ be the successful grid weights; for single-target R@1, $G_q = I_q \cap \mathcal{A}$. Conditional on the training sample and its constant c, define

$$C = \mathbb{E}Y(c), \quad O = \mathbb{E}\max_{a\in\mathcal{A}} Y(a), \quad H = O - C = \Pr(G \neq \varnothing,\ c \notin G). \tag{3}$$

The last equality holds because every query that the constant solves is also solved by the oracle. Each query is therefore *covered* by the constant, *rescuable* by some other weight, or *unreachable* at every weight. Headroom is exactly the rescuable mass. We evaluate the identity fold by fold with each fold's own constant.

What makes a query unreachable? Replaying Equation 2 with stable sorting reproduces all 1,733,970 saved query–weight outcomes. Unreachable queries have one of two causes (Table 1 in Appendix C). In some, the target is absent from the shortlist (16.60% of CIRR and 50.23% of FashionIQ queries). In others, the target is present but outranked at every weight (25.13% and 20.59%). The weight grid rarely matters: it misses a nonempty interval in only 0.17% and 0.12% of queries (117 configuration–query pairs in total). The remaining 11.3% of CIRR queries and 7.4% of FashionIQ queries are rescuable, and they make up all of the headroom (Figure 2b).

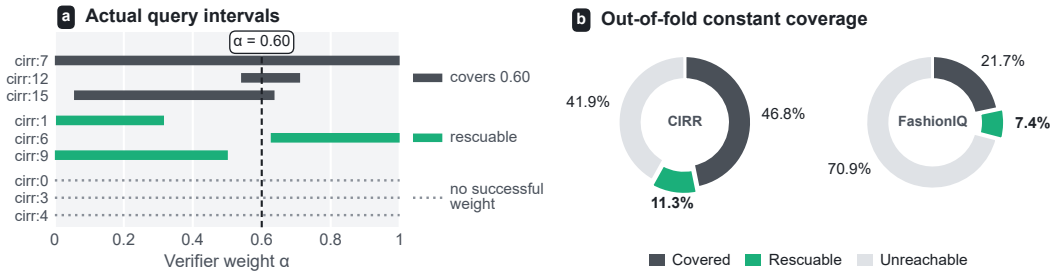

Figure 2: **Exact opportunity accounting.** (a) Feasible intervals of real CIRR queries (InternVL3-2B; first three queries of each class) relative to the weight 0.60. Dotted rows have no successful weight. (b) Share of queries covered by each query's out-of-fold constant, rescuable by another weight, or unreachable, averaged over verifier configurations. The rescuable share is the oracle headroom.

What can a policy know? A deployable policy sees features F computed before the target is known. Let $V(F) = \mathbb{E}[\max_a \mathbb{E}(Y(a) \mid F)]$ be the best value that any F-measurable policy can reach, again conditional on the training sample. A fitted policy π then satisfies

$$O - \mathbb{E}Y(\pi) = \underbrace{O - V(F)}_{\text{information gap}} + \underbrace{V(F) - \mathbb{E}Y(\pi)}_{\text{fitted-policy gap}}. \tag{4}$$

Both terms are nonnegative: the first is lost because F does not reveal the target, the second because the fitted policy uses F imperfectly. They cannot be measured separately, but the decomposition shows which experiment tests which term. Supplying label-derived information asks whether a fixed procedure could exploit information that F lacks (Section 5.1). A feature $g(F)$ computed from F leaves the attainable value unchanged, $V(F, g(F)) = V(F)$, so it can only help the learner and should be compared with other help, such as regularization or fallback, at an equal tuning budget (Section 5.3). Only new evidence about the candidates can enlarge F itself (Section 7).

Takeaway. Headroom is not a property of any learner. It counts the queries that some weight can fix and the constant does not, and retrieval coverage and verifier geometry determine that count.

4 How Much Do Learned Policies Recover?

Calibration must be matched first. Normalization changes what a fixed weight means. Under min–max normalization, weight a scales raw verifier scores by $\frac{a}{1-a}\frac{\text{range}(r_q)}{\text{range}(v_q)}$ relative to retriever scores. A constant normalized weight therefore already adapts to each query's score ranges, and comparing policies under different normalizations mixes calibration with adaptivity (Bruch et al., 2022). Below, the constant, the ridge policy and the candidate policy each select among four normalizations (query min–max, query z-score, global z-score, rank) on the same inner folds; the candidate policy also tunes capacity, temperature and shrinkage.

The candidate policy recovers a few percent of headroom. In its original fixed setup, the ridge policy gains $+0.32$ $[+0.10, +0.53]$ points at CIRR R@1 and $+0.09$ $[-0.09, +0.28]$ at FashionIQ R@1. Under matched calibration its mean gain is at most 0.07 points. The candidate policy gains $+0.61$ and $+0.39$ points at CIRR R@1/R@5 and $+0.69$ and $+0.20$ at FashionIQ R@1/R@10 (Figure 8 and Table 2, Appendix D). Its 95% interval lies above zero in 24 of 42 verifier–metric cells and below zero in none. The two policies differ in inputs, learner and objective at once, so this comparison ranks complete methods. Section 5.2 separates the components.

The gain comes from the posterior, not from tuning. The candidate policy tunes normalization, capacity, temperature and shrinkage. A direct rule without any of these choices does as well. It takes the candidate posterior at a fixed capacity and picks the weight with the highest expected recall. Pooled over configurations, it gains +0.52 points at CIRR R@1 and +0.65 at FashionIQ R@1, within 0.1 point of the tuned policy, and both differences have intervals that include zero (Appendix E).

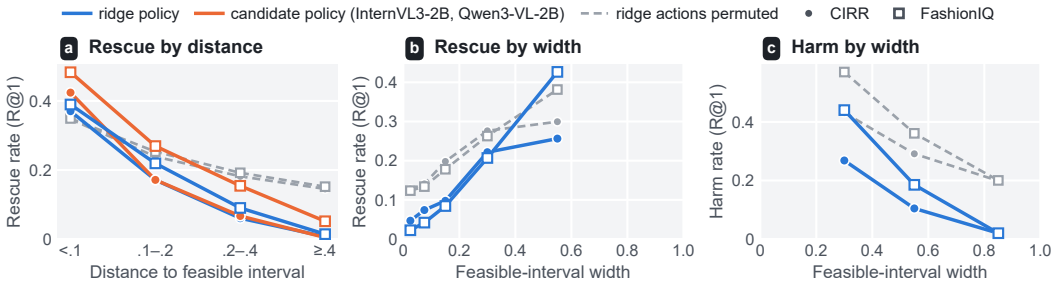

Figure 3: **The ridge policy protects successes but misses distant rescues.** R@1 rescue rate by distance from the constant to the nearest successful weight (a) and by feasible-interval width (b), and harm rate by width (c). Blue: ridge policy; orange: candidate policy (two verifiers); gray: ridge actions permuted within each test fold. Harm conditions on constant successes that some weight would break. Bins with fewer than 100 eligible queries across configurations are omitted.

Net gain is rescue minus harm. Let $r = \Pr(Y(\pi) = 1 \mid G \neq \varnothing, Y(c) = 0)$ be the rate at which a policy rescues rescuable queries, and $h = \Pr(Y(\pi) = 0 \mid Y(c) = 1)$ the rate at which it harms queries that the constant solves. Then

$$\Delta = \mathbb{E}[Y(\pi) - Y(c)] = Hr - Ch, \qquad \rho = \Delta/H \quad (H > 0), \tag{5}$$

where ρ is the recovered share of headroom. The two rates are weighted very differently. On CIRR R@1 the constant covers about four times as many queries as are rescuable ($C \approx 47\%$, $H \approx 11\%$; Figure 2b). One point of harm rate therefore cancels about four points of rescue rate, and a policy must rescue far more often than it harms to gain at all.

Where policies fail. Figure 3 bins rescuable queries by the distance from the constant to the nearest successful weight. As a reference, we permute each fold's ridge actions across its test queries. The permutation keeps how often each weight is chosen but removes any link to the query. Near the constant, the ridge policy rescues 37.0% of CIRR queries versus 34.7% for the reference. At distance 0.4 or more it rescues 0.7% versus 14.5%. Ridge seldom moves far from the constant, so its harm rate stays well below the reference (Figure 3c). The price is the distant rescues. The candidate policy strikes a different balance. At distance 0.4 or more it rescues 0.46% of CIRR queries (reference 6.85%) and 5.15% of FashionIQ queries (3.56%), and its harm rate exceeds its own reference on both datasets (Table 10).

Takeaway. The ridge policy gains by being selective. It avoids harmful moves but rarely finds the weights that rescue a query far from the constant. The candidate policy rescues more and harms more.

5 What Limits Recovery?

Section 3 split the gap into an information gap and a fitted-policy gap. We test three explanations: the decision procedure cannot use target information (Section 5.1), candidate-level features are missing (Section 5.2), or the learner needs an encoding of the binding competitors (Section 5.3).

5.1 Supplied target information closes the gap

We give fixed decision procedures label-derived information that they would not have at deployment (Figure 11, Appendix I). A hint of a successful weight lets the ridge policy recover a median 89% of headroom when exact and 49% with noise SD 0.2, while a random hint recovers nothing (−3%). Mixing the true target into the candidate posterior reaches 99% recovery at mixing strengths of 0.2–0.7. The decision machinery can thus act on target information when it has it. Since the successful weights of every *hypothesized* target follow from the observed scores (Section 3), what a policy lacks is knowledge of which candidate is the target.

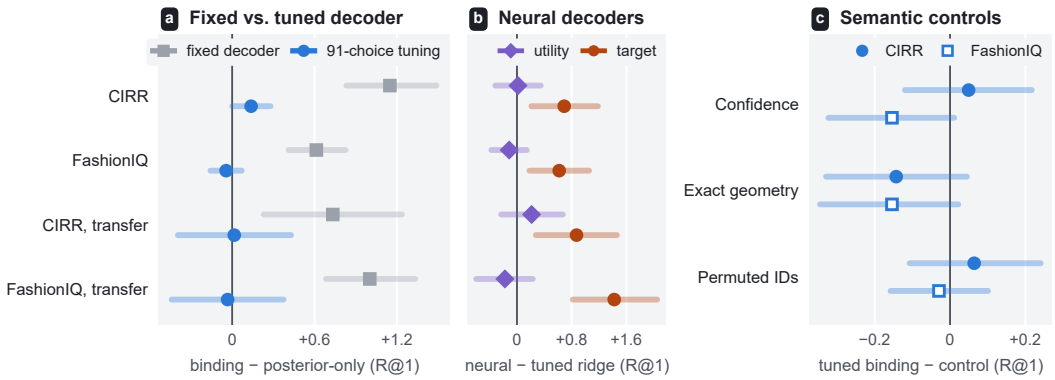

Figure 4: **Under equal tuning, only learning the target helps.** (a) Binding minus posterior-only R@1 with a fixed ridge decoder and with identical 91-choice tuning (97.5% intervals). (b) Neural set decoders minus the tuned posterior-only ridge decoder, trained to predict utilities or to identify the target (98.75% intervals). (c) Tuned binding minus three controls without competition semantics on the original populations (99.1667% Bonferroni intervals). Verifiers are averaged within queries; original-population estimates also average three complete refits.

5.2 Candidate-level features alone do not help

Crossing two inputs (25 query summaries or 1,925 per-candidate features) with two learners (ridge or GBM), all predicting the same 21 weight utilities under shared calibration and equal selection budgets, gives a representation effect of $+0.05$ points on CIRR and -0.02 on FashionIQ, both with intervals that include zero. Selection falls back to the constant in 100 of 100 candidate-ridge and 94 of 100 candidate-GBM fits; without fallback, candidate ridge loses 1.36 and 1.11 points (Appendix F).

5.3 Encoding the competition helps a fixed decoder, not a tuned one

Binding encoding. A target's binding competitors set the endpoints of its feasible interval (Section 3). Auxiliary classifiers, fitted on training folds, predict the target posterior and each candidate's probability of being a lower or upper binding competitor. For each weight, the encoding checks the predicted constraints and averages over hypothesized targets, giving 21 features that estimate whether each weight is safe (Appendix F).

With a fixed decoder, the encoding reduces harm. We add the encoding to a ridge decoder (penalty 10) that already receives the candidate posterior. R@1 rises by $+1.30\,[+0.91, +1.72]$ points on CIRR and $+0.50\,[+0.24, +0.76]$ on FashionIQ. With all models frozen, the gains carry over to the disjoint query populations of Section 6 ($+0.73$ and $+1.00$). The gain comes almost entirely from reduced harm, 1.29 and 0.45 points, while rescue changes by only 0.01 and 0.04 points (Table 18).

Under an equal tuning budget, the advantage disappears. Harm can also be reduced generically, by stronger regularization, by shrinking predicted utilities toward the training mean, or by falling back to the constant when the posterior is uncertain. We therefore give every representation the same 91 tuning choices: the constant, plus ten ridge penalties from 10^{-1} to 10^{8}, each combined with three shrinkage levels or with three fallback thresholds for each of two confidence measures. Auxiliary models are cross-fitted in a fully nested way, so selection never sees labels through the features. The fixed-decoder gain reproduces in this study ($+1.15$ on CIRR, $+0.61$ on FashionIQ). Under matched tuning, however, the binding increment falls to $+0.14\,[0.00, +0.28]$ on CIRR and $-0.04\,[-0.16, +0.07]$ on FashionIQ, and to $+0.02$ and -0.03 on the transfer populations (Figure 4a). Both tuned decoders end close to the constant, with gains between -0.26 and $+0.26$ points (Table 25). Nor does binding beat controls that carry no competition semantics (Figure 4c): sorted posterior probabilities (pure confidence), the exact geometry of every hypothesized target, and binding probabilities retrained after candidate identities are permuted.

What helps is learning the target. The tuned binding decoder does use the channel: permuting candidate identities at test time lowers its recall by 0.12–0.52 points (Table 24). Acting on the posterior directly works better, however. In this nested study, the direct expected-recall rule of Section 4, applied to its own target posterior, gains $+0.72$ and $+0.69$ points and beats the tuned binding decoder in all four panels. A stronger learner shows the same split (Appendix H). A permutation-equivariant set transformer over the 20 candidates, given the same inputs and 91 tuning choices, matches the tuned posterior-only ridge decoder when trained to predict the 21 utilities (+0.01 on CIRR, -0.11 on FashionIQ). Trained to identify the target and deciding by expected recall, the same network beats it by +0.69 and +0.62 points, and by +0.87 and +1.42 on the transfer populations, with all four 98.75% intervals above zero (Figure 4b).

Takeaway. Encodings of the competition add caution, which equal-budget tuning supplies anyway. What helps is target identification: acting on a target posterior, directly or through a network trained to find the target, beats every tuned utility decoder.

6 Does the Pattern Hold Beyond the Development Data?

Prespecified replication. Before scoring two new query populations (3,274 CIRR and 4,539 FashionIQ queries, with images disjoint from the development data), we fixed five predictions and their decision rules. Three predictions were confirmed. Headroom persists (H1), the pooled ridge gain stays within one point (H2), and ridge harms less and rescues fewer distant queries than its permutation reference (H5). The other two were not. H3 predicted a larger ridge gain at R@1 than at a deeper cutoff, and its intervals include zero. H4 bounded the candidate policy's recovery at one tenth of headroom, and the candidate policy exceeded that bound in two of four panels. A fourth verifier, completed after the original analysis, extends H1 to all 16/16 planned cells and changes no verdict (Table 31).

CIRCO and the CIRR test server. On CIRCO (Baldrati et al., 2023), ridge gains $+1.23$ [$+0.18$, $+2.32$] R@1 points and the candidate policy $+2.09$ [$+0.98$, $+3.30$] R@5 points, so small gains are not universal. The estimates are imprecise: CIRCO has 220 queries, 38.2% with the target in the shortlist, and we report single-target recall rather than the official mAP (Appendix J). On the sealed CIRR test set, which returns only aggregate scores, ridge changes R@1 by $+0.34$ and subset R@1 by -0.92 points.

Text two-stage retrieval. To test whether the pattern is specific to composed image retrieval, we applied the same code to five BEIR datasets (Thakur et al., 2021) with 16,506 queries. A BGE-base dense retriever (Xiao et al., 2024) returns the Top-20, and two cross-encoders trained on MS MARCO (Bajaj et al., 2016) act as verifiers. Four predictions were fixed before computing any text score (Appendix K). Oracle headroom is 4.8–11.7 R@1 points and exceeds two points in 5/5 datasets (T1), while ridge gains -0.16 to +0.31 (Figure 13 in Appendix K). The prespecified precision bound on the ridge gain holds in 3/5 (T2); in FiQA and SciDocs the intervals are too wide to meet it. Ridge is again cautious (T3, 5/5): it harms 1.7–20.7% of breakable successes against 23.4–30.8% for its permutation reference, and rescues 0.7–10.4% of distant queries against 11.5–26.6%. The candidate policy gains +0.15 to +1.98 points (T4 holds in 4/5), most on ArguAna (+1.98 points, 23% of headroom), where the cross-encoders alone reach only 15.3–21.8% R@1 against 40.6% for the retriever.

7 Spending Evidence Where the Decision Is Contested

Sections 3–5 locate the bottleneck in knowing which candidate is the target. No function of the fixed scores can supply that knowledge, but new evidence can. We now acquire it only where the fusion decision is contested.

The decision is a duel on the winner path. For R@1, a fusion weight only decides which candidate is ranked first. Let the *winner path* W_q be the candidates that some grid weight ranks first (four in Figure 1); by Section 3, these are the candidates whose feasible interval meets the grid. Moving from the constant's winner c to another winner w changes R@1 by $\mathrm{rel}(w) - \mathrm{rel}(c)$, so only the duels between c and the other winners matter. They are few: scoring every contested path costs 1.6 (CIRR) and 2.0 (FashionIQ) verifier calls per query on average, against 20 for rescoring the shortlist.

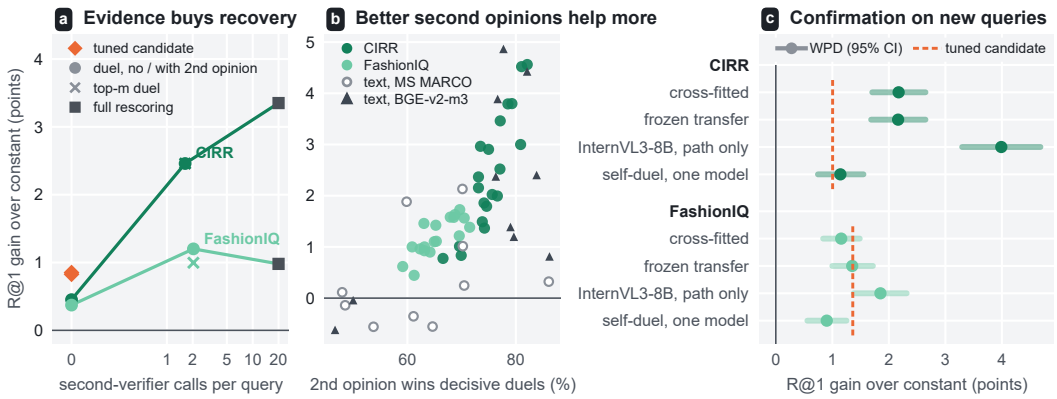

Figure 5: **Targeted evidence recovers what learning cannot.** (a) R@1 gain over the constant against second-verifier calls per query, development populations pooled over primary/second-opinion pairs: the tuned candidate policy, the duel judge without and with a second opinion, a budget-matched duel on the constant's top candidates, and full rescoring with both verifiers. (b) WPD gain against how often the second opinion ranks the relevant duelist first, for every development and text pair; text second opinions are an MS MARCO cross-encoder or BGE-reranker-v2-m3. (c) Replication populations with 95% intervals: cross-fitted and frozen-transfer WPD, the prospective path-only InternVL3-8B run and the one-model self-duel; dashed: tuned candidate policy.

Winner-path duel (WPD). When $|W_q| > 1$, WPD asks a second verifier to score only the path candidates. A logistic judge estimates the odds that w rather than c is relevant from both verifiers' score differences, the candidates' ranks, and where and how widely w wins along the weight axis. It is trained only on duels that switching would decide, where exactly one of w and c is relevant. WPD switches to the best alternative only if the judge's logit exceeds a threshold chosen on inner folds, which trades rescue against harm directly (Equation 5). It remains a fusion-weight policy: it outputs a weight at which the chosen candidate ranks first, so its gain is bounded by the headroom. The protocol was fixed on the development populations before any confirmation outcome (Appendix L).

