PARTIAL RERANKING: ACTION STRUCTURE AND CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A reranker can improve average retrieval while damaging queries already served well by a baseline. Disjoint permutation cycles allow favorable and unfavorable changes to be decided separately, using the same predicted relevance and proposed ranking. This paper studies the value of that choice under uncertain relevance. On a locked Flickr30k-derived gallery, separately calibrated partial replacement increases signed reciprocal-rank gain by 0.000930 for BLIP and 0.002255 for ALBEF; both adjusted intervals are positive, and only ALBEF resolves the specified 0.001 practical margin. Same-price contrasts retain the benefit, as does a finite calibration-optimal whole-query comparator. A separate split-calibration selection-and-guard procedure meets practical criteria in four setting–model cells, while COCO BLIP remains sub-practical and VizWiz BLIP loses gain. Equal-cardinality interventions further show that which changes are bound together matters beyond action count. Variation in cycle switching prices, weighted by predicted loss, characterizes predicted binding cost across prices, but reducing that cost does not ensure better calibrated gain. At a common observed loss, label-aware optima over fixed score menus favor partial actions in all six cells; these hindsight gains do not establish deployable improvement. Adding predicted opportunity to a fixed predictor worsens error in all four model–price settings. The evidence separates structural decision value from outcome prediction and identifies calibration as a consequential limit on partial reranking.

1 INTRODUCTION

Retrieval systems often combine an inexpensive ranking with a more expressive reranker. A higher average score does not imply that every replacement is beneficial. A query with an already correct first result can lose substantially when a new model changes that result. This is a longstanding concern in risk-sensitive retrieval (Collins-Thompson, 2009; Wang et al., 2012; Dinçer et al., 2016). The practical decision is therefore not only which ranking scores best, but also which changes should be applied.

A common decision is binary: keep the baseline ranking or accept the proposed ranking for the whole query. A ranking, however, is a structured object. Its permutation relative to the baseline decomposes into disjoint cycles; any subset of those cycles remains a valid permutation. This creates a precise intermediate action space between doing nothing and replacing everything. Favorable and unfavorable changes need not be accepted together.

Figure 1 isolates joint decisions for fixed cycles. The additional choice matters before the target is known: different possible targets can favor different subsets of changes. Once a single target is known, only its containing cycle can change its rank, and the best partial and whole-query actions attain the same label-aware oracle value. The question is therefore whether separate choices improve decisions under uncertain relevance. Predicted relevance can assign the wrong sign to a change, and finite calibration can select different prices for different action spaces. Even accurate probabilities leave no additional predicted value when all cycle utilities agree in sign.

Table 1 summarizes the central evidence. The two fixed model pipelines both benefit on the locked gallery, but the stronger practical conclusion applies only to ALBEF. Neither loss interval establishes lower loss. A price-matched analysis then shows positive action effects for both models, while the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

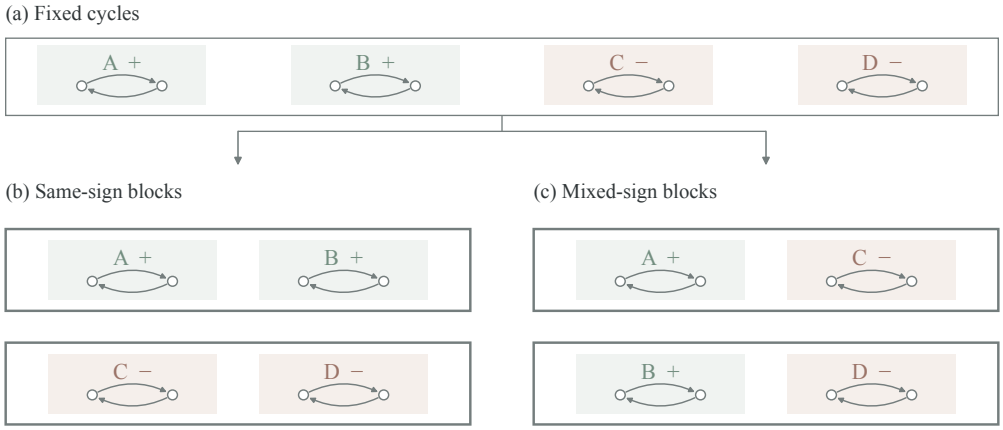

Figure 1: **Binding changes which favorable and unfavorable moves must act together.** (a) A–D denote fixed disjoint cycles; colors and signs encode their predicted utility signs at one price, not magnitudes. (b,c) Each enclosure is one joint apply-or-keep decision. Same-sign and mixed-sign blocks admit four feasible actions each but different subsets. Mixed-block sums and selected outcomes are unspecified. This schematic encodes no measured performance or universal grouping preference; Appendix A gives a numerical example.

Table 1: Main result and its boundary. Locked Flickr30k-derived FINAL: 23,966 image groups, 119,022 queries. Differences are CYCLE − QUERY at separately calibrated prices; all numbers are $\times 10^3$. Intervals adjust the twelve-coordinate primary family. The gain margin is 1.0 and the added-loss margin is 0.2 on this scale.

	Signed gain ΔG		Positive loss ΔL		Gain margin
Model	Mean	Adjusted interval	Mean	Adjusted interval	1.0
BLIP	0.930	$[0.662, 1.192]$	0.019	$[-0.061, 0.104]$	Unresolved
ALBEF	2.255	$[1.927, 2.571]$	0.032	$[-0.051, 0.119]$	Exceeded

price effect is negative for BLIP and positive for ALBEF. This distinction explains why partial replacement can help at the same price yet underperform after separate calibration elsewhere.

The contribution is a controlled account of three links: separate choices can improve retrieval at a common harm price; binding changes utility even at equal action count; and predicted structural value need not survive either outcome prediction or separate calibration. The first link is tested in a locked evaluation, the second by fixed-posterior grouping interventions, and the third by a fixed predictor and a source-fixed minimum-cost pairing. Ordinary grouping controls delimit the useful scope of similarity-based pairing. Grouping, stronger whole-query controls and split-calibration follow-ups are exposed secondary evidence; the original locked comparison remains primary.

2 RELATED WORK

Vision–language retrieval provides both efficient similarity scores and richer matching scores. Contrastive representations (Radford et al., 2021), fine-grained interaction (Yao et al., 2021), and compositional matching (Jiang et al., 2024) improve different parts of this pipeline. BLIP and ALBEF combine contrastive retrieval with image–text matching (Li et al., 2022a; 2021). The present experiments hold their learned parameters and proposed rankings fixed; the intervention is the structure of the final action, not a new representation or reranker.

Risk-sensitive ranking explicitly accounts for degradation relative to a baseline (Wang et al., 2012; Dinçer et al., 2016). Selective prediction and learning to defer instead choose whether to trust a prediction or another decision maker (Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020). Whole-query replacement resembles such a binary gate. Partial permutation replacement differs be-

cause a valid action is assembled within one query, subject to ranking structure. Arbitrarily accepting candidate-level moves can create collisions or invalid rankings. Assignment and constrained-ranking methods provide relevant mathematical tools (Kuhn, 1955; Celis et al., 2018); the cycle construction here is elementary and is not presented as a new general assignment algorithm.

Calibration must also be matched to the decision. Probability calibration (Guo et al., 2017), decision calibration (Zhao et al., 2021), predict-then-optimize (Elmachtoub & Grigas, 2022), and task-based learning (Donti et al., 2017) all distinguish predictive accuracy from downstream utility. Set-valued classification (Sadinle et al., 2019), conformal risk control (Angelopoulos et al., 2024), and learn-then-test (Angelopoulos et al., 2025) supply complementary risk-calibration frameworks. Decision-theoretic treatments further make the action and loss explicit (Kiyani et al., 2025). The finite-sample assumptions of those frameworks do not turn the empirical results below into guarantees under dataset shift (Barber et al., 2023). Candidate-pruning error control (Li et al., 2022b) is also distinct: preserving shortlist recall does not by itself control reciprocal-rank harm after reordering.

3 A Fixed Ranking, Several Action Spaces

For a query x, let $r_0(x, y)$ and $r_a(x, y)$ be the full-gallery ranks of the target image y under the baseline and action a. Define

$$d_a(x, y) = \frac{1}{r_a(x, y)} - \frac{1}{r_0(x, y)}, \qquad G(a) = \mathbb{E}[d_a], \quad L(a) = \mathbb{E}[[-d_a]_+]. \tag{1}$$

Signed gain satisfies $G = G^+ - L$, where $G^+ = \mathbb{E}[[d_a]_+]$. Harm frequency $h(a) = \Pr(d_a < 0)$ measures a different quantity from loss magnitude. All reported means first average captions within an image, then average images. Outside-shortlist queries remain in the denominator and retain the unchanged full-gallery rank.

The reranker proposes a permutation of the baseline top $m = 32$ candidates. Its disjoint nontrivial cycles are $c_1, \ldots, c_K$. Applying a subset of these cycles changes disjoint positions and therefore remains a valid ranking. QUERY accepts all cycles or none; CYCLE can accept each cycle separately. A grouped policy partitions the same cycles into blocks and accepts or rejects each block. This permits a same-cardinality intervention without changing the proposer.

Let $\hat{q}_x(y)$ denote a fixed predicted distribution conditional on the target being inside the shortlist. At a nonnegative harm price λ, each cycle has predicted utility

$$u_c(x; \lambda) = \sum_{y \in c} \hat{q}_x(y) \left[d_c(x, y) - \lambda [-d_c(x, y)]_+ \right]. \tag{2}$$

The empty action has utility zero. CYCLE accepts cycles with positive utility; QUERY accepts when their sum is positive. Fixed deterministic rules handle equality. Both actions use the same scores, probabilities, and price. The population quantity $J = G - \lambda L$ is useful for this matched-price comparison, but a higher J need not mean both higher G and lower L.

