SEMANTIC SUPPORT SHAPES RELIABILITY EVALUATION AND DOWNSTREAM DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic conditional reliability asks whether one predictor harms or corrects another within a specified set of true classes. Existing AUC decompositions and model-update methods do not isolate how semantic support affects reliability-scorer selection. Our support decomposition shows that predicted-label membership resolves every harmed–corrected comparison involving a query boundary, leaving only interior–interior comparisons unresolved. The resulting Semantic-Membership Baseline (SMB), normalized interior evidence, and support-order loss separate support-induced ranking from discrimination between unresolved harmed (H) and corrected (C) cases. We further characterize when semantic queries require incompatible rankings and connect this evaluation distinction to scorer selection, fallback decisions, and downstream utility. Across image classification, medical imaging, audio, intent classification, domain shift, and collaborative occupancy prediction, support-induced ranking is substantial but heterogeneous. Positive interior evidence can coexist with raw conditional AUROC either above or below SMB, and practical reliability scores can be ranked differently by raw AUROC and interior evidence. Support-aware evaluation also has an operational consequence: selecting scorers by interior evidence rather than raw conditional AUROC can change downstream fallback decisions and improve held-out utility.

1 INTRODUCTION

Model updates, information fusion, and alternative inference pipelines routinely revise predictions that are already available. Although such revisions can improve aggregate accuracy, they can also correct previous errors while introducing new ones, motivating work on negative flips, backward-compatible model updates, uncertainty-aware selection, and model deferral (Yan et al., 2021; Schumann et al., 2024; Jitkrittum et al., 2023; Ricci et al., 2026). Applications span image classification, audio and language understanding, language-model updates (Echterhoff et al., 2024), and collaborative semantic perception (Chen et al., 2026). Reliability assessment must distinguish beneficial from harmful changes, beyond aggregate accuracy.

Negative-flip methods reduce update regressions, while uncertainty and selective-prediction methods guide abstention or deferral. Semantic conditional evaluation raises a further issue: predicted-label membership in the query can determine part of the harmed-versus-corrected (H/C) ordering. High conditional AUROC can therefore reflect deterministic support rather than discrimination of unresolved changes. Mixing these effects can change which reliability scorer is selected and which revisions are rejected at deployment.

Let a and b be the initial and revised predictions, and let a label set A define a semantic query. A change is *harmed* (H) when a is correct and b is wrong, and *corrected* (C) for the reverse. Conditional AUROC evaluates their ranking among observations satisfying $Y \in A$, with larger scores indicating harm. The score sees predictor outputs and A; true labels define the evaluation population and outcomes offline. Ranking quality differs from classification accuracy (Xia et al., 2025).

The key difficulty is that semantic conditioning resolves part of the H/C ranking before any reliability evidence is considered. Within H/C observations satisfying $Y \in A$, $a \in A, b \notin A$ forces H, whereas $a \notin A, b \in A$ forces C. These are the *boundary* cases. When both predictions belong to A, membership alone leaves the outcome unresolved; these observations form the *interior* I_A.

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.

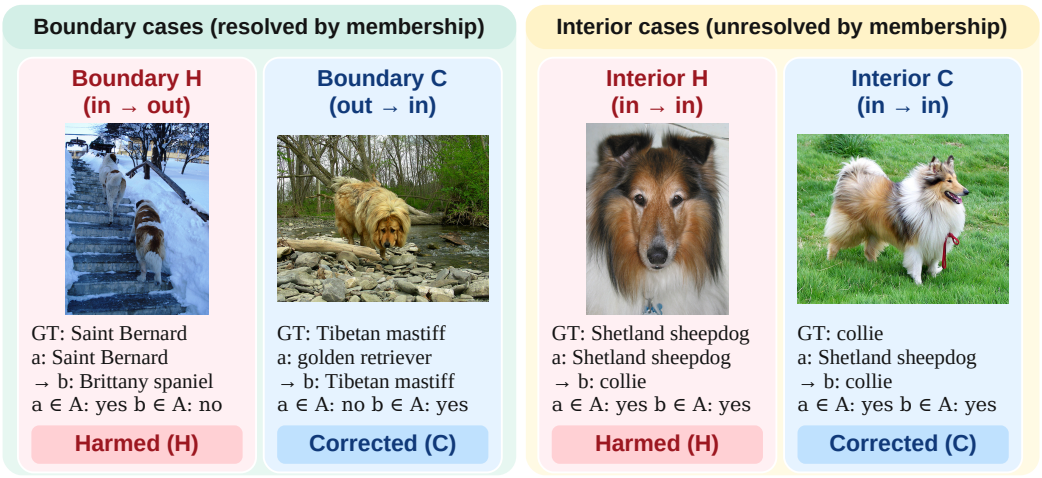


Figure 1: **Semantic membership versus unresolved interior.** Selected ImageNet examples for the registered working-dog query, using frozen ResNet-18→ResNet-50 predictions. Among changed predictions with $Y \in A$ and exactly one prediction correct, membership determines boundary outcomes. The interior examples share the same Shetland-sheepdog-to-collie transition but have opposite outcomes. Ground truth and H/C are offline annotations.

A constant-interior score can therefore obtain high conditional AUROC without distinguishing unresolved H/C outcomes. Figure 1 illustrates this distinction using real ImageNet examples.

General AUC decompositions distinguish within- and cross-group comparisons (Kallus & Zhou, 2019), but do not exploit the deterministic structure induced here by paired predictions and semantic conditioning. *Our starting observation is simple: once the true label is restricted to a semantic set, whether the two predictions fall inside or outside that set can already determine the H/C outcome for boundary cases. This raises a natural question: how much of conditional ranking comes from the reliability score itself, rather than from this semantic support?* We make this structure explicit through the Semantic-Membership Baseline (SMB), which captures ranking available from predicted-label membership alone, and interior evidence, which measures H/C discrimination where membership does not resolve the outcome. For arbitrary scores, support-order loss further quantifies violations of the structurally implied ordering. We also show that different semantic queries can require incompatible rankings of the same observable representation, while a shared posterior with query-dependent readouts can represent the required orders. Together, these results separate what conditional ranking obtains from semantic support from what a reliability score contributes on unresolved cases.

Our contributions are threefold:

- We show how **semantic support structures conditional reliability evaluation**: predicted-label membership can determine part of the H/C ranking, so high conditional AUROC need not imply strong discrimination of unresolved prediction changes.
- We develop the **Semantic-Membership Baseline (SMB), interior evidence, and support-order loss** to separate support-induced ranking from unresolved H/C discrimination, and characterize when different semantic queries impose incompatible ranking requirements.
- We connect this distinction to **downstream decisions**: selecting reliability scorers by interior evidence rather than raw conditional AUROC can change scorer selection, alter which prediction changes are sent to fallback, and improve downstream utility on held-out data. On OrganAMNIST and PathMNIST, query-macro utility improves by 0.73 and 1.88 per 100 changed candidates at 20% fallback; an ImageNet clothing-query case avoids 20 additional net errors at 5% fallback. Complete cross-dataset and configuration results are reported in Appendices P and Q.

2 Related Work

AUC decomposition and evaluation targets. AUROC measures how well a score ranks positive examples above negative ones, with half credit for ties (Fawcett, 2006). Related studies examine correctness judgments, selective-classification metrics, AURC estimation, and AUROC/AUPRC weighting (Ielanskyi et al., 2026; Traub et al., 2024; Zhou et al., 2025; Rabanser & Papernot, 2025; McDermott et al., 2024). Other work decomposes aggregate ranking performance across groups: xAUC studies cross-group ranking disparities (Kallus & Zhou, 2019), while differential label censorship, cluster-AUC, and non-collapsibility analyses show how within- and cross-group contributions can shape an aggregate AUC (Chang et al., 2022; Sudjianto & Liu, 2025; Matos et al., 2026). However, these decompositions do not account for comparisons whose outcomes are already determined by semantic support in paired-prediction evaluation. In our setting, predicted-label membership makes the boundary H/C comparisons deterministic, yielding the closed-form Semantic-Membership Baseline (SMB) and an explicit support-order loss that separate support-induced ranking from unresolved interior discrimination.

Model updates and decisions. Model-update studies reduce negative flips and improve backward compatibility between successive predictors (Yan et al., 2021), while selective classification and cascade deferral use uncertainty or relative prediction loss to decide which predictions to accept, reject, or defer (Geifman & El-Yaniv, 2017; Jitkrittum et al., 2023). Related verifier analyses further connect ranking quality to downstream selection utility (Dorner et al., 2026), while calibration and subpopulation methods study complementary reliability guarantees (Guo et al., 2017; Hebert-Johnson et al., 2018; Sagawa et al., 2020). However, these objectives do not address how semantic conditioning itself can shape the H/C ranking of changes between fixed predictors, or how this structure affects reliability-scorer selection. Our evaluation makes this distinction explicit: H/C ranking measures harmed-versus-corrected discrimination within a semantic query, whereas fallback utility evaluates the resulting deployment decision over all changed candidates, including both-wrong cases that consume budget. Appendix M gives the detailed comparison.

3 Support Structure in Conditional Reliability

3.1 Population, Query, and Observable Support

Let $a(X), b(X)$ be predictions in a finite label set $\mathcal{K}$, observable in X, and let Y be the true label. We work on $D = \{a \neq b\}$. The disjoint outcomes are $H = \{Y = a\}$ (harmed), $C = \{Y = b\}$ (corrected), and $W = \{Y \notin \{a, b\}\}$ (both wrong). The ranking population is $S = H \cup C$. A fixed query $A \subseteq \mathcal{K}$ defines $T_A = \{Y \in A\}$. All probabilities below condition on D and valid labels unless stated otherwise. We require $P(H, T_A), P(C, T_A) > 0$ whenever an AUROC is reported.

A score receives X and A, not the unknown indicator of T_A. Partition the observations by predicted-label membership:

$$\begin{aligned} F_A^H &= \{a \in A, b \notin A\}, \quad F_A^C = \{a \notin A, b \in A\}, \\ I_A &= \{a \in A, b \in A\}, \quad N_A = \{a \notin A, b \notin A\}. \end{aligned} \tag{1}$$

On $S \cap T_A$, F_A^H forces H because $Y \in A$ rules out $b \notin A$ as correct; symmetrically, F_A^C forces C, while N_A has zero mass. Membership leaves the outcome unresolved in I_A. These statements apply to the H/C ranking population; W examples in D can share the same membership patterns.

For independent draws $X_H \sim P(X \mid H, T_A)$ and $X_C \sim P(X \mid C, T_A)$, define

$$\mathrm{AUC}_A(s) = \mathbb{E}[\psi(s(X_H) - s(X_C))], \qquad \psi(t) = \mathbf{1}[t > 0] + \tfrac{1}{2}\mathbf{1}[t = 0]. \tag{2}$$

Set $\alpha_A = P(I_A \mid H, T_A)$, $\beta_A = P(I_A \mid C, T_A)$, and $\pi_A = \alpha_A \beta_A$. Thus π_A is a fraction of *harmed–corrected pairs*, not a fraction of individual examples. Interior AUROC further restricts both draws to I_A and is defined only when $\pi_A > 0$.

3.2 Exact Decomposition and SMB

To separate support-induced ranking from interior discrimination, define a support-respecting readout that places the forced boundary outcomes outside its interior scores:

$$U_A(x) = \begin{cases} 1, & x \in F_A^H, \\ 0, & x \in F_A^C, \\ v(x) \in (0,1), & x \in I_A, \\ \text{undefined}, & x \in N_A. \end{cases} \tag{3}$$

Theorem 1 (Support decomposition)**.** *If $\pi_A > 0$, every readout in Equation 3 satisfies*

$$\boxed{\mathrm{AUC}_A(U_A) = 1 - \pi_A + \pi_A \,\mathrm{AUC}_{I_A}(v).} \tag{4}$$

If $\pi_A = 0$, its conditional AUROC is one, while interior AUROC is undefined.

Proofs and boundary cases are in Appendix A.

Corollary 2 (SMB and normalized evidence)**.** *The constant-interior readout $v = 1/2$ defines the Semantic-Membership Baseline (SMB), with conditional AUROC denoted by $\mathrm{AUC}_{\mathrm{SMB},A}$:*

$$\boxed{\mathrm{AUC}_{\mathrm{SMB},A} = 1 - \frac{\pi_A}{2}.} \tag{5}$$

For $\pi_A > 0$, define

$$\boxed{E_A = \frac{\mathrm{AUC}_A(U_A) - \mathrm{AUC}_{\mathrm{SMB},A}}{1 - \mathrm{AUC}_{\mathrm{SMB},A}} = 2\,\mathrm{AUC}_{I_A}(v) - 1.} \tag{6}$$

Here $E_A \in [-1, 1]$ measures discrimination on the interior. Zero indicates chance ranking for this target; for SMB it follows from a constant score. Zero does not generally imply that a model contains no information. Interior discrimination can still use class or transition priors. Neither E_A nor AUROC establishes calibration or causal attribution.

Every defined singleton query has SMB AUROC one, because distinct predictions cannot both belong to it. For $A = \mathcal{K}$, $\pi_A = 1$ and SMB AUROC is $1/2$. The readout's AUROC lies in $[1 - \pi_A, 1]$. Appendix A gives the finite-sample form and boundary cases.

3.3 Raw Scores Need a Support-Order Loss

An arbitrary learned score need not respect the ordering in Equation 3. Let $\mathcal{V}_A(s)$ denote the support-order loss: the weighted error from violating the membership-implied ordering on pairs outside $I_A \times I_A$, with half error for ties. With $E_A(s) = 2\,\mathrm{AUC}_{I_A}(s) - 1$, partitioning the same pair population yields, for $\pi_A > 0$,

$$\mathrm{AUC}_A(s) = 1 - \frac{\pi_A}{2} + \frac{\pi_A}{2} E_A(s) - \mathcal{V}_A(s), \qquad \mathcal{V}_A(s) \geq 0. \tag{7}$$

SMB AUROC need not dominate raw AUROC: raw AUROC exceeds that of SMB exactly when weighted interior evidence outweighs support-order violations. Positive E_A alone is insufficient when π_A is small or $\mathcal{V}_A$ large, explaining the heterogeneous results in Section 5.

A support projection replaces the boundary scores by zero and one and applies a strictly increasing transform into $(0, 1)$ to the interior. It preserves interior ordering and increases this query's AUROC by $\mathcal{V}_A(s)$. It supplies a diagnostic comparison, not a guarantee of improved selection utility. In particular, substituting raw AUROC into Equation 6 confounds evidence with the support-order loss. Equations 4 and 7 must not be used interchangeably.

4 Query Rankings and Decision Utility

Beyond the support decomposition, we ask whether one fixed score can rank optimally across semantic queries and whether H/C ranking identifies the changes most worth rejecting.

4.1 When a Common Scalar Ranking Is Impossible

A concrete ranking conflict. A scorer observing $Z = (a, b)$ encounters $z_1 = (u, v)$ and $z_2 = (v, u)$, with $u \neq v$ and positive mass in each relevant conditional population. Queries $A = \{u\}$ and $B = \{v\}$ require opposite orders:

Query	$z_1 = (u, v)$	$z_2 = (v, u)$	Required score order
$A = \{u\}$	H	C	$s(z_1) > s(z_2)$
$B = \{v\}$	C	H	$s(z_2) > s(z_1)$

A single query-independent scalar cannot realize both orders.

Formalizing the ranking conflict. Fix the scorer's information Z, either full features or a specified coarsening. On a finite observation space, define the query-conditional mass and harm posterior

$$\mu_Q(i) = P(Z = z_i \mid S, T_Q), \qquad \eta_Q(i) = P(H \mid Z = z_i, S, T_Q).$$

The harmed proportion is $p_Q = \sum_i \mu_Q(i)\eta_Q(i) \in (0, 1)$.

Ranking by η_Q maximizes query AUROC among scores of Z; the regret of $s(Z)$ is

$$\mathrm{Reg}_Q(s) = \mathrm{AUC}_Q(\eta_Q) - \mathrm{AUC}_Q(s).$$

For positive-mass observations with $\eta_Q(i) > \eta_Q(j)$, reversing their optimal order costs

$$d_Q(i, j) = \frac{\mu_Q(i)\mu_Q(j)\big(\eta_Q(i) - \eta_Q(j)\big)}{p_Q(1 - p_Q)}. \tag{8}$$

Each unordered pair contributes zero regret if correct, $d_Q(i, j)$ if reversed, and $d_Q(i, j)/2$ if tied; these nonnegative contributions sum to total regret (Appendix B).

Compatibility across queries. Combining these ordering requirements across queries gives the following criterion.

Theorem 3 (Query-dependent ranking incompatibility)**.** *For finitely many queries and observation atoms, add a directed edge $i \to j$ whenever some query gives both atoms positive mass and $\eta_Q(i) > \eta_Q(j)$. There exists a query-independent scalar score that is Bayes-AUROC-optimal for every query if and only if this graph is acyclic. In particular, if $\eta_A(i) > \eta_A(j)$ and $\eta_B(i) < \eta_B(j)$ on common positive-mass atoms, then every fixed scalar satisfies*

$$\boxed{\mathrm{Reg}_A(s) + \mathrm{Reg}_B(s) \geq \min\{d_A(i, j), d_B(j, i)\} > 0.} \tag{9}$$

The criterion is relative to the specified observation Z. Appendix B proves the result; Appendix B.1 extends it to continuous observations under a positive-pair-measure condition.

4.2 Shared Evidence with Query-Dependent Readouts

Conflicting queries require different orders (Theorem 3), which can be obtained from shared evidence. For features X containing a, b, let

$$r(x) = P(H \mid X = x, S)$$

be the shared H/C posterior before query conditioning.

Within S, $Y = a$ on H and $Y = b$ on C, so conditioning on $Y \in A$ gives

$$R_A(x) = \frac{\mathbf{1}[a \in A]r(x)}{\mathbf{1}[a \in A]r(x) + \mathbf{1}[b \in A]\big(1 - r(x)\big)}, \tag{10}$$

for a positive denominator; otherwise the conditional posterior is undefined. Consistent with Section 3, $R_A(x) = 1$ on F_A^H, $R_A(x) = 0$ on F_A^C, and $R_A(x) = r(x)$ on I_A. Equation 10 thus allows query-dependent orders from one shared posterior, whereas strictly increasing query-dependent calibration of a fixed scalar preserves its order and leaves conflicts unresolved.

4.3 Selection Uses Expected Loss Difference

Conditional AUROC ranks H against C, whereas deployment asks which individual changes to reject to reduce expected loss. All changed candidates in D, including W, therefore compete for the fallback budget.

For nonnegative class weights w, define

$$L_w(\hat{y}, Y) = w(Y)\mathbf{1}[\hat{y} \neq Y].$$

Accepting b instead of a adds loss

$$L_w(b, Y) - L_w(a, Y) = w(a)\mathbf{1}_H - w(b)\mathbf{1}_C. \tag{11}$$

Fallback therefore avoids weighted harm on H, loses a correction on C, and leaves loss unchanged on W. On D, let $p_H = P(H \mid X)$ and $p_C = P(C \mid X)$, with $p_H + p_C \leq 1$ and remaining mass on W. Conditioning Equation 11 on X gives expected fallback gain

$$\Delta_w(X) = w(a)p_H - w(b)p_C. \tag{12}$$

The Bayes policy rejects the top k candidates by Δ_w for an exact quota, or up to k positive-gain candidates in decreasing order for an at-most-k budget, following the relative-error principle in paired-model deferral (Jitkrittum et al., 2023).

Why a conditional posterior and a decision score can rank differently. With query-specific weights $w(y) = \mathbf{1}[y \in A]$, the posterior and fallback gain use the same masses as a ratio and a difference:

$$R_A(x) = \frac{\mathbf{1}[a \in A]p_H}{\mathbf{1}[a \in A]p_H + \mathbf{1}[b \in A]p_C}, \qquad \Delta_w(x) = \mathbf{1}[a \in A]p_H - \mathbf{1}[b \in A]p_C.$$

The ratio measures conditional harm probability; the difference retains the magnitude of expected loss reduction. For example, consider the full-label query $A = \mathcal{K}$ with unit weights:

Candidate	p_H	p_C	$R_{\mathcal{K}}$	Δ_w
x_1	0.09	0.01	0.90	0.08
x_2	0.60	0.30	0.67	0.30

The posterior ranks x_1 higher, but rejecting x_2 yields greater expected gain (0.30 versus 0.08): most of x_1's mass is on W, which the conditional ratio removes.

4.4 Support-Aware Evaluation Protocol

Conditional ranking. For each prespecified query, report total and interior H/C counts, π_A, SMB AUROC, raw and projected AUROC, AUC_{I_A}, E_A, and $\mathcal{V}_A$, retaining undefined values. Compute $E_A = 2\,\mathrm{AUC}_{I_A} - 1$ directly to avoid cancellation at small π_A; fix queries, score orientations, and model selection on development data. Report query-level and aggregate results: projected AUROC and E_A rank methods identically within a query with $\pi_A > 0$, but their macro rankings can differ because projected AUROC weights evidence by query-specific π_A.

Decision utility. Specify the candidate population, loss, and budget; include both-wrong candidates and rank without true query membership. Report avoided harm minus lost corrections with the class weights and denominator explicit.

For uncertainty estimates, resample independent units such as images or scenes rather than induced H–C pairs.

Decision-Oriented Reliability Selection. Let $t = P(H \cup C \mid X, D)$ and $r = P(H \mid X, H \cup C)$. Expected query gain factors as $t\{\mathbf{1}[a \in A]r - \mathbf{1}[b \in A](1 - r)\}$: interior evidence evaluates discrimination in r, whereas fallback utility also depends on outcome mass t. We therefore use interior evidence only to select reliability scorers on validation data, then evaluate the selected scorers by downstream fallback utility; Appendix O gives the full analysis.

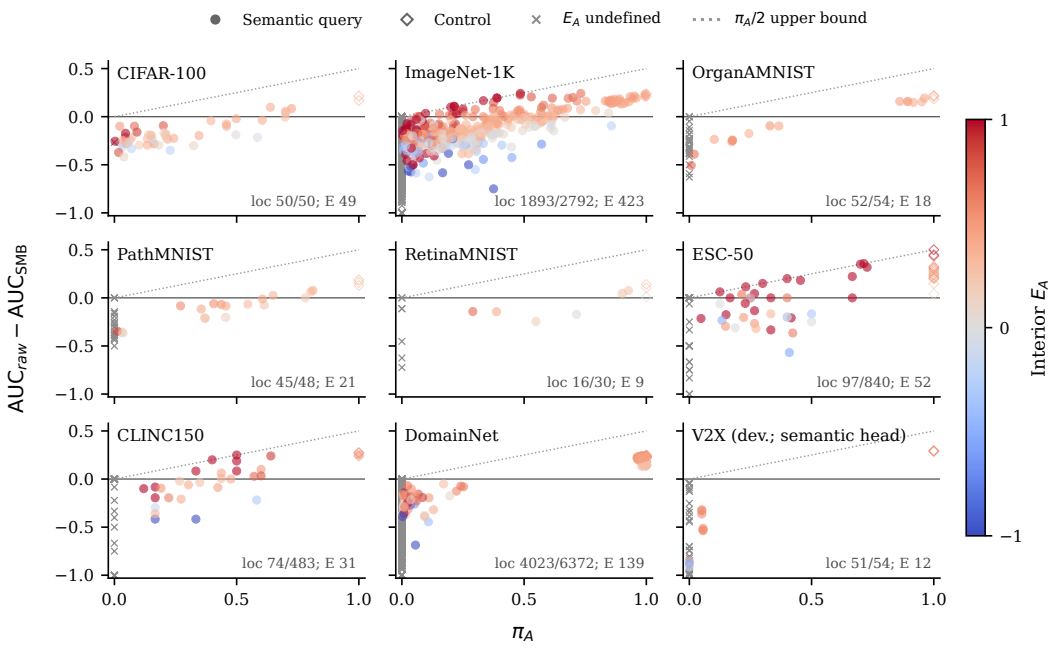


Figure 2: **Interior evidence and aggregate AUROC across semantic sets.** Each point represents a query–fit instance; controls and undefined interior evidence have separate markers. The horizontal axis gives the interior-pair fraction π_A, the vertical axis gives raw conditional AUROC minus SMB AUROC, and color indicates interior evidence E_A. V2X uses raw semantic-head scores on development data (seeds 17, 42, 117); "loc" counts H/C-defined instances and "E" counts defined interiors, including controls. Positive E_A appears on both sides of zero; larger interior pair mass gives it more influence on aggregate ranking.