Development results. Pooled over 20 CIRR and 18 FashionIQ primary/second-opinion pairs, WPD gains +2.46 and +1.20 R@1 points (22% and 15% of headroom), +1.61 and +0.38 more than the tuned candidate policy; all four 95% intervals lie above zero (Figure 5a, Table 40). Three comparisons locate the gain. Without the second opinion, the same judge gains only +0.45 and +0.37. Spending the same calls on the constant's top candidates instead of the winner path changes little (+0.00 and +0.20), so the path mainly sets the budget. Rescoring all 20 candidates with both verifiers gains +3.35 and +0.98; WPD obtains most of that gain on CIRR, and all of it on FashionIQ, with a tenth of the calls. The improvement comes from the new evidence, not from the judge or the exact candidate set. Compared with simply switching to a larger verifier, WPD with an 8B second opinion makes 8–10% of the 8B calls; its R@1 is within the 95% interval of the 8B verifier alone in 5 of 8 development cells, higher in 2 and lower in 1 (post hoc; Appendix L).

Confirmation on new populations. Before any duel outcome on new data, we fixed three predictions: a positive gain (D1), a gain over the tuned candidate policy (D2), and a positive gain for the development model applied unchanged (D3). On the replication populations of Section 6, with two primaries and every other verifier as second opinion, WPD gains +2.17 [+1.71, +2.65] R@1 points on CIRR and +1.16 [+0.84, +1.49] on FashionIQ (20% and 14% of headroom), with the interval above zero in 12 of 12 verifier pairs (D1). Applied unchanged from development, it gains +2.16 and +1.35 (D3; both intervals above zero). It beats the tuned candidate policy on CIRR by +1.17 but not on FashionIQ (−0.20, interval includes zero), so D2 fails there (Figure 5c). A prospective run then tested the deployment path with the strongest development second opinion: InternVL3-8B scored only the winner-path candidates (1.6 and 2.0 calls per query and primary). WPD gained +3.99 [+3.30, +4.68] and +1.85 [+1.38, +2.32], 36% and 22% of headroom. D1 and D3 hold, and WPD also beats the tuned candidate policy on both datasets (+2.98 and +0.49, intervals above zero; not a prespecified test for this run).

The gain follows the information in the second opinion. If WPD works by acquiring evidence, its gain should grow with what the second opinion knows that the primary does not. The text study, where both original verifiers are cross-encoders trained on the same MS MARCO data, provides a test. There WPD gains significantly on only 2/5 datasets (NQ: +1.57 points, 13% of headroom; corrected analysis, Appendix L). As predicted before scoring, a second opinion trained on different data (BGE-reranker-v2-m3) makes the gain significant on 4/5 datasets and raises NQ to +4.65, 40% of headroom. A self-duel, in which the primary VLM itself compares the two duelists in one prompt, needs no second model. On the replication populations it gains +1.14 and +0.90, no more than the tuned candidate policy (Appendix L): asking the same model again adds less than asking another.

8 Related Work

Reranking and score fusion. CIR benchmarks pair a reference image with a textual modification (Liu et al., 2021; Wu et al., 2021; Baldrati et al., 2023), and many systems rerank a shortlist with a stronger multimodal model (Liu et al., 2024; Wen et al., 2026; Du et al., 2026). Linear and rank-based fusion are standard in hybrid retrieval (Cormack et al., 2009; Bruch et al., 2022), and normalization changes what a fusion weight means (Bruch et al., 2022). We hold such pipelines fixed and study how their two scores are combined per query.

Adaptive reranking and routing. Recent methods allocate listwise reranking computation by uncertainty (Yoon et al., 2025), gate reranking on model uncertainty (Dabod et al., 2026), or route each query to a ranking strategy (Genc et al., 2026). Our accounting applies directly to evaluating such methods. Cascades spend expensive rankers on a shrinking candidate set (Wang et al., 2011), and pairwise rerankers compare top candidates directly (Pradeep et al., 2021). WPD is a budgeted cascade of this kind, and asking about the constant's top candidates works about as well as asking about the winner path (Section 7). What it adds is a fusion-specific rule for which duels can change R@1 and when a switch is worth its risk of harm.

Selection, regressions and abstention. Choosing among fixed options per instance is the algorithm selection problem (Rice, 1976), in which the gap between the per-instance best and the single best option measures what selection could gain (Kerschke et al., 2019). Our headroom is this gap for fusion weights. Harm corresponds to the negative flips studied for model updates (Yan et al., 2021), and fallback to the constant is a form of selective prediction (Geifman & El-Yaniv, 2017).

9 Discussion and Conclusion

Oracle headroom in per-query fusion measures opportunity, not learnability. Realizing the opportunity means rescuing queries without harming the many that the constant already solves. Learned policies resolve this trade-off by caution, and features that re-encode the same scores add caution rather than information. The gap is an information gap. Supplied target information closes it, only decoders that learn the target beat equal-budget tuning, and a second opinion spent on the contested winners recovers several times more headroom than any fixed-information policy.

Recommendations. For adaptive fusion, reranking gates and routers: (1) report reachable headroom, rescue and harm, with a permutation reference that keeps the policy's action frequencies; (2) compare adaptive policies under matched calibration and tuning budgets, which here reduced a competitor encoding's CIRR gain from more than one point to $+0.14$; (3) spend extra verification where the fusion decision is contested rather than on the whole shortlist.

Limitations. Our conclusions concern frozen retrieval, Top-20 shortlists, scalar linear fusion and the procedures we evaluate. Bootstrap intervals treat fitted policies as fixed; over 30 refits the candidate policy's gain has an SD of 0.36–0.52 points. WPD needs a second opinion that carries new information, and it saves large-verifier calls rather than dominating the large verifier, which is more accurate in 1 of 8 development cells. It was designed on the development populations and confirmed on new queries from the same benchmarks; on the FashionIQ replication it beats the tuned candidate policy only with the larger second opinion. CIRCO's larger ridge gain shows that the learnability gap varies across benchmarks.

ETHICS STATEMENT

The experiments use public benchmark annotations and model outputs. A single annotated target need not exhaust semantically valid results. Rights to the illustrative benchmark images remain with their owners. No human judgments were collected.

AI USE STATEMENT

We used large language model assistants for writing and polishing the manuscript and for retrieval and discovery of related work. The authors reviewed and revised the assisted text and checked every citation against its primary source. The authors are responsible for all research decisions, sources, analyses, and the final text.

REPRODUCIBILITY STATEMENT

Appendix B specifies scoring, features, fitting and uncertainty. Appendix C gives proofs and boundary cases. Appendices F–H detail the controlled procedures, and Appendices K and L the text study and the duel protocol with its amendments. The supplementary material contains the analysis code, the prespecified protocols, and stored query identities, folds, input hashes and decisions that support score replay and refitting.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated MAchine reading COmprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15338–15347, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Baldrati_Zero-Shot_Composed_Image_Retrieval_with_Textual_Inversion_ICCV_2023_paper.html.

Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. *arXiv preprint arXiv:2210.11934*, 2022. URL https://arxiv.org/abs/2210.11934.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of SIGIR*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Orian Dabod, Amir DN Cohen, and Gabriel Stanovsky. When reranking hurts: Uncertainty-based gating for few-shot reranking. *arXiv preprint arXiv:2606.31087*, 2026.

Enjun Du et al. EviRank: Structured relevance evidence for multimodal image re-ranking. *arXiv preprint arXiv:2608.20886*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4878–4887, 2017.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Ata Cinar Genc, Emir Kaan Korukluoglu, and James Allan. Adaptive re-ranking. *arXiv preprint arXiv:2606.25249*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Pascal Kerschke, Holger H. Hoos, Frank Neumann, and Heike Trautmann. Automated algorithm selection: Survey and perspectives. *Evolutionary Computation*, 27(1):3–45, 2019. doi: 10.1162/evco_a_00242.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2125–2134, 2021. arXiv:2108.04024.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2305.16304.

Andrés Marafioti et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The Expando-Mono-Duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*, 2021.

John R. Rice. The algorithm selection problem. In *Advances in Computers*, volume 15, pp. 65–118. Academic Press, 1976.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. arXiv:2104.08663.

Lidan Wang, Jimmy Lin, and Donald Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2011. doi: 10.1145/2009916.2009934.

Haokun Wen, Xuemeng Song, Haoyu Zhang, Weili Guan, Xiangyu Zhao, et al. UniCVR: From alignment to reranking for unified zero-shot composed visual retrieval. *arXiv preprint arXiv:2604.20318*, 2026.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:1905.12794.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2024. arXiv:2309.07597.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14299–14308, 2021.

Soyoung Yoon, Gyuwan Kim, Gyu-Hwung Cho, and Seung-won Hwang. AcuRank: Uncertainty-aware adaptive computation for listwise reranking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.18512.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2303.15343.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX OVERVIEW

Appendix A walks through real queries with images and is the easiest entry point. Appendix B specifies models, prompts, features, fitting and uncertainty. Appendix C proves the identities of Section 3 and treats boundary cases. Appendix D lists every verifier–metric cell behind Section 4, with refit and partition sensitivity, and Appendix E tests the direct expected-recall rule on every configuration. Appendices F–H detail the controlled comparisons of Section 5, including the neural set decoders, and Appendix I describes the label-derived probes. Appendix J gives the full replication, CIRCO and test-server results, and Appendix K the text retrieval study of Section 6. Appendix L gives the winner-path duel protocol, its amendments and every result behind Section 7.

A A WORKED EXAMPLE

This appendix explains the paper's main quantities on real queries. All examples use CIRR with the InternVL3-2B verifier.

One query, step by step. Figure 6 expands the query of Figure 1.

1. The request asks for a pack with fewer paper towels than the reference.
2. Every candidate in the shortlist has two scores. The retriever ranks candidate A, a Viva pack very similar to the target, first and the target a close second. The verifier strongly dislikes A and prefers B, a single roll, to the target. It scores C, a hamster beside a cardboard tube, highest of all.
3. Changing the verifier weight α moves each candidate's fused score along a straight line, from its retriever score at $\alpha = 0$ to its verifier score at $\alpha = 1$ (Equation 1). The target's line is on top only between $\alpha = 0.003$ and 0.317. Below that range A is first, and above it B overtakes the target. A and B are therefore the target's *binding competitors*: they set the two ends of its feasible interval (Section 3). As α grows further, C and other candidates also pass the target.
4. Every grid weight from 0.05 to 0.30 ranks the target first, so the oracle succeeds. The tuned constant ($\alpha = 0.60$) ranks the target fourth. Depending on the seed, the ridge policy chooses 0.60 or 0.75 and ranks it fourth or seventh.

The query is *rescuable*: it counts toward headroom. The ridge policy does not rescue it, because its successful weights end about 0.3 below the constant, where Figure 3a shows that ridge rarely rescues.

The duel on this query. Only four candidates are ranked first by some grid weight: A at $\alpha = 0$, the target from 0.05 to 0.30, B from 0.35 to 0.70 and C from 0.75 to 1. The constant's winner is B, so the winner-path duel (Section 7) asks a second verifier about these four candidates and nothing else. With Qwen3-VL-8B or Gemma-3-4B as second opinion it switches to the target in all five seeds, with Qwen3-VL-4B in four, with InternVL3-8B and SmolVLM2-2.2B in two, and with the other five second opinions in none. Averaging both verifiers over all 20 candidates does not rescue this query with any second opinion. Figure 1 shows Qwen3-VL-8B, which scores the target highest of the four.

Four kinds of query. Figure 7 shows one real query for each outcome that the accounting distinguishes. To avoid picking favorable examples, each row shows the first query in query-ID order of its type, under the seed-0 decisions of the constant and the ridge policy.

- **Rescued.** The constant ranks the target third; the ridge policy moves to $\alpha = 0.25$, which ranks it first.
- **Harmed.** The constant already ranks the target first; the policy's move to $\alpha = 0.60$ drops it to second.
- **Unreachable.** In the third row the target is in the shortlist, but no weight lifts it above third place. In the fourth row the retriever does not place the target in the Top-20 at all, so no fusion weight can help.

Across all 3,380 queries, the seed-0 decisions of this configuration keep 1,462 successes, rescue 81 queries and harm 50. They miss 328 other rescuable queries; 898 queries are outranked at every weight and 561 lack the target in the shortlist (compare Table 1). The net of 31 rescued queries is

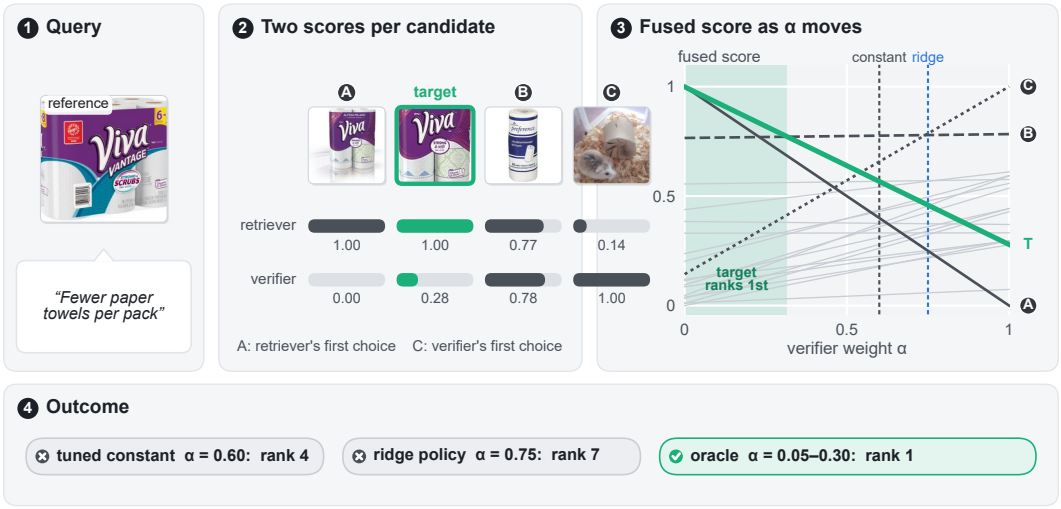

Figure 6: **The query of Figure 1, step by step.** (2) Retriever and verifier scores, min–max normalized within the Top-20 shortlist, for the target and three competitors. (3) Fused score of every shortlist candidate as α varies; gray lines are the other 16 candidates. The target ranks first only in the green interval, whose ends are set by A and B. (4) Rank of the target under each policy.

+0.92 R@1 points for this seed. Averaged over five seeds, the ridge gain for this verifier is +0.60 points (Appendix D).

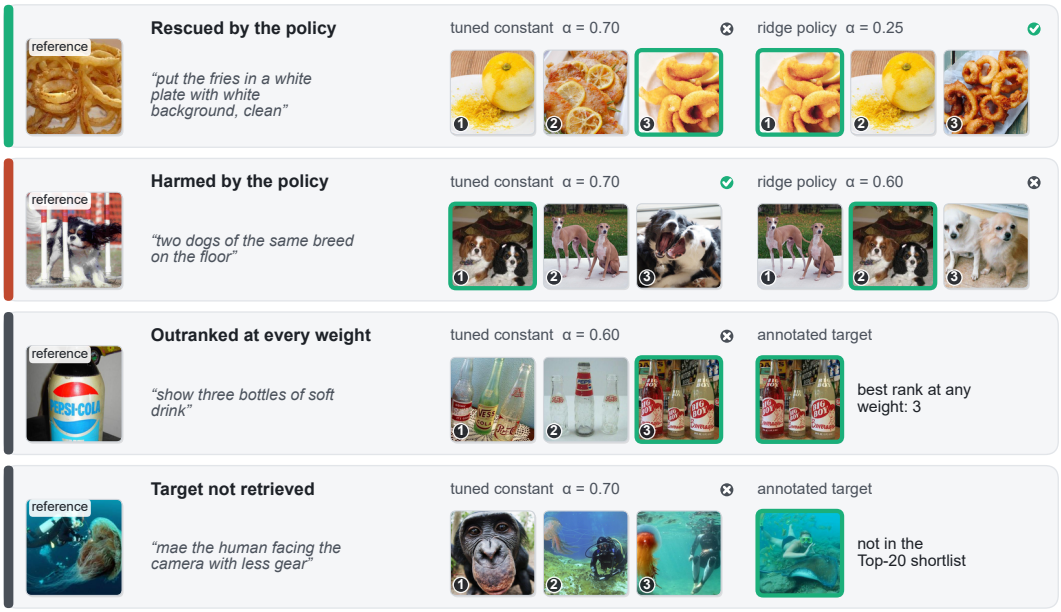

Figure 7: **Four kinds of query** (CIRR, InternVL3-2B, seed 0). Each row shows the reference image, the request and the top three candidates under the tuned constant, and then either under the ridge policy (rows 1–2) or the annotated target (rows 3–4). Green outlines mark the target, and the marks next to each header show whether the target ranks first. Rows are the first query of each type in query-ID order.

B REPRODUCIBLE PROTOCOL AND STATISTICAL ANALYSIS

The main grid contains 21 coefficients from 0 to 1 in steps of .05. Five target-group folds use seeds 0–4; paired cluster resampling uses 10,000 draws and seed 600609. Ridge standardization uses training queries only. The fixed penalty is 10. Nested selection considers X with penalties 1/10/100, XZ with 1/10/100, and Z with 10 using three inner folds. The response subtracts base recall from every action outcome. Test relevance is used only to score saved decisions.

Six CIRR configurations are Qwen3-VL-2B (Bai et al., 2025), InternVL3-2B (Zhu et al., 2025), SmolVLM2-2.2B (Marafioti et al., 2025), LLaVA-OV-0.5B (Li et al., 2024), Gemma-3-4B (Gemma Team, 2025), and Phi-3.5-Vision (whose model card cites Abdin et al., 2024). The panel adds Qwen3-VL-4B and 8B, InternVL3-8B, Idefics3-8B (Laurençon et al., 2024), and LLaVA-OV-7B. FashionIQ includes ten configurations: these eleven except Gemma-3-4B and LLaVA-OV-7B, plus a Qwen3-VL-2B LoRA checkpoint (Hu et al., 2021). Each dataset's analysis population is identical across its models. Family overlap and checkpoint-training differences limit interpretation of model counts.

The first-stage weight was selected on excluded validation pilot target groups, not an independently available full training split. Its frozen values are .15 on CIRR and .25 on FashionIQ. Reference and graph sensitivity change policy-training partitions without repeating this calibration.

Prompt and readout. The first image is a reference. The requested change is: $\{caption\}$. The second image is a candidate result. Does the candidate correctly show the reference with that change applied? Answer with a single word, yes or no. Affirmative and negative sets retain only single-token spellings; empty or overlapping sets are rejected. The score uses the final nonpadding token. Processor overrides, image dimensions, and available software provenance are recorded in the associated data records.

Observable summaries. X contains first-stage top-1 gaps to ranks 2, 3, 6, and the last rank; score standard deviation and mean; maximum score; candidate count; shifted-score entropy; and the top-two gap divided by the full range. Z contains verifier maximum, mean, standard deviation, and shifted-score entropy; top-two and full-range gaps; number tied at the maximum; fraction of distinct scores; base/verifier top-1 disagreement; top-5 overlap; maximum and mean rank displacement; rank correlation; constant-score flag; and the verifier rank of the first-stage top candidate.

Candidate policy. Tables copied unchanged from the frozen analyses call it the candidate-aware policy. Each candidate has 158 features: thirteen score/rank and similarity features, eighty random-projection embedding features, and the complete paired score list plus XZ (65 features). The embedding features include the candidate, differences from reference and text, and elementwise interactions. A fixed Gaussian projection maps 1,024 dimensions to 16 with seed 600609; there is no fitted PCA or label-dependent projection.

Gradient boosting runs 120 iterations with learning rate .06, L2 penalty 10, minimum leaf size 40, and 7 or 15 leaf nodes. An inner group holdout selects capacity, temperature in $\{.5, 1, 2\}$, and shrinkage in $\{0, .25, .5, 1\}$; the selected model is refitted on outer training queries. Predicted candidate probabilities induce expected R@k for each scalar action. Final recall, unshrunk outcomes, and held-out probability Brier scores are retained.

Writing u_j for clipped candidate Bernoulli scores, the candidate policy uses

$$w_j^{(T)} = \frac{\exp\{\mathrm{logit}(u_j)/T\}}{\sum_\ell \exp\{\mathrm{logit}(u_\ell)/T\}}, \qquad \widehat{U}_k(a) = \sum_j w_j^{(T)} \mathbf{1}\{\mathrm{rank}_a(j) \le k\}.$$

Clipping is to $[10^{-6}, 1 - 10^{-6}]$. Shrinkage mixes $\widehat{U}_k$ with mean training utility before maximizing. This model normalizes within the list and has no explicit outside-list class; the binding model of Appendix F adds one. Candidate score ties preserve list order, utility ties choose the smallest action weight, and model-selection ties follow each protocol's declared option order. These are three distinct operations.

B.1 SELECTION AND UNCERTAINTY

The symmetric calibration panel uses identical three-fold target-group inner splits for constants, ridge and the candidate policy. Query min–max, query z-score, training-global z-score and rank normalization are shared choices. Global statistics are fitted on the applicable training subset. Validation pools held-out successes; exact ties favor the constant. The fixed ridge and candidate policies of Sections 2–4, the common-objective factorial and the nested encoding experiment are separate procedures.