The predicted opportunity for separating cycles is

$$\widehat{A}(x; \lambda) = \sum_c [u_c]_+ - [\sum_c u_c]_+ = \frac{1}{2} \left(\sum_c |u_c| - \left| \sum_c u_c \right| \right) \geq 0. \tag{3}$$

The opportunity is zero when cycle utilities share a sign and positive when whole-query binding cancels favorable and unfavorable changes. Equation 3 is an accounting identity for the predicted decision problem, not a guarantee about the realized target.

Cardinality is not geometry. Figure 1(b,c) separates action count from action composition. Section 6 pairs the same cycles using fixed rules, retaining the same odd singleton. Equal block counts do not imply identical feasible sets, accepted-query fractions, or moved-candidate counts. A richer action set's optimal predicted value also differs from the performance of an imperfect predictor selecting from that set (Boyd & Vandenberghe, 2004).

4 CALIBRATION AND A LOCKED EVALUATION

Operating points. For a specified policy path, CAL selects the most permissive operating point satisfying the empirical corrected loss criterion. With n image groups and a top-m change, the loss is bounded by $B_m = 1 - 1/m$. The implemented criterion is

$$\widehat{R}^{+}(a) = \frac{\sum_{i=1}^{n} L_i(a) + B_m}{n+1} \leq \rho. \tag{4}$$

The principal comparison uses $\rho = 0.002$, full CAL of 2,500 image groups, and the original posterior temperature. On fixed QUERY/CYCLE paths, increasing price only removes a target's fixed proposed move, so positive reciprocal-rank loss is non-increasing. Strict ties are rejected and the baseline endpoint applies no moves. Exact calibration and rank checks verify implementation, not exchangeability or arbitrary comparator paths (Appendix B). The correction is not a confidence interval, and reported bootstrap intervals do not establish distribution-free safety. In particular, a shortlist-conditional distribution does not predict whether the target is in the shortlist.

Separate actions from prices. Let M_{ab} be either G or L for action space $a \in \{Q, C\}$ at the price selected by $b \in \{Q, C\}$. The symmetric two-factor decomposition is

$$\begin{aligned} A_M &= \tfrac{1}{2}[(M_{CQ} - M_{QQ}) + (M_{CC} - M_{QC})], \\ P_M &= \tfrac{1}{2}[(M_{QC} - M_{QQ}) + (M_{CC} - M_{CQ})], \qquad T_M = M_{CC} - M_{QQ} = A_M + P_M. \end{aligned} \tag{5}$$

Both same-price contrasts are retained. This is an exact decomposition of four measured cells, not a uniquely identified causal mediation analysis.

Models, data, and roles. The locked confirmation uses BLIP and ALBEF, three fixed source-fitted probability heads per model, and a Flickr30k TRAIN-derived gallery (Young et al., 2014) of 28,966 images. CAL A and B contain 1,250 image groups each, the diagnostic subset contains 2,500, and FINAL contains 23,966 groups with 119,022 caption queries. The full gallery remains fixed across roles. The diagnostic subset fixes the analysis before FINAL; FINAL does not select temperatures, heads, prices, or grouping variants. Three heads measure sensitivity within a fixed pipeline and are averaged within each image, not counted as independent replications.

The gallery excludes overlap detectable in the reconstructed local evaluation history. Incomplete older exposure records and pretrained-model data prevent a global-unseen claim. Two additional settings delimit transport: a VizWiz-Captions VAL evaluation (Gurari et al., 2020) and a fixed 30,000-image COCO TRAIN gallery (Lin et al., 2014). COCO has known training-dataset overlap and is a large-gallery stress test, not independent external validation. These scopes are kept distinct throughout.

Controls and uncertainty. The locked comparison set contains 424 rows per model: 324 action–price cross cells, 99 controls, and the unchanged baseline. Comparators include margin and source-trained whole-query gates, augmented gates, fitted randomized mixtures, and a least-ambiguous-set-inspired protected-ranking control. The protected control uses separate set and risk calibration roles. All rules are fixed before FINAL; higher-gain controls with higher loss are not declared inferior or superior without a matched criterion.

The primary family consists of twelve coordinates: two models, three components in Equation 5, and two metrics, G and L. Paired image-group bootstrap intervals use 3,999 draws and Bonferroni-adjusted percentile endpoints (Efron, 1979). Captions and heads are not resampled as independent observations. Intervals condition on the gallery, fitted heads, and calibration choices. The specified practical margins are 0.001 for added signed gain and 0.0002 for added loss. Secondary analyses have their own stated families rather than silently expanding the primary one.

5 PARTIAL ACTIONS HELP, BUT THE PRICE CAN REVERSE THE RESULT

Figure 2 shows that the same-price action effect is positive for both pipelines. For BLIP, the action contribution is 0.001237, partially offset by a price contribution of -0.000307. For ALBEF, the corresponding values are 0.001874 and 0.000381. Thus, positive total gain is not merely a consequence

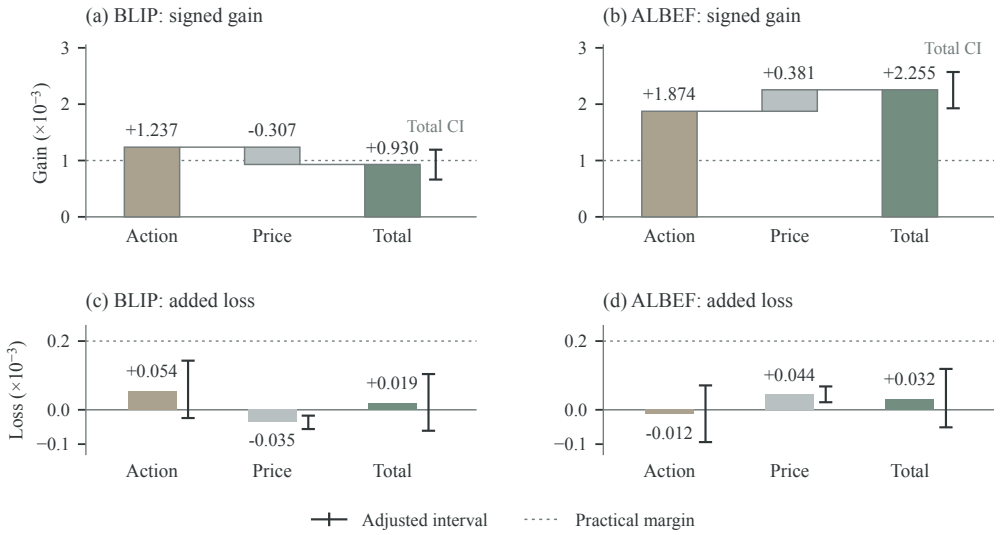

Figure 2: Primary Flickr evaluation: action, price, and total effects. Upper waterfalls add action and price point estimates; separate total bars show their sum. Lower bars show added-loss effects. Capped whiskers beside bars show adjusted image-group-bootstrap intervals from the twelve-coordinate family: total gain above and all loss components below. No intervals are summed. Dotted lines mark gain/loss margins of $1.0/0.2$ ($\times 10^{-3}$). Price opposes the BLIP action effect and reinforces ALBEF, without establishing lower loss. Appendix C retains all twelve component means and intervals.

of choosing a more aggressive price for partial replacement. At either original price, the descriptive gain contrast remains positive.

Both price origins separately have positive gain contrasts: approximately 0.001236–0.001238 for BLIP and 0.001870–0.001878 for ALBEF. The descriptive intervals appear in Appendix C, Table 9. Averaging therefore hides no unfavorable origin. Loss remains unresolved; same-price action effects are distinct from separately calibrated policy comparisons.

An all-query analysis distinguishes changed rankings from changed outcomes. At the original QUERY prices, rankings differ on $57.789\%/60.776\%$ of BLIP/ALBEF image-macro query mass, but target ranks differ on only $2.250\%/3.489\%$. Recovered benefit and avoided harm outweigh added harm and forgone benefit; harmful changes remain. At zero price the rankings coincide. Appendix E gives the exact four-channel accounting and intervals adjusted over a separate family of 60 contrasts.

Absolute results provide context for these small differences. BLIP's baseline MRR is 0.595912, rising to 0.625951 with QUERY and 0.626881 with CYCLE. ALBEF starts at 0.468504 and reaches 0.528074 and 0.530330. ALBEF's larger improvement does not imply a higher final MRR than BLIP. The primary CYCLE losses are 0.001439 and 0.001804, respectively. All twelve primary head-level policy rows remain below the nominal loss budget, but this does not extend to every tested operating point: 51 of 424 ALBEF comparison set rows exceed their corresponding nominal budgets, versus zero of 424 BLIP rows.

ALBEF's adjusted gain interval exceeds 0.001; BLIP's crosses that margin. Both added-loss upper bounds are below 0.0002, but both intervals include zero. These conditional comparisons establish neither equivalence nor population risk control.

A reversal on VizWiz. On VizWiz VAL, ALBEF's total gain difference is 0.001892, with interval $[0.001205, 0.002605]$. Its loss difference is -0.000120, with an interval including zero. BLIP instead has a gain difference of -0.000692 and a loss difference of -0.000365: less gain accompanies less harm. For BLIP, the matched-price action contribution is still positive, 0.001580, while the price contribution is -0.002272. A different fixed calibration-role split changes the total gain differ-

Table 2: Transport boundaries for separately calibrated CYCLE − QUERY, $\times 10^3$. VizWiz intervals adjust four contrasts; COCO intervals use its twelve-coordinate action–price family. Different galleries, exposure scopes, and uncertainty families are not pooled.