5 Experiments

We ask four questions: (1) how semantic support shapes conditional reliability-score comparisons; (2) whether selecting scorers by interior evidence rather than raw conditional AUROC changes scorer selection and downstream fallback utility; (3) whether support-aware deployment improves utility without refitting the scorer; and (4) how these gains depend on validation support and experimental scope. Predictor pairs and semantic queries are fixed throughout. Validation selects the scorer or deployment rule, while TEST labels are used only to measure the resulting ranking and decision utility.

5.1 Support Structure and Reliability-Score Comparisons

Across nine tasks, positive interior evidence appears with both positive and negative Raw-minus-SMB gaps (Figure 2). This pattern agrees with Equation 7: the positive interior contribution $\pi_A E_A/2$ competes with support-order loss. **Therefore, even when a reliability score discriminates interior H/C cases well, its aggregate conditional AUROC need not increase.** The complete cross-task decomposition is reported in Appendices L and J.

We then ask whether this distinction changes which reliability scores appear better. Among the practical scores, 18.7% of pairwise score comparisons reverse on average when moving from raw conditional AUROC to interior evidence, and in 37.3% of the 1614 evaluable query–fit cases, the two criteria identify completely different best-score sets. This disagreement remains after requiring stronger interior support on ImageNet: among the 369 cases with at least ten TEST interior H and C observations, 13.2% of pairwise score comparisons reverse, and 32.2% have no overlap between the best scores identified by the two criteria. **Thus, raw conditional AUROC and interior evidence can lead to different practical scorer rankings, and this disagreement persists under stronger**

Table 1: **Fallback utility after scorer selection.** Selected retrospective cases, all using ResNet-18→ResNet-50. Fallback rejects the highest-risk revised predictions and restores the corresponding original predictions. Utility is query harms avoided minus query corrections lost per 100 changed candidates; intervals are pointwise paired 95% CIs. Medical results average all validation-eligible queries; ImageNet uses one 49-class query (split 117). Complete results are in Appendix P.

Dataset / scope	Seed / split	Budget	Raw-selected	Interior-selected	ΔU [95% CI]
OrganAMNIST (3/6 eligible)	17	20%	9.487	10.213	+0.726 [+0.398,+1.078]
PathMNIST (5/6 eligible)	17	20%	1.875	3.750	+1.875 [+1.084,+2.390]
ImageNet (clothing)	117	5%	-0.132	0.071	+0.203 [+0.101,+0.345]

interior-support requirements. Appendix F.1 reports the number of evaluable cases, tied rankings, and results under additional support thresholds.

5.2 Interior-Aware Scorer Selection and Fallback Utility

We next test whether the evaluation difference changes an actual decision. Interior-aware validation selects a frozen reliability scorer by

$$\widehat{m}_A^{\mathrm{int}} = \arg\max_{m \in \mathcal{M}} \widehat{E}_A^{\mathrm{VAL}}(s_m), \qquad \widehat{s}_A(X) = s_{\widehat{m}_A^{\mathrm{int}}}(X).$$

The seven query-agnostic candidates are Diff-Prob, entropy change, margin change, revised-model entropy, semantic logistic, outcome H/C log odds, and outcome net harm. They comprise four fixed prediction-derived heuristics and three frozen learned readouts; their definitions, features, and fitting procedures are detailed in Appendix P. The comparator uses the same candidate menu, validation data, and tie rule, but selects the scorer with the highest raw conditional AUROC. All scorer heads remain frozen; only the validation criterion changes the selected scorer.

On TEST, the selected scorer ranks all changed predictions, including W and out-of-query cases. At budget q, the highest-risk $\lceil qN \rceil$ changes fall back from the revised prediction b to the original prediction a. This gives $+1$ utility for an avoided query H, -1 for a lost query C, and zero otherwise, while every fallback consumes budget. TEST labels are used only to evaluate realized utility; full protocol details are in Appendix P.

In the cases shown in Table 1, changing the validation criterion changes scorer selection and downstream utility. For PathMNIST seed 17, raw conditional AUROC selects outcome H/C log odds over entropy change (0.852 vs. 0.842), whereas interior evidence reverses the choice (0.846 vs. 0.850). At the same 20% TEST budget (109/544 candidates), this change improves avoided H/lost C from 36/47 to 55/19, raising net utility from -11 to 36. Validation/TEST interior supports are 24/17 and 137/160.

A similar change occurs for ImageNet clothing: raw conditional AUROC selects revised-model entropy, whereas interior evidence selects Diff-Prob. At the same 5% TEST budget (493 fallbacks), this change improves avoided H/lost C from 6/19 to 13/6, raising net utility from -13 to $+7$. The small number of validation interior H/C observations (3 H/16 C) motivates the support-sensitivity analysis in Section 5.4. **These results show that selection by interior evidence can change the scorer and improve downstream fallback utility on held-out prediction changes.** Additional domain-shift and exploratory audio cases are reported in Appendix Q.

5.3 Support-Aware Deployment Ablation

We separately ask whether support structure can improve deployment when the reliability scorer itself is fixed. Unlike Section 5.2, which changes validation-time scorer selection, this experiment keeps Diff-Prob fixed and changes only its deployment ranking.

Across seven dataset–model configurations from the preceding scorer-selection experiments, we compare raw Diff-Prob, support-projected Diff-Prob, and SMB at the same 20% fallback budget.

Table 2: **Support-aware deployment at a 20% fallback budget.** Mean query utility per 100 changed candidates. Raw ranks by Diff-Prob alone; SMB ranks by support tiers alone; Proj. applies support projection to Diff-Prob. $\Delta U = \mathrm{Proj.} - \mathrm{Raw}$, with pointwise paired 95% bootstrap CIs. Q is eligible/planned queries, requiring validation interior H/C support ≥ 1 per side. CIFAR uses ResNet-20→ResNet-56; medical tasks use ResNet-18→ResNet-50.

Configuration	Q	Raw	SMB	Proj.	ΔU [95% CI]
CIFAR-100 (17)	18/20	0.19	0.60	0.60	+0.41 [+0.27,+0.56]
OrganAMNIST (17)	3/6	9.67	8.15	12.32	+2.65 [+2.06,+3.42]
OrganAMNIST (42)	2/6	10.37	4.20	10.61	+0.24 [-0.47,+1.33]
OrganAMNIST (117)	4/6	8.91	9.66	12.64	+3.73 [+3.28,+4.38]
PathMNIST (17)	5/6	4.12	4.52	6.25	+2.13 [+0.46,+3.25]
PathMNIST (42)	5/6	0.95	3.74	4.58	+3.63 [+1.66,+4.75]
PathMNIST (117)	5/6	6.57	6.39	7.78	+1.21 [-0.42,+2.53]

Table 3: **Robustness to validation support.** PathMNIST seed 17 at 20% fallback. Utility is reported per 100 changed candidates. Increasing the minimum validation interior H/C support preserves the advantage of interior-aware scorer selection. Complete threshold results are reported in Appendix P.

Min. VAL H/C support	Eligible queries	Raw-selected	Interior-selected	ΔU [95% CI]
≥ 1	5/6	1.875	3.750	+1.875 [+1.084,+2.390]
≥ 5	5/6	1.875	3.750	+1.875 [+1.084,+2.390]
≥ 10	3/6	1.348	4.228	+2.880 [+1.622,+3.676]

All changed TEST candidates compete for the same quota. The projection ranks candidates by their predicted-label membership in query A: a-only ($a \in A, b \notin A$), interior ($a, b \in A$), neither ($a, b \notin A$), and b-only ($a \notin A, b \in A$), while preserving Diff-Prob order within each non-neutral tier. SMB uses the same support tiers but is constant within each tier. Thus, the comparison isolates the effect of adding support structure to a fixed scorer rather than refitting or reselecting it.

Support-projected Diff-Prob yields strictly positive utility gains over raw Diff-Prob in five of seven configurations; the remaining two intervals include zero, and none are strictly negative (Table 2). Projected Diff-Prob also has higher point-estimate utility than SMB in all six medical configurations. **These results show that adding support structure to a fixed reliability scorer can improve downstream fallback utility without refitting or reselecting the scorer.**

5.4 Validation Support and Robustness

We finally test whether the decision benefit persists under stronger validation support. Table 3 repeats PathMNIST seed 17 selection while increasing the minimum validation interior H/C support from one to five and ten observations per side.

At support ≥ 1, interior-aware selection improves utility from 1.875 to 3.750, with $\Delta U = +1.875$ $[+1.084, +2.390]$. The result is unchanged at support ≥ 5. At support ≥ 10, three eligible queries remain, and interior-aware selection improves utility from 1.348 to 4.228, with $\Delta U = +2.880$ $[+1.622, +3.676]$. **The downstream benefit of interior-aware scorer selection therefore persists under substantially stronger validation support.** Complete results across datasets, queries, support thresholds, and budgets are reported in Appendix P.

6 Conclusion

Our support-aware framework separates support-induced ranking from discrimination among unresolved prediction changes. Our framework uses the Semantic-Membership Baseline (SMB) to capture support structure and interior evidence to evaluate and select reliability scorers beyond that structure. **Crucially, selection by interior evidence can change the selected scorer and improve downstream fallback utility on held-out prediction changes.** Together, these results connect support-aware reliability evaluation to practical decision making beyond metric-level differences.

REFERENCES

Patrick Bilic, Patrick Ferdinand Christ, et al. The liver tumor segmentation benchmark (lits). *CoRR*, abs/1901.04056, 2019.

Trenton Chang, Michael W. Sjoding, and Jenna Wiens. Disparate censorship & undertesting: A source of label bias in clinical machine learning. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 343–390. PMLR, 2022. URL https://proceedings.mlr.press/v182/chang22a.html.

Cheng Chen, Hao Huang, and Saurabh Bagchi. Vision-only gaussian splatting for collaborative semantic occupancy prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):2796–2804, 2026. doi: 10.1609/aaai.v40i4.37269. URL https://ojs.aaai.org/index.php/AAAI/article/view/37269.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f4d943fad13060f97dee47e5200367a8-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A model update strategy for compatible LLM evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7320–7332. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.430. URL https://aclanthology.org/2024.findings-emnlp.430/.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010. URL http://dx.doi.org/10.1016/j.patrec.2005.10.010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c699878945bbc2380ea9f421337bea93-Abstract-Conference.html.

Wittawat Jitkrittum, Neha Gupta, Aditya K Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 9891–9906. Curran Associates, Inc., 2023. doi: 10.52202/075280-0431. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1f09e1ee5035a4c3fe38a5681cae5815-Paper-Conference.pdf.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/73e0f7487b8e5297182c5a711d20bf26-Paper.pdf.

Jakob Nikolas Kather, Johannes Krisam, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 01 2019.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP-IJCNLP*, pp. 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, et al. DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6):100512, 2022a. doi: 10.1016/j.patter.2022.100512.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. Preprint.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022b.

João Matos, Ben Van Calster, Richard D. Riley, Paula Dhiman, and Gary S. Collins. Collapsibility of performance metrics in clinical predictive AI. arXiv preprint arXiv:2608.30568, 2026. URL https://arxiv.org/abs/2608.30568.

Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4df3510ad02a86d69dc32388d91606f8-Abstract-Conference.html.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00149.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of ACM Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Stephan Rabanser and Nicolas Papernot. What does it take to build a performant selective classifier? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/71460926102fade443ea7ec89ae8a73a-Abstract-Conference.html.

Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Mitigating negative flips via margin preserving training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8721–8730, 2026. doi: 10.1609/aaai.v40i11.37825. URL https://ojs.aaai.org/index.php/AAAI/article/view/37825.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Raphael Schumann, Elman Mansimov, Yi-An Lai, Nikolaos Pappas, Xibin Gao, and Yi Zhang. Backward compatibility during data updates by weight interpolation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2846–2861. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.174. URL https://aclanthology.org/2024.eacl-long.174/.

Agus Sudjianto and Alice J. Liu. Decomposing global AUC into cluster-level contributions for localized model diagnostics. arXiv preprint arXiv:2508.07495, 2025. URL https://arxiv.org/abs/2508.07495.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

Guoxuan Xia, Olivier Laurent, Gianni Franchi, and Christos-Savvas Bouganis. Towards understanding why label smoothing degrades selective classification and how to fix it. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9dc5accb1e4f4a9798eae145f2e4869b-Abstract-Conference.html.

X. Xu, F. Zhou, et al. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38(8):1885–1898, 2019.

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*, pp. 14299–14308, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B. Blaschko. A novel characterization of the population area under the risk coverage curve (AURC) and rates of finite sample estimators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79226–79253. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25y.html.

A Proofs and Boundary Cases

A.1 Support Posterior and Pair Decomposition

Given $X = x, S$, the mutually exclusive events $Y = a(x)$ and $Y = b(x)$ have probabilities $r(x)$ and $1 - r(x)$. Intersecting with T_A multiplies these probabilities by $\mathbf{1}[a(x) \in A]$ and $\mathbf{1}[b(x) \in A]$. Dividing the first intersection probability by their sum proves Equation 10. A zero denominator means that the conditioning event has zero probability at x; there is no identified query posterior there. On positive conditional support, the posterior is one on F_A^H, zero on F_A^C, and $r(x)$ on I_A.

Theorem 1 requires no calibration of v. On $H \cap T_A$, the observation belongs to F_A^H or I_A with probabilities $1 - \alpha_A$ and α_A. On $C \cap T_A$, it belongs to F_A^C or I_A with probabilities $1 - \beta_A$ and β_A. Independence of the two draws gives four pair masses:

	F_A^C	I_A
F_A^H	$(1-\alpha_A)(1-\beta_A)$	$(1-\alpha_A)\beta_A$
I_A	$\alpha_A(1-\beta_A)$	$\alpha_A\beta_A$.

(13)

The first three cells contribute their entire mass to support-respecting AUROC. The fourth contributes $\pi_A \operatorname{AUC}_{I_A}$. This proves the theorem including interior ties. If $\pi_A = 0$, the fourth cell is absent rather than an undefined AUROC multiplied by zero.

Setting $v = 1/2$ gives interior AUROC $1/2$ whenever defined and proves Equation 5. Subtracting that equation from the decomposition gives $\pi_A(\operatorname{AUC}_{I_A} - 1/2)$; dividing by the structural headroom $\pi_A/2$ proves Equation 6. It also proves $E_A \in [-1, 1]$, attenuation of within-query method differences by π_A, and equality of projected and interior method rankings when $\pi_A > 0$.

For a singleton query, $a \neq b$ makes I_A empty. Its conditional AUROC is one if both H and C exist, while interior AUROC and E are undefined. For the full label universe, every changed prediction is interior, so $\pi_A = 1$ and SMB AUROC is $1/2$. An empty query, or a query missing H or C, has undefined conditional AUROC. Because interior AUROC lies between zero and one, the support-respecting range is $[1 - \pi_A, 1]$. SMB AUROC is not a universal ceiling or floor for arbitrary raw AUROC.

A.2 Finite Samples and Arbitrary Raw Scores

For counts $n_H, n_C > 0$, let n_{HI}, n_{CI} be the corresponding interior counts. Of the $n_H n_C$ pairs in the empirical tie-aware AUROC, exactly $n_{HI} n_{CI}$ are interior. Therefore

$$\widehat{\pi}_A = \frac{n_{HI} n_{CI}}{n_H n_C} \tag{14}$$

and the empirical identity holds exactly. This combinatorial count does not make these pairs independent sampling units. Images share pairs, and voxels can additionally share scenes and temporal context.

For an arbitrary score s, let $B_{uv}(s)$ be AUROC within a positive/negative pair cell and let w_{uv} be its product mass from the table above. Omit cells with zero mass. Define

$$\mathcal{V}_A(s) = \sum_{(u,v) \neq (I_A, I_A)} w_{uv}\{1 - B_{uv}(s)\} \geq 0. \tag{15}$$

Summing all cells gives $\operatorname{AUC}_A(s) = 1 - \pi_A + \pi_A \operatorname{AUC}_{I_A}(s) - \mathcal{V}_A(s)$ for positive π_A, which is Equation 7. When $\pi_A = 0$, the expression is $1 - \mathcal{V}_A(s)$. A strictly order-preserving interior projection removes only $\mathcal{V}_A$. When $\pi_A > 0$, normalizing the raw aggregate contrast instead gives

$$\frac{\operatorname{AUC}_A(s) - \operatorname{AUC}_{\mathrm{SMB},A}}{1 - \operatorname{AUC}_{\mathrm{SMB},A}} = E_A(s) - \frac{2\mathcal{V}_A(s)}{\pi_A}. \tag{16}$$

It need not lie in $[-1, 1]$ and does not isolate interior evidence. Our numerical readout uses $1/4 + (1/2)\sigma(s)$ on the interior to keep even saturated logits strictly separated from boundary values. AUC computation gives exact half credit to any floating-point ties. The decomposition refers to the actual scores supplied to that computation.

B Full Ranking Incompatibility Proof

The premise is realizable in paired prediction. Take equiprobable observations with predictions (u, v) and (v, u), where $u \neq v$, and let either true label have probability $1/2$ at each observation. Queries $\{u\}$ and $\{v\}$ then have opposite deterministic harm rankings. Every common score has the sum of their AUROCs equal to one, while each structural query readout attains one.

For a query Q, write $\mu_i = \mu_Q(i)$, $\eta_i = \eta_Q(i)$ and $p = p_Q$. The positive and negative masses at atom i are $\mu_i\eta_i/p$ and $\mu_i(1-\eta_i)/(1-p)$. For an unordered pair with $\eta_i > \eta_j$, the two possible oriented contributions are

$$A_{ij} = \frac{\mu_i\mu_j\eta_i(1-\eta_j)}{p(1-p)}, \qquad B_{ij} = \frac{\mu_i\mu_j\eta_j(1-\eta_i)}{p(1-p)}. \tag{17}$$

Ranking i above j contributes A_{ij}, reversing them contributes B_{ij}, and tying contributes $(A_{ij} + B_{ij})/2$. Their difference is $d_Q(i, j) > 0$. Equal-posterior pairs impose no ordering constraint, and diagonal pairs always tie. Thus, with $\ell_{ij} = 0, 1, 1/2$ for correct order, reversal, and tie,

$$\mathrm{Reg}_Q(s) = \sum_{\eta_Q(i) > \eta_Q(j)} d_Q(i, j)\ell_{ij}(s), \tag{18}$$

where every unequal-posterior unordered pair appears once and only positive-mass pairs are included. Ordering by η_Q makes every term zero, establishing Bayes optimality.

For simultaneous optimality, every positive coefficient in every query requires strict score order on its edge. A directed cycle contradicts transitivity of real-number order. Conversely, a finite acyclic graph has a vertex without an incoming edge: following incoming edges indefinitely otherwise repeats a vertex. Remove that vertex and repeat. The resulting topological sequence admits strictly decreasing scores, satisfying every edge and making every term of Equation 18 zero. This proves both directions of Theorem 3. Cycles spanning more than two queries can obstruct compatibility even without a direct two-query reversal.

For the reversal in Equation 9, $s_i > s_j$ incurs $d_B(j, i)$, $s_i < s_j$ incurs $d_A(i, j)$, and a tie incurs half their sum. Each case is at least their minimum. Other regret terms are nonnegative. Consequently the larger query regret is at least half the displayed bound. A randomized query-independent ordering obeys the same sum bound in expectation. Query-dependent monotone transforms cannot change the strict pair order or its ties.

In the paired two-atom construction from the main text, each query has atom masses $1/2, 1/2$, prevalence $1/2$, and a posterior difference of one. Both coefficients equal one. An ordering contributes AUROCs $(1, 0)$, $(0, 1)$, or $(1/2, 1/2)$, proving their sum is always one. The construction includes predictions in the observation and is therefore realizable within the paired setting rather than only as abstract conditional laws.

B.1 Positive Pair Measure for Continuous Observations

Let the two conditional observation laws have densities f_A, f_B with respect to a common probability measure λ; their equally weighted mixture is one choice. Let $\eta_Q(z)$ be the conditional harm posterior and $p_Q \in (0, 1)$. Define

$$k_Q(z, z') = \frac{f_Q(z)f_Q(z')|\eta_Q(z) - \eta_Q(z')|}{p_Q(1-p_Q)} \tag{19}$$

and the oriented reversal region

$$\mathcal{D} = \{(z, z') : \eta_A(z) > \eta_A(z'),\ \eta_B(z) < \eta_B(z')\}. \tag{20}$$

If $c = \int_{\mathcal{D}} \min\{k_A, k_B\}\, d\lambda(z)d\lambda(z') > 0$, every measurable common score obeys $\mathrm{Reg}_A(s) + \mathrm{Reg}_B(s) \geq c$.

To prove the claim, group (z, z') and (z', z) in each AUC integral as in the finite proof. Integrating over positive posterior differences counts each unequal pair once. On $\mathcal{D}$, query A contributes $k_A\ell$; swapping variables in query B's reversed region gives $k_B(1-\ell)$, including $\ell = 1/2$ at score ties. Their sum is pointwise at least $\min(k_A, k_B)$. All other regret contributions are nonnegative, so integration proves the result. A reversal at two isolated points with zero pair measure gives no positive bound.

Table 4: Fixed 20% fallback budget on all changed CIFAR candidates, seed 17. NetGain uses query-weighted loss per changed candidate. Ratio and delta use the same posterior. Intervals are paired image bootstraps; boundary ties use expected random selection. No contrast excludes zero.

Pair	Query	Ratio NetGain	Delta NetGain	Delta minus ratio [95% CI]
R20→R56	all	0.03062	0.03909	0.00847 [-0.0027, 0.0151]
R20→R56	animals	0.03127	0.03257	0.00130 [-0.0065, 0.0073]
R20→R56	plants	0.01824	0.01824	0.00000 [0.0000, 0.0000]
R20→R56	vehicles	0.00586	0.00586	0.00000 [0.0000, 0.0000]
R20→VGG13	all	0.00534	0.00534	0.00000 [-0.0067, 0.0069]
R20→VGG13	animals	0.02068	0.02201	0.00133 [-0.0026, 0.0048]
R20→VGG13	plants	0.00467	0.00467	0.00000 [0.0000, 0.0000]
R20→VGG13	vehicles	0.00600	0.00600	0.00000 [0.0000, 0.0000]

C Utility Derivation and Policy Scope

Equation 11 follows by considering the only three possible outcomes on D. If $Y = a$, accepting b introduces loss $w(a)$. If $Y = b$, accepting b removes loss $w(b)$. Otherwise both predictions incur $w(Y)$, so their difference is zero. Conditional expectation gives Equation 12. Under an exact rejection quota, exchanging a rejected point of smaller expected gain with a retained point of larger gain cannot lower total expected gain. Repeating the exchange gives the largest-Δ_w rejection rule, with arbitrary randomization at ties. For an at-most quota, candidates with negative expected gain remain accepted, and zero-gain candidates are optional.

Let $m = P(S \mid X)$ and $r = P(H \mid X, S)$. Then $\Delta_w = m\{w(a)r - w(b)(1 - r)\}$. For $m > 0$, the factor cancels from an unconstrained sign decision but not in general from budgeted ranking. For $w = \mathbf{1}_A$, the query ratio and delta therefore target different decisions. Cases with zero ratio denominator require a separately specified full-population policy; setting them to 0.5 is a diagnostic convention, not an identified posterior.

Table 4 reports all fixed CIFAR query contrasts at the primary budget. Policies see predicted labels and scores, not $Y \in A$. The weight on the unknown true class is read only to evaluate the loss. Bootstrap utility shares boundary ties fractionally, giving the expected policy under random tie breaking. Ratios for plants and vehicles can have zero difference at this budget because the policies yield the same net benefit; this does not prove equal rankings or equality at every budget. Raw policy CSVs additionally retain avoided H, lost C, rejected W, and improvements divided by all 5,000 images.

C.1 Downstream Decision Utility

We additionally test whether H/C ranking translates into downstream fallback utility. All changed predictions, including both-wrong cases, compete for a fixed 20% fallback budget; utility is harmful changes avoided minus useful corrections lost.

The preferred readout varies by task and configuration (Figure 3). ImageNet favors query ratio in the evaluated comparisons, DomainNet shows mixed behavior, and most remaining intervals include zero. The V2X counterfactual study likewise shows that improving interior AUROC need not improve fallback gain (Table 5); Appendix H details its protocol.

These results show that aggregate conditional ranking, interior discrimination, and downstream decision utility capture distinct properties of a reliability score.