Paired bootstrap weights are shared across policies and verifiers; decision events are constructed before averaging seeds and verifiers within queries. The primary panel reports 95% intervals and the nominal two-one-sided equivalence criterion using 90% intervals (Lakens, 2017). Margins 0.5, 1 and 2 points and the smallest satisfied margin are reported together. The independent H2/CIRCO C2 rules use 95% intervals inside one-point bounds. Bootstrap samples with a zero recovery denominator are counted as undefined. Counts of pointwise-positive cells do not receive an independent-discovery interpretation.

Target folds group by target image. Reference folds group by reference identity. Graph folds use deterministic unweighted modularity communities of the physical reference–target graph on CIRR and connected components on FashionIQ; only within-community CIRR queries remain. Their shared retrieval gallery is unchanged. The 20-partition sensitivity uses seeds 0–19. The CIRR retained-population and graph-split changes are reported separately.

Thirty training-group bootstrap refits use seed 710907 on outer seed 0, fold 0 for CIRR/InternVL3-2B and FashionIQ/Qwen3-VL-2B. Duplicated groups retain their identities in inner folds. Each replicate repeats calibration, tuning and fitting, then predicts the same untouched outer fold. The resulting spread measures training-sample sensitivity conditional on that test fold.

B.2 STUDY STATUS AND REPRODUCIBILITY

The main-panel analyses are exploratory. The independent hypotheses were fixed before scoring their new populations. Subsequent factorial and encoding protocols were fixed before their own outcomes, on previously examined populations. These stages retain separate estimands and multiplicity families. Query identities, candidate ordering, data hashes, folds, selected settings and prediction records are retained. The nested study replays 2,399,320 decision outcomes with zero discrepancies, checks every nested group boundary and matches independent ridge solutions. The figures include illustrative benchmark crops; the original image collections remain with their providers.

C EXACT ACCOUNTING AND MATHEMATICAL BOUNDARIES

Feasible intervals and endpoints. For candidate $j \neq t$, put $b_j = d_j^r$ and $u_j = d_j^v - d_j^r$. If $u_j > 0$, the constraint is $a \geq -b_j/u_j$ or $a > -b_j/u_j$ according to the tie order. If $u_j < 0$, it is the corresponding upper bound. If $u_j = 0$, the entire domain is allowed exactly when b_j satisfies the required weak or strict inequality; otherwise feasibility is empty. Intersect these half-lines and $[0, 1]$. At any finite endpoint, all active inequalities must be satisfied, including strict ones. Equal numerical lower and upper bounds define a singleton only if both endpoints are included; otherwise the set is empty. A missing target is infeasible. Under the flat-verifier fallback, the target is either first for all weights or for none; evaluate the retriever order rather than the affine formula.

Continuous versus finite-grid feasibility. Only $G_q = I_q \cap \mathcal{A}$ is relevant to the evaluated actions. A nonempty open interval or an off-grid singleton may miss every grid point. Endpoint inclusion is tested explicitly. Stable sorting reproduces all saved single-target R@1 actions; Table 1 separates missing targets, continuous infeasibility and grid misses.

Counterexample for R@2. A target at index 0 with scores $(.4, .4)$ against $(1, 0)$ and $(0, 1)$ has R@2 feasible set $[0, .4] \cup [.6, 1]$ when ties favor the target. Thus the R@1 single-interval statement does not extend to R@k. An exact construction sorts target–competitor crossings, evaluates every open segment and boundary, and merges accepted adjacent pieces.

Multiple relevant targets. For R@1 with relevant set $\mathcal{T}_q$, the success set is $I_q^{\text{multi}} = \bigcup_{t \in \mathcal{T}_q \cap K} I_{q,t}$, with each component retaining its endpoint rules. Localization is distance to the union's grid intersection, not to its convex hull. Binary hit-at-k permits exact rescue–harm accounting; graded average precision requires its own utility analysis.

Headroom and rescue–harm identities. For any common finite action set and binary utility, $\{Y(c) = 1\} \subseteq \{\max_a Y(a) = 1\}$. Subtracting their probabilities proves Equation 3. Likewise,

$$Y(\pi) - Y(c) = \mathbf{1}\{Y(\pi) = 1, Y(c) = 0\} - \mathbf{1}\{Y(\pi) = 0, Y(c) = 1\}.$$

Taking expectations proves Equation 5. If the harm rate is conditioned on weight-sensitive successes, its multiplier is that smaller population mass, not total constant coverage. Empty conditional events contribute zero mass; recovery is undefined when $H = 0$. The identities are evaluated under each fold's training-fitted constant before averaging.

Information-conditioned value. Condition on training sample T, and let F denote new-query information, $Y(a)$ bounded measurable utility and $\mathcal{A}$ finite. A fixed tie order gives a measurable maximizer of the conditional means. Every F-measurable policy has value at most $V(F)$, while

$$\max_a \mathbb{E}[Y(a) \mid F] \leq \mathbb{E}[\max_a Y(a) \mid F].$$

Taking expectations proves $V(F) \leq O$; adding and subtracting $V(F)$ gives Equation 4. A measurable partition $B(F)$ restricts available policies, hence $V(B(F)) \leq V(F)$. For a training-fitted deterministic encoder g, $\sigma(F, g(F)) = \sigma(F)$ and therefore $V(F, g(F)) = V(F)$. The equality excludes encoders that read the new target label or extra inputs. A finite-sample lookup or partition optimum does not bound unrestricted population learning.

Calibration scope. Scalar intervals fix normalization. To compare policies choosing different normalizations, either synchronize their calibration or use the joint action $(\text{normalization}, a)$. A scalar distance between unlike schemes is not a common localization metric.

Table 1: Disjoint sources of finite-grid failure in every original configuration. Entries are query counts: target absent (A), present but continuously infeasible (I), continuous interval missed by the grid (M), and grid reachable (R). Their sum is the query count. These are descriptive counts, not independent replications.

Data	Verifier	n	A	I	M	R
CIRR	Gemma-3-4B	3380	561	968	9	1842
CIRR	Idefics3-8B	3380	561	875	10	1934
CIRR	InternVL3-2B	3380	561	891	7	1921
CIRR	InternVL3-8B	3380	561	713	4	2102
CIRR	LLaVA-OV-0.5B	3380	561	1007	2	1810
CIRR	LLaVA-OV-7B	3380	561	807	5	2007
CIRR	Phi-3.5-Vision	3380	561	883	7	1929
CIRR	Qwen3-VL-2B	3380	561	836	4	1979
CIRR	Qwen3-VL-4B	3380	561	732	5	2082
CIRR	Qwen3-VL-8B	3380	561	752	4	2063
CIRR	SmolVLM2-2.2B	3380	561	879	6	1934
FIQ	Idefics3-8B	4539	2280	948	2	1309
FIQ	InternVL3-2B	4539	2280	939	6	1314
FIQ	InternVL3-8B	4539	2280	877	6	1376
FIQ	LLaVA-OV-0.5B	4539	2280	1027	8	1224
FIQ	Phi-3.5-Vision	4539	2280	957	6	1296
FIQ	Qwen3-VL-2B	4539	2280	896	3	1360
FIQ	Qwen3-VL-4B	4539	2280	864	9	1386
FIQ	Qwen3-VL-8B	4539	2280	929	1	1329
FIQ	Qwen3-VL-2B LoRA	4539	2280	915	5	1339
FIQ	SmolVLM2-2.2B	4539	2280	992	8	1259

D COMPLETE RECOVERY RESULTS AND STABILITY

Figure 8 shows every cell of the matched-calibration comparison of Section 4 and Table 2 summarizes it; Figure 9 shows absolute R@1 for the two verifiers of the controlled studies. The following tables list every verifier–metric cell of the primary comparison and of the matched-calibration comparison. Gain intervals are 95%; equivalence columns use 90% intervals. Cells share queries, so they are not independent replications.

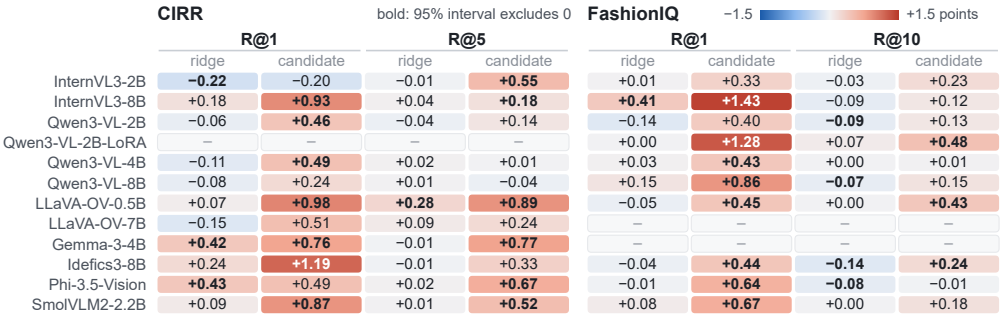

Figure 8: **The candidate policy's gains survive matched calibration.** Gain over the selected constant (points) for every verifier configuration and metric. Both policies choose among the same normalizations. Bold entries have paired 95% intervals that exclude zero.

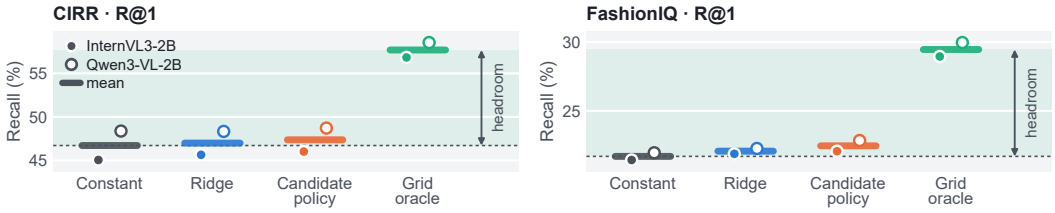

Figure 9: **Large headroom, small realized gains.** R@1 of the tuned constant, the two learned policies of Section 2 (ridge and candidate policies) and the per-query oracle. All four choose among the same 21 weights on query min–max normalized scores. Dots are individual verifiers (InternVL3-2B, Qwen3-VL-2B), bars are their means, and arrows mark oracle headroom.

Table 2: Matched calibration, averaged over verifier configurations. C: R@k (%) of the constant with min–max or selected normalization. Δ: mean gain over the selected constant (points). Above/below: cells whose candidate-minus-constant 95% interval lies above/below zero. Equiv.: cells whose 90% interval lies within ± 1 point (Lakens, 2017).

Cell group	Cells	Min–max C	Selected C	Ridge Δ	Candidate Δ	Above/below	Equiv.
CIRR R@1	11	46.80	47.08	+0.07	+0.61	7/0	5
CIRR R@5	11	72.81	72.90	+0.04	+0.39	6/0	9
FashionIQ R@1	10	21.68	21.76	+0.04	+0.69	8/0	7
FashionIQ R@10	10	45.36	45.39	−0.04	+0.20	3/0	10

Table 3: CIRR R@1: complete primary panel. Eq. lists accepted margins.

Model	Base	Full	Const.	Δ XZ [95%]	Δ nested [95%]	Eq.
Qwen3-VL-2B	36.72	45.00	48.37	-0.04 [-0.46, +0.38]	-0.12 [-0.50, +0.26]	0.5,1.0,2.0
InternVL3-2B	36.72	38.79	45.05	+0.60 [+0.00, +1.18]	+0.49 [-0.08, +1.05]	2.0
SmolVLM2-2.2B	36.72	38.31	46.36	+0.39 [-0.17, +0.95]	+0.22 [-0.28, +0.73]	1.0,2.0
LLaVA-OV-0.5B	36.72	28.20	41.57	-0.25 [-0.87, +0.36]	-0.32 [-0.89, +0.25]	1.0,2.0
Gemma-3-4B	36.72	34.14	42.28	+0.51 [-0.05, +1.09]	+0.44 [-0.07, +0.99]	1.0,2.0
Phi-3.5-Vision	36.72	38.82	44.97	+0.24 [-0.46, +0.94]	+0.14 [-0.52, +0.81]	1.0,2.0
Qwen3-VL-4B	36.72	48.82	50.89	+0.41 [-0.12, +0.95]	+0.49 [+0.01, +0.98]	1.0,2.0
Qwen3-VL-8B	36.72	46.80	49.73	+0.59 [+0.05, +1.13]	+0.60 [+0.11, +1.12]	2.0
InternVL3-8B	36.72	49.11	51.53	+0.29 [-0.23, +0.82]	-0.03 [-0.48, +0.43]	1.0,2.0
Idefics3-8B	36.72	36.72	45.31	+0.44 [-0.16, +1.04]	+0.27 [-0.30, +0.83]	1.0,2.0
LLaVA-OV-7B	36.72	43.25	48.72	+0.32 [-0.29, +0.95]	+0.17 [-0.37, +0.72]	1.0,2.0

Table 4: CIRR R@5: complete primary panel. Eq. lists accepted margins.

Model	Base	Full	Const.	Δ XZ [95%]	Δ nested [95%]	Eq.
Qwen3-VL-2B	64.11	72.28	73.32	+0.08 [-0.32, +0.48]	+0.32 [-0.04, +0.70]	0.5,1.0,2.0
InternVL3-2B	64.11	66.89	71.75	+0.03 [-0.45, +0.53]	-0.11 [-0.49, +0.29]	0.5,1.0,2.0
SmolVLM2-2.2B	64.11	67.40	71.84	-0.20 [-0.67, +0.28]	-0.16 [-0.57, +0.25]	1.0,2.0
LLaVA-OV-0.5B	64.11	58.11	67.76	+0.03 [-0.50, +0.57]	+0.01 [-0.45, +0.48]	0.5,1.0,2.0
Gemma-3-4B	64.11	66.15	70.62	-0.38 [-0.85, +0.09]	-0.51 [-0.90, -0.12]	1.0,2.0
Phi-3.5-Vision	64.11	68.25	72.30	-0.41 [-0.87, +0.05]	+0.08 [-0.34, +0.51]	1.0,2.0
Qwen3-VL-4B	64.11	74.62	75.19	+0.09 [-0.26, +0.47]	+0.08 [-0.27, +0.44]	0.5,1.0,2.0
Qwen3-VL-8B	64.11	73.22	75.68	-0.29 [-0.66, +0.08]	-0.27 [-0.56, +0.02]	1.0,2.0
InternVL3-8B	64.11	75.44	76.45	-0.17 [-0.48, +0.14]	-0.01 [-0.27, +0.24]	0.5,1.0,2.0
Idefics3-8B	64.11	66.80	72.36	-0.17 [-0.61, +0.26]	-0.10 [-0.42, +0.20]	1.0,2.0
LLaVA-OV-7B	64.11	70.47	73.65	-0.15 [-0.64, +0.33]	-0.04 [-0.44, +0.36]	1.0,2.0

Table 5: FIQ R@1: complete primary panel. Eq. lists accepted margins.

Model	Base	Full	Const.	Δ XZ [95%]	Δ nested [95%]	Eq.
Qwen3-VL-2B	19.01	18.20	21.96	+0.31 [-0.14, +0.76]	+0.32 [-0.12, +0.75]	1.0,2.0
InternVL3-2B	19.01	15.64	21.44	+0.45 [+0.01, +0.89]	+0.30 [-0.08, +0.68]	1.0,2.0
SmolVLM2-2.2B	19.01	13.06	21.08	-0.17 [-0.59, +0.25]	-0.32 [-0.67, +0.04]	1.0,2.0
LLaVA-OV-0.5B	19.01	9.83	20.18	-0.13 [-0.55, +0.27]	-0.21 [-0.54, +0.11]	0.5,1.0,2.0
Phi-3.5-Vision	19.01	14.17	21.28	-0.24 [-0.68, +0.18]	-0.45 [-0.82, -0.11]	1.0,2.0
Qwen3-VL-4B	19.01	18.35	23.04	+0.12 [-0.33, +0.56]	+0.02 [-0.37, +0.41]	0.5,1.0,2.0
Qwen3-VL-8B	19.01	17.05	22.25	+0.04 [-0.38, +0.46]	+0.03 [-0.32, +0.39]	0.5,1.0,2.0
InternVL3-8B	19.01	17.78	22.56	+0.24 [-0.25, +0.72]	+0.15 [-0.31, +0.62]	1.0,2.0
Idefics3-8B	19.01	13.84	21.27	+0.36 [-0.08, +0.79]	+0.33 [-0.11, +0.76]	1.0,2.0
Qwen3-VL-2B-LoRA	19.01	13.59	21.77	-0.05 [-0.42, +0.33]	-0.11 [-0.45, +0.24]	0.5,1.0,2.0

Table 6: FIQ R@10: complete primary panel. Eq. lists accepted margins.

Model	Base	Full	Const.	Δ XZ [95%]	Δ nested [95%]	Eq.
Qwen3-VL-2B	41.88	44.75	46.22	-0.27 [-0.54, -0.02]	+0.01 [-0.20, +0.22]	0.5,1.0,2.0
InternVL3-2B	41.88	42.76	45.40	-0.25 [-0.55, +0.04]	-0.23 [-0.43, -0.04]	0.5,1.0,2.0
SmolVLM2-2.2B	41.88	41.29	44.85	-0.44 [-0.74, -0.15]	-0.11 [-0.26, +0.03]	1.0,2.0
LLaVA-OV-0.5B	41.88	39.68	43.81	-0.03 [-0.38, +0.31]	-0.07 [-0.32, +0.18]	0.5,1.0,2.0
Phi-3.5-Vision	41.88	43.09	45.23	-0.37 [-0.63, -0.09]	-0.16 [-0.35, +0.03]	1.0,2.0
Qwen3-VL-4B	41.88	44.44	45.90	-0.17 [-0.41, +0.07]	-0.11 [-0.29, +0.05]	0.5,1.0,2.0
Qwen3-VL-8B	41.88	44.81	45.99	-0.05 [-0.30, +0.19]	-0.07 [-0.26, +0.12]	0.5,1.0,2.0
InternVL3-8B	41.88	44.57	45.90	-0.06 [-0.30, +0.18]	-0.10 [-0.29, +0.09]	0.5,1.0,2.0
Idefics3-8B	41.88	42.08	44.53	-0.22 [-0.53, +0.10]	-0.08 [-0.28, +0.11]	0.5,1.0,2.0
Qwen3-VL-2B-LoRA	41.88	41.93	45.78	-0.07 [-0.33, +0.18]	+0.03 [-0.12, +0.18]	0.5,1.0,2.0

Table 7: Symmetric calibration for every cell. C is the recall of the inner-fold-calibrated constant; differences are relative to it, with paired 95% group intervals.