Setting	Model	ΔG	ΔL
VizWiz	BLIP	$-0.692\ [-1.494, 0.069]$	$-0.365\ [-0.614, -0.156]$
	ALBEF	$1.892\ [1.205, 2.605]$	$-0.120\ [-0.252, 0.005]$
COCO	BLIP	$0.982\ [0.772, 1.195]$	$0.126\ [0.034, 0.215]$
	ALBEF	$1.357\ [1.081, 1.637]$	$0.058\ [-0.055, 0.167]$

ence to 0.001988. This is direct evidence that the sign after calibration cannot be inferred from the action-space inclusion alone. All 36 BLIP policy rows in this exposed VizWiz comparison exceed the nominal 0.002 loss budget; this adverse result is retained.

A larger-gallery boundary. The COCO stress test contains 25,000 FINAL image groups and 119,268 queries. ALBEF's total gain difference is approximately 0.001357, with adjusted interval $[0.001081, 0.001637]$; its added-loss upper bound is 0.000167. BLIP's gain difference is 0.000982, with interval $[0.000772, 0.001195]$, but its loss difference is positive, 0.000126, with interval $[0.000034, 0.000215]$. The latter exceeds the specified added-loss margin at the upper endpoint. The small-loss conclusion from Flickr therefore does not transfer unchanged to BLIP on COCO. No newly fitted model or policy is introduced to remove this boundary.

Table 2 retains the boundaries: VizWiz BLIP has a negative gain estimate with an interval spanning zero and lower positive RR loss; COCO BLIP gains with higher loss and unresolved added-loss tolerance. ALBEF has favorable gain in both settings without established lower loss. These fixed pipelines are not pooled into one success rate. Known COCO training-data overlap and exposed VizWiz outcomes also preclude treating either boundary as universally unseen validation.

6 BINDING MATTERS; PREDICTING ITS VALUE REMAINS DIFFICULT

A same-cardinality intervention. The original cycles are partitioned into pairs using CANONICAL, SIMILAR-RATIO, CROSS-RATIO, and HASH-RANDOM rules. Pairings depend only on frozen predicted cycle statistics, not targets or the price subsequently applied. All partitions have the same number of blocks and the same odd-cycle singleton. This controls action count while changing which gains and losses must be accepted together. The comparison is post-confirmation secondary because the original FINAL outcomes had already been examined.

SIMILAR − CROSS produces positive signed-gain intervals in all four model–price cells. The mean gains range from 0.000419 to 0.000472, with approximately 25% of query actions differing. Price-utility differences are positive with adjusted intervals above zero in three of four cells; BLIP at the QUERY price remains unresolved. Every loss interval includes zero. These effects are smaller than the 0.001 practical-gain margin and do not establish a universal ordering of grouping rules. The differences nevertheless rule out a purely cardinality-based explanation for this particular intervention.

Ordinary groupings delimit transport. The external comparison holds the posterior, proposer, atomic cycles, singleton, and original QUERY price fixed. SIMILAR exceeds CANONICAL in three of four adjusted ΔJ intervals, but no HASH comparison excludes zero (Table 3). CROSS alone would therefore overstate the evidence for a generally useful grouping recipe. The first COCO TRAIN gallery remains a known-media stress test.

Predicted geometry across prices. Write a cycle's predicted signed gain and positive loss as (g_c, l_c). A block loses predicted utility $C_B(\lambda) = \sum_{c\in B}[g_c - \lambda l_c]_+ - [\sum_{c\in B}(g_c - \lambda l_c)]_+$. When $g_c \geq 0$ and $l_c > 0$, define its switching price $\tau_c = g_c/l_c$ and $\bar{\tau}_B = \sum_c l_c\tau_c / \sum_c l_c$. Positive-part integration gives

$$\int_0^\infty C_B(\lambda)\, d\lambda = \frac{1}{2}\sum_{c\in B} l_c(\tau_c - \bar{\tau}_B)^2. \tag{6}$$

Table 3: External SIMILAR − ordinary grouping at the fixed QUERY price. $\Delta J \times 10^3$ with eight-contrast family adjusted intervals: four setting–model cells and two controls. Already exposed evaluations; no pooled source effect.

Setting	Model	CANONICAL	HASH
VizWiz	BLIP	$0.918\ [0.219, 1.800]$	$0.145\ [-0.349, 0.534]$
	ALBEF	$1.019\ [0.243, 1.755]$	$0.416\ [-0.115, 0.997]$
COCO	BLIP	$0.136\ [-0.187, 0.427]$	$-0.014\ [-0.231, 0.172]$
	ALBEF	$0.487\ [0.076, 0.898]$	$0.246\ [-0.067, 0.544]$

Table 4: Separately calibrated MIN-AREA − SIMILAR, RR units $\times 10^3$. Gain intervals adjust eighteen contrasts: six cells times $\Delta G, \Delta L$, and absolute MIN loss. Both policies use the same calibration groups and criterion. Last column is MIN's absolute loss point estimate; 2.000 is the nominal budget, not a population guarantee.

Setting	Model	ΔG [adjusted interval]	MIN L
Flickr	BLIP	$-0.059\ [-0.153, 0.030]$	1.475
	ALBEF	$-0.369\ [-0.516, -0.224]$	1.787
VizWiz	BLIP	$-0.149\ [-0.366, 0.065]$	2.603
	ALBEF	$0.098\ [-0.156, 0.355]$	1.580
COCO	BLIP	$0.011\ [-0.046, 0.068]$	1.788
	ALBEF	$0.085\ [0.00355, 0.171]$	1.499

Binding cost depends on both threshold separation and loss mass; unweighted ratio proximity alone need not minimize it. The identity characterizes predicted decisions, rather than realized gain at a specified risk budget. A finite source-fixed interval handles signed gains and zero losses without deleting cycles (Appendix D).

Reducing predicted cost is insufficient. A deterministic minimum-weight pairing, MIN-AREA, minimizes that finite-interval cost using the same posterior, proposer, and odd singleton. In the six-cell comparison, all 18 headwise differences in realized integrated utility favor MIN-AREA over SIMILAR, but only four individual descriptive 95% intervals exclude zero. All same-price ΔJ intervals at the original QUERY price include zero. The head-specific price intervals are not pooled, and neither result replaces evaluation at $\rho = .002$.

With separate calibration on the same 2,500 image groups, ALBEF gains 0.000084826 on COCO but loses 0.000368521 on Flickr; the 18-contrast family resolves both directions (Table 4). VizWiz BLIP's absolute loss is 0.002603240, above the nominal budget. MIN has lower gain than CYCLE in five of six point estimates. The stored-price crossing further separates composition and price: Flickr ALBEF's $+0.000072571$ composition term and -0.000441092 price term sum to its adverse total. All six composition-gain intervals in the distinct 24-contrast family include zero, including COCO ALBEF. These exposed, calibration-conditional comparisons support a calibration boundary, not a uniformly better pairing; Appendix D retains both families and all comparator paths.

Incremental prediction from the opportunity statistic. A separate fixed prediction test asks whether $\widehat{A}$ improves prediction of realized CYCLE − QUERY price utility. M0 is a ridge predictor using seven prespecified score and action summaries. M1 adds $\widehat{A}$; a matched placebo adds a label-free, image-group-permuted version. All fits use ridge coefficient 10^{-3}. The exposed source development bank is split by a deterministic image-group hash into 70% training and 30% validation subsets. No evaluation outcomes tune the feature set, coefficient, or cutoffs.

M1 worsens MSE in all four FINAL settings (Figure 3, Table 5): approximately 4.6–5.5% for BLIP and 1.7–1.8% for ALBEF, with all four adjusted intervals positive. The fixed-coefficient error decomposition $\Delta\text{MSE} = \mathbb{E}[(f_1 - f_0)^2] - 2\mathbb{E}[(f_1 - f_0)(Y - f_0)]$ reproduces these adverse differences without refitting. This rejects incremental value for the specified learner and split, not information under every estimator. The separate binding comparison relates predicted and realized differences

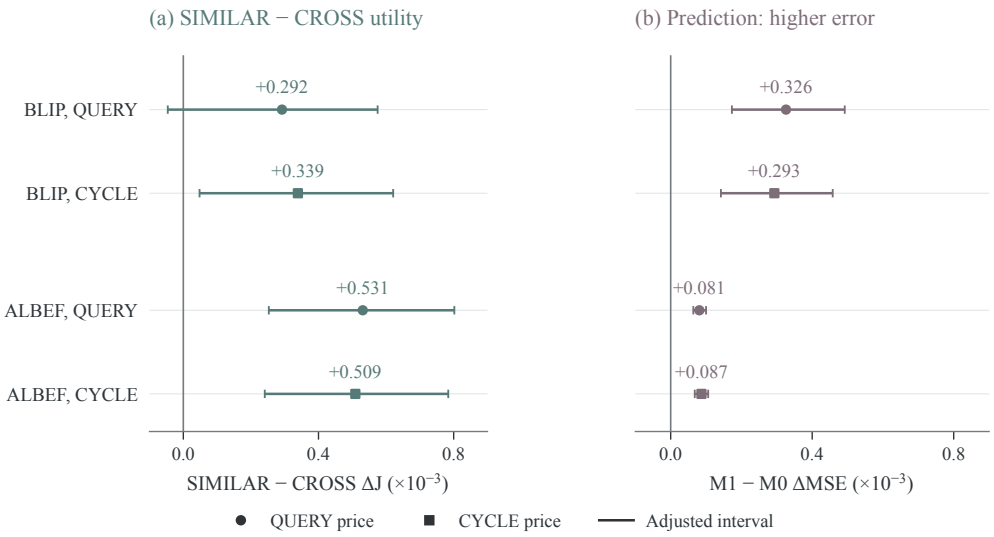

Figure 3: **Binding value and prediction value are different.** (a) SIMILAR − CROSS price utility; (b) direct M1 − M0 prediction-MSE differences after adding $\widehat{A}$. Points show means; horizontal lines are image-bootstrap intervals adjusted over sixteen binding coordinates or four MSE contrasts. Circles and squares distinguish fixed QUERY and CYCLE prices. Positive values favor SIMILAR over CROSS in (a), but indicate worse prediction in (b). The panels retain different estimands; neither establishes causal moderation or irreducible information. Table 5 retains absolute MSE and placebo values.