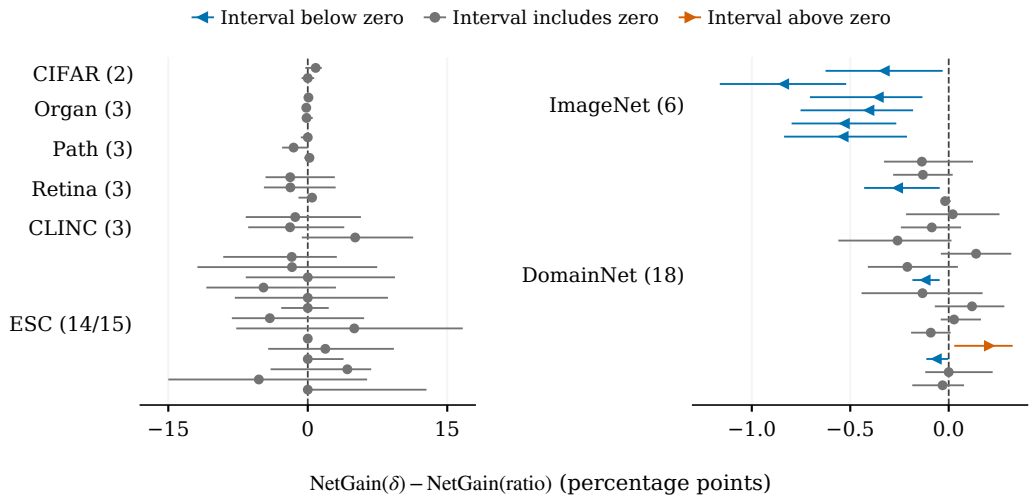


Figure 3: **Decision utility at a fixed 20% fallback budget.** Points show $\mathrm{NetGain}(\delta) - \mathrm{NetGain}(\mathrm{ratio})$ for each available fit on the complete-label query $A = \mathcal{K}$; horizontal lines are paired 95% bootstrap intervals. Negative differences favor ratio and positive differences favor delta. Parentheses give available fit counts (14 of 15 for ESC); the panels use different horizontal scales. All changed predictions compete for the same fallback budget.

Table 5: Negative occupancy probe pilot, seed 17. Interior AUROC is for occupied ambiguity. The 20% rejection quota includes unknown-label candidates; NetGain divides avoided harm minus lost corrections by the valid changed-candidate count. CAL contains two scenes and was reused in residual refinement. No CF variant passes the joint gate against its matched semantic control.

Variant	VAL interior	VAL NetGain	CAL interior	CAL NetGain
semantic	0.80495	0.06973	0.73001	0.05107
loo	0.80688	0.06386	0.74363	0.04998
half	0.81529	0.07000	0.71142	0.04952
counterfactual	0.80591	0.06790	0.72639	0.05107
semantic_residual	0.80405	0.06979	0.73333	0.05102
loo_residual	0.80435	0.06979	0.73026	0.05096
counterfactual_residual	0.80828	0.06937	0.72688	0.05122

D CIFAR-100 Protocol and Reproducibility

The Animals query and its complement reverse harm orderings for the same positive-mass boundary transitions under the four-valued predicted-label membership observation. Table 7 gives empirical regret bounds for this observation, realizing Theorem 3's conflict. The certificate concerns scalars of this specified observation; query-dependent readouts of a shared H/C posterior can adapt the ordering.

The source commit is `786c16252c0fc58ee9adac063f8337cc4a7a497a`. Normalization follows its author's training configuration: means $(0.5070, 0.4865, 0.4409)$ and standard deviations $(0.2673, 0.2564, 0.2761)$, with tensor conversion and no test augmentation. Full official-test top-1 accuracies are 68.83%, 72.63%, and 74.63% for ResNet-20, ResNet-56, and VGG-13-BN. Model states are checked unchanged during inference, and saved logits and checkpoints have recorded SHA-256 hashes. The official archive MD5 is `eb9058c3a382ffc7106e4002c42a8d85`.

A stratified seed of 20260914 selects 50 evaluation images per fine class. Seeds 17, 42, and 117 split the other 50 into 30 reliability TRAIN and 20 VAL images per class. No images overlap within a split. All fitted heads are locked before opening the fixed evaluation results. All reliability heads and scorer selections are evaluated on the fixed local holdout defined above. Frozen GPU inference used one NVIDIA RTX 4090; the complete Slurm job finished in 21 seconds. Reliability heads and bootstrap evaluation ran on CPU.

Table 6: Fixed CIFAR-100 superclass comparison. Raw AUROC averages all 20 queries. SMB has $E_A = 0$; Diff-Prob E_A averages 19/20 defined queries. The raw gap is SMB AUROC minus raw Diff-Prob AUROC. Brackets are paired 95% intervals from 2,000 image bootstraps stratified by fine class; undefined interior queries are excluded in each resample.

Pair	SMB AUROC	Diff-Prob raw	Raw gap [95% CI]	Diff-Prob E_A [95% CI]
R20→R56	0.9195	0.7088	0.2106 [0.1745, 0.2494]	0.3036 [0.1486, 0.4507]
R20→VGG13	0.9177	0.6750	0.2427 [0.2036, 0.2811]	0.3600 [0.1508, 0.4716]

Table 7: Empirical ranking certificates for the four-valued animal-membership observation. Each count and posterior lists a-only / b-only directions. Both directions have positive mass under both queries. The final column bounds the sum of their AUROC regrets for any scalar of this observation.

Pair	Animal counts	Complement counts	Animal η	Complement η	Bound
R20→R56	27/48	46/31	1 / 0	0 / 1	0.02207
R20→VGG13	23/54	47/31	1 / 0	0 / 1	0.02636

Queries. Fine-to-coarse mapping comes from the official labels. All 20 superclass queries are evaluated. Animals unite aquatic mammals, fish, insects, large carnivores, large omnivores/herbivores, medium mammals, non-insect invertebrates, reptiles, and small mammals (45 fine classes, excluding people). Plants unite flowers, fruit/vegetables, and trees (15). Vehicles unite both vehicle groups (10). The animal complement contains 55 classes, and All contains all 100. These are fixed semantic groupings rather than taxonomic claims. All 100 singleton queries are used for boundary checks.

Scores and fitting. Each paired feature vector contains the two 100-way probability vectors, both entropies, both maximum probabilities, both top-two margins, $p_b(a)$, and $p_a(b)$: 208 features. Standardization fits only changed TRAIN candidates. The binary head learns H/C; the multinomial head learns C/H/W from all changed candidates. Logistic regression uses L-BFGS with a 2,000-iteration limit. The regularization search is $C \in \{0.01, 0.1, 1\}$, selected by matching VAL log loss. Every pair, split, and head selects $C = 0.01$.

The fixed higher-means-harm scores are $H(p_b)$, $H(p_b) - H(p_a)$, $\max p_a - \max p_b$, and the margin difference. Fitted readouts are the binary decision function, $\log p_H - \log p_C$, and $p_H - p_C$. Adding SMB, the query posterior ratio, and the query loss difference gives ten methods. SMB uses only label membership; its constant interior contains no learned discrimination. The normalized interior evidence E_A is computed from within-query interior AUROC and is not obtained by normalizing raw conditional AUROC.

Statistical summaries. The 95% intervals use 2,000 paired image bootstraps stratified by true fine class. Interior AUROC is defined for 19 superclass queries in R20→R56 and all 20 in R20→VGG13. Resampling can remove an interior outcome; the macro-average includes 16–19 and 14–20 queries with defined interior H/C comparisons across draws. Weighting E_A by within-query harmed–corrected pair counts gives 0.3478 [0.2031, 0.4807] and 0.2507 [0.1012, 0.4041]. This pair-weighted summary gives greater weight to queries with more interior H–C comparisons.

The primary contrast is fixed to SMB versus Diff-Prob on the superclass macro before prediction inspection. Image bootstrap resamples within each fine class using seed 20260914 and 2,000 draws. The same sampled images are used for both compared scores and both classifier pairs. Counts of defined E queries are retained for every draw. The pair-weighted statistic is $\sum_A n_{HI,A} n_{CI,A} E_A / \sum_A n_{HI,A} n_{CI,A}$, with zero-pair queries contributing zero weight. It differs from the unweighted query macro by emphasizing queries with more interior comparisons. Neither weighting counts pairs as independent experimental units.

The additional ordinary-method witness is selected on seed-17 VAL among entropy, entropy difference, Diff-Prob, margin difference, the binary head, and the two non-query outcome readouts. Eligible reversals require at least five interior H and C examples. Selection maximizes the smaller of the raw-AUROC gap and half the E gap. The selected query and method pair are then fixed on the

local holdout. All rejection fractions $0, 0.05, 0.1, 0.2, 0.5$ are retained, with 0.2 the primary budget. Undefined query ratios are completed at 0.5 for the stated policy stress test.

E Complete CIFAR Query Diagnostics

The following tables preserve empty and sparse interiors. The classifier pairs share the same images, providing an architecture-pair comparison on a common evaluation population. Values are per-query point estimates for the complete fixed query set.

Table 8: R20→R56: all fixed CIFAR queries on the local holdout, Diff-Prob, seed 17. H/C are total conditional counts; IH/IC are interior counts. Raw and projected AUROC are distinct columns. Undefined E is shown as –.

Query	H/C	IH/IC	π_A	SMB	Raw	Proj.	E_A
aquatic mammals	22/34	4/19	0.102	0.949	0.654	0.959	0.184
fish	21/34	8/12	0.134	0.933	0.703	0.952	0.292
flowers	26/20	15/12	0.346	0.827	0.631	0.898	0.411
food containers	23/22	9/12	0.213	0.893	0.666	0.905	0.111
fruit and vegetables	15/21	9/7	0.200	0.900	0.806	0.978	0.778
household electrical devices	13/22	4/5	0.070	0.965	0.633	0.958	-0.200
household furniture	19/32	8/16	0.211	0.895	0.715	0.933	0.359
insects	22/30	6/9	0.082	0.959	0.861	0.988	0.704
large carnivores	18/31	5/6	0.054	0.973	0.713	0.989	0.600
large man-made outdoor things	11/14	5/5	0.162	0.919	0.636	0.961	0.520
large natural outdoor scenes	17/21	7/8	0.157	0.922	0.683	0.916	-0.071
large omnivores and herbivores	23/33	9/6	0.071	0.964	0.748	0.974	0.259
medium mammals	14/27	0/2	0.000	1.000	0.733	1.000	–
non-insect invertebrates	26/41	5/7	0.033	0.984	0.665	0.982	-0.086
people	32/40	22/29	0.498	0.751	0.730	0.848	0.392
reptiles	17/42	4/6	0.034	0.983	0.699	0.979	-0.250
small mammals	27/37	10/15	0.150	0.925	0.732	0.956	0.413
trees	19/31	15/18	0.458	0.771	0.671	0.859	0.385
vehicles 1	11/27	5/12	0.202	0.899	0.734	0.912	0.133
vehicles 2	18/15	4/3	0.044	0.978	0.763	0.996	0.833
animals	190/309	163/261	0.725	0.638	0.722	0.804	0.459
plants	60/72	45/44	0.458	0.771	0.694	0.866	0.415
vehicles	29/42	21/37	0.638	0.681	0.779	0.808	0.398
non animals	204/265	173/219	0.701	0.650	0.706	0.778	0.367
all	394/574	394/574	1.000	0.500	0.713	0.713	0.427

Table 9: R20→VGG13: all fixed CIFAR queries on the local holdout, Diff-Prob, seed 17. H/C are total conditional counts; IH/IC are interior counts. Raw and projected AUROC are distinct columns. Undefined E is shown as –.

Query	H/C	IH/IC	π_A	SMB	Raw	Proj.	E_A
aquatic mammals	26/33	11/17	0.218	0.891	0.656	0.922	0.283
fish	22/39	9/6	0.063	0.969	0.726	0.978	0.296
flowers	18/24	10/14	0.324	0.838	0.537	0.861	0.143
food containers	25/23	12/10	0.209	0.896	0.697	0.920	0.233
fruit and vegetables	16/25	6/9	0.135	0.932	0.637	0.958	0.370
household electrical devices	7/19	1/6	0.045	0.977	0.805	1.000	1.000
household furniture	19/31	11/13	0.243	0.879	0.654	0.929	0.413
insects	7/26	1/4	0.022	0.989	0.890	0.995	0.500
large carnivores	17/32	3/9	0.050	0.975	0.697	0.985	0.407
large man-made outdoor things	12/11	6/4	0.182	0.909	0.614	0.917	0.083

Table 10: R20→R56: all ten readouts, superclass macro mean ± sample standard deviation across the three reliability splits. Fixed heuristics are identical across reliability splits and therefore have zero split variation.

Method	Raw AUROC	Projected AUROC	E_A
SMB	0.9195 ± 0.0000	0.9195 ± 0.0000	0.0000 ± 0.0000
Collaborative entropy	0.6756 ± 0.0000	0.9406 ± 0.0000	0.2076 ± 0.0000
Entropy change	0.7067 ± 0.0000	0.9457 ± 0.0000	0.2591 ± 0.0000
Diff-Prob	0.7088 ± 0.0000	0.9472 ± 0.0000	0.3036 ± 0.0000
Margin change	0.7191 ± 0.0000	0.9511 ± 0.0000	0.3507 ± 0.0000
Semantic logistic	0.6821 ± 0.0127	0.9496 ± 0.0035	0.3740 ± 0.0370
Outcome log odds	0.6680 ± 0.0126	0.9470 ± 0.0038	0.3426 ± 0.0482
Outcome net harm	0.6674 ± 0.0110	0.9477 ± 0.0039	0.3660 ± 0.0424
Query ratio	0.9470 ± 0.0038	0.9470 ± 0.0038	0.3426 ± 0.0482
Query delta	0.8828 ± 0.0110	0.9477 ± 0.0039	0.3660 ± 0.0424

Table 11: R20→VGG13: all ten readouts, superclass macro mean ± sample standard deviation across the three reliability splits. Fixed heuristics are identical across reliability splits and therefore have zero split variation.

Method	Raw AUROC	Projected AUROC	E_A
SMB	0.9177 ± 0.0000	0.9177 ± 0.0000	0.0000 ± 0.0000
Collaborative entropy	0.6718 ± 0.0000	0.9367 ± 0.0000	0.3132 ± 0.0000
Entropy change	0.6687 ± 0.0000	0.9340 ± 0.0000	0.3377 ± 0.0000
Diff-Prob	0.6750 ± 0.0000	0.9358 ± 0.0000	0.3600 ± 0.0000
Margin change	0.6854 ± 0.0000	0.9374 ± 0.0000	0.3762 ± 0.0000
Semantic logistic	0.6589 ± 0.0012	0.9367 ± 0.0021	0.2906 ± 0.0343
Outcome log odds	0.6506 ± 0.0068	0.9361 ± 0.0033	0.2940 ± 0.0155
Outcome net harm	0.6465 ± 0.0056	0.9382 ± 0.0033	0.3254 ± 0.0055
Query ratio	0.9361 ± 0.0033	0.9361 ± 0.0033	0.2940 ± 0.0155
Query delta	0.8589 ± 0.0020	0.9382 ± 0.0033	0.3254 ± 0.0055

Query	H/C	IH/IC	π_A	SMB	Raw	Proj.	E_A
large natural outdoor scenes	15/21	8/9	0.229	0.886	0.537	0.860	-0.222
large omnivores and herbivores	17/34	2/5	0.017	0.991	0.619	0.998	0.800
medium mammals	13/27	1/1	0.003	0.999	0.744	1.000	1.000
non-insect invertebrates	15/40	5/7	0.058	0.971	0.698	0.970	-0.029
people	31/44	25/33	0.605	0.698	0.661	0.815	0.389
reptiles	21/37	5/6	0.039	0.981	0.562	0.983	0.133
small mammals	15/40	6/14	0.140	0.930	0.742	0.947	0.238
trees	17/41	17/24	0.585	0.707	0.492	0.696	-0.039
vehicles 1	13/25	3/10	0.092	0.954	0.794	0.994	0.867
vehicles 2	14/25	3/4	0.034	0.983	0.737	0.989	0.333
animals	153/308	130/254	0.701	0.650	0.692	0.767	0.335
plants	51/90	40/57	0.497	0.752	0.564	0.757	0.020
vehicles	27/50	13/41	0.395	0.803	0.760	0.870	0.343
non animals	187/289	156/242	0.699	0.651	0.647	0.746	0.272
all	340/597	340/597	1.000	0.500	0.668	0.668	0.337

F Practical-Score Rankings and Query-Level Illustrations

Method comparisons use the same query population within a pair. The raw/projected difference records support-order loss. Projected and E method rankings must agree for a fixed query with positive interior pair mass, but their cross-query macros need not agree because the projected macro implicitly weights each interior term by its π_A. Table 12 expands the fish and vehicles queries from the earlier candidate analysis into complete practical-score rankings for both CIFAR pairs at reliability

Table 12: Illustrative rankings of real reliability scores under raw conditional AUROC and interior evidence. Both predictor pairs use reliability split 17 and the shared local holdout. Every available practical score is shown; ranks use unrounded values with average ranks for exact ties. These four fixed examples show how raw conditional AUROC and interior evidence can rank the same practical reliability scores differently.

Method	Raw AUROC	Raw rank	E_A	E_A rank
R20→R56, fish				
outcome_query_ratio	0.9650	1.0	0.4792	3.5
outcome_query_delta	0.8992	2.0	0.4375	5.5
entropy_change	0.7507	3.0	0.5208	1.0
diff_prob	0.7031	4.0	0.2917	8.0
margin_change	0.7003	5.0	0.3750	7.0
collab_entropy	0.6891	6.0	0.2292	9.0
semantic_logistic	0.6485	7.0	0.5000	2.0
outcome_log_odds	0.6120	8.0	0.4792	3.5
outcome_net_harm	0.6050	9.0	0.4375	5.5
R20→R56, vehicles				
outcome_query_ratio	0.8005	1.0	0.3745	7.0
entropy_change	0.7906	2.0	0.4414	3.0
margin_change	0.7816	3.0	0.3977	4.0
outcome_query_delta	0.7800	4.0	0.4414	1.5
diff_prob	0.7791	5.0	0.3977	5.0
outcome_net_harm	0.7406	6.0	0.4414	1.5
semantic_logistic	0.7225	7.0	0.3745	7.0
collab_entropy	0.7176	8.0	0.3205	9.0
outcome_log_odds	0.7159	9.0	0.3745	7.0
R20→VGG13, fish				
outcome_query_ratio	0.9802	1.0	0.3704	2.5
outcome_query_delta	0.8939	2.0	0.3333	4.5
entropy_change	0.7354	3.0	0.1852	8.0
margin_change	0.7284	4.0	0.2963	6.5
diff_prob	0.7261	5.0	0.2963	6.5
collab_entropy	0.7214	6.0	-0.0370	9.0
outcome_log_odds	0.6142	7.0	0.3704	2.5
semantic_logistic	0.6131	8.0	0.4074	1.0
outcome_net_harm	0.6107	9.0	0.3333	4.5
R20→VGG13, vehicles				
outcome_query_ratio	0.8385	1.0	0.1820	6.5
margin_change	0.7748	2.0	0.4146	1.0
outcome_query_delta	0.7652	3.0	0.1520	8.5
diff_prob	0.7600	4.0	0.3433	2.0
collab_entropy	0.7526	5.0	0.3246	3.0
entropy_change	0.7444	6.0	0.2608	4.5
semantic_logistic	0.6770	7.0	0.2608	4.5
outcome_log_odds	0.6504	8.0	0.1820	6.5
outcome_net_harm	0.6489	9.0	0.1520	8.5

Table 13: Practical-score ranking consistency on registered primary queries. N is the number of evaluable instances; correlations exclude constant rankings. D/N and T/N report mean strict-reversal and tie proportions over finite method pairs; ρ is the median Spearman correlation. Top-1 columns report fixed lexicographic tie-breaking (Fixed) and disjoint best-method sets (Disjoint). All percentages are descriptive; V2X uses development data.

Dataset	N	N_τ	Median τ_b [IQR]	ρ	D/N (%)	T/N (%)	Fixed (%)	Disjoint (%)
CIFAR-100	117	114	0.20 [0.06, 0.41]	0.22	30.8	17.5	100.0	82.1
ImageNet-1K	1263	1236	0.39 [0.17, 0.61]	0.47	24.1	19.4	85.1	63.3
OrganAMNIST	15	15	0.29 [0.09, 0.77]	0.34	26.1	7.0	80.0	73.3
PathMNIST	18	18	0.68 [0.43, 0.74]	0.82	18.7	12.5	83.3	55.6
RetinaMNIST	6	6	0.46 [0.40, 0.56]	0.54	24.5	6.9	66.7	33.3
ESC-50	37	36	0.56 [0.27, 0.85]	0.65	9.7	43.4	59.5	18.9
CLINC150	28	26	0.48 [0.34, 0.73]	0.59	12.0	37.7	60.7	25.0
DomainNet	121	120	0.86 [0.18, 0.97]	0.94	14.1	13.6	43.0	28.1
V2X (development)	9	9	0.00 [-0.05, 0.10]	0.00	50.8	1.6	100.0	100.0

split 17. These examples make the ranking changes concrete; the systematic analysis below evaluates the same phenomenon over all eligible primary queries. The systematic analysis below includes all evaluable primary queries rather than selecting individual reversal candidates.

F.1 Systematic Ranking Consistency

Scope and eligibility. We reuse completed per-query metrics and existing scalar prediction caches, with no model training or additional inference. An instance fixes the dataset, predictor pair, reliability or training seed, fold, domain, and semantic query. Within each instance we retain only methods with finite Raw AUROC and E_A, requiring at least two such methods. The nine classification candidates are collaborative entropy, entropy change, Diff-Prob, margin change, the semantic logistic head, outcome log odds, outcome net harm, outcome query ratio, and outcome query delta. SMB, support projections, and other forced boundary-order diagnostics never enter the candidate set. The query ratio and delta are learned posterior readouts that use the query membership; a separate sensitivity analysis removes both to distinguish their contribution from that of the query-agnostic scores. ESC-50 has five to nine available methods per evaluable instance because some reliability heads are unavailable. V2X uses seven existing raw learned scores: semantic, conditional experts, shared anchor, anchor residual, joint ranking, group ranking, and worst-group ranking. These seven heads provide distinct learned reliability rankings on the same V2X development population.

Within-instance statistics. Write $r_i = \mathrm{AUC}_A(s_i)$ and $e_i = E_A(s_i)$ for the m eligible methods in an instance. We compute Kendall τ_b, which accounts for ties, and Spearman ρ, using average ranks for ties. Since $E_A = 2\,\mathrm{AUC}_{I_A} - 1$, interior AUROC produces exactly the same ranking as E_A. Correlations remain undefined if either ranking is constant. Let $N = \binom{m}{2}$, and define

$$D = \sum_{i<j} \mathbf{1}\{(r_i - r_j)(e_i - e_j) < 0\}, \qquad T = \sum_{i<j} \mathbf{1}\{r_i = r_j \text{ or } e_i = e_j\}.$$

Our main pairwise disagreement proportion is D/N, with tie proportion T/N reported separately. Tied pairs remain in this denominator but never count as reversals; the companion statistic $D/(N-T)$ considers only strictly comparable pairs and is undefined when $N = T$. Primary calculations use exact equality of the stored values, with a 10^{-12} tolerance sensitivity reported separately.

Top-1 comparison needs an explicit tie convention. We record the full maximizer sets $B_r = \arg\max_i r_i$ and $B_e = \arg\max_i e_i$. The main text reports $\mathbf{1}\{B_r \cap B_e = \varnothing\}$: the two criteria have no shared best method. We also report the requested single-winner mismatch using the same fixed lexicographic method-name tie-break for both rankings, and mismatch restricted to instances with unique maxima under both criteria. The archived records include both sets, tie indicators, and the upper bound allowing different choices from tied maxima.