Data	Verifier	k	C	Ridge − C	Candidate − C
CIRR	Qwen3-VL-2B	1	48.30	-0.06 [-0.27, $+0.15$]	$+0.46$ [$+0.01$, $+0.91$]
CIRR	Qwen3-VL-2B	5	73.54	-0.04 [-0.12, $+0.05$]	$+0.14$ [-0.11, $+0.38$]
CIRR	InternVL3-2B	1	46.56	-0.22 [-0.39, -0.06]	-0.20 [-0.42, $+0.03$]
CIRR	InternVL3-2B	5	71.87	-0.01 [-0.22, $+0.19$]	$+0.55$ [$+0.16$, $+0.95$]
CIRR	SmolVLM2-2.2B	1	46.62	$+0.09$ [-0.24, $+0.41$]	$+0.87$ [$+0.35$, $+1.39$]
CIRR	SmolVLM2-2.2B	5	72.07	$+0.01$ [-0.15, $+0.17$]	$+0.52$ [$+0.11$, $+0.94$]
CIRR	LLaVA-OV-0.5B	1	41.21	$+0.07$ [-0.26, $+0.40$]	$+0.98$ [$+0.31$, $+1.63$]
CIRR	LLaVA-OV-0.5B	5	68.33	$+0.28$ [$+0.01$, $+0.56$]	$+0.89$ [$+0.46$, $+1.33$]
CIRR	Gemma-3-4B	1	42.74	$+0.42$ [$+0.06$, $+0.79$]	$+0.76$ [$+0.23$, $+1.29$]
CIRR	Gemma-3-4B	5	70.46	-0.01 [-0.30, $+0.28$]	$+0.77$ [$+0.37$, $+1.16$]
CIRR	Phi-3.5-Vision	1	45.11	$+0.43$ [$+0.01$, $+0.86$]	$+0.49$ [-0.08, $+1.05$]
CIRR	Phi-3.5-Vision	5	72.14	$+0.02$ [-0.19, $+0.24$]	$+0.67$ [$+0.28$, $+1.06$]
CIRR	Qwen3-VL-4B	1	51.16	-0.11 [-0.41, $+0.18$]	$+0.49$ [$+0.08$, $+0.91$]
CIRR	Qwen3-VL-4B	5	75.57	$+0.02$ [-0.12, $+0.17$]	$+0.01$ [-0.19, $+0.20$]
CIRR	Qwen3-VL-8B	1	50.84	-0.08 [-0.26, $+0.11$]	$+0.24$ [-0.23, $+0.72$]
CIRR	Qwen3-VL-8B	5	75.61	$+0.01$ [-0.19, $+0.22$]	-0.04 [-0.27, $+0.18$]
CIRR	InternVL3-8B	1	51.26	$+0.18$ [-0.10, $+0.45$]	$+0.93$ [$+0.43$, $+1.43$]
CIRR	InternVL3-8B	5	76.29	$+0.04$ [-0.11, $+0.19$]	$+0.18$ [$+0.02$, $+0.35$]
CIRR	Idefics3-8B	1	45.14	$+0.24$ [-0.16, $+0.63$]	$+1.19$ [$+0.52$, $+1.84$]
CIRR	Idefics3-8B	5	72.07	-0.01 [-0.18, $+0.16$]	$+0.33$ [-0.05, $+0.71$]
CIRR	LLaVA-OV-7B	1	48.89	-0.15 [-0.50, $+0.21$]	$+0.51$ [-0.08, $+1.12$]
CIRR	LLaVA-OV-7B	5	73.99	$+0.09$ [-0.19, $+0.38$]	$+0.24$ [-0.16, $+0.65$]
FashionIQ	Qwen3-VL-2B	1	22.56	-0.14 [-0.30, $+0.02$]	$+0.40$ [-0.04, $+0.82$]
FashionIQ	Qwen3-VL-2B	10	46.19	-0.09 [-0.17, -0.01]	$+0.13$ [-0.09, $+0.35$]
FashionIQ	InternVL3-2B	1	21.76	$+0.01$ [-0.22, $+0.24$]	$+0.33$ [-0.03, $+0.67$]
FashionIQ	InternVL3-2B	10	45.27	-0.03 [-0.11, $+0.05$]	$+0.23$ [-0.03, $+0.49$]
FashionIQ	SmolVLM2-2.2B	1	20.86	$+0.08$ [-0.10, $+0.26$]	$+0.67$ [$+0.30$, $+1.03$]
FashionIQ	SmolVLM2-2.2B	10	44.88	$+0.00$ [-0.05, $+0.05$]	$+0.18$ [-0.03, $+0.40$]
FashionIQ	LLaVA-OV-0.5B	1	20.25	-0.05 [-0.18, $+0.07$]	$+0.45$ [$+0.04$, $+0.84$]
FashionIQ	LLaVA-OV-0.5B	10	43.76	$+0.00$ [-0.16, $+0.17$]	$+0.43$ [$+0.18$, $+0.68$]
FashionIQ	Phi-3.5-Vision	1	21.13	-0.01 [-0.08, $+0.06$]	$+0.64$ [$+0.23$, $+1.04$]
FashionIQ	Phi-3.5-Vision	10	45.27	-0.08 [-0.14, -0.04]	-0.01 [-0.19, $+0.17$]
FashionIQ	Qwen3-VL-4B	1	23.03	$+0.03$ [-0.16, $+0.22$]	$+0.43$ [$+0.10$, $+0.76$]
FashionIQ	Qwen3-VL-4B	10	46.00	$+0.00$ [-0.02, $+0.03$]	$+0.01$ [-0.13, $+0.15$]
FashionIQ	Qwen3-VL-8B	1	22.00	$+0.15$ [-0.11, $+0.40$]	$+0.86$ [$+0.47$, $+1.25$]
FashionIQ	Qwen3-VL-8B	10	46.06	-0.07 [-0.13, -0.02]	$+0.15$ [-0.03, $+0.34$]
FashionIQ	InternVL3-8B	1	22.30	$+0.41$ [$+0.07$, $+0.75$]	$+1.43$ [$+1.01$, $+1.86$]
FashionIQ	InternVL3-8B	10	45.99	-0.09 [-0.19, $+0.01$]	$+0.12$ [-0.03, $+0.26$]
FashionIQ	Idefics3-8B	1	21.71	-0.04 [-0.23, $+0.14$]	$+0.44$ [$+0.06$, $+0.82$]
FashionIQ	Idefics3-8B	10	44.80	-0.14 [-0.22, -0.05]	$+0.24$ [$+0.00$, $+0.48$]
FashionIQ	Qwen3-VL-2B-LoRA	1	21.96	$+0.00$ [-0.20, $+0.19$]	$+1.28$ [$+0.79$, $+1.76$]
FashionIQ	Qwen3-VL-2B-LoRA	10	45.68	$+0.07$ [-0.06, $+0.20$]	$+0.48$ [$+0.25$, $+0.73$]

D.1 Partition and training-sample sensitivity

Twenty partitions give per-partition SDs of 0.09–0.33 points; 13/16 cell–split series retain one sign in at least 18 partitions. Restricting CIRR to the graph-retained population changes five-partition effects by -0.28 to $+0.12$, and changing the split on that population by $+0.05$ to $+0.57$. These separate a population change from a partition change. The 30-refit experiment gives candidate-minus-constant SDs of 0.52 and 0.36 points, with 7/30 and 13/30 positive refits. These spreads are not population confidence intervals.

Table 8: Candidate-aware policy minus the calibrated constant across 20 outer partitions per cell and split. Mean, SD, range, and the number of positive partitions are over partitions; the pooled column averages outcomes over all 20 partitions before the paired 95% group interval.

Data	Verifier	k	Split	Mean	SD	Range	> 0	Pooled [95%]
CIRR	InternVL3-2B	1	target	-0.28	0.23	$[-0.62, +0.03]$	1	-0.28 $[-0.46, -0.10]$
CIRR	InternVL3-2B	1	graph	-0.31	0.21	$[-0.67, +0.03]$	1	-0.31 $[-0.68, +0.04]$
CIRR	InternVL3-2B	5	target	$+0.42$	0.23	$[+0.06, +0.86]$	20	$+0.42$ $[+0.05, +0.79]$
CIRR	InternVL3-2B	5	graph	$+0.66$	0.19	$[+0.31, +0.98]$	20	$+0.66$ $[+0.27, +1.11]$
CIRR	Qwen3-VL-2B	1	target	$+0.60$	0.24	$[+0.27, +0.98]$	20	$+0.60$ $[+0.18, +1.01]$
CIRR	Qwen3-VL-2B	1	graph	$+0.78$	0.33	$[+0.15, +1.34]$	20	$+0.78$ $[+0.25, +1.31]$
CIRR	Qwen3-VL-2B	5	target	$+0.06$	0.16	$[-0.30, +0.36]$	12	$+0.06$ $[-0.13, +0.25]$
CIRR	Qwen3-VL-2B	5	graph	$+0.13$	0.19	$[-0.34, +0.55]$	15	$+0.13$ $[-0.06, +0.33]$
FashionIQ	InternVL3-2B	1	target	$+0.42$	0.22	$[-0.09, +0.84]$	18	$+0.42$ $[+0.09, +0.74]$
FashionIQ	InternVL3-2B	1	graph	$+0.55$	0.26	$[-0.02, +0.93]$	19	$+0.55$ $[+0.21, +0.88]$
FashionIQ	InternVL3-2B	10	target	$+0.21$	0.11	$[-0.02, +0.42]$	19	$+0.21$ $[-0.03, +0.45]$
FashionIQ	InternVL3-2B	10	graph	$+0.25$	0.13	$[+0.07, +0.57]$	20	$+0.25$ $[+0.01, +0.49]$
FashionIQ	Qwen3-VL-2B	1	target	$+0.43$	0.18	$[+0.09, +0.79]$	20	$+0.43$ $[+0.05, +0.81]$
FashionIQ	Qwen3-VL-2B	1	graph	$+0.43$	0.22	$[-0.02, +0.88]$	19	$+0.43$ $[+0.07, +0.79]$
FashionIQ	Qwen3-VL-2B	10	target	$+0.19$	0.14	$[-0.02, +0.59]$	19	$+0.19$ $[+0.01, +0.38]$
FashionIQ	Qwen3-VL-2B	10	graph	$+0.24$	0.09	$[+0.04, +0.40]$	20	$+0.24$ $[+0.05, +0.43]$

Table 9: Training-group bootstrap with complete nested re-selection and refitting. All replicates share one fixed outer-test fold. Spread is the empirical 5th–95th percentile across refits, not a population confidence interval.

Dataset/model	Family	Test n	Mean Δ	SD	90% spread	Positive
CIRR/InternVL3-2B	ridge	648	-0.28	0.43	[-0.94, +0.24]	3/30
CIRR/InternVL3-2B	candidate	648	-0.04	0.52	[-1.25, +0.70]	7/30
FIQ/Qwen3-VL-2B	ridge	895	-0.06	0.48	[-0.84, +0.75]	7/30
FIQ/Qwen3-VL-2B	candidate	895	+0.04	0.36	[-0.45, +0.67]	13/30

D.2 Policy-specific permutation references

Ridge uses 20 within-fold permutations with seed 20260915. The candidate policy uses the exact expectation over permutations of its own saved actions, crosschecked with 1,000 draws, retaining five seeds and two verifiers. The table preserves both favorable and unfavorable action–query associations. Harm conditions on weight-sensitive constant successes.

Table 10: Candidate policy (Section 2) against its own within-fold permutation reference. Rescue uses the original continuous-distance bins and weight-sensitive queries; denominators sum two verifiers after averaging five seeds. Harm conditions on weight-sensitive constant successes. Rates are percent. Nulls are exact expectations under all within-fold permutations, crosschecked with 1,000 draws, and are descriptive references.

Data	Distance/event	Rescuable	Candidate	Own null
CIRR	0.0–0.1	260.8	42.41	7.76
CIRR	0.1–0.2	178.6	17.13	4.20
CIRR	0.2–0.4	215.6	6.68	3.53
CIRR	0.4–1.0	87.4	0.46	6.85
CIRR	Harm (sensitive)	—	7.58	6.27
FIQ	0.0–0.1	224.4	48.31	11.54
FIQ	0.1–0.2	184.2	26.93	6.92
FIQ	0.2–0.4	237.4	15.42	3.51
FIQ	0.4–1.0	58.2	5.15	3.56
FIQ	Harm (sensitive)	—	11.57	6.74

E A DIRECT EXPECTED-RECALL RULE ON EVERY CONFIGURATION

The nested study found that choosing the weight with the highest expected R@1 under the target posterior beat the tuned decoders in all four panels (Table 25). We extended this rule to all 21 panel configurations with a protocol fixed before any of its outcomes. The rule uses the candidate policy's gradient-boosted posterior at its smallest capacity (seven leaves), temperature 1, query min–max scores, and no shrinkage, fallback or inner selection. Features, outer folds, seeds and statistics are the panel's own, and the comparators are the saved matched-calibration constant and tuned candidate policy on identical test queries.

Table 11 pools the configurations within query, and Figure 10 shows every configuration. The direct rule gains +0.52 [+0.23, +0.80] R@1 points on CIRR and +0.65 [+0.45, +0.84] on FashionIQ, with intervals above zero in 7/11 and 7/10 configurations. Its pooled difference from the tuned policy is within 0.1 point in all four cell groups. The tuning of normalization, capacity, temperature and shrinkage therefore contributes little. Most of the candidate policy's gain comes from acting on the posterior itself.

Table 11: Direct expected-recall rule (candidate posterior, seven leaves, temperature 1, no shrinkage, no fallback, no tuning) against the matched constant and the tuned candidate policy. Estimates pool all configurations within query; above/below counts cells whose 95% interval excludes zero.

Cells	Configs	Tuned − constant	Direct − constant	Above/below	Direct − tuned	Above/below
CIRR R@1	11	+0.61 [+0.40, +0.81]	+0.52 [+0.23, +0.80]	7/0	−0.09 [−0.21, +0.03]	0/1
CIRR R@5	11	+0.39 [+0.26, +0.52]	+0.34 [+0.18, +0.50]	3/0	−0.05 [−0.13, +0.03]	0/0
FashionIQ R@1	10	+0.69 [+0.51, +0.87]	+0.65 [+0.45, +0.84]	7/0	−0.05 [−0.11, +0.02]	0/2
FashionIQ R@10	10	+0.20 [+0.11, +0.28]	+0.25 [+0.15, +0.36]	5/0	+0.05 [+0.01, +0.09]	2/0

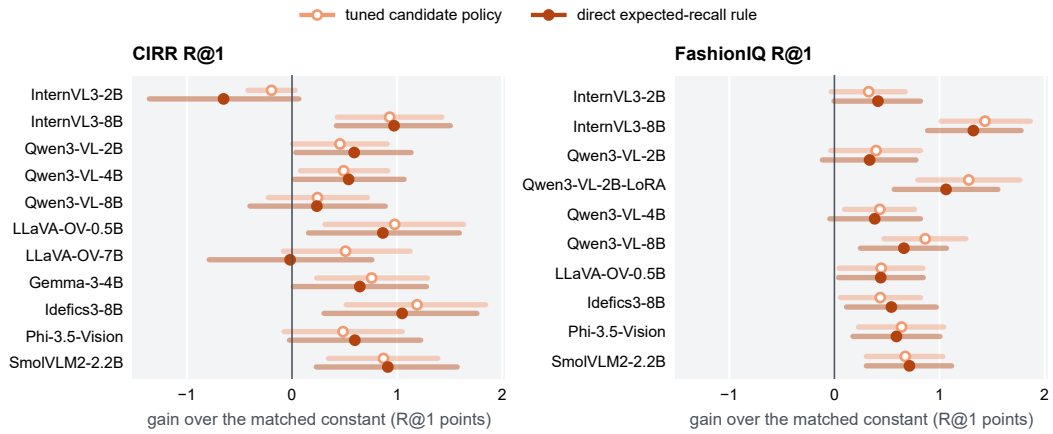

Figure 10: **Tuning adds little to the candidate posterior.** R@1 gain over the matched constant for the tuned candidate policy (hollow) and the direct expected-recall rule (filled), with paired 95% intervals, for every verifier configuration.

F CONTROLLED OBJECTIVES, ENCODINGS AND SELECTION

The experiments in Section 5 are exploratory analyses of identified score and feature records. The factorial and binding specifications were fixed before inspecting their held-out outcomes. The forced-active analysis was added after observing constant fallback and is explicitly a sensitivity analysis. These studies are distinct from the original confirmation hypotheses and their decision rules.

Populations and observable inputs. The original CIRR and FashionIQ cohorts contain 3,380 and 4,539 queries. Each is evaluated with InternVL3-2B and Qwen3-VL-2B; these four cells were selected for aligned feature records and available disjoint-population data, before inspecting controlled-study performance. Candidate features contain 13 scalar and 80 projected interaction coordinates per

candidate, plus 65 shared context coordinates. Flattening the 93 candidate-specific coordinates and retaining the context once gives $20 \times 93 + 65 = 1,925$ coordinates; summary input contains the original 25 XZ coordinates. The shared context contains these summaries.

The flattened arm uses fixed retriever-rank slots rather than shared candidate parameters. Candidate-level rows from one query are correlated, so twenty rows are not twenty independent training examples. Equal penalty grids do not guarantee equally adequate regularization; the extended-search sensitivity in Appendix F.1 addresses this limitation. Feature identity, query order, candidate order, finite values, and input hashes are checked. No retriever or verifier is refit or rescored. All queries, including targets outside the Top-20 list, enter final recall evaluation.

Factorial objective, calibration and selection. For each representation, ridge fits 21 absolute action utilities using training-only standardization and penalty in $\{1, 10, 100\}$. GBM fits the same squared-error objective using uniformly expanded query–action rows, with the action weight as an additional known coordinate. One action-conditioned ensemble shares trees across the 21 outputs; its leaf budgets are $\{7, 15, 31\}$, with 60 iterations, learning rate 0.06, no early stopping, a minimum leaf of 840 expanded rows (at least 40 distinct queries), and regularization 210. This implementation and budget were fixed after training-only timing, before outcome inspection.

Three inner target-group folds select among query min–max, query z-score, training-global z-score, and rank normalization using only the constant's held-out recall. The resulting scheme is shared by every factorial arm. Global normalization parameters are fit on the relevant training rows. Each arm has six candidates: three capacities times shrinkage $\eta \in \{0.5, 1\}$, with predictions $(1-\eta)\bar{Y} + \eta\hat{Y}$, plus the common constant. Inner success counts are pooled across queries; ties prefer the constant, then the fixed capacity and shrinkage order. Selected models refit on the entire outer training set. Five outer target-group folds at seeds 0–4 give 25 fits per cell. The four cells share their query partitions and action space. Equal selection opportunities do not imply equal effective degrees of freedom, computation, or approximation power.

Shrinkage is post-processing of a fitted estimator, not another fit: the original 100 tasks use $4 \times 3 \times 3 \times 100 = 3,600$ inner utility-estimator fits. Global action-utility ties choose the smallest weight, distinct from model-selection and candidate-ranking ties.

The re-evaluated candidate-posterior comparator retains its candidate-level classification objective, 120 boosting iterations, leaf budgets 7/15, temperatures 0.5/1/2, and nonzero shrinkages 0.5/1, plus the constant. It shares the selected calibration but has 13 choices, versus seven in each factorial cell. This comparator and the candidate-policy results of Section 4 are separate from the common-objective GBM cell.

Factorial estimands and fallback sensitivity. Writing M_1, M_2, M_3, M_4 for summary ridge, summary GBM, candidate ridge, and candidate GBM query utilities, respectively, the contrasts are

$$E_{\text{rep}} = \tfrac{1}{2}\{(M_3 - M_1) + (M_4 - M_2)\}, \tag{6}$$

$$E_{\text{learner}} = \tfrac{1}{2}\{(M_2 - M_1) + (M_4 - M_3)\}, \tag{7}$$

$$E_{\text{interaction}} = (M_4 - M_3) - (M_2 - M_1). \tag{8}$$

We also report the representation effect separately under each learner. Candidate ridge selects the constant in 50/50 outer tasks in each dataset; candidate GBM does so in 46/50 CIRR and 48/50 FashionIQ tasks. Identical selected actions give a degenerate conditional contrast.

The forced-active sensitivity chooses the best nonconstant setting using each fold's already saved inner scores. If the original selected setting is adaptive, its decisions are reused; otherwise that training-selected setting is refit on the same outer training data. No outer-test outcome selects a capacity, normalization, or shrinkage. Excluding the constant need not make the predicted action vary across queries, so action variability is reported alongside fallback counts. This diagnostic is post-hoc and cannot replace the primary selected-procedure comparison.

Under forced ridge, replacing summary with candidate input changes R@1 by $-1.36\,[-1.84, -0.88]$ on CIRR and $-1.11\,[-1.48, -0.74]$ on FashionIQ. The corresponding frozen-transfer effects are $-1.24\,[-1.90, -0.56]$ and $-0.83\,[-1.30, -0.35]$. The positive interaction contrasts reflect this ridge penalty alongside the largely constant decisions of the tested GBM; they do not establish that the

GBM learns useful query-specific rescue. Tables 15 and 16 report the contrasts and actual action variation.

Training-only binding predictions. The binding study uses the same four cells and five outer target-group folds at seed 0, with query min–max normalization fixed for all arms. Candidate-level classifiers predict target membership and membership in the lower and upper active competitor sets. Query-level classifiers predict target absence and the two domain-bound sentinels. All six auxiliary classifiers use 60 boosting iterations, seven leaves, learning rate 0.06, minimum leaf 40, regularization 10, and no early stopping. Candidate tasks use the aligned 158-dimensional per-candidate features; query tasks use the 25 summaries. Binding supervision uses the true training target and omits outside-list rows, retaining conflicting lower and upper constraints on infeasible queries. Target and final utility fitting retain outside-list rows. Numerical active-set comparisons use relative/absolute tolerances $10^{-10}/10^{-12}$; domain-sentinel comparisons use $10^{-5}/10^{-8}$. Final action evaluation uses the original stable score tie rule without these supervision tolerances.

Three target-group folds inside each outer training set produce out-of-fold auxiliary features for every downstream training row. The final auxiliary models refit on all outer training rows and predict outer-test features. The downstream ridge has fixed penalty 10 and no shrinkage or model selection, so there is no inner hyperparameter-validation layer into which auxiliary labels could leak. All arms receive the same candidate features and the same cross-fitted target posterior. The candidate-only control zeros the posterior block. Candidate Bernoulli scores are normalized within the list and scaled by the predicted probability of target presence; a separate outside-list channel makes 21 posterior coordinates.

Writing P_q for training target presence, each binding classifier minimizes binary log loss only on rows with $P_q = 1$; its outside-list loss is masked to zero. Target and outside-list classifiers retain all rows. At prediction time there is no outside-probability threshold or hard constant fallback: the candidate posterior is continuously scaled by $1 - \hat{p}_{\mathrm{out}}$, then passed to the fixed decoder. Even if this mass is zero, raw candidate features remain; utility ties still choose the smallest weight.