Table 5: Fixed incremental prediction on FINAL, MSE $\times 10^3$. Positive M1 − M0 differences are adverse. Intervals adjust the four specified contrasts; placebo columns do not define a larger selected family.

Model	Price	M0	M1	Placebo	Δ [adjusted interval]
BLIP	QUERY	5.971	6.297	5.971	0.326 [0.173, 0.492]
	CYCLE	6.384	6.677	6.384	0.293 [0.142, 0.458]
ALBEF	QUERY	4.720	4.802	4.724	0.081 [0.064, 0.100]
	CYCLE	4.753	4.839	4.757	0.087 [0.068, 0.106]

for the same partition contrast; CYCLE − QUERY prediction and SIMILAR − CROSS utility are different estimands.

Where the predicted opportunity can fail. Shortlist-conditioned utility differs from full-population utility. Retrospective accounting uses observed inclusion Z, keeps outside-shortlist queries in the denominator, and records numerical boundary terms. The residual need not isolate probability error. Multiplying all cycle gains and losses by the same positive inclusion mass leaves fixed-price actions unchanged; a support rescaling alone cannot rescue the price-indexed policy path.

7 Calibration, Shortlists, and Actual Cost

Comparison with calibrated gate and mixture controls. The fixed gates and mixtures retain different gain–loss tradeoffs (Appendix E, Table 13). An exposed follow-up gives whole-query selection a finite calibration-optimal menu of same-proposer thresholds and mixtures, at no greater empirical CAL loss than the unchanged CYCLE policy. Flickr CYCLE gains remain 0.000915 for BLIP and 0.001699 for ALBEF; both adjusted intervals are positive. Neither positive-gain comparison establishes no added loss. VizWiz BLIP has a negative gain estimate with an interval spanning zero, alongside lower positive RR loss.

Table 6: Measured serving time, seconds for 3,072 queries (batch 16), averaged over three counterbalanced blocks on the fixed 28,966-image gallery. Baseline is actual contrastive-only retrieval. Resident time includes native inference and policy execution; cold first action is reported separately in Appendix F. Shared-host measurements are not causal speedup estimates.

Model	Baseline	QUERY	CYCLE	Margin	LAC
BLIP	10.240	1188.201	1190.272	1185.251	1187.300
ALBEF	8.870	40.041	40.061	39.315	39.634

A separate procedure selects both action families on 1,250 images and uses another 1,250 to choose each family's own probability of applying the selected policy rather than the baseline. Flickr BLIP/ALBEF, COCO ALBEF and VizWiz ALBEF exceed the .001 gain margin while meeting the .0002 added-loss tolerance and .002 absolute-loss interval criteria. COCO BLIP remains below practical gain; VizWiz BLIP loses gain and has unresolved absolute-budget bounds. These 24-quantity-adjusted results concern selected and guarded policies. A separate label-aware optimization at observed loss .002 favors CYCLE in all six restricted-menu comparisons, including VizWiz BLIP (Appendix G); this hindsight opportunity does not repair the guarded negative result.

Finite calibration support. Twenty fixed calibration-role reallocations are dependent sensitivities, not independent experiments. At 2,500 groups, mean total gain contrasts are 0.001441 for BLIP and 0.002743 for ALBEF; 10th–90th percentiles are $[0.000791, 0.001997]$ and $[0.002050, 0.003836]$. At each of the 36 examined price boundaries, the action change involves one image group. Thus finite calibration composition matters even with unchanged scores.

At 500 image groups, the correction $(31/32)/501 \approx 0.001934$ nearly exhausts the 0.002 budget. Recorded BLIP QUERY/CYCLE gains are $0.013226/0.002649$, versus $0.035575/0.037116$ at 4,000 groups. Action-set inclusion stays fixed while usable operating points change. Evaluation-based choice of the best allocation would erase this limitation.

A fixed shortlist stress test reuses the same scores for $m \in \{8, 16, 32\}$, renormalizes the predicted distribution within each subset, and uses the corresponding bound B_m. Each shortlist receives its own calibration; all queries retain the same full-gallery baseline and denominator. Exact $m = 32$ rank reproduction is checked. At $m = 8$, ALBEF's separately calibrated CYCLE $-$ QUERY gain becomes -0.000460; at $m = 16$, gain rises by 0.003206 but CYCLE loss reaches 0.002076, above the nominal budget. Changing the shortlist changes support, composition, and calibration together. Neither stress result extends the $m = 32$ finding to every shortlist size.

Policy structure is not the dominant serving cost (Table 6): the contrastive baseline is much cheaper, while CYCLE and QUERY have similar resident totals. Cheaper policy arithmetic alone does not establish end-to-end improvement. The three timing blocks are not independent quality replications. A BLIP rank-boundary discrepancy between earlier and recomputed scores precludes exact quality transfer from older FINAL outputs. COCO likewise lacks a complete quality–cost frontier.

What remains outside the evidence. The findings concern fixed pipelines and exposed galleries. Minimum-area optimality concerns predicted cost, not target utility or deployment risk. Grouping comparisons condition on calibration choices. Stored-score analyses establish no model-inference speedup.

8 CONCLUSION

On the Flickr30k-derived gallery, fixed-predictor cycle decisions improve gain at both original prices and beyond the tested calibration-optimal whole-query menu. Split-calibration selection and method-specific guarding meet practical criteria in four cells, but COCO BLIP remains sub-practical and VizWiz BLIP loses gain. Restricted-menu hindsight gains persist at equal observed loss in six cells, but do not establish deployable gains or equal population risk. Binding composition matters, but minimum predicted cost and added opportunity features do not ensure better outcomes. Partial-action opportunity and successful calibration require distinct evidence.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, mathematical formulation and proof drafting, experimental design and implementation, numerical analysis, literature search and reference verification, figure preparation, and manuscript drafting and editing. The authors are responsible for the final content.

ETHICS STATEMENT

The study uses existing image–caption datasets and pretrained retrieval models. Dataset licenses and access conditions remain applicable. VizWiz's distinct capture context does not support claims about individual users or accessibility outcomes. Retrieval degradation remains possible, and the tested empirical budgets do not justify safety-critical deployment.

REPRODUCIBILITY STATEMENT

Appendix B specifies fixed roles, controls and statistical units; Appendix D details the grouping and prediction tests, and Appendix F records serving measurements and shortlist controls. Results exclude incomplete evaluations and retain the limits from prior dataset exposure and differences between repeated computations.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *International Colloquium on Automata, Languages, and Programming*, pp. 28:1–28:15, 2018. doi: 10.4230/LIPIcs.ICALP.2018.28.

Kevyn Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 837–846, 2009. doi: 10.1145/1645953.1646059.

B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. Risk-sensitive evaluation and learning to rank using multiple baselines. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 483–492, 2016. URL https://eprints.gla.ac.uk/119285/.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pp. 417–434, 2020. URL https://arxiv.org/abs/2002.08565.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. ComCLIP: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024. URL https://aclanthology.org/2024.naacl-long.370/.

Shayan Kiyani, George J. Pappas, Aaron Roth, and Hamed Hassani. Decision-theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/kiyani25a.html.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/505259756244493872b7709a8a01b536-Abstract.html.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022a. URL https://proceedings.mlr.press/v162/li22n.html.

Minghan Li, Xinyu Zhang, Ji Xin, Hongyang Zhang, and Jimmy Lin. Certified error control of candidate set pruning for two-stage relevance ranking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 333–345, 2022b. URL https://aclanthology.org/2022.emnlp-main.23/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. URL https://arxiv.org/abs/1609.00451.

Lidan Wang, Paul N. Bennett, and Kevyn Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 761–770, 2012. URL https://www.microsoft.com/en-us/research/publication/robust-ranking-models-via-risk-sensitive-optimization/.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. URL https://arxiv.org/abs/2111.07783.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. URL https://aclanthology.org/Q14-1006/.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22313–22324, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bbc92a647199b832ec90d7cf57074e9e-Abstract.html.

A Definitions and Structural Examples

Valid partial permutations. Let π map baseline positions in the shortlist to positions under the proposed ranking. Its disjoint-cycle decomposition partitions the moved positions. For any subset S of cycles, define π_S to equal π on cycles in S and the identity elsewhere. Every position has exactly one predecessor and one successor under π_S, so π_S is a permutation. Candidate-level acceptance without this closure property need not be valid: accepting only one direction of a swap assigns two candidates to one rank.

Predicted opportunity identity. For fixed cycle utilities u_c, separate selection achieves $\sum_c [u_c]_+$ and all-or-none selection achieves $[\sum_c u_c]_+$. Substituting $[u]_+ = (u + |u|)/2$ yields Equation 3. Nonnegativity follows from the triangle inequality. Equality holds when nonzero utilities share a sign. This is a conditional decision statement for the supplied utility vector, not additional label-aware headroom. Once the single target is known, only its containing cycle can change its rank; applying that cycle or the whole proposer yields the same target rank. Both oracle action spaces therefore attain $\max\{0, 1/r_p(y) - 1/r_0(y)\}$, also under a nonnegative positive-loss price. Replacing conditional expected utility with estimated probabilities introduces a separate error.

Same cardinality, different feasible sets. Consider four disjoint two-candidate cycles with utilities $(0.001, 0.0008, -0.0009, -0.0007)$. Binding the first two and last two gives block utilities $(0.0018, -0.0016)$, hence selected utility 0.0018. Cross binding gives $(0.0001, 0.0001)$, hence 0.0002. Both policies have two binary block decisions and four feasible actions. No confidence or posterior changes. Reversing the assignment of utilities to a fixed grouping can reverse which grouping wins; the example proves possible binding effects, not universal optimality of a named partition.