Coverage and aggregation. The main summary uses the previously registered primary queries; additional queries, all-class controls, and singleton controls remain in the full CSV with their status and exclusion reasons. There are 1614 evaluable instances among 3978 primary instances, with

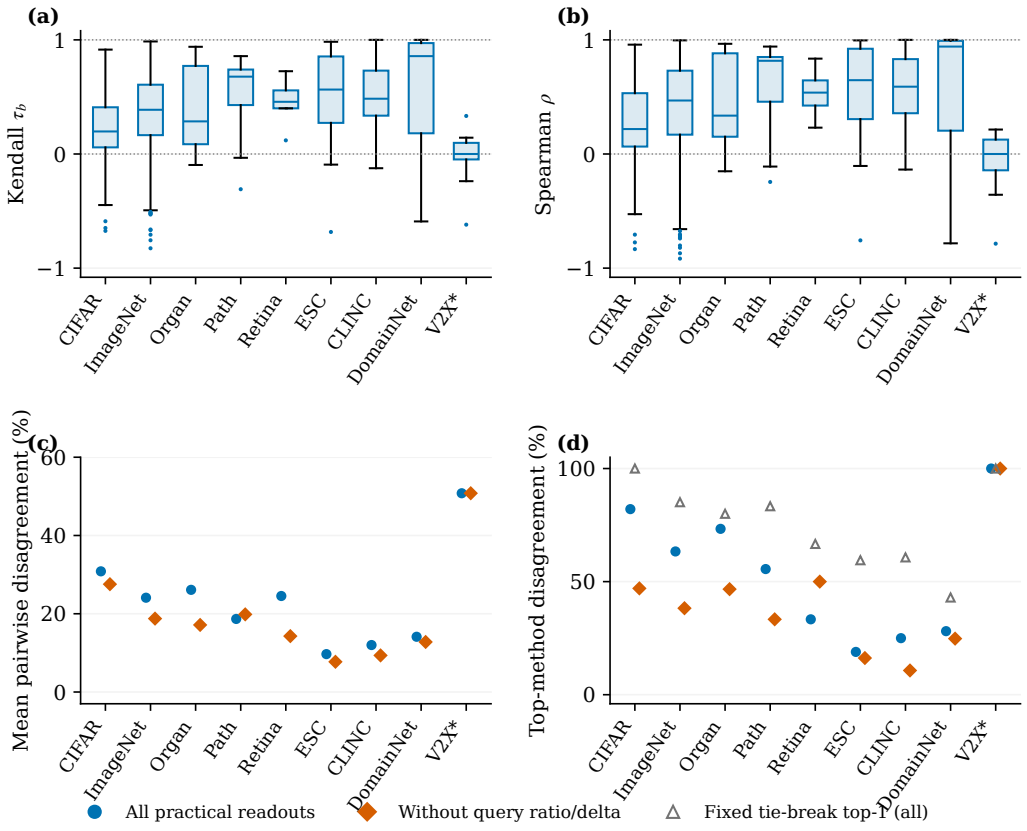


Figure 4: **Ranking consistency between raw conditional AUROC and interior evidence across real reliability scores.** (a–b) Kendall τ_b and Spearman ρ distributions: boxes show quartiles and medians, whiskers extend to observations within 1.5 IQR, and points beyond the whiskers are shown individually. (c) Mean within-instance strict-reversal proportion D/N, including tied pairs in the denominator. (d) Circles and diamonds show the fraction of instances with disjoint best-method sets; open triangles show single-winner mismatch with fixed tie-breaking. Blue uses all practical readouts; orange removes query ratio/delta. Panels (a–b) use 1580 instances with defined correlations, while (c–d) retain all 1614 evaluable instances. Panels (a–b) summarize valid query-level instances; each marker in (c–d) summarizes one dataset. Agreement varies substantially across datasets: several tasks show frequent pairwise reversals and disjoint best-method sets, while others retain higher rank correlation. SMB is excluded from all candidate method sets. V2X* denotes development data.

1580 defined correlation values and 34 constant-ranking instances. In particular, the three additional ImageNet architecture pairs each store only Diff-Prob: their 1185 primary instances cannot compare methods, regardless of interior support. The remaining unavailable instances have missing H/C or interior H/C support; undefined evidence is never replaced by zero. ImageNet contributes 1263 evaluable instances, so the pooled description primarily reflects its overlapping hierarchy queries and repeated reliability splits. Within a dataset, the displayed reversal proportion is the arithmetic mean of instance proportions, rather than a test based on the number of method pairs. Separate files retain predictor-pair and fit-level summaries, as well as a hierarchical macro that first averages queries within a fit, then folds/domains within a seed, seeds within a predictor pair, and predictor pairs within a dataset. Queries share observations, domains share source-trained models, and frozen-model reliability splits reuse evaluation images. No query-level or method-pair-level significance tests are performed, and no pseudo-independent confidence intervals are attached.

Observed alignment. Across evaluable primary instances, median Kendall τ_b is 0.389 (IQR 0.151–0.647) and median Spearman ρ is 0.469 (IQR 0.168–0.767). Mean and median D/N are 23.4% and 22.2%, respectively; mean tie proportion is 19.4%, and mean $D/(N-T)$ is 29.6%. Using a

Table 14: Interior-support sensitivity of practical-score rankings. Each cell reports evaluable N / mean strict pairwise disagreement (%) / disjoint best sets (%). Threshold m requires at least m interior H and m interior C observations. Filtering changes the query composition; empty strata are NA, not zero disagreement.

Dataset	Readouts	$m = 1$	$m = 5$	$m = 10$
CIFAR-100	All	117 / 30.8 / 82.1	81 / 31.9 / 85.2	30 / 29.1 / 93.3
CIFAR-100	No ratio/delta	117 / 27.6 / 47.0	81 / 27.6 / 43.2	30 / 15.7 / 23.3
ImageNet	All	1263 / 24.1 / 63.3	600 / 26.1 / 74.3	369 / 24.6 / 71.0
ImageNet	No ratio/delta	1263 / 18.8 / 38.2	600 / 16.4 / 39.0	369 / 13.2 / 32.2
OrganAMNIST	All	15 / 26.1 / 73.3	13 / 26.9 / 76.9	12 / 26.4 / 75.0
OrganAMNIST	No ratio/delta	15 / 17.1 / 46.7	13 / 16.5 / 46.2	12 / 16.3 / 41.7
PathMNIST	All	18 / 18.7 / 55.6	16 / 18.2 / 56.2	15 / 17.0 / 60.0
PathMNIST	No ratio/delta	18 / 19.8 / 33.3	16 / 18.2 / 31.2	15 / 15.2 / 26.7
RetinaMNIST	All	6 / 24.5 / 33.3	6 / 24.5 / 33.3	5 / 25.6 / 40.0
RetinaMNIST	No ratio/delta	6 / 14.3 / 50.0	6 / 14.3 / 50.0	5 / 12.4 / 40.0
ESC-50	All	37 / 9.7 / 18.9	0 / NA / NA	0 / NA / NA
ESC-50	No ratio/delta	37 / 7.7 / 16.2	0 / NA / NA	0 / NA / NA
CLINC150	All	28 / 12.0 / 25.0	4 / 24.3 / 50.0	0 / NA / NA
CLINC150	No ratio/delta	28 / 9.4 / 10.7	4 / 10.7 / 0.0	0 / NA / NA
DomainNet	All	121 / 14.1 / 28.1	83 / 7.2 / 15.7	74 / 3.5 / 6.8
DomainNet	No ratio/delta	121 / 12.8 / 24.8	83 / 4.6 / 10.8	74 / 2.1 / 5.4
V2X (dev.)	All	9 / 50.8 / 100.0	6 / 54.0 / 100.0	6 / 54.0 / 100.0
V2X (dev.)	No ratio/delta	9 / 50.8 / 100.0	6 / 54.0 / 100.0	6 / 54.0 / 100.0

fixed tie-break, raw AUROC and interior evidence select different top scorers in 1323/1614 instances (82.0%). More conservatively, their sets of tied best scorers have no overlap in 976/1614 instances (60.5%). At least one criterion has a tied maximum in 41.8% of instances; among the 940 instances with two unique maxima, the mismatch proportion is 76.5%. The disjoint-set statistic therefore provides the stricter measure of top-scorer disagreement. Equal weighting of the nine dataset means gives 23.4% pairwise disagreement and 53.3% disjoint maxima, distinct from both pooled-instance and hierarchical summaries.

Removing query ratio/delta raises median τ_b to 0.619 and reduces mean pairwise disagreement to 18.7%, fixed-tie-break mismatch to 51.3%, and disjoint maxima to 37.3%. The 10^{-12} numerical-tie sensitivity gives 23.0% pairwise disagreement and 59.9% disjoint maxima with all readouts. The pattern therefore survives removal of the two query-dependent posterior readouts, but its magnitude depends on the candidate set. The support-threshold analysis shows that ranking disagreement depends partly on interior sample size: for DomainNet, requiring at least ten interior H and ten interior C observations reduces mean disagreement from 14.1% to 3.5%. No ESC-50 instance meets a five-per-outcome interior threshold, and no CLINC150 instance meets a ten-per-outcome threshold. These empty strata retain zero coverage and undefined summaries in the archived sensitivity analysis.

Interpretation. Across the evaluated reliability scores, raw conditional AUROC and interior evidence frequently induce different pairwise rankings and different top-scorer choices. This pattern remains after removing the two query-dependent posterior readouts, showing that the disagreement is not driven solely by query-aware scores. Table 14 further shows how the effect changes as stronger interior-support requirements are imposed. Together, these results demonstrate that aggregate conditional ranking and unresolved-case discrimination can lead to materially different scorer comparisons.

F.2 IMAGENET QUALITATIVE CASE SELECTION

Figure 5 reuses frozen ResNet-18 (ImageNet-1K V1) and ResNet-50 (V2) outputs and reliability fit 17, without retraining, calibration, or changes to score orientation. The registered working-dog query is `wnid_n02103406`, with zero-based class IDs 222–250 and 252. The ranking population requires changed predictions, $Y \in A$, and exactly one correct prediction; interior cases additionally require both predicted labels in A. GT and H/C annotations are used only for offline evaluation, not

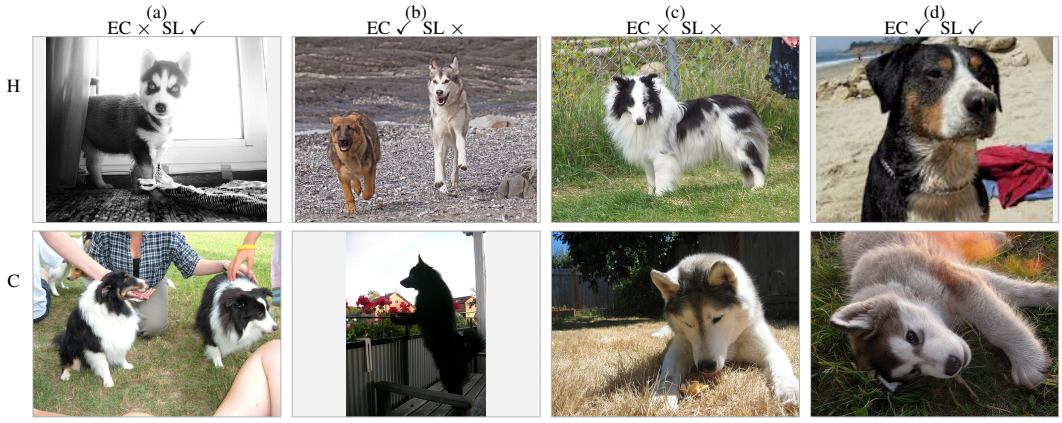


Figure 5: **Real interior H–C ranking examples.** Each column pairs a harmed image (H, top) with a corrected image (C, bottom): (a) entropy change ranks the pair incorrectly and semantic logistic correctly; (b) entropy change ranks it correctly and semantic logistic incorrectly; (c) both rank it incorrectly; (d) both rank it correctly. ResNet-18 to ResNet-50 predictions, fit 17, and the registered working-dog query are fixed; both predictions belong to the query for all eight images. Larger scores indicate harm. Tables 15 and 16 give the case records and complete-query metrics. The four columns visualize all four combinations of correct and incorrect H–C ordering for the two compared scores.

as scoring inputs. Table 15 records all eight displayed images and their original risk scores; Table 16 reports metrics on the complete valid query population. Exact score ties receive half credit.

The case-selection procedure enumerated all 395 primary semantic queries and registered controls across the six existing ImageNet evaluations (two revised predictors, three reliability fits), prioritizing ResNet-18 to ResNet-50. For display eligibility, each interior outcome required at least 20 images, and opposite raw/interior AUROC gaps each had to be at least 0.01; these are presentation filters, not significance criteria or changes to the original evaluation protocol. Fit 17 and the 30-class working-dog query were selected for semantic interpretability, populated boundaries, and a descriptive reversal before viewing the images, without maximizing the ranking gap.

With NumPy seed 20260918, every interior H–C pair received a random priority in stable sample order. The original three pairs (entropy-change wrong/semantic-logistic correct, both wrong, and both correct) were selected in that order using the lowest available priority without repeated images. To complete the four outcomes, we retained those pairs and added the lowest-priority entropy-change correct/semantic-logistic wrong pair excluding all previously selected images. The display order is entropy-change wrong/semantic-logistic correct, entropy-change correct/semantic-logistic wrong, both wrong, and both correct. All 2,403 candidate interior pairs were retained: 541 in the first disagreement direction, 363 in the reverse direction, 600 shared errors, and 899 shared correct orders, with no ties for these two scores. The original JPEG files are embedded directly in LaTeX; all eight photographs use identical 4:3 display frames, with aspect-preserving display crops where suitable and an uncropped fit with white side margins for the portrait in column (b), while the source files remain unchanged. No stretching, retouching, or generative reconstruction is applied.

Table 15: Offline annotations and frozen risk scores for Figure 5. IDs abbreviate `ILSVRC2012_val_<ID>.JPEG`. EC: entropy change; SL: semantic logistic. Both predictions are in the query for every row. Scores are compared within each method, not across scales.

Pair	Type	Image ID	GT	Old a	Revised b	EC	SL
(a)	H	00039381	Eskimo dog	Eskimo dog	Siberian husky	3.1919	-3.0526
(a)	C	00049871	Shetland sheepdog	collie	Shetland sheepdog	3.6474	-3.2306
(b)	H	00029190	Eskimo dog	Eskimo dog	malamute	3.6525	-3.7383
(b)	C	00029723	schipperke	groenendael	schipperke	3.2957	-3.6092
(c)	H	00037075	collie	collie	Shetland sheepdog	3.7538	-3.1797
(c)	C	00030708	Siberian husky	malamute	Siberian husky	4.5635	-2.6202
(d)	H	00010617	EntleBucher	EntleBucher	Appenzeller	3.7740	-1.3983
(d)	C	00017722	malamute	Siberian husky	malamute	2.2956	-3.4899

Table 16: Complete-query metrics for the same frozen configuration: 43 H and 157 C overall, with 27 H and 89 C in the interior. These metrics are computed from all valid query examples, not the eight displayed images. $E_A = 2\,\mathrm{AUC}_{I_A} - 1$.

Score	Raw AUROC	Interior AUROC	E_A	Support-order loss
Entropy change	0.606429	0.525177	0.050354	0.224559
Semantic logistic	0.591172	0.599251	0.198502	0.266183

G Occupancy Data and Earlier Reliability Models

The local Semantic-OPV2V split contains 31 TRAIN, 3 VAL, 2 calibration, and 7 development TEST scenes, with 4,252/655/582/885 complete frames. Each frame contains a $100 \times 100 \times 8$ grid. True labels 1–12 are occupied, 13 is empty, and 0 is ignored. Unknown-label candidates still enter fallback budget ordering but are excluded from scored errors. The occupied query is the set 1–12; road/sidewalk uses labels 5 and 6; built structure uses 1, 2, 9, 10, and 12. We retain the implemented treatment of predicted unknown label 0 and the valid-truth evaluation population.

The frozen global score used in the decomposition is a $48 \to 128 \to 64 \to 1$ head. Inputs contain ego and collaborative probability vectors, their difference, both entropies and their difference, both maximum probabilities, and Jensen–Shannon divergence. TRAIN normalization is fixed. The head was learned from an earlier 186-frame TRAIN cache, with 18 VAL and 12 calibration frames, using AdamW at learning rate 10^{-3}, weight decay 10^{-4}, and batches of 4,096. It has 14,849 parameters. Training balances outcome labels and scenes within each label; a positive affine calibration is inherited and does not change rankings. The full-frame support diagnostic reuses this frozen head under three seeds.

Earlier experiments explored conditional experts, affine alignment, shared anchors, and gated residuals. The conditional experts mix predicted-occupancy-weighted probabilities from two condition-specific heads without ground-truth inputs at inference. Their raw All/occupied AUROCs are 0.6594/0.7957, compared with 0.7913/0.4460 for the global score. Later residual objectives use full-frame TRAIN and therefore have a different training history. These observations motivated the conditional diagnosis. Despite their different training histories, all scores obey the same support decomposition, allowing the structural effect to be evaluated consistently.

For completeness, the projected occupied score averages 0.987960 across the three frozen heads, while interior AUROC averages 0.784519. Table 27 uses seed 17 (0.988129 and 0.787544); Figure 2 retains all three seeds. The SMB and pair fraction do not depend on the head. The seven development scenes provide 885 complete frames for the full-frame V2X support analysis.

H Counterfactual Probe Analysis

The probe pilot freezes cached Gaussian messages and reruns the fusion/backend after whole-collaborator deletion or complementary half-message deletion. It zeros the selected source's opacity through the existing filter while retaining ego messages. Predictions are recomputed rather than

Table 17: V2X support decomposition on all 885 development TEST frames, frozen semantic head, seed 17. Projected AUROC uses the structural readout and must not be interpreted as raw-head AUROC. For Built structure, SMB/projected AUROC values round to one; interior counts reveal the sparse evidence.

Query	π_A	SMB	Projected	Interior	E_A	Interior H/C
All	1.00000	0.50000	0.79459	0.79459	0.58918	248704/396298
Occupied	0.05587	0.97206	0.98813	0.78754	0.57509	12745/43071
Empty	0.00000	1.00000	1.00000	–	–	0/0
Road/sidewalk	0.05096	0.97452	0.98908	0.78576	0.57153	4264/16809
Built structure	1.19e-07	1.00000	1.00000	0.66667	0.33333	4/9

interpolated. With one collaborator, whole-collaborator deletion exactly reproduces ego-only prediction, as verified in the GPU extraction. It adds no new observation beyond the paired baseline. Partial-message deletion supplies a different model response, but does not identify a real-world causal effect on semantic truth.

There are 186/18/12 clean raw TRAIN/VAL/CAL frames across 31/3/2 scenes; 145/15/10 frames have changes. Candidate counts are 139,401/19,169/19,727; occupied interior H/C counts total 10,559/1,025/376. This pilot population differs from the complete-frame support diagnostic. Original ego/collaborative outputs align with zero voxel mismatches. Unknown labels compete for budget but do not contribute to measured error.

The first stage compares equally sized $92 \to 128 \to 64 \to 3$ heads using semantic, deletion-only, half-deletion, or all-probe features through fixed masks. The loss is natural outcome cross-entropy plus ambiguous H/C binary cross-entropy, with the same VAL-loss selection rule. A second development stage freezes the semantic head and trains zero-initialized residuals within predicted occupied-to-occupied cases, comparing semantic-only and probe residuals at equal capacity. This refinement has already observed CAL. Scores outside the designated predicted region remain unchanged.

Table 5 reports the pilot results. The joint gate requires CAL interior AUROC to exceed the matched semantic control by at least 0.01 and NetGain by at least 0.001 at the prespecified 20% rejection budget. No probe variant satisfies both components of the prespecified joint gate. Whole-agent deletion gains 0.01362 interior AUROC but loses 0.00109 NetGain relative to the first-stage semantic control. Because the single-collaborator observation is redundant, that contrast also cannot be attributed to additional counterfactual information. The residual comparison is made against the corresponding semantic residual, not a weaker old head.

The pilot uses one seed and two CAL scenes and is reported as a controlled diagnostic of whether the additional probe features improve the two target metrics. More observations can weakly expand the class of feasible population scores or policies when they may be ignored and their cost is excluded. That elementary inclusion argument guarantees neither strict benefit nor successful finite-sample learning. The probe results therefore separate counterfactual feature augmentation from the support effect studied in the main analysis.

I Locked Cross-Domain Extension Protocol

Completed original fixed matrix. All 60 training configurations, three frozen-model inference configurations and 36 paired evaluations are complete. The latter comprise 51 fit/domain summaries because DomainNet evaluates six domains with each source-trained pair. A separate architecture-family validation adds three frozen checkpoints and three fixed-score pair evaluations (Appendix K). The repository retains machine-readable class/query mappings, data-role counts, SHA-256 hashes, every query/control result and unavailable-head reasons.

Fixed model pairs and training. The three MedMNIST tasks (Yang et al., 2023) pair the official small-image ResNet-18 and ResNet-50 implementations (He et al., 2016), trained from scratch on native 28×28 images. Each run uses 100 epochs, Adam at 10^{-3}, batch size 128, learning-rate factors of 0.1 after epochs 50 and 75, normalization to $[-1, 1]$, and no image augmentation. OrganAMNIST supplies axial organ crops derived from abdominal CT (Bilic et al., 2019; Xu et al., 2019); PathMNIST

supplies colorectal histology patches (Kather et al., 2019); RetinaMNIST supplies five-level fundus severity labels from DeepDRiD (Liu et al., 2022a). The latter is treated as five-class classification to keep the paired-prediction setting fixed.

ESC-50 (Piczak, 2015) pairs AudioSet-pretrained PANNs Cnn6 and Cnn14 (Kong et al., 2020). Audio is mono at 32 kHz. The author frontend uses a 1,024-sample window, hop 320, 64 mel bins, and 50–14,000 Hz range. A 50-class cross-entropy head replaces the AudioSet output; all trainable parameters are fine-tuned for 50 epochs with Adam at 10^{-4} and batch size 16. GPU training uses automatic mixed precision, with the audio spectrogram and log-mel operations kept in FP32 to preserve finite log energies. The author model's training-time spectrogram masking and dropout remain active; evaluation disables them.

CLINC150 (Larson et al., 2019) uses only the 150 in-scope intents and pairs BERT-base-uncased (Devlin et al., 2019) with RoBERTa-base (Liu et al., 2019). Runs use 10 epochs, AdamW at 2×10^{-5}, weight decay 0.01, batch size 32, maximum length 64, 10% warmup then linear decay, and gradient norm clipping at one. All newly trained models use seeds 17, 42 and 117, paired by seed. Classifier checkpoints minimize classifier-validation cross-entropy; the final test set is not evaluated during training.

Predefined primary queries. OrganAMNIST uses left/right femur $\{1, 2\}$, left/right kidney $\{4, 5\}$, left/right lung $\{7, 8\}$, and each complement. PathMNIST uses $\{0, 5, 7\}$ (adipose, smooth muscle, cancer-associated stroma), $\{6, 8\}$ (normal colon mucosa, colorectal adenocarcinoma epithelium), $\{1, 2, 4\}$ (background, debris, mucus), and each complement. These are researcher-defined unions, not an official clinical taxonomy. The lymphocyte class 3 remains included in the appropriate complements. RetinaMNIST uses severity at least 1, 2 or 3; severity at least 4 is a singleton boundary control, not a primary query. ESC-50 uses its five official broad sound groups and CLINC150 its ten official intent domains. Every dataset also includes the full class set and all singletons as controls. Controls are excluded from primary-query macros. No query is selected using predictions or outcomes.

Classifier and reliability isolation. For MedMNIST and CLINC150, the official training split is divided by class into 80% classifier TRAIN, 12% reliability TRAIN and 8% reliability VAL. Official validation selects the classifier and official test supplies final evaluation. A pretraining exact-text audit of CLINC150 removes five training observations overlapping later roles; all 3,000 official validation and 4,500 official test observations remain. Normalized duplicate texts are grouped in internal splitting and resampling.

ESC-50 completes five outer folds for every seed. Fold f is test, fold $f + 1$ modulo five supplies reliability data, and the other three folds supply classifier training/validation. Internal classifier and reliability splits keep original recording identifiers together; reliability uses a 60/40 group split. Four source identifiers (131943, 134049, 209698 and 234879) occur across official folds. Before training, we exclude their conflicting *development* clips from the relevant roles: test sources take precedence over reliability sources, which take precedence over classifier sources. Each official test fold retains all 400 clips. Metrics are computed within folds before equal fold averaging, so no AUROC compares scores from different fold models.

Because the downloaded MedMNIST arrays do not expose patient or scan identifiers, resampling is performed at image level; group identifiers are used whenever they are available.

Frozen ImageNet appendix experiment. ImageNet-1K (Russakovsky et al., 2015) uses torchvision ResNet-18 V1, ResNet-50 V2 and ConvNeXt-T V1 (Liu et al., 2022b), giving two pairs with the same ResNet-18. We verify the official validation archive and devkit checksums, all 50,000 image identities and the ground-truth-to-WordNet mapping. The class order is the sorted WordNet identifier order used by torchvision. Forty images per class form the fixed 40,000-image evaluation set; the other ten are split 6/4 into reliability TRAIN/VAL under three split seeds. The three splits vary reliability TRAIN/VAL assignments while keeping the frozen classifiers and 40,000-image evaluation population fixed. All 395 different proper, nonsingleton descendant label sets in the devkit hierarchy are retained. Query size relationships are descriptive, with no assumed monotonicity or query independence.