Encodings and controls. Let $\hat{p}_t$ denote the posterior on candidate t and $\hat{b}_{s,j}$ normalized predicted membership scores for side $s \in \{L, U\}$, including a domain sentinel ∂. These binding scores are marginal predictions, not a target-conditioned joint posterior. If $W_a(t, j)$ records whether hypothetical target t beats j at action a under the stable tie rule, the learned encoding is

$$e_a^{\mathrm{bind}} = \sum_{t=1}^{20} \hat{p}_t \prod_{s \in \{L,U\}} \frac{\hat{b}_{s,\partial} + \sum_{j \neq t} \hat{b}_{s,j} W_a(t, j)}{\max\{1 - \hat{b}_{s,t}, 10^{-12}\}}.$$

This is a fixed feature transformation, not a calibrated probability model for the joint constraints. The exact-hypothesis control instead supplies

$$e_a^{\mathrm{all}} = \sum_{t=1}^{20} \hat{p}_t \, \mathbf{1}\{t = \arg\max_j s_j(a)\},$$

enumerating all possible targets from observed scores without indexing the true test target. Flat-verifier fallback follows the original retriever ordering in both transformations.

Three controls replace predicted constraints with two nonself competitors for every hypothesized target: a query-seeded random pair, the two highest verifier scores, or the two nearest absolute fused-score margins at weight 0.5. A fourth control supplies a fixed label-free random projection of candidate features. Each active encoding has 21 coordinates. Every decoder has 1,925 candidate, 21 posterior, and 21 encoding slots; unused blocks are zero. Thus active negative controls match the added feature dimension, whereas zero blocks in the shared-posterior baseline add no effective capacity. The common standardizer and ridge predict the same 21 utilities. The true evaluation target and successful-action set are used only after decisions are saved.

Frozen external transfer. After fitting and selecting on each complete original population, we freeze the decoder, auxiliary models, and calibration and evaluate on 3,274 CIRR and 4,539 FashionIQ queries from the previously studied independent populations. Their combined reference, candidate, and target image sets have zero overlap with the corresponding original population in all four

configurations. External labels enter only evaluation. This is a disjoint-population, frozen-model transfer test. Because these populations were analysed before, it is not a new confirmation, and the three-of-five replication record of Section 6 concerns only its original hypotheses.

Metrics, paired uncertainty and sensitivity. Recall, rescue and harm are evaluated on binary query–seed outcomes before averaging seeds and the two verifiers. Rescue and harm masses use all queries. Conditional rescue uses grid-reachable queries missed by the common constant, conditional harm uses constant successes, and distant rescue additionally requires grid distance at least 0.2 (0.4 is retained as sensitivity). Localization is the nearest successful-grid distance among reachable queries. Ratios use summed numerators and denominators; zero denominators yield undefined results rather than selected finite intervals.

We draw 10,000 paired target-group bootstrap samples, using identical query weights for all arms and both verifiers in a dataset. The main intervals are 95%. Three factorial contrasts in each of two datasets and two evaluation regimes form 12 primary comparisons, accompanied by 99.5833% Bonferroni intervals. Binding versus the shared-posterior decoder uses 97.5% intervals for two datasets within each regime; four negative controls across two datasets use 99.375% intervals. Exact-hypothesis comparisons are exploratory. All contrasts are reported regardless of sign. The bootstrap conditions on the saved fitted policies. Five partition-and-refit repetitions quantify factorial sensitivity but do not turn these intervals into uncertainty for the entire training program; the binding study uses one outer partition. The tables retain absolute recall, relative improvement, rescue, harm, localization and denominator masses.

Table 12: Evaluation populations and denominator masses. Headroom equals grid-oracle recall minus the common constant. Distant mass is the percentage of all queries meeting the distant-rescue denominator. Values average verifier-specific and, where applicable, seed-specific events before forming rates.

Experiment	Dataset	n	Groups	Constant R@1	Headroom	Distant %
factorial	CIRR	3380	1775	47.43 [45.68, 49.18]	10.32 [9.54, 11.11]	5.57
factorial	FashionIQ	4539	4539	22.16 [21.05, 23.32]	7.32 [6.74, 7.92]	4.58
transfer	CIRR	3274	2187	49.13 [47.45, 50.80]	10.38 [9.58, 11.19]	5.77
transfer	FashionIQ	4539	4539	27.92 [26.67, 29.15]	7.78 [7.17, 8.40]	4.63
binding	CIRR	3380	1775	46.49 [44.72, 48.27]	11.20 [10.37, 12.05]	5.58
binding	FashionIQ	4539	4539	21.68 [20.58, 22.81]	7.78 [7.16, 8.40]	4.19
binding transfer	CIRR	3274	2187	48.52 [46.86, 50.20]	11.00 [10.16, 11.83]	5.83
binding transfer	FashionIQ	4539	4539	27.13 [25.92, 28.31]	8.55 [7.92, 9.22]	4.17
forced	CIRR	3380	1775	47.43 [45.68, 49.18]	10.32 [9.54, 11.11]	5.57
forced	FashionIQ	4539	4539	22.16 [21.05, 23.32]	7.32 [6.74, 7.92]	4.58
forced transfer	CIRR	3274	2187	49.13 [47.45, 50.80]	10.38 [9.58, 11.19]	5.77
forced transfer	FashionIQ	4539	4539	27.92 [26.67, 29.15]	7.78 [7.17, 8.40]	4.63

Table 13: Cross-fitted original populations: common-objective factorial. R@1, gain (percentage points), relative improvement (percent), and recovered headroom (percent), with paired 95% target-group intervals. Both verifiers are averaged within each query. Constants and oracle headroom are reported in Table 12.

Dataset	Method	R@1	Gain	Relative %	Recovery %
CIRR	Summary Ridge	47.34 [45.59, 49.08]	-0.09 [-0.23, 0.05]	-0.19 [-0.49, 0.10]	-0.89 [-2.26, 0.44]
	Summary GBM	47.41 [45.66, 49.16]	-0.02 [-0.04, 0.00]	-0.04 [-0.09, 0.00]	-0.20 [-0.41, 0.00]
	Candidate Ridge	47.43 [45.68, 49.18]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	Candidate utility GBM	47.43 [45.68, 49.17]	-0.01 [-0.01, 0.00]	-0.01 [-0.03, 0.00]	-0.06 [-0.15, 0.00]
	Candidate posterior GBM	47.62 [45.87, 49.36]	0.18 [-0.05, 0.42]	0.39 [-0.10, 0.89]	1.78 [-0.49, 4.04]
FashionIQ	Summary Ridge	22.21 [21.09, 23.35]	0.04 [-0.11, 0.19]	0.19 [-0.50, 0.87]	0.57 [-1.56, 2.59]
	Summary GBM	22.15 [21.04, 23.31]	-0.01 [-0.02, 0.00]	-0.04 [-0.09, 0.00]	-0.12 [-0.28, 0.00]
	Candidate Ridge	22.16 [21.05, 23.32]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	Candidate utility GBM	22.16 [21.05, 23.32]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	Candidate posterior GBM	22.50 [21.40, 23.64]	0.34 [0.06, 0.63]	1.54 [0.25, 2.89]	4.67 [0.79, 8.51]

Table 14: Cross-fitted original populations: decision outcomes for common-objective factorial. Rescue and harm are percentage-point masses over all queries. Distant rescue is a percentage conditional on a reachable target missed by the constant at distance at least 0.2. Localization is distance to the nearest successful grid action, conditional on grid reachability. All brackets are paired 95% intervals.

Dataset	Method	Rescue	Harm	Distant rescue	Localization
CIRR	Summary Ridge	0.41 [0.33, 0.50]	0.50 [0.40, 0.61]	0.48 [0.17, 0.85]	0.043 [0.039, 0.047]
	Summary GBM	0.01 [0.00, 0.02]	0.03 [0.01, 0.05]	0.00 [0.00, 0.00]	0.042 [0.037, 0.046]
	Candidate Ridge	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.041 [0.037, 0.046]
	Candidate utility GBM	0.00 [0.00, 0.00]	0.01 [0.00, 0.01]	0.00 [0.00, 0.00]	0.042 [0.037, 0.046]
	Candidate posterior GBM	1.14 [0.96, 1.32]	0.95 [0.80, 1.12]	3.83 [2.66, 5.13]	0.043 [0.039, 0.047]
FashionIQ	Summary Ridge	0.69 [0.57, 0.80]	0.64 [0.54, 0.76]	1.64 [1.00, 2.36]	0.064 [0.058, 0.070]
	Summary GBM	0.00 [0.00, 0.01]	0.01 [0.00, 0.02]	0.00 [0.00, 0.00]	0.061 [0.056, 0.067]
	Candidate Ridge	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.061 [0.056, 0.067]
	Candidate utility GBM	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.061 [0.056, 0.067]
	Candidate posterior GBM	1.73 [1.51, 1.95]	1.39 [1.19, 1.59]	13.53 [11.11, 16.05]	0.063 [0.057, 0.069]

Table 15: Factorial contrasts in R@1 points. The first three contrasts in each dataset and evaluation regime form the 12-comparison primary family; the wider intervals use 99.5833% marginal coverage (Bonferroni). Conditional representation effects are explanatory follow-ups; their wider intervals are shown at the same level and are not additional members of the stated 12-comparison family.

Population	Dataset	Contrast	Estimate [95%]	Wider interval
Original	CIRR	representation	0.05 [-0.02, 0.12]	[-0.05, 0.16]
Original	CIRR	learner	0.03 [-0.04, 0.10]	[-0.06, 0.13]
Original	CIRR	interaction	-0.08 [-0.21, 0.06]	[-0.27, 0.12]
Original	CIRR	representation under ridge	0.09 [-0.05, 0.23]	[-0.10, 0.29]
Original	CIRR	representation under gbm	0.01 [-0.01, 0.04]	[-0.02, 0.05]
Original	FashionIQ	representation	-0.02 [-0.09, 0.06]	[-0.12, 0.09]
Original	FashionIQ	learner	-0.03 [-0.10, 0.05]	[-0.13, 0.08]
Original	FashionIQ	interaction	0.05 [-0.10, 0.20]	[-0.17, 0.26]
Original	FashionIQ	representation under ridge	-0.04 [-0.19, 0.11]	[-0.25, 0.18]
Original	FashionIQ	representation under gbm	0.01 [0.00, 0.02]	[-0.00, 0.03]
External	CIRR	representation	0.13 [-0.01, 0.26]	[-0.06, 0.32]
External	CIRR	learner	0.13 [-0.01, 0.26]	[-0.06, 0.32]
External	CIRR	interaction	-0.26 [-0.52, 0.02]	[-0.64, 0.12]
External	CIRR	representation under ridge	0.26 [-0.02, 0.52]	[-0.12, 0.64]
External	CIRR	representation under gbm	0.00 [0.00, 0.00]	[0.00, 0.00]
External	FashionIQ	representation	-0.02 [-0.15, 0.12]	[-0.22, 0.18]
External	FashionIQ	learner	-0.02 [-0.15, 0.12]	[-0.22, 0.18]
External	FashionIQ	interaction	0.03 [-0.24, 0.31]	[-0.36, 0.43]
External	FashionIQ	representation under ridge	-0.03 [-0.31, 0.24]	[-0.43, 0.36]
External	FashionIQ	representation under gbm	0.00 [0.00, 0.00]	[0.00, 0.00]

Table 16: Constant fallback and action variation across the 50 outer tasks per dataset (two verifiers, five folds, five seeds). Selected constant counts concern the primary model-selection outcome. Forced-equal counts mean every forced action equals the reference constant in that task. Distinct actions summarize the median and range across tasks. This distinguishes forcing an estimator from obtaining an adaptive decision rule.

Dataset	Method	Selected constant	Forced equal	Distinct actions
CIRR	Summary Ridge	26/50	0/50	18 [14, 21]
CIRR	Summary GBM	38/50	46/50	1 [1, 4]
CIRR	Candidate Ridge	50/50	0/50	21 [21, 21]
CIRR	Candidate utility GBM	46/50	48/50	1 [1, 5]
CIRR	Candidate posterior GBM	23/50	0/50	13 [7, 21]
FashionIQ	Summary Ridge	13/50	0/50	20 [16, 21]
FashionIQ	Summary GBM	33/50	48/50	1 [1, 2]
FashionIQ	Candidate Ridge	50/50	0/50	21 [21, 21]
FashionIQ	Candidate utility GBM	48/50	49/50	1 [1, 2]
FashionIQ	Candidate posterior GBM	3/50	0/50	18 [12, 21]

Table 17: Learned binding minus each comparator (R@1 points). The wider interval is 97.5% for the two primary learned-binding versus shared-posterior comparisons within each regime. It is 99.375% for the eight negative-control comparisons per regime; the exact-hypothesis comparison is exploratory and shown at the same width. External intervals describe a separate post-hoc transfer analysis.

Population	Dataset	Comparator	Estimate [95%]	Wider interval
Original	CIRR	Shared posterior	1.30 [0.91, 1.72]	[0.85, 1.78]
Original	CIRR	Exact hypotheses	0.16 [-0.12, 0.44]	[-0.24, 0.55]
Original	CIRR	Random competitor	1.42 [1.03, 1.83]	[0.88, 1.99]
Original	CIRR	Top-score competitor	0.70 [0.35, 1.05]	[0.23, 1.18]
Original	CIRR	Margin competitor	0.70 [0.36, 1.04]	[0.20, 1.17]
Original	CIRR	Random projection	1.35 [0.96, 1.76]	[0.80, 1.92]
Original	FashionIQ	Shared posterior	0.50 [0.24, 0.76]	[0.20, 0.79]
Original	FashionIQ	Exact hypotheses	-0.09 [-0.28, 0.10]	[-0.34, 0.18]
Original	FashionIQ	Random competitor	0.45 [0.19, 0.72]	[0.10, 0.83]
Original	FashionIQ	Top-score competitor	0.11 [-0.11, 0.33]	[-0.21, 0.42]
Original	FashionIQ	Margin competitor	0.43 [0.19, 0.68]	[0.09, 0.78]
Original	FashionIQ	Random projection	0.48 [0.23, 0.74]	[0.13, 0.86]
External	CIRR	Shared posterior	0.73 [0.29, 1.17]	[0.23, 1.24]
External	CIRR	Exact hypotheses	-0.58 [-0.89, -0.27]	[-1.01, -0.15]
External	CIRR	Random competitor	0.89 [0.43, 1.34]	[0.21, 1.50]
External	CIRR	Top-score competitor	0.34 [-0.06, 0.75]	[-0.23, 0.90]
External	CIRR	Margin competitor	0.18 [-0.20, 0.56]	[-0.34, 0.71]
External	CIRR	Random projection	0.73 [0.29, 1.17]	[0.11, 1.35]
External	FashionIQ	Shared posterior	1.00 [0.73, 1.29]	[0.68, 1.33]
External	FashionIQ	Exact hypotheses	-0.13 [-0.32, 0.06]	[-0.39, 0.12]
External	FashionIQ	Random competitor	0.90 [0.62, 1.20]	[0.51, 1.32]
External	FashionIQ	Top-score competitor	0.17 [-0.07, 0.40]	[-0.14, 0.47]
External	FashionIQ	Margin competitor	0.31 [0.09, 0.54]	[0.01, 0.63]
External	FashionIQ	Random projection	1.00 [0.73, 1.29]	[0.62, 1.41]

Table 18: Decision-level changes from adding learned binding to the fixed shared-posterior decoder. Rescue change and harm reduction are population percentage-point masses and sum to the R@1 increment. Distant-rescue change is conditional percentage points; localization change is grid-weight distance, with negative values better. Brackets are paired 95% intervals for explanatory secondary outcomes.

Population	Dataset	Rescue change	Harm reduction	Distant change	Localization change
Original	CIRR	0.01 [-0.22, 0.26]	1.29 [0.98, 1.61]	-1.86 [-4.55, 0.82]	-0.002 [-0.005, 0.000]
Original	FashionIQ	0.04 [-0.12, 0.22]	0.45 [0.26, 0.65]	-1.05 [-3.50, 1.46]	-0.003 [-0.006, 0.000]
External	CIRR	-0.23 [-0.50, 0.05]	0.96 [0.63, 1.29]	-4.45 [-7.34, -1.57]	0.000 [-0.002, 0.003]
External	FashionIQ	0.25 [0.09, 0.42]	0.75 [0.53, 0.98]	0.79 [-1.79, 3.36]	-0.007 [-0.010, -0.004]

F.1 Regularization and action-space sensitivity

Extending ridge penalties to $\{0.1, 1, 10, 100, 1000, 10000\}$ with the same shrinkages 0.5/1 improves forced candidate ridge in all four original/transfer panels. Calibration and inner folds remain fixed; original results use seed 0. The largest penalty is selected throughout the forced candidate arm, so this sensitivity does not bracket its optimum. The nested experiment of Appendix G searches up to 10^8.

Table 19: Extended Ridge search, with identical penalties 0.1, 1, 10, 100, 1,000 and 10,000 in both representations and the same two shrinkages. Selected includes the constant; forced excludes it using inner validation only. Original results use one five-fold partition, not the five-seed primary factorial. Intervals are conditional paired 95%.

Regime	Data	Selection	Summary − C	Candidate − C	Candidate − Summary
Original	CIRR	selected	$-0.24\,[-0.53, +0.06]$	$+0.09\,[-0.07, +0.25]$	$+0.33\,[+0.03, +0.62]$
Original	CIRR	forced	$-0.18\,[-0.49, +0.14]$	$-0.56\,[-1.01, -0.12]$	$-0.38\,[-0.85, +0.08]$
Original	FIQ	selected	$+0.00\,[-0.24, +0.24]$	$-0.03\,[-0.13, +0.07]$	$-0.03\,[-0.28, +0.22]$
Original	FIQ	forced	$+0.03\,[-0.24, +0.31]$	$-0.58\,[-0.95, -0.22]$	$-0.62\,[-0.99, -0.24]$
Transfer	CIRR	selected	$-0.18\,[-0.42, +0.06]$	$+0.00\,[+0.00, +0.00]$	$+0.18\,[-0.06, +0.42]$
Transfer	CIRR	forced	$-0.21\,[-0.47, +0.05]$	$-0.43\,[-0.88, +0.03]$	$-0.21\,[-0.67, +0.25]$

Regime	Data	Selection	Summary − C	Candidate − C	Candidate − Summary
Transfer	FIQ	selected	$+0.01\,[-0.26, +0.28]$	$+0.00\,[+0.00, +0.00]$	$-0.01\,[-0.28, +0.26]$
Transfer	FIQ	forced	$+0.01\,[-0.26, +0.28]$	$-0.14\,[-0.53, +0.24]$	$-0.15\,[-0.52, +0.21]$

Table 20: Paired effect of extending the penalty search on forced Ridge, versus the narrow-grid run on exactly the same queries and seed-0 folds. These post-hoc sensitivity contrasts change the tuning budget and do not replace the original matched-budget factorial.

Regime	Data	Summary: wide − narrow	Candidate: wide − narrow
Original	CIRR	$-0.19\,[-0.35, -0.04]$	$+0.89\,[+0.34, +1.43]$
Original	FIQ	$-0.01\,[-0.14, +0.12]$	$+0.52\,[+0.14, +0.89]$
Transfer	CIRR	$+0.21\,[-0.06, +0.50]$	$+1.24\,[+0.66, +1.81]$
Transfer	FIQ	$-0.02\,[-0.10, +0.06]$	$+0.65\,[+0.26, +1.06]$

For the same binding-study posterior, analytical scalar selection maximizes the posterior mass of the candidate ranked first by each action. Direct selection instead takes the largest candidate posterior. A separate fixed RRF control uses $1/(60 + r_j) + 1/(60 + v_j)$ with one-based ranks (Cormack et al., 2009). All use identical candidate lists. The four Direct-minus-Grid comparisons have 98.75% intervals; their small differences concern this posterior. Neither rule reproduces a trained listwise reranker.

Table 21: Paired gains over the constant using frozen predictions. The final column uses 98.75% intervals for the four Direct-minus-Grid comparisons across datasets and regimes; all four include zero. Other intervals are descriptive 95% intervals conditional on the fitted posterior. No equivalence conclusion follows from nonsignificance.