Price is not merely an offset. A threshold on a single query score can omit intermediate partial actions. A constructed pair of changes with $(G, L) = (0.003, 0.0001)$ and $(0.006, 0.002)$ illustrates the distinction: at price $\lambda = 10$, the first has positive utility and the second negative utility. A rule that admits only their combination cannot express the first action alone. This is a representational limitation, not evidence that every scalar threshold or every calibrated whole-query gate is weak. The natural comparisons therefore include both same-price cross cells and strong whole-query controls.

Metric units and risk. The target rank is always the full-gallery rank. Positive gain and positive loss are separately nonnegative; signed gain is their difference. A method can improve signed gain while increasing harm frequency or harm magnitude. The budget in this paper concerns average positive reciprocal-rank loss, not a conditional error rate among accepted queries. A bound on one quantity does not imply a bound on the others.

Mixtures. A randomized policy selects an entire valid permutation according to a fixed coin. Its expected G and L are mixtures of the corresponding metrics. Reciprocal rank of an averaged rank is not used. Randomized mixtures can interpolate between operating points, but any price, mixing weight, and validation rule must be fixed without evaluation-label selection. No new interpolation result is inferred from the observed calibration slack.

Protection and assignment. A set-protected control enforces rank constraints for candidates in a calibrated set. The remaining assignment can be represented with deadlines and checked against an independent heap-based construction. Such a constraint can sharply reduce gain even when its feasibility is exact. Feasibility and coverage of the protected set are distinct from the utility of the resulting ranking, and neither implies that the target is present in the shortlist.

B PROTOCOL, CONTROLS, AND STATISTICAL SCOPE

Table 7: Evaluation scopes. Query counts refer to the evaluated role, not the full caption bank. Rows are different settings, not exchangeable replications.

Setting	Gallery	Eval. images	Eval. queries	Interpretation
Flickr30k-derived	28,966	23,966	119,022	Locked local confirmation
VizWiz VAL	7,459	4,748	20,579	Exposed transport analysis
COCO TRAIN stress	30,000	25,000	119,268	Known-media gallery stress

Fixed paths and calibration assumptions. Before calibration-target access, the posterior, proposal permutation, block partition, score calculation and tie rules are fixed. For baseline rank b, fixed proposal rank r, and the target-containing block's score s, positive RR loss is

$$\ell(\lambda) = \max(1/b - 1/r, 0)\, \mathbf{1}\{s > \lambda\}.$$

Other disjoint cycles cannot change that target rank. Loss is therefore non-increasing and right-continuous in the price, including after caption and image averaging. Fixed scalar thresholds have the same structure; the original scalar controls apply the unmodified matching-score proposal, whereas the newer controls below apply the probability-fitted proposal. Predicted zero loss does not imply zero realized loss. The explicit infinite-price endpoint returns the baseline even for predicted-zero-loss actions.

Calibration chooses the smallest feasible price from exact score boundaries. The primary correction uses 2,500 image groups, not 12,427 caption queries, with bound $31/32$. This correction alone does not establish exchangeability, validate every comparator's assumptions, or guarantee population safety. For a fixed baseline mixture, expected loss is proportional to execution probability; individual independently randomized outcomes need not form a monotone path. The protected-ranking control fixes its proposal on 1,250 set-calibration images and chooses its mixture on another 1,250 risk-calibration images. Its risk correction therefore uses $n = 1250$.

Roles and exposure. The Flickr gallery contains 1,250 calibration-A, 1,250 calibration-B, 2,500 diagnostic, and 23,966 FINAL image groups. Its full caption bank contains 143,868 queries. The first COCO gallery uses the same calibration and diagnostic sizes, followed by 25,000 FINAL groups; the complete bank contains 143,148 queries. Roles are fixed at image level, and captions from one image never become independent role units. Gallery membership is shared across query roles because retrieval uses the same candidate collection.

Available local exposure records exclude known prior evaluation galleries, not all pretraining or historical local use. Both COCO stress galleries have known model-training-media overlap. Additional descriptive analyses use a second 30,000-image gallery and ordinary grouping rules. The tables retain the first gallery's specified estimands rather than pooling the two galleries or choosing a favorable replicate. New minimum-area, separate calibration, and price-cross analyses use already exposed outcomes; the earlier primary's locked design does not transfer to those additions.

Fixed posterior and model parameters. BLIP uses the frozen `Salesforce/blip-itm-large-coco` checkpoint, revision `19502f1e`, with ViT-L/16. ALBEF uses the author's MSCOCO online weights and ViT-B/16, source revision `b9727e43`. Both receive 384×384 bicubic inputs and run FP32 without TF32. ITC is a dot product of normalized projected features; ITM supplies negative and positive raw logits. BLIP tokenization uses dynamic padding without truncation within the checked 512-token limit; ALBEF retains the author's 30-word caption preprocessing and 30-token limit.

Each fixed posterior head is a candidate-wise $5 \to 16 \to 1$ tanh network over ITC, baseline-rank fraction, two ITM logits, and ITM-rank fraction, using feature normalization fitted only on training images and a listwise softmax at temperature one. Three seeds $(0, 1, 2)$ use 200 FP64 optimization steps, learning rate .03, coefficient regularization .01, and Adam moments $(.9, .999)$ with $\epsilon = 10^{-8}$; no head is selected on FINAL. The source training subset contains 4,000 image groups for BLIP and 1,524 for ALBEF. The main proposer sorts the fitted posterior, not raw ITM. Fitted-posterior

ties use image-ID lexicographic order for BLIP and baseline-top32 index for ALBEF. Scalar gates retain their original raw-score proposer, so separate calibration alone does not make their comparison a same-proposer intervention.

Strong comparator comparison set. The 424-row comparison set contains a baseline and the full fixed action–price crosses, together with margin gates, source-trained query gates, source-trained augmented gates, fitted mixtures, and set-protected mixtures. The 324 cross cells span three heads, three posterior variants, three calibration allocations, three budgets, and four action–price combinations. The source-trained gates use only the designated development information. LAC-related controls use a set-calibration role distinct from the risk-calibration role. The primary comparison remains the original posterior, full 2,500-group calibration, and $\rho = .002$; other rows are sensitivity and comparator evidence, not candidates for post-hoc primary selection.

Paired uncertainty. Within each image, caption outcomes are averaged; the three fixed heads are then averaged within the same image-level contrast. Paired bootstrap draws sample those image contrasts with replacement. For the twelve primary coordinates, the percentile probability in each tail is $0.05/(2 \cdot 12)$. The 3,999-draw intervals are approximate and conditional, not exact simultaneous confidence guarantees. The equal-cardinality intervention uses a separate sixteen-coordinate family: two models, two price origins, and G, L, J, h. The incremental predictor uses four adjusted MSE contrasts. Other 95% intervals and calibration-role percentiles are descriptive and are labeled accordingly.

The new external ordinary-grouping family contains eight fixed QUERY-price ΔJ contrasts. The separate-calibration family comprises eighteen contrasts: six cells times MIN $-$ SIMILAR gain, loss, and MIN absolute loss. The price-crossing family comprises twenty-four contrasts: six cells times composition/price contributions to gain/loss. The latter two use 19,999 paired-image draws, averaging the three fixed heads within each image before resampling, with tail probabilities $.05/(2 \cdot 18)$ and $.05/(2 \cdot 24)$. Calibration choices are not resampled. The headwise integrated-area intervals are individual descriptive 95% intervals, not an eighteen-contrast adjustment. No interval converts exposed follow-up design into fresh confirmation.

Calibration boundary diagnostics. Action changes at the examined price boundaries are assessed by image group, rather than by individual caption. All 36 examined boundaries involve one image group. The largest full-calibration slack is 3.318×10^{-5} for BLIP and 6.262×10^{-5} for ALBEF. These observations explain why tiny price changes can be consequential in this finite sample, without establishing that randomized boundary interpolation would improve held-out performance. No interpolation outcome is reported.

C Full Contrasts and Adverse Settings

Table 8: Twelve primary Flickr coordinates, $\times 10^3$, with adjusted intervals. Action and price are the symmetric decomposition, not one arbitrarily chosen reference price.

Model	Component	ΔG	ΔL
BLIP	Action	$1.237\ [0.972, 1.487]$	$0.054\ [-0.024, 0.143]$
	Price	$-0.307\ [-0.379, -0.239]$	$-0.035\ [-0.056, -0.017]$
	Total	$0.930\ [0.662, 1.192]$	$0.019\ [-0.061, 0.104]$
ALBEF	Action	$1.874\ [1.541, 2.186]$	$-0.012\ [-0.094, 0.071]$
	Price	$0.381\ [0.306, 0.462]$	$0.044\ [0.022, 0.068]$
	Total	$2.255\ [1.927, 2.571]$	$0.032\ [-0.051, 0.119]$

Table 9: Direct same-price action contrasts on locked Flickr FINAL, $\times 10^3$. Each row changes only action space at the named price. Descriptive 95% intervals do not extend the twelve-coordinate primary family.

Model	Fixed price origin	ΔG	ΔL
BLIP	QUERY	$1.236\ [1.058, 1.414]$	$0.055\ [-0.001, 0.114]$
	CYCLE	$1.238\ [1.065, 1.412]$	$0.053\ [-0.003, 0.111]$
ALBEF	QUERY	$1.878\ [1.647, 2.102]$	$-0.009\ [-0.064, 0.047]$
	CYCLE	$1.870\ [1.634, 2.095]$	$-0.014\ [-0.072, 0.045]$

Table 9 retains both original price origins. Cross-gallery boundaries appear in Table 2. Both VizWiz calibration allocations are reported in the following paragraph.