Table 18: Primary-query support ranges across all seeds and folds. IH/IC ranges include zeros. Overlapping queries reuse observations, and pair products serve as weights rather than sample sizes. Summary means average queries with defined metrics; SD is variation across seeds with defined E, whose number is reported explicitly. Per-seed contributing fold counts are retained in the result JSON.

Task	IH range	IC range	Mean π	SMB	Raw	SD(E)	Seeds
CLINC150	0–5	0–11	0.363	0.818	0.766	0.116	3/3
ESC-50	0–4	0–11	0.273	0.864	0.784	0.053	3/3
OrganAMNIST	0–820	0–412	0.393	0.804	0.723	0.028	3/3
PathMNIST	1–284	12–204	0.445	0.778	0.661	0.036	3/3
RetinaMNIST	0–66	0–48	0.520	0.740	0.670	0.232	3/3

DomainNet appendix experiment. The official cleaned DomainNet lists (Peng et al., 2019) define Real source training and six separate domain tests. Real training is divided by class into 70/10/12/8% classifier TRAIN, classifier VAL, reliability TRAIN and reliability VAL. ImageNet-pretrained ResNet-18 V1 and ResNet-50 V2 are separately fine-tuned for 20 epochs with SGD at 0.01, momentum 0.9, weight decay 10^{-4}, batch size 64 and cosine decay. Training uses random resized 224×224 crops and horizontal flips; evaluation resizes to 256 and center-crops 224. The source-trained classifier pair and reliability heads remain fixed for all six target-domain evaluations. No target-domain labels are used to fit either component.

The fixed named subsets, each paired with its complement, are: cat/dog/horse/cow/sheep/elephant; airplane/bicycle/bus/car/train/truck; apple/banana/grapes/pear/pineapple/watermelon; and cello/drums/guitar/piano/trumpet/violin. These are declared examples, not exhaustive animal, transport, fruit or instrument taxonomies. Only protocol-required images are extracted; hashes retain the provenance of task-owned archives removed after validation.

Shared diagnostics and uncertainty. The common interface contains aligned sample IDs, true class IDs, paired logits, data roles, and available source/domain/fold identifiers. Frozen metadata defines the class order and queries. We reuse the CIFAR probability features and fixed score orientations. Binary H/C and three-outcome H/C/W logistic heads choose $C \in \{0.01, 0.1, 1\}$ using reliability-validation cross-entropy. Missing required fitting outcomes produce an explicit unavailable-head reason. No final-test metric selects classifiers, heads, queries or model pairs.

Every query retains H/C/W and interior H/C counts, π_A, SMB, raw and projected AUROC, interior AUROC, E_A, and the support-order loss. Projection uses a bounded rank transform for interior values, preserving their order and ties even when logits saturate. Absent H or C makes conditional AUROC undefined; absent interior H or C makes E_A undefined. Neither missing jobs nor undefined statistics are assigned zero.

The two primary contrasts are SMB AUROC minus raw Diff-Prob AUROC and Diff-Prob interior E_A. Each fit uses 2,000 paired, class-stratified bootstrap resamples of observations or available source groups. Each draw takes the mean over queries still defined in that draw; its contributing-query range and valid replicate count are reported. The pair-weighted E_A sensitivity gives more weight to queries with more interior H–C comparisons. Overlapping queries and pairs reuse observations. Three training-seed summaries are descriptive means and standard deviations; per-fit intervals are not relabeled as seed-mean intervals. For ESC-50, fold summaries are averaged within each seed and missing contributing folds are disclosed. All changed examples, including W, compete for the fixed 20% main fallback budget before truth is used to measure utility.

I.1 Support and Locked Classifier Accuracy

I.2 ImageNet: Hierarchical Queries at Larger Scale

Both frozen pairs meet the joint macro criterion. R18→R50 has SMB/raw AUROC 0.9178/0.7107, with gap 0.2071 [0.1865, 0.2268] and $E = 0.2654$ [0.2080, 0.3593]. R18→ConvNeXt-T gives 0.9182/0.7249, gap 0.1933 [0.1692, 0.2116] and $E = 0.3283$ [0.2552, 0.3969]. These fixed-score

Table 19: Accuracy (%) of the locked classifiers: seed mean [minimum, maximum]. ESC-50 averages all five official test folds within each seed. ImageNet uses shared frozen predictions, so changing the reliability split cannot change classifier accuracy. A larger architecture is not assumed to improve accuracy; all fixed pairs are retained.

Task / pair	Domain	Classifier a	Classifier b
CLINC150	all	95.9 [95.8, 96.0]	96.5 [96.5, 96.6]
DomainNet	clipart	38.8 [38.5, 39.1]	45.3 [45.1, 45.4]
DomainNet	infograph	14.9 [14.7, 15.0]	20.5 [20.5, 20.6]
DomainNet	painting	38.2 [38.0, 38.4]	44.7 [44.4, 44.9]
DomainNet	quickdraw	4.0 [3.9, 4.1]	5.1 [4.4, 5.5]
DomainNet	real	77.5 [77.4, 77.5]	83.1 [83.0, 83.2]
DomainNet	sketch	25.2 [25.0, 25.3]	31.4 [31.1, 31.8]
ESC-50	all	86.3 [86.0, 86.6]	93.7 [93.5, 94.0]
ImageNet-1K (ConvNeXt-T)	all	69.7 [69.7, 69.7]	82.5 [82.5, 82.5]
ImageNet-1K (R50)	all	69.7 [69.7, 69.7]	80.7 [80.7, 80.7]
OrganAMNIST	all	93.1 [92.9, 93.4]	91.0 [90.5, 91.9]
PathMNIST	all	89.7 [88.8, 91.3]	89.7 [88.8, 90.3]
RetinaMNIST	all	53.0 [50.0, 54.8]	50.6 [47.5, 52.2]

Table 20: ImageNet-1K primary-query summaries. All locked pairs and domains are retained; undefined queries remain in the accompanying CSV. ImageNet repeats reliability splits of the same frozen classifiers; DomainNet repeats three independent source-domain training seeds.

Pair	Domain	Defined interior	SMB	Raw	E
R18→ConvNeXt-T	all	207/395	0.918	0.725	0.328
R18→R50	all	214/395	0.918	0.711	0.265

contrasts are identical across the three reliability splits because the splits share the same classifiers and test images.

Of 395 hierarchy queries, raw/interior comparisons are defined in 362/214 queries for R50 and 355/207 for ConvNeXt-T. Among defined interiors, the smaller H/C count is at most five for 122/214 and 119/207 queries. Both pairs have 34 negative-E queries, and raw AUROC exceeds SMB AUROC in 40 and 43 queries, respectively. The query-level counts show that the macro advantage coexists with heterogeneous individual-query behavior.

Figure 6 retains every hierarchy query with a defined ordinate. Descriptive Spearman correlations of query size with π_A are 0.630/0.620 for R50/ConvNeXt-T, compared with 0.251/0.114 for E_A. Individual queries need not follow a monotone relation. Fine-grained queries frequently lack interior comparisons. Because WordNet queries overlap, these correlations describe the observed hierarchy rather than treating queries as independent samples.

At the fixed 20% all-change fallback budget, every one of the six fitted ImageNet comparisons has a strictly negative expected-delta-minus-ratio utility interval. Thus the ratio policy outperforms the estimated expected-delta policy in these comparisons. This compares two readouts of the fitted outcome head, not SMB versus Diff-Prob, and does not contradict the oracle expected-loss identity.

I.3 DomainNet: Source-Fitted Models Across Six Domains

All 18 domain–seed evaluations have positive macro interior-evidence intervals; painting at seed 17 and quickdraw at seed 117 also meet the joint criterion. Their SMB-minus-raw AUROC gaps are 0.0409 [0.0090, 0.0733] and 0.1234 [0.0572, 0.1901]. Across all 18 domain–seed evaluations, interior evidence remains positive, while the SMB–raw gap varies across domains and seeds. This separates the persistence of interior discrimination from the direction of the aggregate Raw-minus-SMB contrast under domain shift.

Raw comparisons are defined for 7–8 of eight queries and interior comparisons for 4–8. In quickdraw at seeds 42 and 117, every small semantic subset lacks interior support; the interior macro therefore

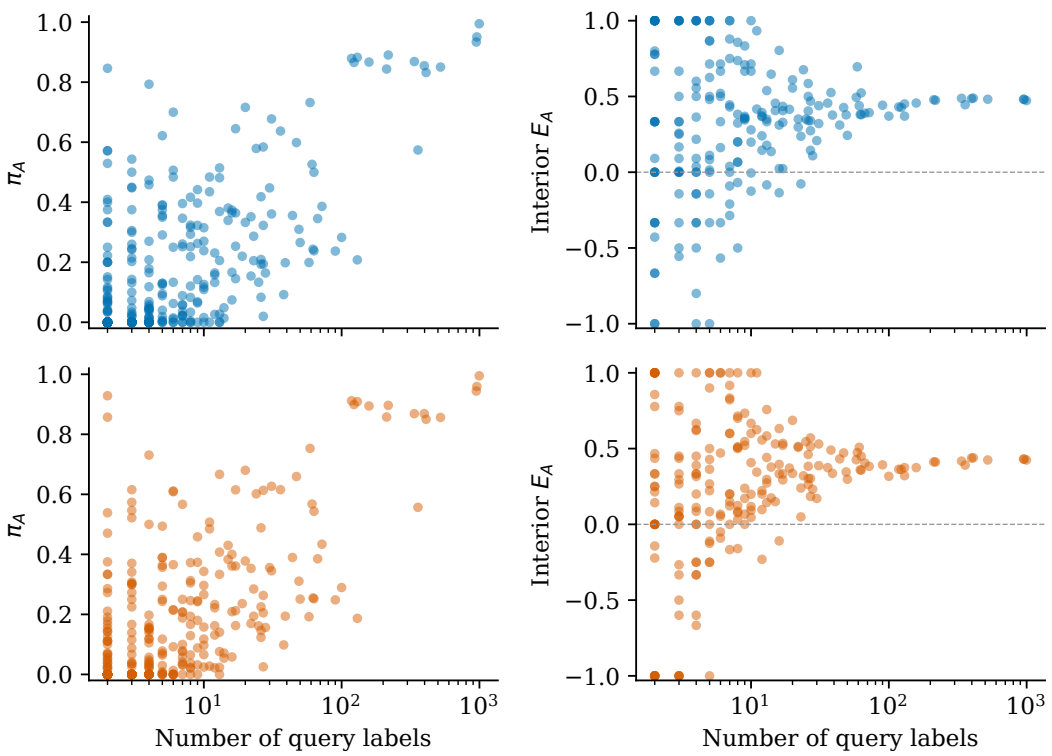


Figure 6: All 395 prespecified ImageNet hierarchy queries, with undefined ordinates omitted rather than assigned zero. Top: R18→ConvNeXt-T; bottom: R18→R50. Left: interior pair fraction π_A versus query cardinality; right: Diff-Prob interior evidence. Each dot is a query on the same fixed evaluation images, and hierarchy queries may overlap. Repeated reliability splits do not change these fixed-score quantities.

Table 21: DomainNet primary-query summaries. All locked pairs and domains are retained; undefined queries remain in the accompanying CSV. ImageNet repeats reliability splits of the same frozen classifiers; DomainNet repeats three independent source-domain training seeds.

Pair	Domain	Defined interior	SMB	Raw	E
R18→R50	clipart	7–8/8	0.744	0.741	0.367
R18→R50	infograph	5–7/8	0.734	0.695	0.314
R18→R50	painting	7–8/8	0.749	0.730	0.305
R18→R50	quickdraw	4–5/8	0.728	0.614	0.310
R18→R50	real	8/8	0.734	0.744	0.504
R18→R50	sketch	7/8	0.751	0.756	0.486

summarizes the four complements. At the fixed fallback budget, one expected-delta-minus-ratio interval is positive (painting, seed 117), three are negative (painting, seed 17; quickdraw, seeds 42/117), and fourteen include zero, showing domain- and seed-dependent utility differences.

I.4 Per-Fit Uncertainty and Undefined Comparisons

The following intervals retain every completed fit. An undefined comparison is marked by a dash; the machine-readable query tables retain its explicit reason. Valid bootstrap counts report how many resamples contain defined interior comparisons.

Table 22: Every new fit: paired 95% bootstrap intervals for SMB AUROC minus raw Diff-Prob AUROC and interior E. Fold indices start at zero. IQ is the range of defined interior queries across resamples; B is the number of finite E replicates out of 2,000. Each interval averages queries with defined comparisons for one frozen fitted model.

Task/pair	Domain	Seed/fold	Gap [95% CI]	E [95% CI]	IQ; B
OrganAMNIST	all	17/0	0.101 [0.051, 0.170]	0.501 [0.385, 0.581]	4–5; 2000
OrganAMNIST	all	42/0	0.095 [0.062, 0.142]	0.444 [0.310, 0.547]	4–5; 2000
OrganAMNIST	all	117/0	0.044 [0.007, 0.083]	0.468 [0.360, 0.559]	5–5; 2000
PathMNIST	all	17/0	0.081 [0.028, 0.133]	0.396 [0.259, 0.527]	6–6; 2000
PathMNIST	all	42/0	0.138 [0.079, 0.192]	0.378 [0.194, 0.479]	5–6; 2000
PathMNIST	all	117/0	0.130 [0.069, 0.194]	0.327 [0.146, 0.504]	5–6; 2000
RetinaMNIST	all	17/0	0.176 [0.032, 0.328]	0.110 [-0.223, 0.457]	2–2; 2000
RetinaMNIST	all	42/0	0.095 [-0.043, 0.309]	0.560 [0.108, 0.909]	2–2; 2000
RetinaMNIST	all	117/0	-0.061 [-0.173, 0.057]	0.241 [0.031, 0.444]	2–2; 2000
CLINC150	all	17/0	0.053 [-0.056, 0.217]	0.415 [0.111, 0.762]	3–9; 2000
CLINC150	all	42/0	0.054 [-0.060, 0.178]	0.406 [0.139, 0.737]	5–10; 2000
CLINC150	all	117/0	0.051 [-0.056, 0.167]	0.611 [0.287, 0.818]	4–9; 2000
ESC-50	all	17/0	0.248 [0.000, 0.624]	0.222 [-1.000, 1.000]	0–2; 1738
ESC-50	all	17/1	-0.115 [-0.250, 0.000]	1.000 [1.000, 1.000]	0–1; 1154
ESC-50	all	17/2	0.084 [-0.071, 0.262]	0.167 [-0.667, 1.000]	0–3; 1989
ESC-50	all	17/3	0.327 [-0.241, 0.733]	1.000 [1.000, 1.000]	0–2; 1713
ESC-50	all	17/4	0.005 [-0.178, 0.228]	0.792 [0.250, 1.000]	1–4; 2000
ESC-50	all	42/0	0.169 [-0.070, 0.407]	0.409 [-0.333, 1.000]	0–4; 1994
ESC-50	all	42/1	-0.210 [-0.475, 0.000]	1.000 [1.000, 1.000]	0–2; 1639
ESC-50	all	42/2	0.088 [-0.171, 0.345]	0.667 [0.000, 1.000]	0–3; 1986
ESC-50	all	42/3	0.154 [-0.156, 0.511]	0.056 [-0.840, 1.000]	0–3; 1989
ESC-50	all	42/4	-0.035 [-0.201, 0.353]	1.000 [1.000, 1.000]	0–3; 1865
ESC-50	all	117/0	0.195 [-0.150, 0.530]	1.000 [1.000, 1.000]	0–1; 1741
ESC-50	all	117/1	-0.096 [-0.278, 0.133]	1.000 [1.000, 1.000]	0–1; 1316
ESC-50	all	117/2	0.091 [-0.125, 0.463]	0.667 [0.000, 1.000]	0–3; 1917
ESC-50	all	117/3	0.194 [-0.006, 0.496]	0.500 [0.000, 1.000]	0–2; 1993
ESC-50	all	117/4	0.099 [-0.200, 0.594]	0.444 [-0.750, 1.000]	0–3; 1932
ImageNet-1K R50	all	17/0	0.207 [0.187, 0.227]	0.265 [0.208, 0.359]	167–201; 2000
ImageNet-1K CT	all	17/0	0.193 [0.169, 0.212]	0.328 [0.255, 0.397]	162–195; 2000
ImageNet-1K R50	all	42/0	0.207 [0.187, 0.227]	0.265 [0.208, 0.359]	167–201; 2000
ImageNet-1K CT	all	42/0	0.193 [0.169, 0.212]	0.328 [0.255, 0.397]	162–195; 2000
ImageNet-1K R50	all	117/0	0.207 [0.187, 0.227]	0.265 [0.208, 0.359]	167–201; 2000
ImageNet-1K CT	all	117/0	0.193 [0.169, 0.212]	0.328 [0.255, 0.397]	162–195; 2000
DomainNet	clipart	17/0	0.023 [-0.025, 0.072]	0.168 [0.019, 0.590]	5–8; 2000
DomainNet	infograph	17/0	0.015 [-0.041, 0.070]	0.281 [0.142, 0.475]	4–5; 2000
DomainNet	painting	17/0	0.041 [0.009, 0.073]	0.297 [0.156, 0.451]	6–7; 2000
DomainNet	quickdraw	17/0	0.030 [-0.046, 0.078]	0.357 [0.109, 0.511]	4–5; 2000
DomainNet	real	17/0	-0.001 [-0.037, 0.035]	0.485 [0.305, 0.642]	6–8; 2000
DomainNet	sketch	17/0	-0.011 [-0.048, 0.031]	0.539 [0.329, 0.674]	4–7; 2000
DomainNet	clipart	42/0	-0.012 [-0.057, 0.037]	0.479 [0.283, 0.649]	5–7; 2000
DomainNet	infograph	42/0	0.058 [-0.028, 0.141]	0.304 [0.053, 0.565]	4–7; 2000
DomainNet	painting	42/0	0.001 [-0.031, 0.034]	0.301 [0.126, 0.500]	5–7; 2000
DomainNet	quickdraw	42/0	0.188 [-0.007, 0.291]	0.294 [0.249, 0.340]	4–4; 2000
DomainNet	real	42/0	-0.018 [-0.054, 0.016]	0.582 [0.394, 0.708]	6–8; 2000
DomainNet	sketch	42/0	-0.010 [-0.050, 0.033]	0.482 [0.275, 0.667]	4–7; 2000
DomainNet	clipart	117/0	-0.002 [-0.048, 0.049]	0.454 [0.237, 0.637]	5–7; 2000
DomainNet	infograph	117/0	0.045 [-0.049, 0.107]	0.357 [0.147, 0.644]	4–7; 2000
DomainNet	painting	117/0	0.014 [-0.015, 0.045]	0.316 [0.193, 0.608]	6–8; 2000
DomainNet	quickdraw	117/0	0.123 [0.057, 0.190]	0.280 [0.232, 0.329]	4–4; 2000
DomainNet	real	117/0	-0.011 [-0.049, 0.029]	0.447 [0.253, 0.625]	5–8; 2000
DomainNet	sketch	117/0	0.003 [-0.037, 0.050]	0.437 [0.161, 0.645]	4–7; 2000

J Query-Level Structural Distributions and Aggregate Gaps

J.1 Matched-Query Contribution Accounting

Figure 7(a,b) uses the registered primary queries only, unlike Figure 2, which also retains secondary unions and controls. For each retained query we compute $\pi_A E_A/2$ before averaging, using the same

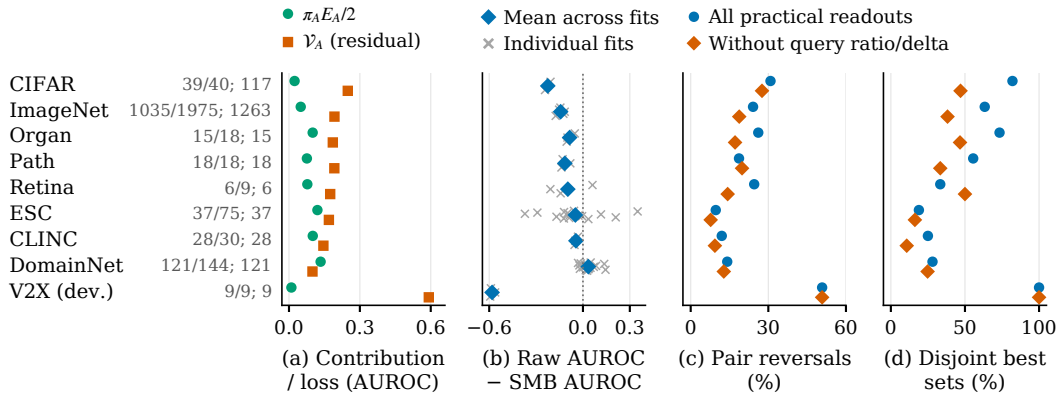


Figure 7: **Interior contributions and method rankings across tasks.** (a) Interior contribution $\pi_A E_A/2$ and support-order loss $\mathcal{V}_A$, averaged across fits; the loss is computed as a residual of the stored metrics. (b) Their difference, equal to raw conditional AUROC minus SMB AUROC; diamonds show means across fits and crosses show individual fits. (c) Fraction of practical-score pairs whose order reverses between raw conditional AUROC and interior evidence E_A. (d) Fraction of evaluations with disjoint best-method sets. In (c,d), blue includes all practical readouts and orange excludes query-dependent ratio/delta readouts. Counts beside task names give valid/planned query–fit records for (a,b), followed by valid ranking instances for (c,d); V2X uses development data.

Table 23: Coverage and matched-query contribution means for Figure 7(a,b). Counts partition planned primary records into missing H/C, defined H/C but missing interior evidence, and jointly valid records. All three means use the same valid queries within each fit and then weight fits equally; violations are decomposition residuals.

Dataset	Planned	H/C NA	E_A NA	Valid	$\pi_A E_A/2$	$\mathcal{V}_A$	Gap
CIFAR-100	40	0	1	39	0.024	0.249	-0.225
ImageNet	1975	187	753	1035	0.050	0.192	-0.143
OrganAMNIST	18	0	3	15	0.100	0.185	-0.085
PathMNIST	18	0	0	18	0.076	0.192	-0.116
RetinaMNIST	9	1	2	6	0.078	0.174	-0.097
ESC-50	75	29	9	37	0.120	0.169	-0.049
CLINC150	30	0	2	28	0.101	0.146	-0.045
DomainNet	144	3	20	121	0.134	0.099	0.034
V2X (dev.)	9	0	0	9	0.010	0.592	-0.582

jointly defined query set for all three terms. Table 23 separates missing H/C from missing interior evidence. Its fit means are averaged without weighting by the number of H–C pairs; ESC folds, training seeds, and DomainNet domains remain distinct fitted evaluations. The five ImageNet pairs use one fixed-score evaluation each. V2X averages its three raw semantic heads over the three primary queries on development data; Table 27 reports the seed-17 occupied query with separate raw and projected columns. The cached values determine $\mathcal{V}_A$ as $\pi_A E_A/2 - (\mathrm{AUC}_A - \mathrm{AUC}_{\mathrm{SMB},A})$. Equation 7 makes this an exact decomposition of the observed Raw-minus-SMB gap into interior contribution and support-order loss. Query records and per-fit means are archived in `data/main_revision/`.

J.2 Full Query Geometry and Controls

Query-level geometry. Figure 2 shows every locked query, including controls, without selecting on the observed contrast. Positive E_A occurs on both sides of zero: ImageNet has 243 positive-evidence semantic instances with raw AUROC below SMB AUROC and 83 above; DomainNet has 35 and 72. At small π_A, even strong evidence contributes at most $\pi_A/2$ above SMB AUROC before support-order loss is deducted. Therefore, a positive interior contribution can coexist with a negative

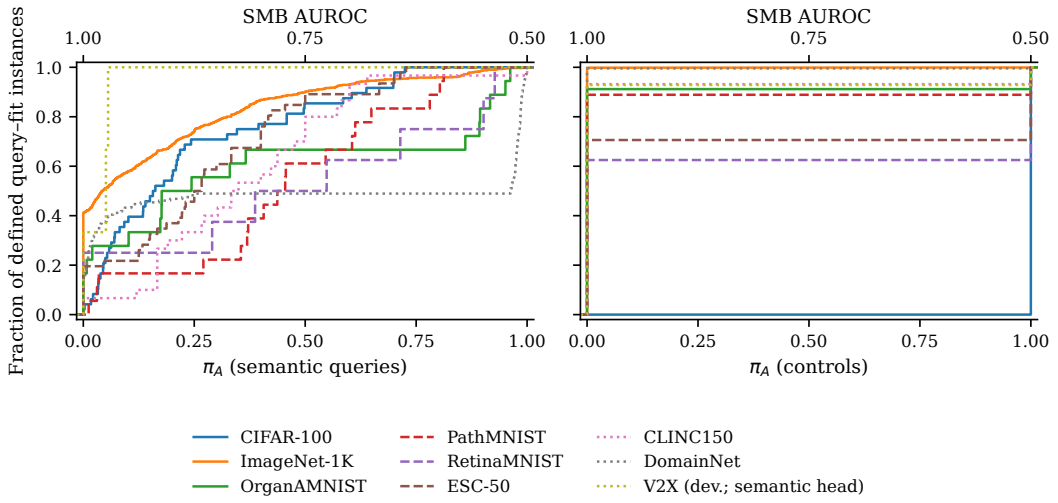


Figure 8: **Where structural contribution occurs.** Empirical CDFs of the interior pair fraction across all locatable query–fit instances, separately for semantic queries (left) and controls (right). The top axis maps π_A to SMB AUROC exactly; no line is fitted. Table 24 reports coverage, including omitted undefined locations. Curves weight each retained instance equally, not each observation or pair. CIFAR controls contain only the archived full-class query; control sets differ across tasks.