Regime	Data	Direct − C	Grid − C	RRF − C	Direct − Grid
Original	CIRR	$+0.84\,[+0.34, +1.35]$	$+0.86\,[+0.35, +1.37]$	$-2.97\,[-3.82, -2.16]$	$-0.01\,[-0.18, +0.15]$
Original	FIQ	$+0.79\,[+0.36, +1.22]$	$+0.66\,[+0.25, +1.07]$	$-1.13\,[-1.63, -0.65]$	$+0.13\,[-0.03, +0.30]$
Transfer	CIRR	$+0.55\,[+0.02, +1.09]$	$+0.47\,[-0.05, +1.00]$	$-3.82\,[-4.68, -2.95]$	$+0.08\,[-0.11, +0.26]$
Transfer	FIQ	$+1.21\,[+0.75, +1.69]$	$+1.11\,[+0.66, +1.56]$	$-1.53\,[-2.04, -1.02]$	$+0.10\,[-0.09, +0.29]$

G EQUAL-BUDGET TUNING AND SEMANTIC CONTROLS

Question and timing. Does competition encoding improve decisions beyond generic conservative routing and regularization? This exploratory protocol was fixed before its new fits on previously studied populations. Scores, Top-20 candidates, query min–max calibration and the 21 actions are fixed. We run 60 outer fits (two datasets, two verifiers, three seeds and five folds), plus four full-training transfers to the previously studied disjoint external populations. No verifier inference is repeated.

Equal selection opportunities. Each representation receives 91 choices: the training-optimal constant and ten ridge penalties from 10^{-1} to 10^{8}, each combined with three utility shrinkages $(0.25, 0.5, 1)$, three maximum-posterior gate quantiles, or three negative-entropy gate quantiles $(0.25, 0.5, 0.75)$. Gate thresholds use decoder-training cross-fitted predictions; rejected queries return to the training-optimal constant. Selection maximizes pooled inner held-out R@1, with deterministic ties preferring the constant. This matches selection opportunities, not effective feature dimension or auxiliary supervision cost.

Nested fitting. Three selection folds sit inside each outer training set. Each selection-training subset uses a further three-fold auxiliary crossfit to train its decoder; auxiliary models fitted only on that subset predict its selection-validation rows. These valid outer-training auxiliary predictions are then used for the final decoder; final auxiliary models fit all outer-training rows. Test labels never enter prediction or selection. All splits respect target groups. The original six auxiliary classifiers and their capacities are retained.

Semantic controls. All six representations contain the same candidate features and target posterior. Their additional blocks are zero (posterior-only), learned binding, exact hypothetical-target geometry, sorted posterior probabilities (confidence), or one of two prespecified identity permutations of the learned binding probabilities. The same per-query permutation is applied to both binding sides, preserving their joint probability pairs and domain-sentinel mass while breaking alignment to candidate scores. Both training and testing are permuted before refitting the control. This preserves auxiliary supervision but not all encoding covariance properties. Exact geometry and its analytical argmax require no true test target.

Frozen-program checks. Separately, the selected and fixed-penalty binding decoders receive the same identity permutations at test time, freezing the target posterior, remaining features and routing thresholds. This tests dependence of an already fitted program, not the usefulness of an encoding after retraining. Encoding-shift magnitudes are recorded because distribution shift remains an alternative explanation. Swapping the two side factors is an algebraic invariance check: their product is commutative, so it is not a semantics-destroying intervention.

Uncertainty and estimands. We form decision events before averaging verifiers and refits within each query, then use 10,000 paired target-group bootstrap replicates. Primary contrasts use 97.5% intervals for two datasets within each original/transfer family. Three secondary contrasts per dataset use 99.1667% intervals; frozen interventions are exploratory. These intervals condition on fitted predictions. Individual refit effects are reported separately and are not additional independent datasets. Non-significance does not establish equivalence.

Table 22: Binding minus posterior-only R@1 in percentage points. Fixed comparison: ridge penalty 10, no routing. Selected comparison: identical 91-choice portfolios. Intervals are 97.5%.

Population	Fixed decoder	Equal-budget selection
Original CIRR	$+1.15\,[+0.83, +1.49]$	$+0.14\,[+0.00, +0.28]$
Original FashionIQ	$+0.61\,[+0.41, +0.83]$	$-0.04\,[-0.16, +0.07]$
Transfer CIRR	$+0.73\,[+0.23, +1.24]$	$+0.02\,[-0.40, +0.43]$
Transfer FashionIQ	$+1.00\,[+0.68, +1.33]$	$-0.03\,[-0.44, +0.37]$

Table 23: Selected binding minus semantic controls, R@1 points, 99.1667% intervals. Permuted denotes the mean of two separately retrained identity controls.

Population	Confidence	Exact geometry	Permuted
Original CIRR	$+0.05\,[-0.12, +0.22]$	$-0.14\,[-0.33, +0.04]$	$+0.06\,[-0.11, +0.24]$
Original FashionIQ	$-0.15\,[-0.32, +0.01]$	$-0.15\,[-0.35, +0.02]$	$-0.03\,[-0.16, +0.10]$
Transfer CIRR	$+0.02\,[-0.45, +0.52]$	$-0.21\,[-0.57, +0.12]$	$+0.18\,[-0.38, +0.75]$
Transfer FashionIQ	$-0.11\,[-0.59, +0.37]$	$-0.11\,[-0.47, +0.26]$	$+0.10\,[-0.29, +0.51]$

Table 24: Complete-refit primary effects (seeds 0/1/2 for original data) and frozen identity-permutation effects relative to the intact selected binding program. Perturbation intervals are exploratory 95%; all values are R@1 points.

Population	Refit effects	Frozen perm. 1	Frozen perm. 2
Original CIRR	+0.30, +0.13, -0.01	$-0.12\,[-0.24, -0.01]$	$-0.19\,[-0.30, -0.07]$
Original FashionIQ	-0.18, -0.04, +0.09	$-0.14\,[-0.22, -0.07]$	$-0.16\,[-0.25, -0.08]$
Transfer CIRR	+0.02	$-0.44\,[-0.71, -0.17]$	$-0.15\,[-0.43, +0.12]$
Transfer FashionIQ	-0.03	$-0.52\,[-0.74, -0.30]$	$-0.42\,[-0.65, -0.20]$

Table 25: Absolute performance and decision accounting. Rescue and harm are full-population masses; distant rescue is conditional on grid-reachable constant failures at distance at least 0.2. Localization is mean distance on reachable queries; recovery is net gain divided by oracle headroom. R@1, gain and masses are points; distant rescue and recovery are percentages.

Population / policy	R@1	Gain	Rescue	Harm	Distant	Loc.	Recovery
Original CIRR / Constant	46.67	0.00	0.00	0.00	0.00	0.04	0.00
Original CIRR / Analytic posterior	47.40	0.72	2.61	1.89	7.98	0.06	6.58
Original CIRR / Posterior, selected	46.46	-0.21	0.84	1.05	1.30	0.04	-1.92
Original CIRR / Binding, selected	46.60	-0.07	1.07	1.14	1.99	0.04	-0.67
Original FashionIQ / Constant	21.69	0.00	0.00	0.00	0.00	0.06	0.00
Original FashionIQ / Analytic posterior	22.38	0.69	2.13	1.44	14.50	0.10	8.88
Original FashionIQ / Posterior, selected	21.57	-0.11	0.70	0.81	2.56	0.06	-1.47
Original FashionIQ / Binding, selected	21.53	-0.16	0.78	0.94	3.89	0.07	-2.03
Transfer CIRR / Constant	48.52	0.00	0.00	0.00	0.00	0.04	0.00
Transfer CIRR / Analytic posterior	48.99	0.47	2.49	2.02	8.38	0.06	4.31
Transfer CIRR / Posterior, selected	48.26	-0.26	0.81	1.07	0.26	0.04	-2.36
Transfer CIRR / Binding, selected	48.27	-0.24	1.27	1.51	4.71	0.05	-2.22
Transfer FashionIQ / Constant	27.13	0.00	0.00	0.00	0.00	0.05	0.00
Transfer FashionIQ / Analytic posterior	28.24	1.11	2.70	1.59	11.35	0.09	13.02
Transfer FashionIQ / Posterior, selected	27.40	0.26	1.02	0.76	2.11	0.05	3.09
Transfer FashionIQ / Binding, selected	27.36	0.23	1.53	1.30	8.18	0.06	2.71

Interpretation. Original CIRR has a positive pointwise 95% interval, but its corrected interval touches zero; the other three primary intervals include zero. Corrected contrasts against confidence, exact geometry and identity-permuted refits also include zero. Frozen perturbations often reduce recall. Together these results distinguish program dependence from a stable incremental encoding advantage under matched selection. Three refits characterize selected training variability; the existing external populations provide transfer rather than untouched confirmation.

H NEURAL SET DECODERS

Protocol. The protocol was fixed before any neural-decoder outcome and reuses everything of the nested study except the decoder: outer and inner folds, the cached fully nested auxiliary posteriors, the routing thresholds, the 91 tuning choices and the paired bootstrap. Each candidate is a token of its 93 candidate features and auxiliary target probability, concatenated with the 65 shared context features and the outside-list probability. A linear embedding is followed by two pre-LayerNorm Transformer encoder layers without positional encoding (four heads, dropout 0.1), which makes the model permutation-equivariant over candidates. The *utility head* pools the tokens and predicts the 21 action utilities with squared error, the objective of the ridge decoders. The *target head* predicts a distribution over the 20 candidates and "outside the list" with cross-entropy and decides by expected R@1. Training uses AdamW (learning rate 10^{-3}, batch 128, 60 epochs, no early stopping). The ten capacity settings, the analogue of the ten ridge penalties, cross width $\{32, 64\}$ with weight decay $\{10^{-4}, \ldots, 1\}$. Smoke tests verified permutation invariance and equivariance, reproducible training,

and that the expected-recall decoder applied to the auxiliary posterior reproduces the nested study's direct rule exactly.

Results. Table 26 lists all four panels. The primary contrasts use 98.75% intervals for two contrasts in two datasets. The target head is also close to the direct expected-recall rule on the auxiliary posterior. It is lower on the original populations (-0.25 [-0.71, $+0.22$] and -0.19 [-0.54, $+0.16$], 95% intervals) and higher on the transfer populations ($+0.14$ [-0.40, $+0.68$] and $+0.57$ [$+0.14$, $+1.00$]).

Table 26: Neural set decoders under the nested 91-choice budget. Gains over the constant (R@1 points) for the tuned posterior-only ridge decoder, the direct expected-recall rule on the auxiliary posterior, and the two tuned neural decoders; the last two columns are the primary contrasts with 98.75% (Bonferroni) intervals.

Population	Ridge	Direct	NSP-utility	NSP-posterior	Utility − ridge	Posterior − ridge
Original CIRR	−0.21	+0.72	−0.20	+0.48	+0.01 [−0.32, +0.35]	+0.69 [+0.21, +1.18]
Original FashionIQ	−0.11	+0.69	−0.23	+0.50	−0.11 [−0.37, +0.14]	+0.62 [+0.18, +1.05]
Transfer CIRR	−0.26	+0.47	−0.05	+0.61	+0.21 [−0.23, +0.67]	+0.87 [+0.28, +1.46]
Transfer FashionIQ	+0.26	+1.11	+0.09	+1.69	−0.18 [−0.59, +0.23]	+1.42 [+0.82, +2.05]

I LABEL-DERIVED DIAGNOSTIC PROBES

Both probes in Figure 11 use evaluation labels to measure diagnostic sufficiency for fixed decision procedures.

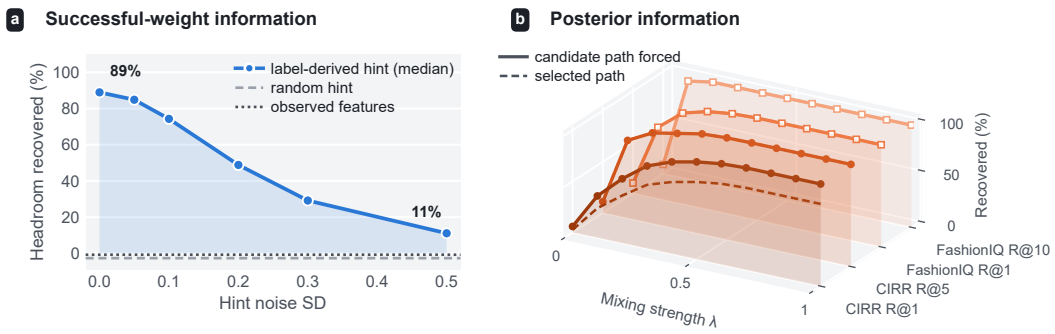

Figure 11: **Label-derived information closes the gap.** (a) Headroom recovered by the ridge policy when a hint of a successful weight, with Gaussian noise of the given SD, is added to its features. Random hints and observed features alone serve as controls. (b) Recovery when the true target is mixed into the candidate policy's posterior with strength λ, where $\lambda = 1$ supplies the target. Solid curves enforce the candidate path; dashed curves keep training-time selection. Both probes use evaluation labels and serve as diagnostics.

Successful-weight hint. For informative queries, use the median successful grid weight; otherwise draw a uniform hint. Add Gaussian error, clip to $[0, 1]$, and encode it with 21 radial features of bandwidth 0.1 centered on the action grid. Random hints and observed features alone provide controls. Table 27 reports every dataset–metric panel.

Posterior injection. The candidate Bernoulli scores supplied to the fitted machinery become $p_\lambda = (1 - \lambda)\hat{p} + \lambda\,\mathrm{onehot}(t)$ on an eleven-point grid, before clipping and logit/temperature normalization (Appendix B). Consequently, the implemented objective is not a linear mixture of expected utilities. On outside-list queries the target vector is zero; at $\lambda = 1$ clipping yields uniform scores and every action still misses. The procedure does not infer target absence at deployment.

Eight cells (two datasets, two verifiers, two metrics) retain both the forced-candidate path and the path chosen by inner selection. Without injection, recovery is 5.9–10.9% and gains range from $+0.24$ to $+0.77$. On the forced path, 99% recovery occurs at λ of 0.2–0.7, where posterior top-1 accuracy

Table 27: Median share of oracle headroom (%) recovered by the fixed ridge policy when a label-derived hint of a successful weight, with Gaussian error of the given SD, is added to its features. The random column replaces the hint by a uniform draw; the last column uses the observed features only.

Cell group	Cells	0	.05	.1	.2	.3	.5	random	XZ only
CIRR R@1	11	90	87	74	50	31	16	2	4
CIRR R@5	11	85	83	72	47	26	8	-5	-3
FashionIQ R@1	10	87	84	73	46	27	15	-2	1
FashionIQ R@10	10	90	87	80	54	30	4	-4	-7

is 0.45–0.83. This 99% index was chosen after the fact; the prespecified threshold (the smallest injection with an interval above zero) was already met at zero injection in all four panels. One panel with free selection never reaches 99%. The endpoint $\lambda = 1$ checks that supplied information is used; it is not an accuracy that a learner could be expected to reach.

J INDEPENDENT PROTOCOLS AND EXTERNAL RESULTS



Figure 12: **Replication and CIRCO.** Gains over the constant with 95% intervals. The shaded band marks the prespecified one-point bound. Gains stay small on the two new populations but are larger on CIRCO.

The independent protocols were frozen before scoring previously unused populations, including sampling, verifier coverage, thresholds and inference. We report their verdicts under the original rules. The controlled studies of Section 5 reuse the new CIRR and FashionIQ populations for transfer and do not count as further confirmations.

Coverage and criteria. The original record confirms H1, H2 and H5, with H1 limited to 12 of 16 planned cells from three of four verifiers, because the fourth verifier's scoring was incomplete. H3 and H4 remain unconfirmed: H3 intervals include zero; candidate recovery exceeds one tenth of headroom in two of four panels. H2/CIRCO C2 require a 95% interval wholly inside $(-1, +1)$, preserving their original criterion.

Completed record. We completed the fourth verifier (Phi-3.5-Vision) after the original analysis with the frozen scoring settings, and reran the frozen analysis unchanged; the original outputs are archived. With 4 verifiers, H1 covers 16/16 cells (minimum headroom 2.2 points) and the verdicts are unchanged: H1, H2, H5 confirmed, H3, H4 not (Table 31). The pooled ridge R@1 gain is +0.35 [+0.02, +0.69] on CIRR and +0.13 [−0.11, +0.37] on FashionIQ. The other tables of this appendix keep the original three-verifier record.

Table 28: The three populations used for confirmation. None was used in any fusion-policy analysis before its protocol was frozen. The first stage is the frozen SigLIP-L/384 retriever with the composition weight transferred from the validation pilot, never re-selected. Target-in-top-20 bounds what any fusion weight can achieve: queries outside it are misses under every action.

Population	Queries	Gallery	Target in top 20	First-stage R@1
CIRR train subset	3,274	2,302	86.0%	36.4
FashionIQ train subset	4,539	3,645/6,182/5,261	54.7–58.2%	20.9–26.6
CIRCO validation	220	123,403	38.2%	5.9

Table 29: Every hypothesis fixed before the confirmation data were scored, with the quantity that decided it. A rule the estimate violates is reported as not confirmed whatever the interval width; underpowered marks a rule the data satisfy but cannot certify at the pre-registered precision.

	Rule	Observed	Outcome
H1	headroom ≥ 2 points in every cell	smallest 2.2 over 12/16 cells	supported on evaluated cells
H2	pooled ridge 95% interval inside $(-1, +1)$, gain below a fifth of headroom	all four panels	confirmed
H3	R@1 gain exceeds the deeper-cutoff gain, interval above zero	CIRR $+0.43$ $[-0.01, +0.88]$, FashionIQ $+0.24$ $[-0.09, +0.58]$	not confirmed
H4	candidate-aware gain positive and below a tenth of headroom	positive in 4/4; share 4.4–13.2%, over a tenth in 2/4	not confirmed
H5	harm and far-rescue rates below the permuted null	harm 7.4% vs 24.3% (CIRR)	confirmed
C1	CIRCO headroom ≥ 2 points in every cell	smallest 3.0 over 8 cells	confirmed
C2	pooled CIRCO ridge 95% interval inside $(-1, +1)$ at R@1	$+1.23$ $[+0.18, +2.32]$, width 2.1	not confirmed
C3	CIRCO candidate-aware gain below a tenth of headroom	R@1 $+0.70$ $[-0.18, +1.57]$ vs 0.73; R@5 $+2.09$ $[+0.98, +3.30]$ vs 0.68	underpowered (R@1), not confirmed (R@5)

Table 30: CIRCO by verifier. Transfer applies the constant and the ridge fitted on the whole CIRR validation population without refitting. The last column counts the queries whose recall still depends on the weight at R@1; every CIRCO interval rests on that many queries out of 220.

Verifier	Headroom R@1	Headroom R@5	Transfer R@1	Informative
InternVL3-2B	7.5	8.6	-0.91	39
Qwen3-VL-2B	5.0	3.0	$+0.00$	45
LLaVA-OV-0.5B	5.6	10.0	-1.36	27
Phi-3.5-Vision	10.8	5.5	$+1.82$	50

Table 31: Replication verdicts under the unchanged prespecified rules: the original record (3 verifiers) and the completed record (4 verifiers, fourth verifier scored after the original analysis).

Hyp.	Rule	Original	Completed
H1	headroom ≥ 2 in all 16 cells	confirmed	confirmed
H2	pooled ridge 95% interval inside (-1, +1) and pooled gain $<$ 20% of mean headroom in all four panels	confirmed	confirmed
H3	pooled R@1 gain minus deeper-cutoff gain has a 95% interval above zero in both datasets	not confirmed	not confirmed
H4	pooled candidate gain positive in ≥ 3 of 4 panels, no panel interval entirely below zero, every gain $<$ 10% of mean headroom	not confirmed	not confirmed
H5	pooled fixed-ridge harm rate below permuted and far (≥ 0.4) rescue rate below permuted, both 95% intervals below zero, in each dataset	confirmed	confirmed

Table 32: Complete original H2 scorecard: all four pooled Ridge-minus-constant panels and their 95% intervals, copied from the frozen confirmation analysis. The H1 coverage remains 12 of 16 planned cells.

Panel	Ridge gain [95%]	Headroom
cirr R1	$+0.44\,[+0.08, +0.79]$	11.34
cirr R5	$+0.01\,[-0.26, +0.28]$	6.23
fiq R1	$+0.11\,[-0.16, +0.37]$	7.68
fiq R10	$-0.13\,[-0.33, +0.07]$	3.25

Table 33: Complete original H5 scorecard on both independent populations. Rates and differences are percentage points; intervals and nulls follow the prespecified confirmation procedure.

Data	Event	Ridge	Ridge null	Difference [95%]
CIRR	harm	7.35	24.33	$-16.97\,[-17.80, -16.12]$
CIRR	far rescue	0.00	12.62	$-12.62\,[-13.42, -11.80]$
FIQ	harm	8.39	25.64	$-17.24\,[-18.17, -16.29]$
FIQ	far rescue	4.05	13.79	$-9.75\,[-12.42, -6.68]$

CIRCO. The frozen retriever transfers its pilot composition weight. Only 84/220 annotated targets enter the Top-20 list, and 27–50 queries per verifier are weight-sensitive. Ridge gains $+1.23$ [$+0.18$, $+2.32$] R@1 points; candidate R@5 gains $+2.09$ [$+0.98$, $+3.30$]. Re-scoring the same actions with any annotated relevant image changes ridge gains by -1.18 to $+3.09$. These are recall analyses, not official multi-ground-truth mAP. The single-target geometry and all CIRR-to-CIRCO transfer intervals are retained below.