The VizWiz action–price decomposition retains both calibration allocations. For BLIP, full calibration gives action 0.001580, price -0.002272, and total -0.000692; split calibration gives 0.001820, 0.000168, and 0.001988. For ALBEF, the corresponding triples are $(0.001920, -0.000028, 0.001892)$ and $(0.002060, 0.000316, 0.002376)$. A favorable action effect can therefore coexist with an unfavorable total. The eight fixed residual intervals all include zero, so a particular direction of probability error is not established.

The 20 fixed calibration-role reallocations hold role size constant while changing composition. For 2,500 groups, BLIP's mean contrast is 0.001441 and ALBEF's is 0.002743. At 1,250 groups, the means are 0.001753 and 0.002644. Reusing the same evaluation images makes these dependent sensitivity measurements. Selecting the most favorable reallocation after observing evaluation outcomes would invalidate their intended interpretation.

D BINDING AND INCREMENTAL PREDICTION DETAILS

Exact grouping rules. Nontrivial cycles are sorted lexicographically by their sorted baseline-position members. When the count is odd, the last canonical cycle is the same singleton in every arm. CANONICAL pairs adjacent cycles in canonical order. SIMILAR-RATIO pairs adjacent cycles in ascending gain/loss order; CROSS-RATIO pairs opposite ends of that order. Ratios compare exact rational representations of the stored binary floating-point values; $0/0$ is assigned zero and positive gain over zero loss is assigned $+\infty$. Canonical cycle identity breaks ratio ties. HASH-RANDOM pairs adjacent cycles after sorting their SHA256 hashes of the fixed namespace, query identity and canonical members, using namespace `P04_CYCLE_BINDING_20260923_V1`. Thus no random-seed draw or operating price changes the partition. The common singleton is excluded before pairing, preserving the same number of available blocks. Target outcomes do not define these grouping rules.

Table 10: Equal-cardinality SIMILAR − CROSS intervention, $\times 10^3$. Price utility $J = G - \lambda L$ uses the respective frozen original price. Intervals adjust sixteen coordinates; all four loss intervals include zero.

Model	Price	ΔG	ΔL	ΔJ [adjusted interval]
BLIP	QUERY	0.419	0.023	0.292 $[-0.046, 0.575]$
	CYCLE	0.426	0.016	0.339 $[0.048, 0.621]$
ALBEF	QUERY	0.472	−0.014	0.531 $[0.253, 0.802]$
	CYCLE	0.454	−0.013	0.509 $[0.241, 0.784]$

The pairing rules are constructed without evaluation targets. Similar and crossed ratio orderings use the same atomic cycle summaries; hash-random and canonical pairings provide fixed alternatives. An odd singleton is selected identically before pairing. Prices are applied after the partition is fixed. The comparison changes action geometry while controlling the number of block decisions, without preserving the feasible set or introducing a new probability model.

Table 5 reports the complete fixed M0/M1/placebo comparison. M0's seven features are posterior margin, entropy, cycle count, predicted whole-query signed gain and positive loss, moved-candidate fraction, and normalized mean absolute rank displacement. Normalization fitted only on training images and fixed selection fractions $.10, .25, .50$ do not use FINAL labels. The internal source development bank was already exposed, and the selection evaluation does not replace the adverse primary MSE comparison.

The placebo permutation operates by image group within the evaluation role so that captions are not treated as independent observations. The feature uses the entire evaluation role without labels and is unavailable to a predictor serving one query at a time. This diagnostic checks whether an additional matched feature creates the same apparent behavior as the unpermuted opportunity statistic.

For any fixed price, the realized difference is averaged across all image groups, including groups whose target lies outside the top 32. Predicted shortlist-conditioned contributions are multiplied by the observed inclusion indicator in retrospective accounting. Selection differences between FP64 arithmetic and exact arithmetic are recorded as numerical boundary terms rather than erased. The remaining residual is an accounting discrepancy that can combine probability error, support conditioning, and finite precision. No causal interpretation follows from a nonzero residual alone.

The two mechanism tests answer different questions. Equal-cardinality binding establishes that different partitions of fixed changes can have different realized utility. The ridge test asks whether the predicted opportunity supplies useful incremental information to one specified predictive family. The positive binding result and the adverse ridge result concern different estimands: realized utility under changed partitions and incremental prediction value for a specified learner.

Finite-interval binding and minimum-area pairing. For $g \geq 0$ and $l > 0$, direct integration gives $\int_0^\infty [g - \lambda l]_+ d\lambda = g^2/(2l)$. Subtracting the block integral from the sum of cycle integrals yields

$$\frac{1}{2}\left(\sum_{c\in B}\frac{g_c^2}{l_c} - \frac{(\sum_{c\in B} g_c)^2}{\sum_{c\in B} l_c}\right) = \frac{1}{2}\sum_{c\in B} l_c(\tau_c - \bar{\tau}_B)^2,$$

which proves Equation 6 by expanding the weighted squared deviations. For a pair, the cost is $l_i l_j(\tau_i - \tau_j)^2/[2(l_i + l_j)]$. The identity shows that predicted binding cost depends on both loss mass and threshold separation; realized target utility is assessed empirically.

For $0 \le a < b < \infty$ and $l \ge 0$, define $F(g, l) = ([g - al]_+^2 - [g - bl]_+^2)/(2l)$ for $l > 0$, and $F(g, 0) = (b - a)[g]_+$. The integrated block cost is $\sum_c F(g_c, l_c) - F(\sum_c g_c, \sum_c l_c)$, valid also for signed gains and zero losses. Thus no zero loss enters a ratio or is replaced by an artificial positive weight. Strict acceptance at positive utility fixes actions at ties; a zero-utility endpoint has no effect on the predicted integral. Direct evaluation at price breakpoints verifies the formula. The fixed interval starts at zero and ends at the image-macro weighted source-development 95th percentile of finite positive gain/loss events, separately for each model/head. Target mapping metadata supplies grouping identities, not outcome-dependent interval selection. Head-specific area units are not interchangeable.

MIN-AREA solves a minimum-weight perfect pairing on at most sixteen nontrivial cycles, after excluding the common odd singleton. A bitmask dynamic program with deterministic lexicographic ties uses finite-interval pair costs. The construction changes neither posterior nor proposer. Predicted cost reduction is mechanical; realized integrated utility is the empirical test. All 18 headwise point contrasts favor MIN over SIMILAR, but only four individual descriptive intervals exclude zero. The same-price MIN $-$ SIMILAR ΔJ intervals at the original QUERY price all include zero. The original primary remains unchanged.

Separate calibration and fixed-price crossing. All eleven policies retain full calibration (2,500 image groups), $\rho = .002$, exact strict thresholds, baseline fallback, and original denominator. Existing QUERY/CYCLE and scalar ranks are reproduced before evaluation of the new pairing. MIN's absolute-loss intervals in the eighteen-contrast family are $[.001226, .001737]$ (Flickr BLIP), $[.001537, .002059]$ (Flickr ALBEF), $[.001849, .003445]$ (VizWiz BLIP), $[.000994, .002298]$ (VizWiz ALBEF), $[.001526, .002065]$ (COCO BLIP), and $[.001251, .001751]$ (COCO ALBEF). Only two upper endpoints fall below .002; these remain conditional empirical intervals, not deployment guarantees.

Table 4 retains every source–model cell. COCO ALBEF's gain is 0.000084826, with an eighteen-contrast-adjusted lower endpoint 0.00000355; Flickr ALBEF loses 0.000368521, while VizWiz BLIP's absolute loss remains 0.002603240. The eleven policies include the baseline, QUERY, CYCLE, four original pairings, minimum-area pairing, and three scalar gates. MIN has lower gain than CYCLE in five of six point estimates. The scalar gates retain their raw-score proposer, whereas the pairings and QUERY/CYCLE use fitted posterior scores; the full comparison is not a pure-partition intervention.

The 2×2 crossing uses the fixed MIN/SIMILAR pairings at both stored separately calibrated prices, without refitting. Flickr ALBEF's composition term $+0.000072571$ and price term -0.000441092 sum to its adverse total. The price term's interval in the 24-contrast family is $[-0.000550660, -0.000339780]$. COCO ALBEF has identical stored prices, hence zero price term; its composition interval in the twenty-four-contrast family is $[-0.000000222, 0.000171294]$, including zero. The 18- and 24-contrast families answer their separately specified comparisons and both condition on the realized calibration sample and fixed heads. Neither supplies causal attribution or fresh confirmation.

Table 11: Fixed MIN/SIMILAR pairing crossed with both stored separately calibrated prices. Symmetric gain components, $\times 10^3$, sum to the direct difference after separate calibration. Intervals adjust twenty-four gain/loss components, separately from the eighteen contrasts in Table 4. Zero price effects reflect exact equal stored prices, not inferred equivalence.

Setting	Model	Composition ΔG	Price ΔG
Flickr	BLIP	-0.011 $[-0.092, 0.072]$	-0.048 $[-0.098, -0.003]$
	ALBEF	0.073 $[-0.038, 0.184]$	-0.441 $[-0.551, -0.340]$
VizWiz	BLIP	0.086 $[-0.071, 0.252]$	-0.236 $[-0.402, -0.089]$
	ALBEF	0.098 $[-0.156, 0.366]$	0 $[0, 0]$
COCO	BLIP	0.024 $[-0.035, 0.082]$	-0.012 $[-0.026, -0.001]$
	ALBEF	0.085 $[-0.000222, 0.171]$	0 $[0, 0]$

The crossing adds no fit or new calibration. The 24-contrast family leaves every composition-gain interval crossing zero, while price-gain intervals are negative in the four cells with differing prices. These finite exposed contrasts explain the measured decomposition without identifying a general causal price mechanism. The previously favorable integrated area and COCO result in the eighteen-contrast family cannot replace the unfavorable Flickr result or VizWiz risk boundary.