Table 24: Coverage of the query-level figures. L/N denotes locatable/total query–fit instances; E counts defined interior evidence. Below/above count semantic instances with strictly positive E_A and raw AUROC below/above SMB AUROC. Queries, seeds, folds and domains may share observations; the counts report query–fit coverage. All control aliases remain in the export.

Task	Semantic L/N	E	Below	Above	Control L/N	E
CIFAR-100	48/48	47	36	4	2/2	2
ImageNet-1K	717/790	421	243	83	1176/2002	2
OrganAMNIST	18/18	15	9	6	34/36	3
PathMNIST	18/18	18	14	4	27/30	3
RetinaMNIST	8/9	6	4	2	8/21	3
ESC-50	46/75	37	12	11	51/765	15
CLINC150	30/30	28	11	12	44/453	3
DomainNet	141/144	121	35	72	3882/6228	18
V2X (dev.)	9/9	9	7	0	42/45	3

Raw-minus-SMB gap when support-order loss is larger. Appendix J separates semantic-query and control distributions of π_A.

Figure 8 displays empirical distributions of π_A, not a fitted structural identity. Its secondary axis is the deterministic map $1 - \pi_A/2$. Semantic queries and full-class/singleton controls are separated because the latter concentrate at the endpoints. The distributions summarize the fixed query–fit collection with defined H and C. Query definitions and the number of evaluated instances differ by dataset, as specified in the protocol.

The export contains 10,723 instances after verifying and removing 5,684 repeated fixed-score reliability-split records from CIFAR/ImageNet. It retains both classifier pairs, CIFAR's four predefined semantic unions, all new-task primary queries and controls, and the archived V2X queries. New training seeds, ESC folds and DomainNet domains remain separate; no scores are pooled across fitted classifiers. V2X uses three fitted versions of the existing semantic head on the same development population. Here occupied denotes the query $A = \{1, \ldots, 12\}$; semantic head denotes the reliability score, not a different occupancy model. The 18 archived query names per seed include three semantic unions and 15 controls (including the empty/class-13 alias). Thus "loc 51/54; E 12" counts 51 H/C-defined query–seed instances out of 54, with interior evidence defined for 12; these

instances share the underlying evaluation data. For occupied at seed 17, raw AUROC is 0.45960, SMB is 0.97206, and projected AUROC is 0.98813. For the occupied query, only $\pi_A = 0.05587$ of H–C pairs lie in the unresolved interior. Within that interior, the score has positive evidence $E_A = 0.57509$, giving projected AUROC 0.98813. The raw score nevertheless falls to AUROC 0.45960 because its support-order loss is $\mathcal{V}_A = 0.52853$. This is a concrete example of Equation 7: strong interior discrimination can coexist with poor aggregate conditional ranking when support-order loss dominates. Repeated query aliases and exact plotted overlaps are retained and counted in the accompanying audit JSON.

For Figure 2, V2X raw AUROC is recomputed from its archived semantic-head scores with the same masks used for interior evidence. Order-preserving rank projection avoids numerical tie changes introduced by the legacy score transform; the resulting tiny differences from archived diagnostics are recorded in the audit, without changing any reported four-decimal result. Each exported row verifies the SMB AUROC and raw-gap identity to 10^{-10}. When $\pi_A = 0$, E_A is undefined and the raw gap is $-\mathcal{V}_A$; it is plotted with a gray cross, never with imputed zero evidence. Absent H or C makes the location undefined. None of these descriptive plots adds a significance test or changes the frozen primary comparisons.

K IMAGENET ARCHITECTURE-FAMILY VALIDATION

Controlled comparison. We extend the two frozen ImageNet pairs in Appendix I with R18→Swin-T, R18→ViT-B/16, and Swin-T→Swin-B (Liu et al., 2021; Dosovitskiy et al., 2021). All five pairs use the identical 40,000 evaluation images, class order, ground-truth mapping, 395 nontrivial WordNet queries, full-class control, and 1,000 singleton controls. The original CNN results are imported without recomputation. The three added models are frozen torchvision 0.20.1 `IMAGENET1K_V1` checkpoints, with each weight's official preprocessing. Swin-T and Swin-B resize the shorter side to 232 and 238 pixels with bicubic interpolation, respectively; ViT-B/16 uses 256 pixels with bilinear interpolation. All use a central 224×224 crop and the corresponding official normalization. Inference retains the original automatic mixed precision setting, and the exported $50{,}000 \times 1{,}000$ logits pass finite-value, label, and image-alignment checks. The classifier accuracies in Table 25 use only the fixed 40,000-image evaluation set.

Diff-Prob remains the old predictor's maximum softmax probability minus the revised predictor's maximum softmax probability, with larger values indicating harm. The remaining 10 images per class retain their 6/4 reliability TRAIN/VAL divisions for seeds 17, 42, and 117. No backbone or reliability head is fitted in this extension, so fixed-score results are identical across those reliability splits and are evaluated once. The comparison changes the frozen predictor architecture while keeping the evaluation images, semantic queries, score definition, and bootstrap procedure fixed.

Support, evidence, and violations. SMB AUROC reaches 0.918, 0.920, and 0.906 on the three added pairs, while raw Diff-Prob AUROC reaches 0.737, 0.737, and 0.681. Their macro interior evidence is 0.414, 0.348, and 0.240, respectively. Both the SMB-minus-raw AUROC and interior-evidence bootstrap intervals are positive in every added pair (Table 26). Thus the stronger reversal occurs in the evaluated CNN–Transformer and Transformer–Transformer comparisons. The higher SMB AUROC comes from deterministic semantic support, whereas Diff-Prob additionally provides positive discrimination within the unresolved interior. Equation 7 separates the weighted contribution of that evidence from the support-order violation penalty. All 6,980 exported query/control records retain missing outcomes, and the applicable support and raw identities agree to within 2.23×10^{-16}.

Figure 9 shows the query-level relationship across all five frozen ImageNet predictor pairs. Positive interior evidence appears both above and below the zero-gap line across the evaluated architecture families, consistent with Equation 7: the interior contribution is weighted by π_A and can be offset by support-order loss.

Missing support and statistical scope. Of 395 primary queries, H/C ranking is defined for 350, 365, and 356 queries in the added pairs; interior evidence is defined for only 204, 214, and 196. Undefined quantities are never replaced by zero. The paired bootstrap reuses the original 2,000 class-stratified image resamples and averages queries defined in each resample. For Swin-T→Swin-B, the number of interior-defined queries falls from 196 in the observed sample to a median of 163

Table 25: Frozen ImageNet architecture comparison on the same 40,000 images and 395 primary queries. Accuracies are percentages; $|D|$ counts changed predictions. SMB AUROC and raw AUROC average H/C-defined queries, whereas E_A averages interior-defined queries (Q_I). R18→R50 is CNN/CNN, R18→ConvNeXt-T is CNN/modern ConvNet, the next two are CNN/Transformer, and the last is Transformer/Transformer. Existing CNN results are reused; no new reliability head is fitted.

Pair	Acc. a	Acc. b	$\|D\|$	SMB	Raw	E_A	$\bar{\pi}_A$	Q_I/395
R18→R50	69.73	80.74	9858	0.918	0.711	0.265	0.164	214/395
R18→ConvNeXt-T	69.73	82.47	9933	0.918	0.725	0.328	0.164	207/395
R18→Swin-T	69.73	81.33	9912	0.918	0.737	0.414	0.165	204/395
R18→ViT-B/16	69.73	81.07	10170	0.920	0.737	0.348	0.161	214/395
Swin-T→Swin-B	81.33	83.45	4329	0.906	0.681	0.240	0.187	196/395

Table 26: Architecture-family macro contrasts and point-estimate query counts. Intervals use 2,000 class-stratified paired image resamples and average queries with defined comparisons in each resample. $\Delta_{\mathrm{SMB}} = \mathrm{AUC}_{\mathrm{SMB}} - \mathrm{AUC}_{\mathrm{raw}}$. N_{rev} counts SMB AUROC>raw AUROC with $E_A > 0$; N_{raw} counts raw AUROC>SMB AUROC with $E_A > 0$. Counts exclude controls and are not per-query significance tests. Two positive marginal intervals do not constitute a simultaneous 95% confidence region.

Pair	Δ_{SMB} [95% CI]	E_A [95% CI]	Pair-weighted E_A	N_{rev}	N_{raw}
R18→R50	0.207 [0.187, 0.227]	0.265 [0.208, 0.359]	0.426	127	40
R18→ConvNeXt-T	0.193 [0.169, 0.212]	0.328 [0.255, 0.397]	0.477	116	43
R18→Swin-T	0.181 [0.157, 0.198]	0.414 [0.329, 0.466]	0.509	120	50
R18→ViT-B/16	0.182 [0.161, 0.199]	0.348 [0.277, 0.420]	0.498	124	47
Swin-T→Swin-B	0.226 [0.201, 0.255]	0.240 [0.162, 0.345]	0.331	100	39

per resample (range 140–183). Each bootstrap macro therefore averages the queries whose interior H/C comparison is defined in that resample. Pair-weighted E_A uses $n_{H,I}n_{C,I}$ weights; overlapping hierarchy queries reuse pairs and are not disjoint populations. The three added pairs contain 120, 124, and 100 queries, respectively, with positive interior evidence and SMB AUROC above raw AUROC; they also contain 50, 47, and 39 positive-evidence queries with the opposite aggregate ordering.

Semantic disagreement profiles. Figure 10 shows query size versus π_A and the empirical distribution of π_A over H/C-defined queries. For R18→R50, R18→ConvNeXt-T, R18→Swin-T, R18→ViT-B/16, and Swin-T→Swin-B, descriptive Spearman correlations between query size and π_A are 0.630, 0.620, 0.647, 0.627, and 0.652. The four R18-based pairs have similar mean π_A values (0.161–0.165) and strongly correlated profiles on pairwise common-defined queries (0.905–0.925). Swin-T→Swin-B has higher mean π_A (0.187) but lower median (0.022, versus 0.035–0.047 for the R18-based pairs). On common-defined queries, its mean π_A exceeds those of the R18-based pairs by 0.020–0.027, yet the empirical distribution functions cross. The crossing distributions show how architecture changes the semantic disagreement profile, while π_A measures conditional H–C pair mass rather than a fraction of individual changed predictions.

Controlled architecture-family validation. To assess dependence on the predictor architecture, we keep ImageNet images, semantic queries, score orientations, support projection, and the paired image-bootstrap procedure fixed while changing the frozen predictor pair. The comparison adds R18→Swin-T, R18→ViT-B/16, and Swin-T→Swin-B to the existing R18→R50 and R18→ConvNeXt-T pairs. Each checkpoint uses its official preprocessing; no new reliability head or backbone is trained. Fixed-score quantities are evaluated once on the shared holdout; the reliability splits vary TRAIN/VAL assignments while sharing the same frozen TEST predictions. Section 5.1 summarizes the results, with full protocols and query-level coverage in Appendix K.

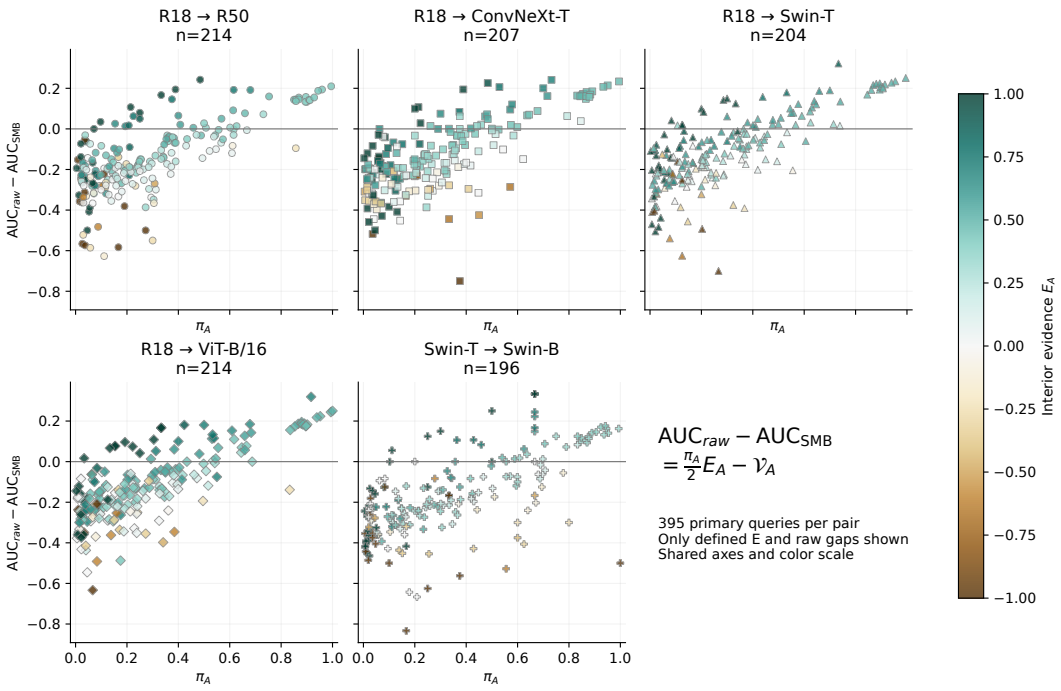


Figure 9: **Semantic-membership contribution and interior evidence across frozen ImageNet predictor pairs.** Each point represents one semantic query A, with interior-pair fraction π_A on the horizontal axis, raw conditional AUROC minus SMB AUROC on the vertical axis, and interior evidence E_A indicated by color. The same qualitative relationship appears across CNN–CNN, CNN–ConvNeXt, CNN–Transformer, and Transformer–Transformer predictor pairs.

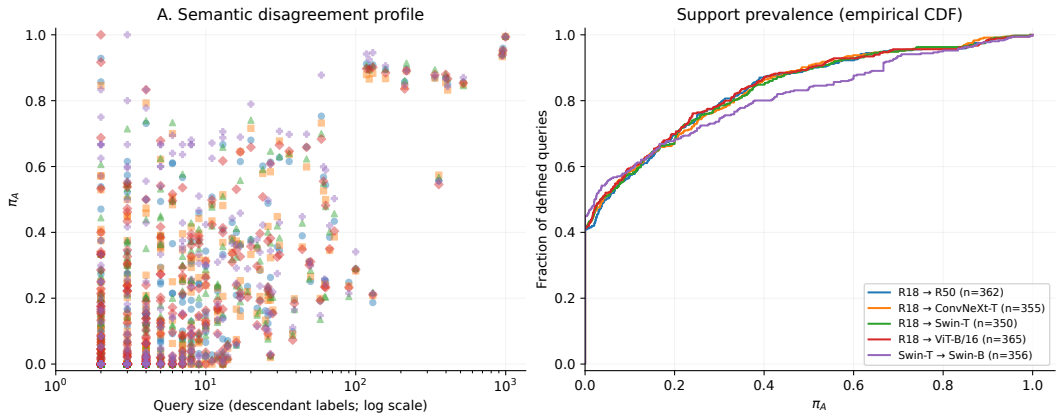


Figure 10: **Semantic disagreement profiles across frozen architecture pairs.** Left: query size versus π_A, using distinct markers for the five pairs. Right: empirical distributions over each pair's H/C-defined primary queries, with their denominators in the legend. Undefined H/C queries and full/singleton controls are excluded; defined $\pi_A = 0$ queries remain. The size axis is logarithmic, and each query has equal weight. The correlations are descriptive and do not assume query independence or a monotone relation.

L Additional Cross-Task Summaries

Table 27 summarizes the historical macro-averages over defined queries. SMB exceeds raw Diff-Prob on CIFAR-100, ImageNet-1K, OrganAMNIST and PathMNIST despite positive interior evidence; the sign varies across other tasks and fits. The V2X row uses the seed-17 occupied query and reports raw

Table 27: **Structural contribution across tasks and predictor pairs.** Historical macro-averages over defined queries; raw/interior denominators can differ. SMB denotes the Semantic-Membership Baseline; Raw and E_A use Diff-Prob except for V2X. For ImageNet, ranges summarize all five frozen architecture pairs; complete pairwise results are shown in Figure 9 and Appendix K. V2X reports the occupied query on 885 development frames, seed 17: Raw and projected columns use the same frozen semantic head and E_A uses the same head. Q_I/Q gives interior-defined queries out of those planned.

Task / predictor pair	SMB	Raw	Proj.	E_A	Q_I/Q
CIFAR R20→R56	0.919	0.709	—	0.304	19/20
CIFAR R20→VGG13	0.918	0.675	—	0.360	20/20
OrganAMNIST	0.804	0.723	—	0.471	5/6
PathMNIST	0.778	0.661	—	0.367	6/6
RetinaMNIST	0.740	0.670	—	0.304	2/3
RetinaMNIST (seed 117)	0.543	0.604	—	0.241	2/3
ESC-50	0.864	0.784	—	0.662	1–4/5
CLINC150 BERT→RoBERTa	0.818	0.766	—	0.477	9–10/10
ImageNet-1K (5 pairs)	0.906–0.920	0.681–0.737	—	0.240–0.414	196–214/395
DomainNet (6 domains)	0.728–0.751	0.614–0.756	—	0.305–0.504	4–8/8
V2X occupied (dev.)	0.972	0.460	0.988	0.575	1/1

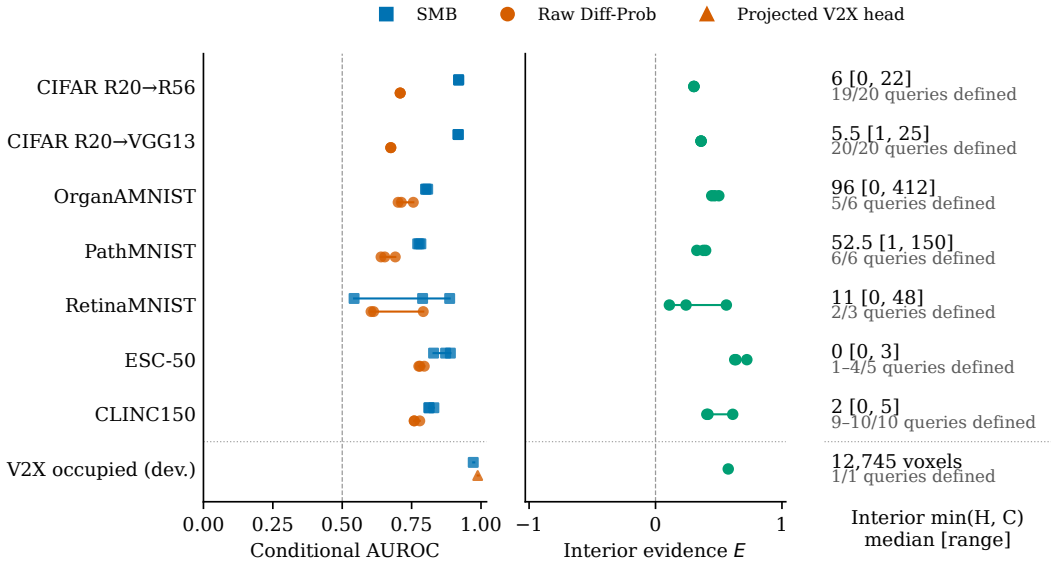


Figure 11: **Structural contribution, interior evidence and available support.** Left: SMB AUROC and raw Diff-Prob AUROC (V2X: projected semantic head). Middle: Diff-Prob E_A (V2X: the same semantic head, occupied query, seed 17). Points are seed macro-averages over defined queries; segments span seed minima/maxima, not CIs. ESC averages five separately evaluated folds per seed; CIFAR repeats splits of fixed classifiers. Right: median [min, max] of $\min(n_{H,I}, n_{C,I})$ over primary queries/fits, including zeros, and defined/planned interior queries. Counts report observed evaluation units; V2X counts voxels. Raw/interior macro denominators may differ. Appendix intervals retain missingness.

and projected scores separately. Table 33 provides common-query metrics with explicit TEST-support thresholds.

Table 28: Primary-query conditional AUROC and interior evidence. Raw and E use Diff-Prob, except V2X E uses its semantic head; V2X raw is omitted and its projected AUROC is 0.9881. Trained tasks report mean $\pm$ descriptive SD across three training seeds, not confidence intervals; ESC first averages five separately evaluated folds within each seed. DomainNet entries span the six domain means. CIFAR/ImageNet splits share predictions. Q_I is defined/planned interior queries per fit (ranges retain missingness); aggregate and interior macros can have different denominators. "Reversal" denotes the stronger macro contrast, not every query. Medical/DomainNet pairs are R18→R50, audio Cnn6→Cnn14, and intent BERT→RoBERTa. Intervals and H/C counts are in the appendix.

Task / Pair	Domain	SMB	Raw	Interior E	Q_I	Interpretation
CIFAR R20→R56	Images	0.919	0.709	0.304	19/20	Reversal
CIFAR R20→VGG13	Images	0.918	0.675	0.360	20/20	Reversal
OrganAMNIST	CT anatomy	0.804 ± 0.004	0.723 ± 0.029	0.471 ± 0.028	5/6	Reversal
PathMNIST	Histology	0.778 ± 0.006	0.661 ± 0.027	0.367 ± 0.036	6/6	Reversal
RetinaMNIST	Retina	0.740 ± 0.177	0.670 ± 0.106	0.304 ± 0.232	2/3	Seed variation
ESC-50	Audio	0.864 ± 0.031	0.784 ± 0.010	0.662 ± 0.053	1–4/5	Sparse H/C
CLINC150	Intent	0.818 ± 0.010	0.766 ± 0.012	0.477 ± 0.116	9–10/10	Gap uncertain
ImageNet R18→R50	Images	0.918	0.711	0.265	214/395	Reversal
ImageNet R18→ConvNeXt-T	Images	0.918	0.725	0.328	207/395	Reversal
DomainNet	Six domains	0.728–0.751	0.614–0.756	0.305–0.504	4–8/8	Domain variation
V2X occupied	Occupancy	0.972	–	0.575	1/1	Development

M Detailed Related-Work Comparison

Table 29: Closest objectives and the additional restriction studied here.

Analysis	Main target	Relation to this paper
Group/cluster AUC and non-collapsibility (Kallus & Zhou, 2019; Chang et al., 2022; Sudjianto & Liu, 2025; Matos et al., 2026)	Within/cross-group ranking	General pair partition; cross-group terms need not be perfect
Model updates (Yan et al., 2021)	Negative flips	Same H/C outcomes; training reduces regression
Cascade deferral (Jitkrittum et al., 2023)	Relative prediction error	Selection follows loss difference, not conditional AUROC alone
This study	H/C within $Y \in A$	Forced support baseline, interior discrimination, query order constraints

N FOUNDATION-MODEL VALIDATION

We additionally examine whether the support effect identified in the main experiments extends to foundation-model updates. These experiments use naturally occurring prediction changes between different model configurations rather than constructing synthetic harmed or corrected examples. We consider two LLM settings and a VLM extension. The completed evaluations comprise two LLM settings and an 8,192-image VLM experiment, reported below.

N.1 LLM MODEL UPGRADE: QWEN3-4B TO QWEN3-8B

We first treat a model-scale upgrade from Qwen3-4B to Qwen3-8B as a paired-predictor update. The task is CLINC150 in-scope intent classification with 150 candidate labels and ten semantic domain queries. Predictions use full-candidate answer-sequence likelihoods including EOS, rather than free-form generation or calibrated class posteriors. Each model configuration was evaluated on 5,700 text instances, with 4,500 instances retained in the final test population. The test accuracy increases from 48.53% for Qwen3-4B to 61.62% for Qwen3-8B. Among the prediction changes, 275 are harmed cases (H), in which the smaller model is correct and the larger model is wrong, and 864 are corrected cases (C), in which the larger model corrects an error made by the smaller model.