Table 34: CIRCO structural reanalysis at single-target R@1. All 220 queries; unchanged min–max OOF policies. Coverage, oracle, headroom and gain are percentage points. Width is the median continuous-interval width conditional on a nonempty interval. These descriptive quantities do not explain a causal distribution effect.

Verifier	Constant	Grid oracle	Headroom	Ridge Δ	Width
InternVL3-2B	11.09	18.64	7.55	$+1.73$	0.689
Qwen3-VL-2B	17.73	22.73	5.00	$+1.64$	0.613
LLaVA-OV-0.5B	8.91	14.55	5.64	-0.91	0.432
Phi-3.5-Vision	13.73	24.55	10.82	$+2.45$	0.471

Table 35: Complete CIRR-to-CIRCO frozen Ridge transfer: all eight metric cells and original paired 95% intervals. Every interval includes zero. These are single-target recall differences, not official multi-ground-truth mAP.

Verifier	Metric	Ridge minus constant [95%]
InternVL3-2B	R@1	$-0.91\,[-2.73, +0.91]$
InternVL3-2B	R@5	$-0.45\,[-1.82, +0.91]$
Qwen3-VL-2B	R@1	$+0.00\,[-1.82, +1.82]$
Qwen3-VL-2B	R@5	$+0.91\,[-0.91, +2.73]$
LLaVA-OV-0.5B	R@1	$-1.36\,[-3.64, +0.45]$
LLaVA-OV-0.5B	R@5	$-2.27\,[-5.45, +0.91]$
Phi-3.5-Vision	R@1	$+1.82\,[-0.91, +5.00]$
Phi-3.5-Vision	R@5	$-0.91\,[-2.27, +0.00]$

Sealed test. Only the constant and summary ridge were submitted. The constant reaches 47.33 R@1 and ridge 47.66, a change of $+0.34$ points. Subset recall changes from 76.19 to 75.28 (-0.92 points). Six of eight reported differences are negative, one positive and one zero; R@50 is structurally identical because only the Top-20 ordering changes. Both point estimates meet the predeclared T1/T2 tolerances. The server returns aggregate scores only, so statistical equivalence cannot be evaluated.

K TEXT TWO-STAGE RETRIEVAL

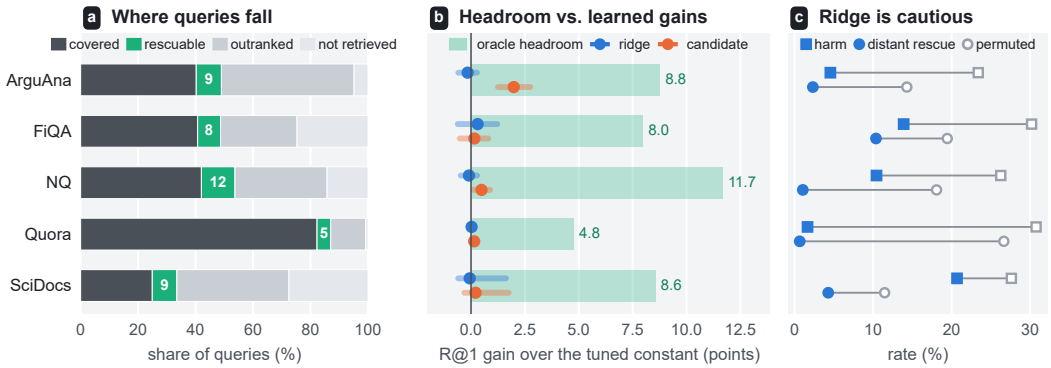

Figure 13: **The same gap in text two-stage retrieval.** Five BEIR datasets, averaged over two cross-encoder verifiers. (a) Share of queries covered by the tuned constant, rescuable by another weight, or unreachable. (b) Oracle headroom against the gains of the ridge policy and the matched-calibration candidate policy, with 95% intervals. (c) Ridge harm rate and distant-rescue rate (grid distance ≥ 0.4) against the within-fold permutation of its own actions.

Protocol. The protocol, including four predictions, was fixed before any text score, ranking or outcome was computed. We used the test splits of ArguAna, FiQA, NQ, Quora and SciDocs from the official BEIR distribution (Thakur et al., 2021), keeping every query with at least one relevant document (qrel score above zero). As in BEIR evaluation, a document whose identifier equals the query's is removed from its candidate list. The first stage is BGE-base-en-v1.5 (Xiao et al., 2024) with normalized embeddings and exact inner-product search. The two verifiers are the MS MARCO cross-encoders MiniLM-L6 and BGE-reranker-base, read as raw logits. Actions, tie rules, the 25 query summaries, the primary constant and ridge policies, the matched-calibration panel code, the folds and the bootstrap are the paper's, applied unchanged. Queries that share a relevant document form one group for folds and resampling. Per-candidate features follow the construction of Appendix B, with text substitutes for the image features: document length, query–document token overlap and fixed random projections of the dense embeddings.

Two differences from composed image retrieval. First, the cross-encoders alone are weaker than the retriever alone on four of the five datasets, with NQ the exception (columns α=0 and α=1 of Table 36). The fusion weight still matters: headroom over the tuned constant is 4.8–11.7 points. Second, relevant documents are shared across queries in SciDocs. One group there contains 55% of the queries, so the group bootstrap has few effective units and its intervals are wide and skewed.

Results. Table 36 reports every dataset–verifier cell, and Table 37 reports the four predictions. Headroom is at least two points everywhere (T1). The ridge gain is within one point with the required precision on ArguAna, NQ and Quora but not on FiQA and SciDocs, where the upper interval ends exceed one point; their point estimates are small. Ridge harms less and rescues fewer distant queries than its permutation reference on all five datasets (T3). Figure 14 shows the rescue profile by distance. The candidate policy's gain is positive and below a tenth of headroom except on ArguAna, where it recovers 23% of headroom.

Table 36: Text two-stage retrieval, every dataset and cross-encoder (R@1, five seeds). In list: share of queries with a relevant document in the Top-20. α=0 and α=1 rank by the retriever or the cross-encoder alone. R@1 values are %; headroom and gains are points with paired 95% group-bootstrap intervals. Pooled rows average the two cross-encoders within query.

Dataset	Verifier	n	In list	α=0	α=1	Constant	Oracle	Headroom	Ridge Δ	Candidate Δ
ArguAna	MiniLM-L6	1,406	95.2	40.6	21.8	40.6	49.0	8.4	−0.46 [−1.01, +0.09]	+1.98 [+1.14, +2.82]
ArguAna	BGE-reranker-base	1,406	95.2	40.6	15.3	39.9	49.0	9.1	+0.14 [−0.50, +0.78]	+1.99 [+1.09, +2.93]
pooled								8.8	−0.16 [−0.59, +0.28]	+1.98 [+1.25, +2.75]
FiQA	MiniLM-L6	648	75.3	39.2	36.0	39.3	48.0	8.7	+1.60 [+0.55, +2.72]	−0.25 [−1.08, +0.56]
FiQA	BGE-reranker-base	648	75.3	39.2	34.6	42.1	49.4	7.3	−0.99 [−2.41, +0.40]	+0.56 [−0.31, +1.42]
pooled								8.0	+0.31 [−0.62, +1.22]	+0.15 [−0.52, +0.80]
NQ	MiniLM-L6	3,452	85.8	35.8	40.0	40.6	52.1	11.5	+0.03 [−0.49, +0.56]	+0.24 [−0.27, +0.75]
NQ	BGE-reranker-base	3,452	85.8	35.8	37.4	43.5	55.4	11.9	−0.23 [−0.78, +0.31]	+0.73 [+0.16, +1.30]
pooled								11.7	−0.10 [−0.48, +0.28]	+0.49 [+0.09, +0.89]
Quora	MiniLM-L6	10,000	99.2	81.8	73.8	82.6	87.6	5.0	+0.05 [−0.10, +0.19]	+0.04 [−0.12, +0.21]
Quora	BGE-reranker-base	10,000	99.2	81.8	58.2	82.0	86.5	4.5	+0.01 [−0.10, +0.11]	+0.25 [+0.07, +0.43]
pooled								4.8	+0.03 [−0.06, +0.12]	+0.15 [+0.02, +0.27]
SciDocs	MiniLM-L6	1,000	72.5	24.3	21.9	25.2	32.9	7.7	−0.14 [−0.98, +2.54]	−0.80 [−1.29, +0.69]
SciDocs	BGE-reranker-base	1,000	72.5	24.3	20.9	24.7	34.1	9.4	+0.02 [−1.12, +1.45]	+1.22 [+0.37, +3.20]
pooled								8.6	−0.06 [−0.59, +1.62]	+0.21 [−0.31, +1.75]

Table 37: The four predictions fixed before any text score was computed, per dataset. Harm is conditional on constant successes that some weight breaks; distant rescue is conditional on rescuable queries at grid distance ≥ 0.4; both are ridge / within-fold permutation reference (%). Candidate recovery is the pooled matched-calibration gain as a percentage of headroom.

Dataset	T1	T2	Harm	Distant rescue	T3	Cand. recovery (%)	T4
ArguAna	yes	yes	4.6 / 23.4	2.3 / 14.3	yes	22.6	no
FiQA	yes	no	13.9 / 30.2	10.4 / 19.4	yes	1.9	yes
NQ	yes	yes	10.4 / 26.3	1.1 / 18.1	yes	4.2	yes
Quora	yes	yes	1.7 / 30.8	0.7 / 26.6	yes	3.1	yes
SciDocs	yes	no	20.7 / 27.6	4.3 / 11.5	yes	2.5	yes

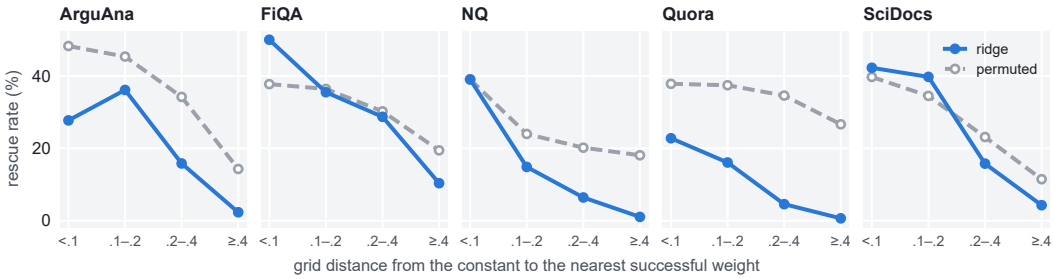

Figure 14: **Ridge misses distant rescues in text retrieval too.** R@1 rescue rate of the ridge policy and of the within-fold permutation of its actions, by grid distance from the tuned constant to the nearest successful weight, pooled over the two cross-encoders and five seeds.

L WINNER-PATH DUEL

Protocol. The method was designed by exploration on the development populations (CIRR and FashionIQ validation) and fixed before any duel outcome on the confirmation populations was computed. For each query, the winner path W_q is computed from the primary scores over the 21 weights. The constant's winner c is the candidate ranked first by the cross-fitted min–max constant. When $|W_q| > 1$, a second verifier scores the path candidates. For each alternative $w \in W_q \setminus \{c\}$, a standardized logistic regression ($C = 1$) estimates $P(\mathrm{rel}(w) = 1 \mid \text{exactly one of } w, c \text{ relevant})$. Its inputs are the differences $w - c$ of the min–max and raw retriever and primary-verifier scores and of the raw second-verifier score, the retriever and primary-verifier ranks of w and c, the share of weights at which w wins, the distance from the constant weight to the nearest such weight, and $|W_q|$. WPD switches to the alternative with the largest logit if it exceeds a threshold $\tau \in \{-1, -0.5, 0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3\}$, chosen on three inner group folds by pooled net R@1 change, with ties preferring larger τ. The chosen weight is any weight at which the new winner ranks

first. Cross-fitted results use five outer group folds and five seeds. On the replication populations, a frozen transfer additionally fits the constant, judge and threshold on the full development population of the same verifier pair. Prespecified tests: the pooled gain over the constant has a 95% interval above zero (D1); WPD beats the tuned candidate policy (D2); the frozen transfer gain is above zero (D3).

Amendment and disclosure. The first implementation labelled a duel as a rescue whenever w was relevant, even when c was relevant too. With several relevant documents per query (FiQA, Quora and SciDocs; 16–20% of labelled text duels) switching then changes nothing, yet such pairs pushed the judge towards switching. We corrected the labels to $\mathrm{rel}(w) - \mathrm{rel}(c)$ before any replication outcome. Single-target populations (CIRR, FashionIQ, ArguAna) are unaffected. The first text run gave pooled gains of ArguAna -0.12, FiQA +0.08, NQ +1.69, Quora -0.34, SciDocs -1.26 points, and all its outputs are kept. The corrected text results are therefore not untouched confirmation. The same amendment added a prespecified secondary variant that may abstain entirely when no threshold helps on the inner folds; it changes results by at most a few hundredths of a point (Table 40).

Prospective path-only scoring. An addendum written before any of its scores existed tested the deployment path on the replication populations. InternVL3-8B, the strongest development second opinion, scored only the winner-path candidates of the InternVL3-2B and Qwen3-VL-2B primaries (union per query), with the frozen prompt, image size and scorer. Unscored candidates stay missing, so variants that need all 20 second-opinion scores are not computed. WPD with this second opinion gains +3.99 [+3.30, +4.68] on CIRR and +1.85 [+1.38, +2.32] on FashionIQ; frozen transfer from development, where InternVL3-8B had scored every candidate, gains +3.79 [+3.08, +4.50] and +2.02 [+1.53, +2.50]. Each primary needs 1.6 (CIRR) and 2.0 (FashionIQ) second-verifier calls per query; scoring the union of both primaries' paths took 2.3 and 2.8 calls per query.

Second amendment: two mechanism tests. Before any self-duel or BGE-reranker-v2-m3 score existed, we added two tests of whether the gain comes from the second opinion's information. In the *self-duel*, the primary VLM itself compares two path candidates in one prompt (reference image, requested change, option A, option B; answer A or B), in both orders. The pair score, half the difference of the two A-minus-B log-probability gaps, replaces the second-verifier difference in the judge; everything else is unchanged. Each alternative winner costs two prompts with three images each. The design was explored on the development populations and frozen before the replication analysis, which is its confirmation. On development it gains +0.92 [+0.46, +1.37] on CIRR and +0.58 [+0.30, +0.85] on FashionIQ (+1.67 (InternVL3-2B) and +0.17 (Qwen3-VL-2B) on CIRR, and +0.55 (InternVL3-2B) and +0.61 (Qwen3-VL-2B) on FashionIQ), with 1.9 and 2.4 prompts per query. On the replication populations it gains +1.14 [+0.74, +1.55] and +0.90 [+0.56, +1.25] (+1.69 (InternVL3-2B) and +0.60 (Qwen3-VL-2B) on CIRR, and +0.68 (InternVL3-2B) and +1.12 (Qwen3-VL-2B) on FashionIQ); frozen transfer gains +1.28 [+0.78, +1.78] and +0.97 [+0.67, +1.28]. For text, BGE-reranker-v2-m3, trained on different data from the two MS MARCO cross-encoders, scored only the path candidates and served as second opinion for both primaries. The prediction was larger text gains than with an MS MARCO second opinion. WPD gains +2.39 [+1.61, +3.20] on ArguAna, +2.64 [+1.42, +3.90] on FiQA, +4.65 [+3.90, +5.35] on NQ, +1.01 [+0.79, +1.22] on Quora and -0.33 [-0.74, +0.40] on SciDocs; the interval lies above zero in 4/5 datasets, and WPD beats the tuned candidate policy in 3/5. The text datasets had been analysed before, so this is a mechanism test rather than a confirmation.

Cost-matched comparison (post hoc). Is a small primary with WPD a better use of a large verifier than fusing the large verifier alone? On the development populations we compared, on identical queries, folds and seeds, each 2B primary alone, each 8B verifier alone (fused with the retriever at its own cross-fitted constant), the primary with WPD using the 8B verifier as second opinion, and the primary with full rescoring by both (Figure 15, Table 41). WPD calls the 8B verifier 1.6–2.1 times per query instead of 20 (8–10% of its calls, or 33–35% of the 8B-alone compute when every call is weighted by parameter count, a rough proxy that ignores differences in image tokens). Its R@1 is within the 95% interval of the 8B verifier alone in 5 of 8 cells, higher in 2 (Qwen3-VL-2B with InternVL3-8B on FashionIQ, +0.98 [+0.25, +1.68]; Qwen3-VL-2B with Qwen3-VL-8B on FashionIQ, +1.28 [+0.57, +2.00]) and lower in 1 (InternVL3-2B with InternVL3-8B on CIRR, -1.92 [-3.16, -0.67]). Full rescoring by both verifiers is the most accurate option on CIRR, at the highest

cost. WPD therefore trades a small, dataset-dependent difference in accuracy for a large cut in large-verifier calls; it does not dominate the larger verifier.

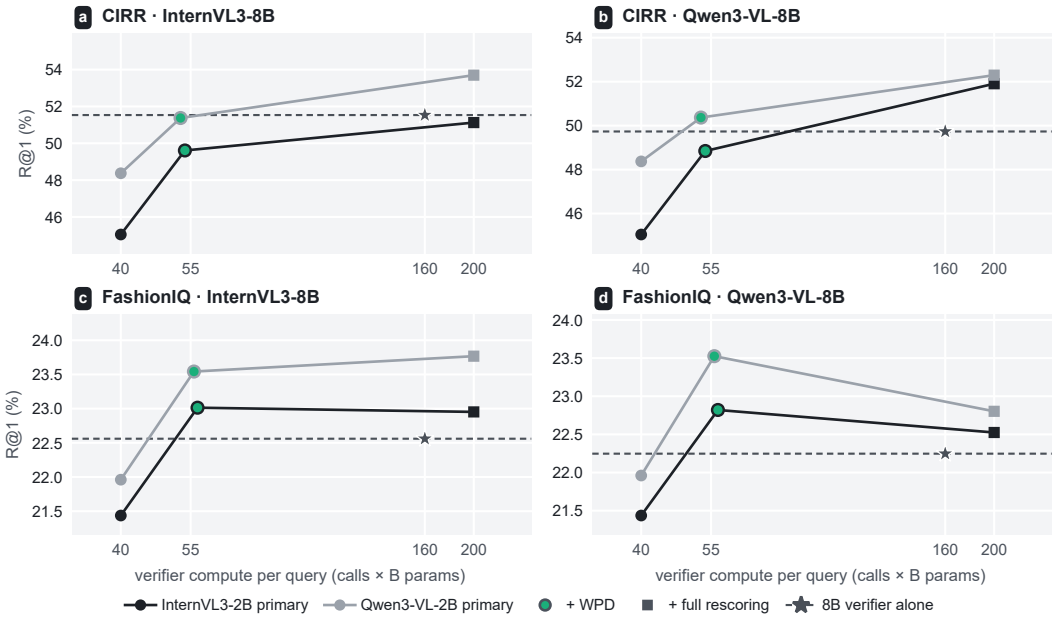

Figure 15: **WPD against the larger verifier alone, at matched cost.** Development R@1 against verifier compute per query (calls times billions of parameters, log scale). Lines connect a 2B primary alone, the primary with WPD using the 8B verifier as second opinion, and the primary with full rescoring by both verifiers; the star and dashed line mark the 8B verifier fused alone with the retriever. Post hoc; Table 41 gives the intervals.

Results. Table 38 lists every development pair, Table 39 every confirmation and mechanism pair, and Table 40 pools the pairs within each population. The gain depends strongly on the second opinion: the recovered share of headroom ranges over 7–39% on CIRR and 6–23% on FashionIQ. With the two MS MARCO cross-encoders, WPD gains +1.57 [+1.01, +2.13] points on NQ and +0.28 [+0.13, +0.43] on Quora, and −0.35 to +0.66 on the other three datasets; D1 holds in 2/5 datasets and D2 in 1/5.

Table 38: Winner-path duel on the development populations, every primary and second-opinion verifier (R@1 points, five seeds, paired 95% intervals). Recovery is the gain as a percentage of headroom; calls are second-verifier calls per query. Full rescoring averages both verifiers over all 20 candidates (20 calls).