E TARGET-RANK CHANGES AND STRONGER WHOLE-QUERY CONTROLS

Action disagreement and target outcomes. The all-query comparison retains all 119,022 queries, 23,966 images and three fixed heads per model, including outside-shortlist targets. At zero price, QUERY and CYCLE produce identical rankings. At the QUERY price, rankings differ on 57.789%/60.776% of image-macro query mass for BLIP/ALBEF, but target ranks differ on only 2.250%/3.489%. The corresponding CYCLE-price rates are 57.411%/60.861% for rankings and 2.214%/3.520% for target ranks. Outside-shortlist mass, 10.217%/17.251%, remains in every denominator.

Let R, V, H, F denote nonnegative RR mass from recovered benefit, avoided harm, added harm and forgone benefit, respectively. A target can occupy only its baseline or proposal rank. At a common price,

$$\Delta G = R + V - H - F, \quad \Delta L = H - V, \quad \Delta J = R + (1+\lambda)V - (1+\lambda)H - F.$$

At the QUERY price the four masses, $\times 10^3$, are $(2.057, .174, .229, .767)$ for BLIP and $(3.330, .231, .222, 1.461)$ for ALBEF. At the CYCLE price the masses are $(2.047, .178, .230, .756)$ and $(3.329, .239, .225, 1.474)$. Thus recovered benefit coexists with added harm. These rounded components reproduce the original same-price gains up to rounding, not an independent gain estimate or causal mediation.

Every evaluated posterior is strictly ordered and positive; all nontrivial-cycle predicted gains and losses are positive. No zero-price ranking disagreements or exact-versus-floating selection differences occur. Exact arithmetic finds no zero nontrivial-cycle utility or whole-query sum at either original price. Near-boundary checks at $10^{-12}, 10^{-10}, 10^{-8}$ leave all strict selection masks unchanged. These checks use the original scores and targets, without new inference or fitting.

Table 12: Same-price target-channel signed gain contributions, $\times 10^3$. Intervals adjust over a separate family of 60 contrasts: two models, two prices, four channels' G, L, J contributions and three direct totals. Counts and heads are not independent replications.

Model	Price	Channel	Signed gain [adjusted interval]
BLIP	QUERY	Recovered	$2.057[1.840, 2.290]$
		Avoided harm	$.174[.123, .235]$
		Added harm	$-.229[-.327, -.156]$
		Forgone	$-.767[-.904, -.627]$
	CYCLE	Recovered	$2.047[1.828, 2.280]$
		Avoided harm	$.178[.124, .239]$
		Added harm	$-.230[-.325, -.159]$
		Forgone	$-.756[-.897, -.617]$
ALBEF	QUERY	Recovered	$3.330[3.029, 3.661]$
		Avoided harm	$.231[.172, .297]$
		Added harm	$-.222[-.301, -.163]$
		Forgone	$-1.461[-1.693, -1.256]$
	CYCLE	Recovered	$3.329[3.029, 3.658]$
		Avoided harm	$.239[.181, .307]$
		Added harm	$-.225[-.306, -.162]$
		Forgone	$-1.474[-1.701, -1.253]$

The 9,999 paired-image draws average fixed heads within image and use Bonferroni percentile tails $.05/(2 \cdot 60)$. CAL and heads are not resampled. This exposed family conditions on the same gallery and calibration choices and does not replace the original family of 12 primary coordinates.

Whole-query gates using the same proposal. The source-fitted ridge and margin gates use CYCLE's probability-fitted proposal; the earlier gates in Table 13 use the unmodified matching-score proposal. The 326 ridge coordinates comprise an intercept, 160 candidate features, 32 fitted probabilities, 32 proposal ranks and 101 cycle summaries at price 1. FIT-only standardization of the 325 non-intercept coordinates, image-balanced weights and ridge penalty .01 fit targets $g = d_p(x, y)$ (signed RR change) and $\ell = \max\{-g, 0\}$ (positive RR loss), leaving the intercept unpenalized. The score is $s(x) = \widehat{g}(x) - \min\{1, \max\{0, \widehat{\ell}(x)\}\}$: only the loss prediction is clipped, separately from

Table 13: Strong fixed comparators on locked FINAL, $\rho = .002$. Entries are image-macro G/L, $\times 10^3$; fixed heads are averaged. All rows use full calibration except the protected mixture, whose separate set/risk split is part of the control. No evaluation-based best-row selection or simultaneous superiority claim is made.

Policy	BLIP G/L	ALBEF G/L
QUERY	$30.039/1.420$	$59.570/1.773$
CYCLE	$30.968/1.439$	$61.826/1.804$
Margin gate	$26.628/1.996$	$48.838/2.443$
Source-trained query gate	$25.800/1.073$	$59.700/2.091$
Source-trained augmented gate	$24.057/1.269$	$44.677/1.787$
Fitted randomized mixture	$4.823/1.584$	$9.618/1.718$
Set-protected mixture	$12.419/0.236$	$28.447/0.289$

the positive-part target definition. Neither predicted gain nor score is clipped; the score can lie outside $[0, 1]$ and is not a probability. The margin gate uses the top-two fitted-probability difference. These six ridge fits preserve the probability heads and the separate opportunity predictor.

SOURCE FIT retains 20,012 queries/4,000 images for BLIP and 7,860/1,524 for ALBEF. Full CAL uses 2,500 image groups and strict whole-query score thresholds with the same correction and baseline fallback. Models precede score calculation, and thresholds precede the new evaluation. Earlier evaluation exposure remains. A separate family of 12 quantities comprises four gain contrasts, four loss contrasts and four gate absolute losses; 9,999 paired-image draws condition on fixed heads and CAL choices. Only ALBEF versus margin establishes lower loss. BLIP versus margin exceeds the

Table 14: CYCLE − same-proposal whole-query gates, $\times 10^3$, with intervals adjusted over the separate family of 12 quantities. Last column is gate absolute positive loss.

Model	Gate	ΔG	ΔL	Gate L
BLIP	Source ridge	$3.351[2.680, 4.001]$	$-.046[-.230, .131]$	$1.485[1.260, 1.723]$
	Margin	$6.957[6.089, 7.881]$	$-.010[-.258, .243]$	$1.449[1.239, 1.664]$
ALBEF	Source ridge	$12.701[11.671, 13.697]$	$-.136[-.372, .089]$	$1.940[1.676, 2.230]$
	Margin	$14.450[13.242, 15.591]$	$-.372[-.632, -.093]$	$2.177[1.909, 2.449]$

.0002 added-loss tolerance at its upper endpoint. ALBEF margin's point loss exceeds .002; source ridge's upper endpoint also exceeds that budget. Lower harm frequency does not imply lower loss magnitude.

A finite calibration-optimal whole-query menu. For each fixed head, the menu includes every strict threshold on the QUERY gain/loss ratio, the source ridge and fitted-probability margin, plus baseline and full proposal. Equal scores enter together. Calibration maximizes expected signed gain over their convex hull, subject to expected loss no greater than the original CYCLE empirical CAL loss. An optimum mixes at most two vertices; ties minimize loss, then use a fixed policy ordering. The CYCLE operating point is unchanged. All mixtures precede evaluation, without new fitting. Optimality concerns this finite CAL menu, not every learned selector or evaluation risk.

Table 15: CYCLE − calibration-optimal whole-query mixture, $\times 10^3$. Flickr intervals adjust over four contrasts; the four external cells adjust over a separate family of eight contrasts. Absolute losses are point estimates. All six exposed settings are shown.

Setting	Model	ΔG	ΔL	CYCLE L	Mixture L
Flickr	BLIP	$.915[.692, 1.136]$	$.006[-.064, .079]$	1.439	1.433
	ALBEF	$1.699[1.381, 2.019]$	$-.037[-.121, .047]$	1.804	1.842
COCO	BLIP	$.826[.614, 1.038]$	$.074[-.013, .164]$	1.839	1.765
	ALBEF	$1.244[.989, 1.505]$	$-.030[-.133, .072]$	1.535	1.565
VizWiz	BLIP	$-.785[-1.642, .029]$	$-.380[-.639, -.145]$	2.505	2.885
	ALBEF	$1.492[.684, 2.312]$	$-.106[-.289, .085]$	1.584	1.690

All 2,500 CAL image groups and original evaluation denominators remain included. Each family uses 9,999 paired-image draws after fixed-head averaging; CAL, heads and galleries are not resampled. None of the six cells establishes both positive gain and no added loss. VizWiz BLIP has lower loss but unresolved gain and both point losses above .002. These conditional families do not replace the primary test or establish equal evaluation risk.

Separate selection and method-specific guarding. A different exposed procedure selects both families on 1,250 images and guards on another 1,250. Flickr/COCO use their original CAL halves; VizWiz uses its existing set/risk partition. Each of three heads maximizes SELECT gain over at most two vertices at empirical loss at most .002. CYCLE uses all nonnegative cycle-score price boundaries and both endpoints; WHOLE uses the preceding three-path union. All 36 choices precede GUARD metrics. Posterior, proposal and ridge remain fixed.

For method $j \in \{C, W\}$, let r_j be GUARD positive loss after averaging captions and fixed heads within image. With $n = 1250$, $B = 31/32$ and $\rho = .002$, the independent baseline-thinning probability is

$$q_j = \min\{1, \max\{0, ((n+1)\rho - B)/(nr_j)\}\}.$$

Zero loss with nonnegative allowance uses $q_j = 1$; negative allowance uses zero. All twelve observed probabilities are strictly between zero and one; empirical guarded loss is .0012266, giving corrected criterion .002. Each method uses its own probability. Evaluation integrates head choice, selected mixture and baseline coin analytically, rather than averaging ranks. Equal guard criteria do not match evaluation risk. Exchangeability and selection/guard independence are not established by role separation on exposed galleries.