Table 30 reports the conditional reliability results using Diff-Prob. The reported ranking metrics are unweighted macro-averages over the ten domain queries; H/C counts refer to the full test population. SMB reaches an AUROC of 0.9492, whereas the raw Diff-Prob AUROC is 0.7494. Within the unresolved interior population, the AUROC is 0.5289, corresponding to normalized interior evidence

$$E_A = 2\ \mathrm{AUC}_{I_A} - 1 = 0.0578.$$

A paired bootstrap analysis with 2,000 resamples grouped by normalized utterance gives a 95% interval of $[-0.154,\ 0.270]$ for E_A, which includes zero. Across resamples, the macro averages seven to ten queries with defined interior evidence (median nine). Thus, the large SMB–raw gap is accompanied by near-chance Diff-Prob discrimination within the unresolved interior.

This result concerns reliability-score interpretation rather than the quality of the model upgrade itself. The larger model substantially improves overall test accuracy and produces considerably more corrected than harmed predictions. Rather, the result concerns the interpretation of the *reliability score*: much of the conditional H–C ranking available in this setting can already be obtained from the semantic support structure induced by the paired predictions and the query.

N.2 LLM REVIEW UNDER SPARSE H/C SUPPORT

We next study a different type of update in which Qwen3-8B first produces an answer and subsequently reviews its own initial prediction. Unlike the model-scale comparison above, the review stage changes only 84 predictions in the 4,500-instance test population. These changes comprise 10 harmed cases, 11 corrected cases, and 63 both-wrong cases (W).

The resulting conditional evaluation is support-limited. Only 3 of the 10 semantic queries contain sufficient H/C observations to define the raw conditional AUROC, and none contains sufficient

Table 30: **Foundation-model LLM experiments.** The first setting compares Qwen3-4B with Qwen3-8B and contains sufficient support for both raw and interior evaluation. The second setting compares the initial Qwen3-8B answer with a review stage; its sparse H/C support makes interior evidence undefined. Ranking metrics average over evaluable queries (10/10 for the upgrade and 3/10 for review Raw and SMB); review interior coverage is 0/10. Undefined quantities are reported as NA rather than imputed.

Update	H	C	SMB	Raw	Interior	E_A
Qwen3-4B → Qwen3-8B	275	864	0.9492	0.7494	0.5289	0.0578
Qwen3-8B → Review	10	11	1.0000	0.4667	NA	NA

interior H/C support to define E_A. We therefore do not aggregate these sparse queries into an interior-evidence estimate.

This case shows how sparse H/C events affect conditional reliability evaluation. Even when a prediction-update mechanism produces observable changes, the number and semantic distribution of H/C events may be insufficient to estimate the unresolved interior ranking. Our evaluation protocol exposes this lack of support explicitly rather than replacing undefined interior quantities with zero or interpreting a small number of evaluable queries as representative of the full query set.

The large number of both-wrong cases also reinforces the distinction between H/C ranking and downstream decision utility discussed in Section 5. Both-wrong changes do not enter the H–C AUROC population, but they remain relevant when a reliability score is used to decide whether an updated prediction should be accepted or rejected.

N.3 Expanded VLM Evaluation

We evaluate the support–interior decomposition on 8,192 ImageNet TEST images, extending the initial 256-image study to a larger H/C population. The sample extends the same fixed ID-hash ordering, independently of labels and prediction outcomes, and includes the original 256 images. Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct evaluate the same images over the full 1,000-class output space using complete candidate-sequence likelihoods including EOS. Model revisions and image-resolution settings are fixed; inference uses BF16 candidate batches of four and two, respectively, without additional training or recalibration. The predictor pair produces 3,426 prediction changes: 981 harmed (H), 849 corrected (C), and 1,596 wrong-to-wrong (W) transitions. The corresponding top-1 accuracies are 57.18% and 55.57%, respectively.

Table 31 reports the conditional H/C ranking results. Across all three reliability scores, support projection changes aggregate ranking while preserving discrimination in the unresolved interior. For Diff-Prob, the contrast is

$$\text{Raw AUROC: } 0.7444 \longrightarrow 0.9321, \qquad \text{Interior AUROC: } 0.6927 \longrightarrow 0.6927.$$

Entropy change and margin change exhibit the same pattern: their projected Raw AUROCs both reach 0.9311, while their Interior AUROCs remain 0.6897 and 0.6880, respectively.

The constant-interior control makes the separation explicit: its Interior AUROC is 0.5000, yet its Raw AUROC remains 0.8929. Thus, semantic support structure alone yields high aggregate conditional AUROC without interior H/C discrimination. At the same time, the raw reliability scores retain interior evidence, with Interior AUROC ranging from 0.6880 to 0.6927.

This contrast agrees with Equation 7, $\mathrm{AUC}_A(s) = 1 - \pi_A/2 + \pi_A E_A(s)/2 - \mathcal{V}_A(s)$. For each query, support projection removes the support-order loss $\mathcal{V}_A(s)$ while preserving π_A and $E_A(s)$, and hence the interior evidence contribution $\pi_A E_A(s)/2$. The constant-interior control additionally sets $E_A = 0$, leaving the SMB value $1 - \pi_A/2$. The intervention therefore isolates changes in support-induced ordering from changes in interior discrimination.

The expanded sample increases the number of primary queries with defined Raw AUROC from 28 to 256 and with defined Interior AUROC from 19 to 121, out of the same 395 registered queries. The shared pattern across Diff-Prob, entropy change, and margin change shows that this separation occurs across all three evaluated score constructions. **High aggregate conditional AUROC does not by**

Table 31: **Expanded VLM evaluation on 8,192 TEST images.** Raw and Interior AUROC are unweighted macro-averages over registered primary queries for which the respective H/C comparison is defined; the coverage columns report these denominators out of 395 queries. Raw evaluates the complete conditional H/C population of each query, while Interior restricts it to cases unresolved by semantic membership. Support projection preserves interior ranking; the constant-interior control preserves support-induced ordering with tied interior scores.

Score	Variant	Raw	Interior	Raw cov.	Interior cov.
Diff-Prob	raw	0.7444	0.6927	256/395	121/395
	projected	0.9321	0.6927	256/395	121/395
	constant interior	0.8929	0.5000	256/395	121/395
Entropy change	raw	0.7317	0.6897	256/395	121/395
	projected	0.9311	0.6897	256/395	121/395
	constant interior	0.8929	0.5000	256/395	121/395
Margin change	raw	0.7313	0.6880	256/395	121/395
	projected	0.9311	0.6880	256/395	121/395
	constant interior	0.8929	0.5000	256/395	121/395

itself imply strong interior H/C discrimination. The expanded VLM results provide empirical evidence for treating semantic support structure and interior evidence as distinct components of conditional ranking.

N.4 Implications for the Cross-Task Results

The completed LLM and VLM experiments extend the main analysis to fixed-label intent classification and ImageNet classification. In the Qwen3-4B $\rightarrow$ Qwen3-8B update, SMB AUROC is high while the interior Diff-Prob ranking is close to chance and its bootstrap interval includes zero. In the review experiment, insufficient H/C and interior support determines which quantities are estimable.

The VLM experiment further separates support ordering from interior evidence through matched raw, projected, and constant-interior scores (Appendix N.3). Together with the heterogeneous results in the main experiments, these observations reinforce the distinction among three questions:

$$\boxed{\text{aggregate conditional ranking} \neq \text{interior discrimination} \neq \text{decision utility.}}$$

A high conditional AUROC need not imply strong discrimination among unresolved interior cases, while a lack of estimable interior evidence need not imply that the underlying model update is ineffective. The appropriate interpretation therefore depends jointly on the support structure, the available interior population, and the downstream decision objective.

O From Conditional Evidence to Fallback Decisions

Connecting conditional ranking to fallback decisions. We next test whether differences in conditional reliability ranking translate into utility under a fixed fallback budget. For a query A, every changed prediction in $D = \{a \neq b\}$ enters the budget pool, including W and cases whose true class is outside A. Labels determine the realized gain $g_A = \mathbf{1}_H\mathbf{1}[a \in A] - \mathbf{1}_C\mathbf{1}[b \in A]$ only after the policy has selected candidates. Filtering this pool by $Y \in A$ would give the policy unavailable information. We evaluate three linked effects: semantic support changes measured conditional ranking, the validation objective can change which scorer is selected, and the selected scorer can change held-out fallback utility. The experiments below evaluate these effects separately so that metric changes, scorer selection, and deployment utility are not conflated.

Separating ranking correction from decision utility. Consider $A = \{0, 1\}$ and four changed observations:

	a	b	Y	s	g_A
x_1	0	1	0	3	1
x_2	2	0	0	2	-1
x_3	0	2	1	1	0
x_4	0	2	0	0	1

The conditional H/C population consists of x_1, x_2, x_4; raw AUROC is $1/2$. Support projection puts the a-only observations x_3, x_4 above the interior x_1, and puts b-only x_2 last, so conditional AUROC becomes 1. If raw order breaks ties within the a-only tier, x_3 takes the single fallback slot, reducing realized utility from 1 to 0. With uniform tie breaking within that tier, expected utility is $1/2$, still below 1. The example uses no true labels to form the ranking. The example shows why projected AUROC and fallback utility must be evaluated separately: correcting support order can improve conditional ranking without preserving the same top-budget decisions.

Decision-relevant factorization. Fix observables X containing both predictions, and condition throughout on D. Define $t(X) = P(H \cup C \mid X, D)$ and $r(X) = P(H \mid X, H \cup C)$ when $t > 0$; assign any $r \in [0, 1]$ when $t = 0$. Writing $\alpha = \mathbf{1}[a \in A]$ and $\beta = \mathbf{1}[b \in A]$, the expected query gain is

$$d_A(X) = t(X)\{\alpha r(X) - \beta[1 - r(X)]\}. \tag{21}$$

On the interior this is $t(2r - 1)$; on a-only support it is tr; on b-only support it is $-t(1 - r)$; on neither support it is zero. Membership determines some signs, but does not determine the ordering of positive gains across regions. Interior ranking evaluates discrimination in r among unresolved cases, while budgeted fallback decisions additionally depend on the outcome mass t and the calibration of the estimated gain.

Proposition 4 (Conditional decision regret)**.** *For a fixed batch of n observable candidates and exact quota $0 \le k \le n$, let B^* maximize $\sum_{i \in B} d_i$ over sets of size k, and let $\widehat{B}$ maximize the corresponding sum of plug-in gains from $\widehat{t}_i, \widehat{r}_i \in [0, 1]$. Assume each d_i is the expected gain conditional on the available batch information (for example, under independent sampling with sufficient X_i). Then*

$$\sum_{i \in B^*} d_i - \sum_{i \in \widehat{B}} d_i \le 2k \max_i \left\{ |\widehat{t}_i - t_i| + t_i(\alpha_i + \beta_i)|\widehat{r}_i - r_i| \right\}.$$

Proof. Set $f_i(r) = (\alpha_i + \beta_i)r - \beta_i$. For $r \in [0, 1]$, $|f_i(r)| \le 1$ and $|f_i(\widehat{r}) - f_i(r)| = (\alpha_i + \beta_i)|\widehat{r} - r|$. Expanding $\widehat{t}_i f_i(\widehat{r}_i) - t_i f_i(r_i)$ as $(\widehat{t}_i - t_i) f_i(\widehat{r}_i) + t_i[f_i(\widehat{r}_i) - f_i(r_i)]$ bounds each gain error by the displayed expression, denoted ϵ_i. Insert the two plug-in sums into the regret. Their difference is nonpositive by optimality of $\widehat{B}$; the remaining two sums are bounded by $2k \max_i \epsilon_i$. For $k = 0$ both sums vanish, and when $t_i = 0$ the arbitrary definition of r_i has no effect on d_i. □

The bound connects errors in the estimated H/C posterior and outcome mass to regret under the fallback rule. We therefore measure realized fallback utility directly in the experiments below.

Interior training as an ablation. We fit a shared query-agnostic linear scorer $s_\theta(X)$ using

$$\mathcal{L}_I(\theta) = \frac{1}{|\mathcal{Q}_{\text{train}}|} \sum_{A \in \mathcal{Q}_{\text{train}}} \frac{1}{|\mathcal{P}_A|} \sum_{(h,c) \in \mathcal{P}_A} \log(1 + \exp[-s_\theta(h) + s_\theta(c)]) + 0.05\|\theta\|_2^2,$$

where $\mathcal{P}_A$ contains TRAIN interior H–C pairs and only TRAIN-supported queries enter $\mathcal{Q}_{\text{train}}$. The matched raw counterpart uses all query H–C pairs with identical features, penalty and optimizer; TRAIN-supported query sets can differ, so this ablation changes both pair composition and query coverage. Each query uses all pairs up to 4,000, otherwise 4,000 pairs sampled with replacement using a fixed seed. This objective isolates training for interior H/C discrimination; the separate three-outcome head models the additional H/C/W outcome mass needed by the decision readouts. A separately fitted three-outcome logistic head supplies $\widehat{p}_H, \widehat{p}_C, \widehat{p}_W$. Its query ratio and gain readouts use identical probabilities, isolating the effect of retaining outcome mass.

Table 32: **Direct decision consequences at a 20% fallback budget.** Values are utility differences per 100 changed TEST candidates with paired 95% bootstrap intervals. P–R compares projected and raw Diff-Prob; Sel compares interior-selected and raw-selected scorers; Learn compares the projected interior-trained and raw-trained rankers. Q gives validation-eligible/planned queries.

Configuration	Q	P–R	Sel	Learn
CIFAR (17)	18/20	+0.41 [+0.27,+0.56]	+0.01 [-0.05,+0.09]	+0.00 [+0.00,+0.00]
Organ (17)	3/6	+2.65 [+2.06,+3.42]	+0.00 [+0.00,+0.00]	+0.09 [-0.23,+0.30]
Organ (42)	2/6	+0.24 [-0.47,+1.33]	+0.00 [+0.00,+0.00]	-0.24 [-0.89,+0.36]
Organ (117)	4/6	+3.73 [+3.28,+4.38]	+0.00 [-0.11,+0.15]	-0.45 [-0.80,-0.06]
Path (17)	5/6	+2.13 [+0.46,+3.25]	+0.00 [+0.00,+0.00]	-1.36 [-2.35,-0.44]
Path (42)	5/6	+3.63 [+1.66,+4.75]	-0.31 [-0.63,-0.04]	-0.08 [-1.11,+0.40]
Path (117)	5/6	+1.21 [-0.42,+2.53]	+0.88 [+0.48,+1.21]	+0.59 [+0.18,+0.99]

Selection and deployment protocol. The study evaluates seven fixed configurations across CIFAR-100, OrganAMNIST, and PathMNIST using the same selection and fallback protocol. TRAIN fits the heads and feature scaler; validation chooses a scorer separately for each query by raw AUROC or interior AUROC, with fixed candidate order for scorer-selection ties. The candidate set comprises Diff-Prob, entropy change, margin change, the raw-trained ranker and the interior-trained ranker. All are query-agnostic scorers; the two selection rules receive the same query information. Selected names are saved before TEST evaluation. $E_A = 2\,\mathrm{AUC}_{I_A} - 1$ selects a scorer; it is never assigned as an individual sample score.

Budgets are $5/10/20/50\%$ of all changed TEST candidates, rounded upward. A deployable projection extension orders a-only, interior, neither, and b-only tiers, preserving raw order within each non-neutral tier; all within-tier ties use a frozen hash of the sample identifier. SMB instead ties within each tier. Neither support is absent from the conditional metric but must be handled explicitly in deployment. The primary loss uses unit H/C weights, with $w_H/w_C = 0.5, 2$ as sensitivity checks. Raw and interior selection are compared both directly and with the same projection. The two learned heads likewise receive the same projection in their decision comparison. The query ratio and expected-gain readouts are separate query-aware comparisons; the realized-label oracle is an analysis-only bound.

Query support and uncertainty. Decision utility is averaged over queries that satisfy the validation-support requirement. We repeat the analysis with minimum interior validation support of 1, 5, and 10 H and C examples to test whether scorer selection depends on sparse validation evidence. Metric comparisons use matched TEST query sets at the corresponding TEST-support thresholds. We use 300 paired image-bootstrap resamples. Each resample uses the same sampled images across overlapping queries and recomputes the fallback quota and TEST ranking while keeping the validation-selected scorer fixed. The reported intervals therefore quantify TEST-sampling variation for the fixed selection rule. Classifier seeds remain separate configurations, and medical-image intervals use images as the resampling unit because patient identifiers are unavailable.

All seven configurations. We test validation-based scorer selection and matched raw/interior training on seven configurations across three datasets. Every changed TEST prediction, including both-wrong cases, competes for the same fallback quota; ground-truth query membership is used only to evaluate realized gain after selection. Raw and interior validation criteria select different scorers on 14/42 eligible query–configuration instances. We next measure whether these changed selections alter held-out fallback utility. At the primary 20% budget, projection improves held-out utility with a strictly positive interval in 5/7 configurations and has no strictly negative interval. Validation-based scorer selection and interior-only training show configuration-dependent effects, including the positive PathMNIST seed-117 result. The experiments separate three operational mechanisms: support correction at deployment, validation-based scorer selection, and training for interior discrimination. Among them, support correction gives the most consistent held-out utility improvement in this study. Appendix O supplies the full protocol, a projection counterexample and the calibration assumptions used in the decision-regret bound.

Complete result availability. The archive retains per-query avoided H, lost C, selected W, oracle gaps, and weighted-utility results at all four budgets. The common-query tables below evaluate

all metrics on matched TEST query sets, while the decision tables use the queries that satisfy the prespecified validation-support rule.

Table 33: Common-query Diff-Prob metrics. m is minimum TEST interior support per side; Q is query count. Undefined rows are retained.

Configuration	m	Q	Raw	SMB	Proj.	Interior	E_A
CIFAR (17)	1	19	0.708	0.915	0.944	0.652	0.304
CIFAR (17)	5	15	0.713	0.901	0.937	0.673	0.347
CIFAR (17)	10	4	0.691	0.818	0.890	0.700	0.400
Organ (17)	1	5	0.664	0.763	0.866	0.750	0.501
Organ (17)	5	4	0.708	0.705	0.833	0.724	0.447
Organ (17)	10	4	0.708	0.705	0.833	0.724	0.447
Organ (42)	1	5	0.665	0.769	0.857	0.722	0.444
Organ (42)	5	4	0.681	0.714	0.823	0.694	0.389
Organ (42)	10	4	0.681	0.714	0.823	0.694	0.389
Organ (117)	1	5	0.708	0.760	0.870	0.734	0.468
Organ (117)	5	5	0.708	0.760	0.870	0.734	0.468
Organ (117)	10	4	0.719	0.722	0.846	0.721	0.442
Path (17)	1	6	0.691	0.772	0.860	0.698	0.396
Path (17)	5	6	0.691	0.772	0.860	0.698	0.396
Path (17)	10	5	0.704	0.729	0.834	0.710	0.421
Path (42)	1	6	0.640	0.777	0.837	0.689	0.378
Path (42)	5	5	0.638	0.734	0.804	0.642	0.284
Path (42)	10	5	0.638	0.734	0.804	0.642	0.284
Path (117)	1	6	0.653	0.783	0.863	0.663	0.327
Path (117)	5	5	0.661	0.744	0.839	0.683	0.367
Path (117)	10	5	0.661	0.744	0.839	0.683	0.367

Coverage across support thresholds. The complete archive retains all budgets, validation-support thresholds, and query-availability states. Configurations with no eligible query remain unavailable rather than entering the macro average.

Table 34: Mean query utility per 100 changed candidates at 20% budget on validation-eligible queries. Raw and P denote Diff-Prob before and after support projection; RSel and ISel are the scorers selected by raw and interior validation AUROC. Gain uses the estimated expected-gain readout, and Oracle gives the realized-label upper bound for reference.

Configuration	Raw	P	SMB	RSel	ISel	Gain	Oracle
CIFAR (17)	0.19	0.60	0.60	0.18	0.19	0.86	1.34
Organ (17)	9.67	12.32	8.15	10.87	10.87	12.91	20.03
Organ (42)	10.37	10.61	4.20	10.37	10.37	2.87	20.02
Organ (117)	8.91	12.64	9.66	9.06	9.06	12.27	20.03
Path (17)	4.12	6.25	4.52	3.53	3.53	-4.89	16.69
Path (42)	0.95	4.58	3.74	1.30	0.99	2.86	16.18
Path (117)	6.57	7.78	6.39	9.28	10.17	12.37	15.38

Absolute decision utility and baselines.

Projection-tie sensitivity. A separate robustness check retains exact ties within the forced H and C support tiers, resolving them by the same frozen sample-ID hashes. At the 20% budget, 5/7 configurations still have positive pointwise intervals against raw Diff-Prob. The archived per-configuration values retain the corresponding effect magnitudes under this alternative tie treatment.

P MULTI-DATASET VALIDATION SELECTION AND DECISION UTILITY

We next connect the evaluation effect to model selection: does choosing a scorer by interior validation AUROC lead to a different scorer and different held-out fallback decisions than choosing by raw conditional AUROC? The fixed-Diff-Prob experiment in Appendix O separately evaluates support projection at deployment.

Predictor pairs and selection protocol. The selection study covers 24 configurations: ImageNet ResNet-18→ResNet-50/ConvNeXt-T and CIFAR-100 ResNet-20→ResNet-56/VGG13 with three reliability splits each; OrganAMNIST, PathMNIST and RetinaMNIST with three classifier seeds each; and CLINC150 BERT→RoBERTa with three seeds. ImageNet reliability splits share frozen classifiers and the same 40,000 TEST images. Each query selects the scorer with the highest validation raw or interior AUROC from the same seven candidates, with fixed method order resolving ties. The heads and their regularization choices remain frozen. This experiment uses a fixed seven-score candidate menu; the earlier five-score ablation in Appendix O is reported separately.

Candidate scorers. Let p_a, p_b be the original and revised classifiers' probability vectors, $h(p) = -\sum_k p_k \log p_k$, and $m(p) = p_{(1)} - p_{(2)}$ the top-two probability margin. Diff-Prob is $\max_k p_a(k) - \max_k p_b(k)$; entropy change is $h(p_b) - h(p_a)$; margin change is $m(p_a) - m(p_b)$; revised-model entropy is $h(p_b)$. The learned readouts use both full probability vectors, both entropies, maxima and margins, and the cross-probabilities $p_b(a), p_a(b)$, standardized using changed reliability-TRAIN candidates. Semantic logistic is a binary logistic head trained on H/C cases (H=1, C=0), returning its linear decision function. A separate multinomial logistic head trained on all changed reliability-TRAIN candidates predicts C/H/W probabilities; outcome H/C log odds returns $\log p_H - \log p_C$, and outcome net harm returns $p_H - p_C$. Both heads minimize L2-regularized cross-entropy, with inverse regularization $C \in \{0.01, 0.1, 1\}$ selected by the corresponding reliability-VAL log loss; logarithms use probabilities clipped below at 10^{-12}. All seven scorers are query-agnostic: neither their inputs nor their training targets depend on the semantic query. In the scorer-selection experiment, heads, standardization and regularization are frozen; Raw-selected and Interior-selected share exactly this menu and differ only in the validation selection criterion.

Fallback population, utility and intervals. All changed TEST candidates, including W and true classes outside the query, compete for $\lceil qN \rceil$ fallbacks at $q = 5/10/20/50\%$. TEST labels compute realized query gain $\mathbf{1}_H\mathbf{1}[a \in A] - \mathbf{1}_C\mathbf{1}[b \in A]$ after selection; utility is reported per 100 changed candidates. The primary comparison deploys each selected scorer directly and averages equally over queries with at least one validation-interior H and C; thresholds 5 and 10 assess support sensitivity. For the original fixed-validation results, 300 paired TEST resamples recompute the quota and preserve hashed sample-ID ties and dependence across overlapping queries. Resampling units are images for image datasets and normalized-utterance source groups for CLINC150. The paired intervals quantify TEST-sampling variation with the trained heads and validation selections fixed.

P.1 HELD-OUT CONSEQUENCES OF VALIDATION SELECTION

Changes in selected errors. For OrganAMNIST seed 17, the kidney query selects outcome net harm under raw validation and entropy change under interior validation. Both policies send 285 of 1,423 changed candidates to fallback at 20% budget. Avoided H rises from 79 to 110, with 10 C lost under each policy, increasing net gain from 69 to 100. Validation interior support is 10 H/4 C and TEST support is 132 H/75 C. For PathMNIST seed 17, the complement of background/debris/mucus has 24 H/17 C in validation and 137 H/160 C on TEST; its net gain rises from -11 to 36 at the same 109-candidate quota. In both cases, changing only the validation criterion changes the deployed scorer and increases the number of harmful updates intercepted at the same fallback budget.