Data	Primary	Second opinion	WPD gain	Rec. (%)	No second opinion	Full rescoring	Calls
CIRR	InternVL3-2B	InternVL3-8B	+4.56 [+3.67, +5.48]	39	+0.85 [+0.31, +1.39]	+6.08 [+4.92, +7.26]	1.70
CIRR	InternVL3-2B	Qwen3-VL-4B	+4.52 [+3.58, +5.48]	38	+0.85 [+0.31, +1.39]	+6.84 [+5.64, +8.05]	1.70
CIRR	InternVL3-2B	Qwen3-VL-8B	+3.80 [+2.86, +4.73]	32	+0.85 [+0.31, +1.39]	+6.86 [+5.59, +8.16]	1.70
CIRR	InternVL3-2B	Qwen3-VL-2B	+3.79 [+2.93, +4.65]	32	+0.85 [+0.31, +1.39]	+5.64 [+4.52, +6.78]	1.70
CIRR	InternVL3-2B	LLaVA-OV-7B	+3.46 [+2.59, +4.36]	29	+0.85 [+0.31, +1.39]	+4.72 [+3.56, +5.88]	1.70
CIRR	InternVL3-2B	SmolVLM2-2.2B	+2.96 [+2.11, +3.81]	25	+0.85 [+0.31, +1.39]	+3.47 [+2.37, +4.61]	1.70
CIRR	InternVL3-2B	Phi-3.5-Vision	+2.91 [+2.09, +3.72]	25	+0.85 [+0.31, +1.39]	+3.53 [+2.29, +4.77]	1.70
CIRR	InternVL3-2B	Gemma-3-4B	+2.37 [+1.56, +3.17]	20	+0.85 [+0.31, +1.39]	+2.18 [+1.08, +3.29]	1.70
CIRR	InternVL3-2B	Idefics3-8B	+2.15 [+1.26, +3.04]	18	+0.85 [+0.31, +1.39]	+2.79 [+1.62, +3.98]	1.70
CIRR	InternVL3-2B	LLaVA-OV-0.5B	+0.78 [+0.01, +1.53]	7	+0.85 [+0.31, +1.39]	+0.30 [−0.87, +1.44]	1.70
CIRR	Qwen3-VL-2B	InternVL3-8B	+3.00 [+2.20, +3.80]	29	+0.05 [−0.40, +0.52]	+5.33 [+4.21, +6.41]	1.57
CIRR	Qwen3-VL-2B	Qwen3-VL-4B	+2.52 [+1.76, +3.29]	25	+0.05 [−0.40, +0.52]	+4.41 [+3.29, +5.52]	1.57
CIRR	Qwen3-VL-2B	LLaVA-OV-7B	+2.02 [+1.22, +2.82]	20	+0.05 [−0.40, +0.52]	+3.89 [+2.70, +5.06]	1.57
CIRR	Qwen3-VL-2B	Qwen3-VL-8B	+1.99 [+1.20, +2.79]	20	+0.05 [−0.40, +0.52]	+3.92 [+2.80, +5.12]	1.57
CIRR	Qwen3-VL-2B	Idefics3-8B	+1.86 [+1.06, +2.66]	18	+0.05 [−0.40, +0.52]	+2.29 [+1.07, +3.49]	1.57
CIRR	Qwen3-VL-2B	SmolVLM2-2.2B	+1.79 [+1.02, +2.57]	18	+0.05 [−0.40, +0.52]	+2.70 [+1.51, +3.87]	1.57
CIRR	Qwen3-VL-2B	InternVL3-2B	+1.49 [+0.76, +2.22]	15	+0.05 [−0.40, +0.52]	+2.32 [+1.21, +3.46]	1.57
CIRR	Qwen3-VL-2B	Phi-3.5-Vision	+1.37 [+0.68, +2.04]	13	+0.05 [−0.40, +0.52]	+0.83 [−0.33, +1.94]	1.57
CIRR	Qwen3-VL-2B	LLaVA-OV-0.5B	+1.01 [+0.38, +1.65]	10	+0.05 [−0.40, +0.52]	−0.82 [−2.04, +0.36]	1.57
CIRR	Qwen3-VL-2B	Gemma-3-4B	+0.83 [+0.21, +1.46]	8	+0.05 [−0.40, +0.52]	−0.26 [−1.43, +0.91]	1.57
FashionIQ	InternVL3-2B	Qwen3-VL-4B	+1.73 [+1.27, +2.21]	23	+0.26 [−0.00, +0.53]	+1.94 [+1.40, +2.51]	2.09

Data	Primary	Second opinion	WPD gain	Rec. (%)	No second opinion	Full rescoring	Calls
FashionIQ	InternVL3-2B	InternVL3-8B	+1.58 [+1.06, +2.08]	21	+0.26 [−0.00, +0.53]	+1.52 [+1.02, +2.03]	2.09
FashionIQ	InternVL3-2B	Qwen3-VL-8B	+1.38 [+0.86, +1.91]	18	+0.26 [−0.00, +0.53]	+1.09 [+0.50, +1.68]	2.09
FashionIQ	InternVL3-2B	Qwen3-VL-2B	+1.22 [+0.76, +1.67]	16	+0.26 [−0.00, +0.53]	+1.43 [+0.88, +1.99]	2.09
FashionIQ	InternVL3-2B	Qwen3-VL-2B-LoRA	+1.00 [+0.59, +1.42]	13	+0.26 [−0.00, +0.53]	+1.01 [+0.53, +1.51]	2.09
FashionIQ	InternVL3-2B	Idefics3-8B	+0.96 [+0.60, +1.34]	13	+0.26 [−0.00, +0.53]	+0.51 [−0.04, +1.06]	2.09
FashionIQ	InternVL3-2B	Phi-3.5-Vision	+0.90 [+0.45, +1.35]	12	+0.26 [−0.00, +0.53]	+0.97 [+0.42, +1.53]	2.09
FashionIQ	InternVL3-2B	LLaVA-OV-0.5B	+0.62 [+0.29, +0.96]	8	+0.26 [−0.00, +0.53]	−0.19 [−0.72, +0.36]	2.09
FashionIQ	InternVL3-2B	SmolVLM2-2.2B	+0.45 [+0.11, +0.79]	6	+0.26 [−0.00, +0.53]	+0.06 [−0.45, +0.59]	2.09
FashionIQ	Qwen3-VL-2B	Qwen3-VL-4B	+1.63 [+1.11, +2.15]	20	+0.48 [+0.15, +0.82]	+1.64 [+1.03, +2.25]	1.98
FashionIQ	Qwen3-VL-2B	InternVL3-8B	+1.58 [+1.00, +2.17]	20	+0.48 [+0.15, +0.82]	+1.81 [+1.19, +2.43]	1.98
FashionIQ	Qwen3-VL-2B	Qwen3-VL-8B	+1.56 [+1.03, +2.12]	20	+0.48 [+0.15, +0.82]	+0.84 [+0.23, +1.46]	1.98
FashionIQ	Qwen3-VL-2B	Qwen3-VL-2B-LoRA	+1.46 [+0.97, +1.94]	18	+0.48 [+0.15, +0.82]	+1.50 [+0.90, +2.09]	1.98
FashionIQ	Qwen3-VL-2B	Idefics3-8B	+1.42 [+0.86, +1.98]	18	+0.48 [+0.15, +0.82]	+1.41 [+0.72, +2.11]	1.98
FashionIQ	Qwen3-VL-2B	SmolVLM2-2.2B	+1.11 [+0.65, +1.59]	14	+0.48 [+0.15, +0.82]	+0.68 [+0.04, +1.32]	1.98
FashionIQ	Qwen3-VL-2B	InternVL3-2B	+1.10 [+0.58, +1.63]	14	+0.48 [+0.15, +0.82]	+0.90 [+0.31, +1.52]	1.98
FashionIQ	Qwen3-VL-2B	Phi-3.5-Vision	+1.00 [+0.56, +1.45]	12	+0.48 [+0.15, +0.82]	+0.46 [−0.14, +1.08]	1.98
FashionIQ	Qwen3-VL-2B	LLaVA-OV-0.5B	+0.93 [+0.41, +1.43]	12	+0.48 [+0.15, +0.82]	+0.04 [−0.61, +0.71]	1.98

Table 39: Winner-path duel on the confirmation and mechanism populations, every primary and second-opinion pair (R@1 points, five seeds, paired 95% intervals). Recovery is the gain as a percentage of headroom; transfer applies the development model unchanged; calls are second-verifier calls per query for this primary; for the self-duel, pairwise prompts per query.

Population	Data	Primary	Second opinion	WPD gain	Rec. (%)	Transfer	Calls
Replication	CIRR	InternVL3-2B	LLaVA-OV-0.5B	+1.39 [+0.75, +2.03]	12	+1.13 [+0.49, +1.79]	1.70
Replication	CIRR	InternVL3-2B	Phi-3.5-Vision	+2.30 [+1.47, +3.13]	20	+2.78 [+1.87, +3.70]	1.70
Replication	CIRR	InternVL3-2B	Qwen3-VL-2B	+4.16 [+3.25, +5.09]	36	+3.60 [+2.79, +4.44]	1.70
Replication	CIRR	Qwen3-VL-2B	InternVL3-2B	+2.27 [+1.46, +3.08]	21	+2.08 [+1.26, +2.89]	1.54
Replication	CIRR	Qwen3-VL-2B	LLaVA-OV-0.5B	+0.96 [+0.36, +1.55]	9	+1.22 [+0.51, +1.94]	1.54
Replication	CIRR	Qwen3-VL-2B	Phi-3.5-Vision	+1.95 [+1.15, +2.75]	18	+2.17 [+1.34, +3.01]	1.54
Replication	FashionIQ	InternVL3-2B	LLaVA-OV-0.5B	+0.58 [+0.22, +0.95]	7	+0.86 [+0.51, +1.23]	2.04
Replication	FashionIQ	InternVL3-2B	Phi-3.5-Vision	+1.08 [+0.65, +1.54]	14	+1.70 [+1.06, +2.36]	2.04
Replication	FashionIQ	InternVL3-2B	Qwen3-VL-2B	+1.69 [+1.19, +2.19]	21	+2.12 [+1.48, +2.75]	2.04
Replication	FashionIQ	Qwen3-VL-2B	InternVL3-2B	+1.26 [+0.71, +1.79]	14	+1.23 [+0.62, +1.87]	1.86
Replication	FashionIQ	Qwen3-VL-2B	LLaVA-OV-0.5B	+0.99 [+0.52, +1.45]	11	+0.84 [+0.35, +1.32]	1.86
Replication	FashionIQ	Qwen3-VL-2B	Phi-3.5-Vision	+1.35 [+0.79, +1.90]	15	+1.37 [+0.77, +1.98]	1.86
Prospective	CIRR	InternVL3-2B	InternVL3-8B	+4.51 [+3.58, +5.47]	40	+4.34 [+3.40, +5.30]	1.70
Prospective	CIRR	Qwen3-VL-2B	InternVL3-8B	+3.46 [+2.60, +4.33]	32	+3.24 [+2.28, +4.17]	1.54
Prospective	FashionIQ	InternVL3-2B	InternVL3-8B	+1.98 [+1.39, +2.56]	25	+2.34 [+1.70, +3.00]	2.04
Prospective	FashionIQ	Qwen3-VL-2B	InternVL3-8B	+1.72 [+1.12, +2.32]	20	+1.70 [+1.10, +2.29]	1.86
Self-duel, dev.	CIRR	InternVL3-2B	same model	+1.67 [+0.94, +2.39]	14	–	2.00
Self-duel, dev.	CIRR	Qwen3-VL-2B	same model	+0.17 [−0.34, +0.67]	2	–	1.82
Self-duel, dev.	FashionIQ	InternVL3-2B	same model	+0.55 [+0.19, +0.91]	7	–	2.53
Self-duel, dev.	FashionIQ	Qwen3-VL-2B	same model	+0.61 [+0.22, +1.00]	8	–	2.35
Self-duel, repl.	CIRR	InternVL3-2B	same model	+1.69 [+1.08, +2.29]	15	+2.05 [+1.38, +2.72]	2.01
Self-duel, repl.	CIRR	Qwen3-VL-2B	same model	+0.60 [+0.13, +1.08]	6	+0.52 [−0.21, +1.23]	1.77
Self-duel, repl.	FashionIQ	InternVL3-2B	same model	+0.68 [+0.27, +1.09]	9	+0.71 [+0.40, +1.04]	2.44
Self-duel, repl.	FashionIQ	Qwen3-VL-2B	same model	+1.12 [+0.61, +1.64]	13	+1.23 [+0.77, +1.72]	2.20
Text	ArguAna	BGE-reranker-base	MiniLM-L6	−0.36 [−0.92, +0.21]	−4	–	1.72
Text	ArguAna	MiniLM-L6	BGE-reranker-base	+0.11 [−0.14, +0.37]	1	–	1.40
Text	FiQA	BGE-reranker-base	MiniLM-L6	−0.56 [−1.33, +0.15]	−8	–	1.57
Text	FiQA	MiniLM-L6	BGE-reranker-base	+1.88 [+0.65, +3.18]	22	–	1.37
Text	NQ	BGE-reranker-base	MiniLM-L6	+1.01 [+0.39, +1.62]	9	–	1.61
Text	NQ	MiniLM-L6	BGE-reranker-base	+2.13 [+1.20, +3.08]	19	–	1.28
Text	Quora	BGE-reranker-base	MiniLM-L6	+0.32 [+0.13, +0.51]	7	–	0.99
Text	Quora	MiniLM-L6	BGE-reranker-base	+0.25 [+0.03, +0.46]	5	–	0.62
Text	SciDocs	BGE-reranker-base	MiniLM-L6	−0.56 [−1.67, +0.00]	−6	–	1.65
Text	SciDocs	MiniLM-L6	BGE-reranker-base	−0.14 [−0.72, +1.74]	−2	–	1.49
Text, v2-m3	ArguAna	BGE-reranker-base	BGE-reranker-v2-m3	+2.38 [+1.52, +3.31]	26	–	1.72
Text, v2-m3	ArguAna	MiniLM-L6	BGE-reranker-v2-m3	+2.40 [+1.37, +3.46]	28	–	1.40
Text, v2-m3	FiQA	BGE-reranker-base	BGE-reranker-v2-m3	+1.39 [−0.19, +2.99]	19	–	1.57
Text, v2-m3	FiQA	MiniLM-L6	BGE-reranker-v2-m3	+3.89 [+2.38, +5.56]	45	–	1.37
Text, v2-m3	NQ	BGE-reranker-base	BGE-reranker-v2-m3	+4.43 [+3.51, +5.38]	37	–	1.61
Text, v2-m3	NQ	MiniLM-L6	BGE-reranker-v2-m3	+4.87 [+3.91, +5.80]	42	–	1.28
Text, v2-m3	Quora	BGE-reranker-base	BGE-reranker-v2-m3	+0.81 [+0.54, +1.09]	18	–	0.99
Text, v2-m3	Quora	MiniLM-L6	BGE-reranker-v2-m3	+1.20 [+0.90, +1.50]	24	–	0.62
Text, v2-m3	SciDocs	BGE-reranker-base	BGE-reranker-v2-m3	−0.62 [−1.83, −0.10]	−7	–	1.65
Text, v2-m3	SciDocs	MiniLM-L6	BGE-reranker-v2-m3	−0.04 [−0.49, +1.41]	−1	–	1.49

Table 40: Winner-path duel pooled over primary/second-opinion pairs within each population (R@1 points, paired 95% intervals). WPD is the protocol method; the abstaining variant may never switch. The last two columns compare WPD with the tuned candidate policy and give the frozen transfer from development.

Population	Data	Pairs	Headroom	WPD gain	Abstaining	WPD − candidate	Transfer
Development	CIRR	20	11.0	+2.46 [+1.98, +2.93]	+2.46 [+1.98, +2.93]	+1.61 [+1.21, +2.00]	–
Development	FashionIQ	18	7.8	+1.20 [+0.91, +1.50]	+1.20 [+0.91, +1.50]	+0.38 [+0.09, +0.66]	–
Replication	CIRR	6	11.1	+2.17 [+1.71, +2.65]	+2.17 [+1.71, +2.65]	+1.17 [+0.77, +1.57]	+2.16 [+1.69, +2.65]
Replication	FashionIQ	6	8.3	+1.16 [+0.84, +1.49]	+1.16 [+0.84, +1.49]	−0.20 [−0.51, +0.10]	+1.35 [+1.00, +1.72]
Prospective (InternVL3-8B)	CIRR	2	11.1	+3.99 [+3.30, +4.68]	+3.99 [+3.30, +4.68]	+2.98 [+2.34, +3.61]	+3.79 [+3.08, +4.50]
Prospective (InternVL3-8B)	FashionIQ	2	8.3	+1.85 [+1.38, +2.32]	+1.85 [+1.38, +2.32]	+0.49 [+0.04, +0.94]	+2.02 [+1.53, +2.50]
Self-duel, development	CIRR	2	11.0	+0.92 [+0.46, +1.37]	+0.92 [+0.46, +1.37]	+0.07 [−0.34, +0.46]	–
Self-duel, development	FashionIQ	2	7.8	+0.58 [+0.30, +0.85]	+0.58 [+0.30, +0.85]	−0.25 [−0.55, +0.06]	–
Self-duel, replication	CIRR	2	11.1	+1.14 [+0.74, +1.55]	+1.14 [+0.74, +1.55]	+0.14 [−0.24, +0.51]	+1.28 [+0.78, +1.78]
Self-duel, replication	FashionIQ	2	8.3	+0.90 [+0.56, +1.25]	+0.90 [+0.56, +1.25]	−0.46 [−0.79, −0.13]	+0.97 [+0.67, +1.28]
Text	ArguAna	2	8.8	−0.12 [−0.45, +0.21]	−0.14 [−0.48, +0.19]	−2.43 [−3.14, −1.74]	–
Text	FiQA	2	8.0	+0.66 [−0.09, +1.44]	+0.71 [−0.02, +1.45]	+0.71 [−0.19, +1.62]	–
Text	NQ	2	11.7	+1.57 [+1.01, +2.13]	+1.57 [+1.01, +2.13]	+1.04 [+0.50, +1.57]	–
Text	Quora	2	4.8	+0.28 [+0.13, +0.43]	+0.28 [+0.13, +0.43]	+0.09 [−0.08, +0.25]	–
Text	SciDocs	2	8.6	−0.35 [−0.71, +0.60]	−0.35 [−0.71, +0.60]	−0.86 [−2.42, −0.31]	–
Text, BGE-v2-m3 2nd opinion	ArguAna	2	8.8	+2.39 [+1.61, +3.20]	+2.39 [+1.61, +3.20]	+0.09 [−0.71, +0.88]	–
Text, BGE-v2-m3 2nd opinion	FiQA	2	8.0	+2.64 [+1.42, +3.90]	+2.64 [+1.42, +3.90]	+2.69 [+1.59, +3.83]	–
Text, BGE-v2-m3 2nd opinion	NQ	2	11.7	+4.65 [+3.90, +5.35]	+4.65 [+3.90, +5.35]	+4.12 [+3.43, +4.80]	–
Text, BGE-v2-m3 2nd opinion	Quora	2	4.8	+1.01 [+0.79, +1.22]	+1.01 [+0.79, +1.22]	+0.81 [+0.60, +1.02]	–
Text, BGE-v2-m3 2nd opinion	SciDocs	2	8.6	−0.33 [−0.74, +0.40]	−0.33 [−0.74, +0.40]	−0.84 [−2.57, −0.28]	–

Table 41: Cost-matched comparison on the development populations (post hoc). R@1 (%) of the small primary alone, the large verifier alone, the primary with WPD using the large verifier as second opinion, and the primary with full rescoring by both verifiers. All use the same queries, folds and seeds, and each fuses with the retriever at a cross-fitted constant weight. Large calls: large-verifier calls per query for WPD (20 for the other two large-verifier rows). The last column is paired 95% intervals.

Data	Primary	Large verifier	Primary	Large	Primary + WPD	Full	Large calls	WPD − large alone
CIRR	InternVL3-2B	InternVL3-8B	45.05	51.53	49.61	51.12	1.70	−1.92 [−3.16, −0.67]
CIRR	InternVL3-2B	Qwen3-VL-8B	45.05	49.73	48.85	51.91	1.70	−0.89 [−2.18, +0.38]
CIRR	Qwen3-VL-2B	InternVL3-8B	48.37	51.53	51.37	53.70	1.57	−0.16 [−1.42, +1.12]
CIRR	Qwen3-VL-2B	Qwen3-VL-8B	48.37	49.73	50.37	52.30	1.57	+0.63 [−0.69, +1.93]
FashionIQ	InternVL3-2B	InternVL3-8B	21.44	22.56	23.01	22.95	2.09	+0.45 [−0.15, +1.06]
FashionIQ	InternVL3-2B	Qwen3-VL-8B	21.44	22.25	22.82	22.52	2.09	+0.57 [−0.10, +1.24]
FashionIQ	Qwen3-VL-2B	InternVL3-8B	21.96	22.56	23.54	23.77	1.98	+0.98 [+0.25, +1.68]
FashionIQ	Qwen3-VL-2B	Qwen3-VL-8B	21.96	22.25	23.53	22.80	1.98	+1.28 [+0.57, +2.00]