Table 16: Guarded CYCLE − guarded WHOLE, $\times 10^3$. Intervals adjust over 24 quantities: six settings times gain, added loss and both absolute losses. The last two columns are absolute-loss upper endpoints, not point losses.

Setting	Model	ΔG	ΔL	$U_{L,C}$	$U_{L,W}$
Flickr	BLIP	$4.187[3.874, 4.502]$	$.042[-.048, .132]$	1.093	1.032
	ALBEF	$3.457[3.233, 3.673]$	$.033[-.027, .093]$	1.634	1.595
COCO	BLIP	$.257[.169, .347]$	$.011[-.029, .048]$	1.282	1.269
	ALBEF	$3.096[2.852, 3.343]$	$.094[.008, .184]$	1.607	1.485
VizWiz	BLIP	$-2.772[-3.356, -2.231]$	$-.238[-.371, -.116]$	2.117	2.445
	ALBEF	$1.682[1.231, 2.108]$	$-.029[-.146, .084]$	1.588	1.611

Table 17: Absolute guarded positive loss with intervals adjusted over 24 quantities ($\times 10^3$) and unscaled execution probabilities.

Setting	Model	CYCLE L	WHOLE L	q_C	q_W
Flickr	BLIP	$.957[.833, 1.093]$	$.916[.800, 1.032]$	.466864	.345393
	ALBEF	$1.437[1.251, 1.634]$	$1.403[1.228, 1.595]$	.625899	.574597
COCO	BLIP	$1.140[1.010, 1.282]$	$1.129[1.003, 1.269]$	.398134	.395652
	ALBEF	$1.372[1.159, 1.607]$	$1.278[1.093, 1.485]$	.869877	.773473
VizWiz	BLIP	$1.659[1.270, 2.117]$	$1.897[1.444, 2.445]$	.465396	.546229
	ALBEF	$1.138[.766, 1.588]$	$1.166[.795, 1.611]$	.577111	.554957

The 9,999 paired-image draws condition on SELECT, GUARD, fixed heads and galleries. Practical criteria were gain lower bound above .001, added-loss upper bound at most .0002, and both absolute-loss upper bounds at most .002. Four named cells meet all three; COCO BLIP does not attain practical gain and VizWiz BLIP loses gain. No positive-gain cell establishes no added loss. COCO ALBEF has strictly positive added loss despite meeting the tolerance.

The larger procedure effects substantially involve guard probabilities. Flickr BLIP's unthinned CYCLE/WHOLE gains are $.035163/.035407$, but the larger CYCLE probability reverses the sign. VizWiz BLIP reverses oppositely. Exact point accounting separates selected-policy and probability terms: Flickr BLIP $-.000098877 + .004286087 = .004187211$; VizWiz BLIP $.001072110 - .003844382 = -.002772272$. These terms are not action-only effects or causal mediation. The

correction is not a confidence bound; known-media COCO and previously exposed VizWiz are not new populations. GUARD checking verified saved image losses and the probability formula, not an independent target-to-rank calculation. Technical access to compressed VizWiz records included both roles, although selection calculations used SELECT rows only.

F SERVING COST AND SHORTLIST STRESS

Table 18: Actual serving measurements on the fixed Flickr gallery. Resident totals process 3,072 queries; brackets give the range across three blocks. Cold first action is a process-cold measurement with an unflushed operating-system cache. All values are seconds.

Model	Policy	Resident mean [range]	Cold first action mean
BLIP	Baseline	10.240 [10.152, 10.339]	523.153
	QUERY	1188.201 [1172.812, 1196.516]	563.644
	CYCLE	1190.272 [1182.632, 1201.334]	564.284
	Margin	1185.251 [1180.555, 1194.325]	564.170
	LAC	1187.300 [1177.672, 1200.963]	564.808
ALBEF	Baseline	8.870 [8.814, 8.939]	157.927
	QUERY	40.041 [39.877, 40.330]	248.592
	CYCLE	40.061 [39.952, 40.224]	251.247
	Margin	39.315 [39.179, 39.524]	252.169
	LAC	39.634 [39.404, 39.880]	253.621

Each serving measurement uses the same ordered inputs and a batch size of 16, producing 192 batches. Baseline-only retrieval executes the contrastive path without the reranking forward pass. The other arms include their actual native matching computation and final action construction. The resident total is measured directly rather than composed from cached microbenchmarks. Process loading and first-action time are measured separately; the two columns are not additive components of an individual-request latency.

The three timing blocks use a fixed counterbalanced order on a shared GPU host. Differences of a few tenths of a second do not establish causal acceleration. Repeated cost measurement records one BLIP rank-boundary discrepancy between earlier and newly computed scores at a fixed query and retains its action/quality sensitivity. Policy implementations agree on the aligned inputs, but the BLIP rank-boundary difference prevents exact transfer of the earlier evaluation quality values. The timings compare model inference cost with policy computation cost.

Table 19: ALBEF shortlist stress: each m uses its own calibration and bound B_m. Signed gain and positive loss are $\times 10^3$, averaged across the fixed heads. The whole-query and partial policies use their respective calibrated prices.

m	QUERY G	CYCLE G	ΔG	QUERY L	CYCLE L
8	51.915	51.456	-0.460	1.919	1.759
16	57.626	60.832	3.206	1.901	2.076
32	59.570	61.826	2.255	1.773	1.804

The shortlist stress uses cached native scores, not additional model training or duplicated GPU inference. The $m = 8$ and $m = 16$ posterior distributions are normalized subsets of the frozen top-32 distribution. Full-gallery baseline ranks and outside-shortlist outcomes are unchanged. Each m has its own calibration path and $B_m = 1 - 1/m$; a separate evaluation at the fixed $m = 32$ price isolates price sensitivity. The comparison is conditional on preserving the original $m = 32$ ranks.

The negative $m = 8$ contrast and the $m = 16$ loss above the nominal budget are retained. Neither the larger gain at $m = 16$ nor the lower loss at $m = 8$ constitutes uniform dominance. Changing shortlist size changes target support, posterior normalization, the proposed permutation, and calibration opportunities. The resulting comparison is a deployment stress test, not a clean intervention on cycle topology alone.

G RESTRICTED-MENU OPPORTUNITY AT EQUAL OBSERVED LOSS

This exposed follow-up separates opportunities in fixed score menus from the operating points selected by calibration. The posterior, fitted proposer, top-32 shortlist, strict tie rule, source-trained ridge and margin scores, gallery, original image/caption denominators and three fitted heads remain fixed. CYCLE contains the full nonnegative-price path, the baseline and full fitted proposal. WHOLE contains the union of the QUERY, SOURCE_FITTED and MARGIN_FITTED threshold paths, plus the same baseline and full proposal.

For each head and method, a mixture of at most two menu vertices maximizes observed signed RR gain G subject to observed positive RR loss $L \leq .002$. Ties prefer lower loss and then a fixed lexicographic order. Captions are averaged within image and the three fixed heads are averaged without treating heads as independent observations. Since G is already signed, loss is not subtracted again. All 36 original point optima (six cells, three heads and two methods) attain $L = 1/500$: the original-sample losses are equal, although equality was not imposed and future risks need not match.

Table 20: Label-aware restricted-menu optima. All gains, losses and intervals are multiplied by 10^3. Both menus attain observed loss 2 in every original cell. Intervals adjust a separate six-contrast family. These are hindsight quantities, not selected-policy performance.

Gallery	Model	CYCLE G	WHOLE G	ΔG	Interval
Flickr	BLIP	35.458211	34.466772	.991439	[.389639,1.539689]
Flickr	ALBEF	64.405671	62.304421	2.101250	[1.157552,2.993785]
COCO	BLIP	20.360018	19.796771	.563247	[.202155,.900598]
COCO	ALBEF	30.139082	28.884078	1.255003	[.665371,1.750180]
VizWiz	BLIP	38.344425	36.160350	2.184075	[.257354,3.604318]
VizWiz	ALBEF	44.001803	41.439818	2.561985	[.660461,3.841362]

Each of 3,999 paired image-bootstrap draws reoptimizes both menus, retaining zero-effect and outside-shortlist queries in the denominator. Seeds follow table order from 2609261545 through 2609261550; each tail uses $.05/(2 \cdot 6)$. These intervals are approximate bootstrap diagnostics of optimized empirical functionals, not exact simultaneous guarantees. Selection optimism and nonregularity remain. All differences are positive, but only Flickr ALBEF's entire interval exceeds the existing .001 practical reference; COCO BLIP's interval lies below it.

The restricted oracle uses evaluation labels and cannot be deployed as a calibration-only policy. In particular, the earlier guarded VizWiz BLIP difference remains $\Delta G = -.002772272$ and $\Delta L = -.000238425$, with unresolved absolute-loss budget bounds. Its positive oracle difference of $.002184075$ establishes a menu opportunity, not recovery by the evaluated selection procedure or proof that calibration error alone caused the negative result. An unrestricted target-aware whole-query oracle can match the partial action's single-target outcome; no superiority over that oracle is claimed. COCO remains a known-media stress setting rather than a third untouched source replication.

Table 21: Descriptive gain shortfall to each policy family's oracle at that policy's own observed headwise loss, multiplied by 10^3. Columns use different risk levels; differences between columns are not risk-matched effects or causal guarding penalties.

Gallery	Model	Original CYCLE	Guarded CYCLE	Guarded WHOLE
Flickr	BLIP	.395696	11.143370	14.030167
Flickr	ALBEF	.469077	16.197363	16.819533
COCO	BLIP	.102014	7.504688	7.285169
COCO	ALBEF	.102712	2.738241	4.007358
VizWiz	BLIP	.236406	12.868152	9.853888
VizWiz	ALBEF	.927681	11.373490	10.316585

Original VizWiz BLIP loss is $.002505292$, above the common ceiling; its own-risk comparison therefore remains separate.