For the 49-class ImageNet clothing query (`wnid_n03051540`), interior validation selects Diff-Prob instead of the raw-selected entropy score. On TEST, Diff-Prob also has higher interior AUROC (0.673 versus 0.621). At the 5% fallback budget, both policies rank the same pool of 9,858 changed predictions, of which 493 receive fallback. Validation contains 3 interior H and 16 interior C; the support-threshold analysis below evaluates the sensitivity of this selection to that validation size. For CIFAR-100 `vehicles_2`, ResNet-20→ResNet-56, split 117, interior selection changes revised-model entropy to Diff-Prob at 20% budget. The gain is obtained at the same fallback budget, with

avoided harms increasing from 3 to 9 while lost corrections remain at two. The corresponding utility difference is $+0.391$ $[0.031, 0.782]$ per 100 changed candidates.

Validation-support sensitivity. Support sensitivity separates sparse selection cases from effects supported by larger interior validation populations. The OrganAMNIST kidney query has 10 H and 4 C in validation, while the ImageNet clothing and CIFAR vehicle examples are removed by the five-per-side threshold; the CIFAR vehicle interior contains one H and two C. PathMNIST provides a higher-support case: the complement of background/debris/mucus contains 24 H and 17 C in validation and 137 H and 160 C on TEST, and its net gain rises from -11 to 36 under the same 109-candidate quota. This result shows that the observed decision benefit is not restricted to the sparsest validation interiors. PathMNIST seed 17 retains all five eligible queries and utilities $1.875/3.750$ at minimum-side support five. At support ten, the three retained queries yield $1.348/4.228$, with difference $+2.880$ $[1.622, 3.676]$. The next experiment varies the validation sample itself while fixing TEST predictions.

P.2 Replication Across Validation Samples, Models, and Datasets

We next test whether validation-based scorer selection behaves similarly when the validation sample, predictor pair, and dataset change. We repeat scorer selection over 50 independently resampled validation sets while keeping each configuration's TEST predictions fixed. The resulting experiment directly measures how validation-sample variation changes the selected scorer and its held-out fallback utility. The matrix spans natural images, medical images, intent classification, domain shift, and audio, using the same seven-score selection rule and 20% fallback budget throughout. Across predictor seeds, domains, and folds, 56 configurations contain the complete seven-score menu.

Changing the validation sample while holding TEST fixed. Each draw resamples the validation groups with replacement and reruns both raw-AUROC and interior-AUROC scorer selection on exactly the same draw. Because TEST predictions never change, variation in ΔU comes from changes in the validation sample and the scorer selected from it. Each draw samples the original number of validation source groups with replacement, retaining every observation in a sampled group. Image datasets use image groups, ESC-50 uses source recordings and CLINC150 uses normalized utterances. The two criteria share the same draw; all queries share its weights, preserving their overlap. Weighted AUROC, the original method tie order and minimum interior support one determine the selected scorers and eligible queries before TEST utility is computed. Classifiers, heads, normalization and head hyperparameters stay frozen; the head hyperparameters were selected using the original validation pool. DomainNet selections use real-domain validation and transfer unchanged to its six TEST domains. Every TEST fallback pool includes W and out-of-query outcomes, and every draw uses the original hashed sample-ID tie rule.

Table 35 reports mean paired TEST utilities over draws with defined query macros. Its intervals are the 2.5th and 97.5th percentiles of the paired utility differences across validation draws, with TEST fixed. They measure variation in validation-based scorer selection; the original result tables resample TEST while fixing the selected scorers. B reports how many of the 50 validation draws contain at least one eligible query, and Q reports the range of eligible-query counts across draws. A second analysis fixes each configuration's query set to the intersection of validation-eligible queries across all 50 draws; its per-draw results and both selectors' individual utility intervals are archived.

Table 35: **Validation-sample replication at fixed TEST predictions.** All available seven-score configurations, with predictor/head seed in parentheses. B is defined/50 validation draws; Q is the eligible/planned query range. Utilities and paired differences are per 100 changed TEST candidates. Intervals describe validation-resampling variation. Positive ΔU means that the scorer selected by interior validation AUROC achieves higher TEST fallback utility than the scorer selected by raw validation AUROC. R18/R50 and R20/R56 denote ResNet pairs; CNeXt is ConvNeXt-T and RoB. is RoBERTa. ESC-50 seed 17, fold 1 has five archived candidates and is unavailable for this seven-score replication.

Configuration	B	Q	Raw	Interior	ΔU	95% interval
ImageNet R18/CNeXt (17)	50/50	53–71/395	-0.6727	-0.6761	-0.0034	[-0.0198, +0.0155]
ImageNet R18/CNeXt (42)	50/50	47–63/395	-0.7263	-0.7289	-0.0026	[-0.0238, +0.0236]
ImageNet R18/CNeXt (117)	50/50	46–67/395	-0.7055	-0.7125	-0.0070	[-0.0448, +0.0154]
ImageNet R18/R50 (17)	50/50	48–64/395	-0.6746	-0.6820	-0.0073	[-0.0611, +0.0314]
ImageNet R18/R50 (42)	50/50	53–72/395	-0.6268	-0.6309	-0.0041	[-0.0409, +0.0193]
ImageNet R18/R50 (117)	50/50	43–64/395	-0.7267	-0.7485	-0.0218	[-0.0591, +0.0022]
CIFAR-100 R20/R56 (17)	50/50	11–18/20	0.1901	0.1935	+0.0033	[-0.0530, +0.0509]

Configuration	B	Q	Raw	Interior	ΔU	95% interval
CIFAR-100 R20/R56 (42)	50/50	10–17/20	0.1704	0.1425	-0.0279	[-0.0697, +0.0254]
CIFAR-100 R20/R56 (117)	50/50	10–17/20	0.2075	0.1816	-0.0259	[-0.0651, +0.0246]
CIFAR-100 R20/VGG13 (17)	50/50	10–14/20	0.0791	0.0541	-0.0250	[-0.0810, +0.0299]
CIFAR-100 R20/VGG13 (42)	50/50	9–16/20	0.0480	0.0725	+0.0245	[-0.0328, +0.1012]
CIFAR-100 R20/VGG13 (117)	50/50	10–16/20	0.0783	0.0971	+0.0188	[-0.0356, +0.0609]
OrganAMNIST R18/R50 (17)	50/50	3/6	9.5798	9.3643	-0.2155	[-1.6866, +0.7262]
OrganAMNIST R18/R50 (42)	48/50	0–2/6	8.5856	8.3989	-0.1866	[-2.7845, +0.0000]
OrganAMNIST R18/R50 (117)	49/50	0–4/6	9.1009	9.1846	+0.0837	[-1.2314, +1.1147]
PathMNIST R18/R50 (17)	50/50	5/6	2.6037	2.8706	+0.2669	[-2.2748, +1.9605]
PathMNIST R18/R50 (42)	50/50	5/6	0.9695	0.8939	-0.0756	[-0.5935, +0.4962]
PathMNIST R18/R50 (117)	50/50	4–5/6	7.4429	6.9679	-0.4750	[-3.9752, +2.1333]
RetinaMNIST R18/R50 (17)	49/50	0–2/3	-4.5822	-4.8421	-0.2599	[-3.6792, +2.8302]
RetinaMNIST R18/R50 (42)	44/50	0–2/3	6.2872	7.0943	+0.8071	[-3.6332, +2.8037]
RetinaMNIST R18/R50 (117)	50/50	2/3	1.1204	1.1019	-0.0185	[+0.0000, +0.0000]
CLINC150 BERT/RoB. (17)	45/50	0–3/10	0.5781	0.5781	+0.0000	[+0.0000, +0.0000]
CLINC150 BERT/RoB. (42)	34/50	0–1/10	1.2739	1.2739	+0.0000	[+0.0000, +0.0000]
CLINC150 BERT/RoB. (117)	43/50	0–3/10	0.6369	0.6369	+0.0000	[+0.0000, +0.0000]
DomainNet clipart R18/R50 (17)	50/50	4–8/8	0.7843	0.7872	+0.0030	[-0.0023, +0.0405]
DomainNet infograph R18/R50 (17)	50/50	4–8/8	-0.2021	-0.1988	+0.0033	[-0.0118, +0.0144]
DomainNet painting R18/R50 (17)	50/50	4–8/8	0.6732	0.6737	+0.0005	[-0.0089, +0.0275]
DomainNet quickdraw R18/R50 (17)	50/50	4–8/8	0.0767	0.0756	-0.0011	[-0.0120, +0.0042]
DomainNet real R18/R50 (17)	50/50	4–8/8	0.5586	0.5611	+0.0025	[-0.0038, +0.0154]
DomainNet sketch R18/R50 (17)	50/50	4–8/8	0.0936	0.0937	+0.0001	[-0.0113, +0.0051]
DomainNet clipart R18/R50 (42)	50/50	4–7/8	0.3043	0.3108	+0.0065	[-0.0015, +0.0425]
DomainNet infograph R18/R50 (42)	50/50	4–7/8	-0.0376	-0.0382	-0.0006	[-0.0060, +0.0074]
DomainNet painting R18/R50 (42)	50/50	4–7/8	0.5485	0.5660	+0.0176	[-0.0070, +0.2160]
DomainNet quickdraw R18/R50 (42)	50/50	4–7/8	0.0009	0.0012	+0.0003	[-0.0004, +0.0027]
DomainNet real R18/R50 (42)	50/50	4–7/8	0.4775	0.4983	+0.0208	[+0.0000, +0.2276]
DomainNet sketch R18/R50 (42)	50/50	4–7/8	0.1941	0.2039	+0.0098	[-0.0032, +0.1222]
DomainNet clipart R18/R50 (117)	50/50	4–6/8	0.4592	0.4585	-0.0007	[-0.0124, +0.0158]
DomainNet infograph R18/R50 (117)	50/50	4–6/8	-0.1368	-0.1357	+0.0012	[-0.0039, +0.0140]
DomainNet painting R18/R50 (117)	50/50	4–6/8	0.4377	0.4390	+0.0013	[-0.0237, +0.0337]
DomainNet quickdraw R18/R50 (117)	50/50	4–6/8	0.1918	0.1917	-0.0001	[-0.0034, +0.0024]
DomainNet real R18/R50 (117)	50/50	4–6/8	0.9331	0.9322	-0.0009	[-0.0194, +0.0100]
DomainNet sketch R18/R50 (117)	50/50	4–6/8	0.2214	0.2233	+0.0018	[-0.0045, +0.0102]
ESC-50 fold 2 Cnn6/Cnn14 (17)	22/50	0–1/5	-3.4483	-3.4483	+0.0000	[+0.0000, +0.0000]
ESC-50 fold 3 Cnn6/Cnn14 (17)	41/50	0–2/5	-3.5965	-3.2038	+0.3927	[-1.6949, +4.2373]
ESC-50 fold 4 Cnn6/Cnn14 (17)	41/50	0–2/5	-1.3843	-1.2195	+0.1648	[+0.0000, +2.7027]
ESC-50 fold 5 Cnn6/Cnn14 (17)	0/50	0/5	NA	NA	NA	NA
ESC-50 fold 1 Cnn6/Cnn14 (42)	22/50	0–1/5	1.5437	1.8868	+0.3431	[+0.0000, +3.5849]
ESC-50 fold 2 Cnn6/Cnn14 (42)	20/50	0–1/5	-1.5385	-1.5385	+0.0000	[+0.0000, +0.0000]
ESC-50 fold 3 Cnn6/Cnn14 (42)	0/50	0/5	NA	NA	NA	NA
ESC-50 fold 4 Cnn6/Cnn14 (42)	28/50	0–1/5	0.0000	0.0000	+0.0000	[+0.0000, +0.0000]
ESC-50 fold 5 Cnn6/Cnn14 (42)	31/50	0–1/5	-1.1885	-1.1885	+0.0000	[+0.0000, +0.0000]
ESC-50 fold 1 Cnn6/Cnn14 (117)	38/50	0–1/5	4.3694	5.6604	+1.2910	[+0.0000, +1.8868]
ESC-50 fold 2 Cnn6/Cnn14 (117)	0/50	0/5	NA	NA	NA	NA
ESC-50 fold 3 Cnn6/Cnn14 (117)	0/50	0/5	NA	NA	NA	NA
ESC-50 fold 4 Cnn6/Cnn14 (117)	38/50	0–2/5	3.6011	6.3366	+2.7355	[+0.0000, +5.2632]
ESC-50 fold 5 Cnn6/Cnn14 (117)	0/50	0/5	NA	NA	NA	NA
ESC-50 fold 1 Cnn6/Cnn14 (17)	NA	NA	NA	NA	NA	five-score archive

Changing validation samples. Resampling the validation population changes which scorer is selected while leaving TEST predictions fixed. For PathMNIST seed 17, interior-based selection yields higher TEST utility in 35 of 50 validation draws, ties in two, and lower utility in 13. Its mean utility increases from 2.604 to 2.871 ($\Delta U = +0.267$). All five eligible queries remain present in every draw, so this change is driven by scorer selection rather than query-set composition. Across configurations, Table 35 reports the corresponding variation induced solely by changing the validation sample.

Changing predictor pairs and datasets. On CIFAR-100, changing the predictor pair changes the effect of the validation criterion. For ResNet-20→VGG13, splits 42 and 117 yield positive mean differences of $+0.0245$ and $+0.0188$, with interior selection improving TEST utility in 34/50 and 38/50 validation draws. The ResNet-20→ResNet-56 pair exhibits a different selection pattern under the same protocol. The same selection effect also appears outside the original image-classification settings. For DomainNet painting (seed 42), mean TEST utility increases from 0.5485 to 0.5660 under interior-based selection. For ESC-50 seed 117, fold 4, it increases from 3.601 to 6.337 across the 38 validation draws with eligible queries. These cases extend the observed selection consequence to domain-shifted images and audio.

Replication outcome. Validation-based scorer selection produces measurable changes in held-out fallback utility across natural images, medical images, domain shift, intent classification, and audio. The 50-draw resampling experiment further shows that these changes recur when the validation sample itself is varied while TEST predictions remain fixed. Their magnitude depends on the predictor pair and dataset, demonstrating that the operational consequence is configuration-dependent rather than tied to a single validation split. Together with the fixed-validation experiments, these results extend the scorer-selection effect across validation samples, model pairs, and data modalities.

P.3 Complete Configuration Results and Deployment Contrasts

Table 36: **Complete seven-score validation-selection comparison at a 20% fallback budget.** Raw and Interior denote TEST utility obtained from the scorers selected by raw and interior validation AUROC, respectively. $\Delta U = \text{Interior} - \text{Raw}$ is utility per 100 changed candidates; Q gives validation-eligible/planned queries. All 24 configurations are reported.

Configuration	Q	Raw	Interior	ΔU	[95% CI]
ImageNet R18/R50 (17)	70/395	-0.537	-0.546	-0.009	[-0.026,+0.001]
ImageNet R18/R50 (42)	82/395	-0.479	-0.485	-0.007	[-0.021,+0.005]
ImageNet R18/R50 (117)	69/395	-0.571	-0.575	-0.004	[-0.017,+0.010]
ImageNet R18/CNeXt (17)	76/395	-0.546	-0.554	-0.008	[-0.018,-0.000]
ImageNet R18/CNeXt (42)	70/395	-0.575	-0.585	-0.010	[-0.025,+0.005]
ImageNet R18/CNeXt (117)	77/395	-0.552	-0.550	0.002	[-0.008,+0.013]
CIFAR R20/R56 (17)	18/20	0.195	0.195	0.000	[-0.060,+0.056]
CIFAR R20/R56 (42)	19/20	0.185	0.103	-0.082	[-0.123,-0.010]
CIFAR R20/R56 (117)	18/20	0.214	0.163	-0.051	[-0.114,+0.029]
CIFAR R20/VGG (17)	14/20	0.043	0.067	0.024	[-0.033,+0.086]
CIFAR R20/VGG (42)	16/20	0.046	0.104	0.058	[-0.013,+0.117]
CIFAR R20/VGG (117)	17/20	0.086	0.094	0.008	[-0.057,+0.075]
Organ R18/R50 (17)	3/6	9.487	10.213	0.726	[+0.398,+1.078]
Organ R18/R50 (42)	2/6	10.371	10.371	0.000	[+0.000,+0.000]
Organ R18/R50 (117)	4/6	9.057	9.057	0.000	[-0.113,+0.178]
Path R18/R50 (17)	5/6	1.875	3.750	1.875	[+1.084,+2.390]
Path R18/R50 (42)	5/6	1.260	0.992	-0.267	[-0.573,+0.000]
Path R18/R50 (117)	5/6	8.073	6.532	-1.541	[-2.478,-0.384]
Retina R18/R50 (17)	2/3	-4.717	-4.717	0.000	[+0.000,+0.000]
Retina R18/R50 (42)	2/3	5.607	8.411	2.804	[-2.804,+7.477]
Retina R18/R50 (117)	2/3	0.926	0.926	0.000	[+0.000,+0.000]
CLINC BERT/RoB. (17)	3/10	0.676	0.676	0.000	[+0.000,+0.000]
CLINC BERT/RoB. (42)	1/10	1.274	1.274	0.000	[+0.000,+0.000]
CLINC BERT/RoB. (117)	3/10	0.637	0.637	0.000	[-0.637,+0.637]

Configuration coverage. Table 36 reports all 24 configurations under the same seven-score candidate menu, 20% fallback budget, and minimum validation-support rule. The complete matrix shows that changing the validation criterion can improve, preserve, or reduce downstream utility depending on the predictor configuration. This selection experiment is distinct from the fixed-Diff-Prob deployment experiment in Table 2, which isolates the effect of support projection on a fixed scorer.

Separating validation selection from deployment-time projection. To isolate the two mechanisms, we also apply the same support projection to the scorers chosen by both validation criteria. Under this matched deployment rule, the remaining difference reflects scorer selection rather than projection itself. The resulting contrasts are reported separately from the fixed-scorer projection experiment and are never added to its utility gains.

Reproduction records. The archive retains all configurations, budgets, support thresholds, validation selections, per-query decision outcomes, bootstrap arrays, and model/prediction hashes. The

validation-resampling records additionally store all 50 resampled validation draws and their selected scorers, query coverage, and TEST utilities.

Q Additional Decision Cases: DomainNet and ESC-50

Extending validation selection to domain shift and audio. We apply the same seven-score validation-selection protocol to DomainNet and ESC-50, testing whether the scorer-selection effect also appears under visual domain shift and in a different modality. The study covers DomainNet ResNet-18→ResNet-50 across three seeds and six TEST domains, and ESC-50 Cnn6→Cnn14 across three seeds and five official outer folds: 18 DomainNet and 15 ESC-50 configurations. DomainNet fits and selects on real-domain development data and transfers each choice to all TEST domains. ESC-50 preserves source-recording separation within each fold. Both selectors share the seven query-agnostic candidates, frozen heads and fixed method tie order defined in Appendix P.

Decision protocol and coverage. All changed TEST predictions compete for the same fallback quota, including both-wrong and out-of-query outcomes. DomainNet uses image-level paired resampling, while ESC-50 resamples source recordings to preserve recording-level dependence. All scorer candidates, tie rules, budgets, and validation-support thresholds match Appendix P.

Held-out image and audio utility. Interior-based validation changes the deployed scorer in both visual domain-shift and audio settings. For DomainNet painting (seed 42), TEST utility increases from 0.61004 to 0.61396 per 100 changed candidates. For ESC-50 seed 117, fold 4, utility increases from 1.316 to 3.947 at the same 20% fallback budget. The latter corresponds to three additional harmful prediction changes intercepted without increasing the eight-fallback quota. Table 37 reports all 33 original configurations. The validation-resampling analysis in Appendix P.2 repeats selection across 50 validation samples using the same frozen candidate scores.

Sensitivity to validation support. We increase the minimum number of interior H and C validation examples to distinguish effects supported by larger validation populations from selections driven by sparse interiors. In the DomainNet seed-42 case, the vehicle and fruit queries each have one validation-interior H and three C, the instrument query has four H and two C, and each complement query has at least 313 H and 847 C. Threshold five retains the four complement queries and yields zero macro difference. The ESC-50 example is supported by a smaller validation interior: the domestic and urban queries contain one H each and one and two C, respectively. The support-threshold analysis therefore separates this sparse audio case from the higher-support results reported for PathMNIST.

Table 37: **Complete DomainNet and ESC-50 validation-selection results at a 20% fallback budget.** Raw and Interior denote TEST utility from scorers selected by raw and interior validation AUROC. $\Delta U = \text{Interior} - \text{Raw}$; Q gives validation-eligible/planned queries. DomainNet uses three seeds across six TEST domains, and ESC-50 uses three seeds across five outer folds. ESC-50 seed 17, fold 1 uses the five scorers available in its archive because its three-outcome training population lacks a required outcome.

Configuration	Q	Raw	Interior	ΔU	[95% CI]
DN clipart (117)	6/8	0.4141	0.4096	-0.0045	[-0.0158,+0.0068]
DN infograph (117)	6/8	-0.1451	-0.1451	+0.0000	[-0.0051,+0.0051]
DN painting (117)	6/8	0.3654	0.3716	+0.0062	[+0.0000,+0.0139]
DN quickdraw (117)	6/8	0.1841	0.1845	+0.0004	[+0.0000,+0.0017]
DN real (117)	6/8	0.8938	0.8955	+0.0017	[-0.0084,+0.0151]
DN sketch (117)	6/8	0.1944	0.1982	+0.0039	[-0.0026,+0.0104]
DN clipart (17)	8/8	0.6665	0.6682	+0.0017	[-0.0102,+0.0136]
DN infograph (17)	8/8	-0.1804	-0.1753	+0.0050	[-0.0038,+0.0164]
DN painting (17)	8/8	0.5241	0.5276	+0.0034	[-0.0103,+0.0161]
DN quickdraw (17)	8/8	0.0548	0.0526	-0.0023	[-0.0052,+0.0003]
DN real (17)	8/8	0.4593	0.4631	+0.0038	[-0.0102,+0.0166]
DN sketch (17)	8/8	0.0615	0.0644	+0.0029	[-0.0049,+0.0107]

Configuration	Q	Raw	Interior	ΔU	[95% CI]
DN clipart (42)	7/8	0.2911	0.2970	+0.0059	[-0.0059,+0.0176]
DN infograph (42)	7/8	-0.0213	-0.0241	-0.0028	[-0.0071,+0.0000]
DN painting (42)	7/8	0.6100	0.6140	+0.0039	[-0.0040,+0.0144]
DN quickdraw (42)	7/8	0.0040	0.0044	+0.0004	[-0.0007,+0.0018]
DN real (42)	7/8	0.5735	0.5793	+0.0059	[-0.0059,+0.0190]
DN sketch (42)	7/8	0.2401	0.2458	+0.0056	[-0.0023,+0.0124]
ESC (117, fold 1)	1/5	3.7736	5.6604	+1.8868	[-1.8182,+5.6604]
ESC (117, fold 2)	0/5	—	—	—	—
ESC (117, fold 3)	0/5	—	—	—	—
ESC (117, fold 4)	2/5	1.3158	3.9474	+2.6316	[+0.0000,+6.5789]
ESC (117, fold 5)	0/5	—	—	—	—
ESC (17, fold 1)	2/5	-3.2609	-3.2609	+0.0000	[+0.0000,+0.0000]
ESC (17, fold 2)	1/5	-3.4483	-3.4483	+0.0000	[+0.0000,+0.0000]
ESC (17, fold 3)	2/5	-3.3898	-4.2373	-0.8475	[-3.5088,+2.5424]
ESC (17, fold 4)	2/5	-1.3514	-1.3514	+0.0000	[+0.0000,+0.0000]
ESC (17, fold 5)	0/5	—	—	—	—
ESC (42, fold 1)	1/5	1.8868	1.8868	+0.0000	[+0.0000,+0.0000]
ESC (42, fold 2)	1/5	-1.5385	-1.5385	+0.0000	[+0.0000,+0.0000]
ESC (42, fold 3)	0/5	—	—	—	—
ESC (42, fold 4)	1/5	0.0000	0.0000	+0.0000	[+0.0000,+0.0000]
ESC (42, fold 5)	1/5	-1.7544	-1.7544	+0.0000	[+0.0000,+0.0000]

Complete configuration coverage. Table 37 reports every DomainNet domain/seed and ESC-50 fold/seed configuration under the same decision protocol. Across this matrix, the validation criterion changes utility in both directions, while the highlighted DomainNet and ESC-50 cases show that the scorer-selection consequence extends beyond the original image-classification datasets. The validation-resampling analysis separately measures how these selections vary when the validation sample changes.