WHEN CONTEXT HELPS: RECOGNIZER-RELATIVE BENEFIT INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual context is not intrinsically helpful: the same surrounding evidence can reduce one recognizer's loss and increase another's. We study what can transfer when this per-example benefit target changes with the recognizer and no target benefit labels are available. The answer is not a universal gate. A matched source router trained directly on routed NLL transfers a substantially worse benefit ordering than explicit benefit supervision, including on unseen supervised CNNs. More importantly, under one frozen CLIP-trained decoder and a fixed target benefit event, the target recognizer's own label-free response relations are the highest-AUC evidence origin in all 20 non-source target–dataset cases across four datasets. This is not built into the target definition: the interface contains neither the ground-truth label nor the target correct-class probability, and the decoder and normalization are learned only on the source. Finally, target responses can remain predictively useful where the frozen source interpretation fails: structural relations improve target-supervised prediction on ImageNet-9 while the same source-learned augmentation degrades it. These separations identify a *recognizer-relative benefit interface*: portable response coordinates need not imply a portable decoder or decision rule.

1 INTRODUCTION

Visual context can disambiguate an object, reinforce an already correct prediction, or amplify a mistake (Zhang et al., 2020; Xiao et al., 2021). The unresolved transfer question is not simply whether context is useful, but whether a rule learned to predict *one recognizer's* incremental benefit can be reused for another recognizer without target benefit labels. Let A_m denote recognizer m's reduction in correct-class log loss when context is added and $H_m = \mathbf{1}[A_m > 0]$. For the same object, context, and label, H_m can change with m. Replacing the recognizer therefore changes both the observed responses and the prediction target itself. This is a model-relative value-transfer problem rather than ordinary category or domain transfer.

Three transfer questions are easily conflated. *Objective portability* asks whether optimizing the source routing loss produces the most reusable score. *Evidence portability* asks which recognizer's responses best expose a fixed target benefit event. *Decoder portability* asks whether a source-learned interpretation of those responses remains valid on the target recognizer. We isolate them rather than treating "a transferable router" as one indivisible object. In particular, our evidence-origin experiment fixes the target event, observations, source decoder, and source normalization, and changes only the recognizer that instantiates a common set of response relations.

Self-origin dominance is not built into the definition of benefit. Although H_t depends on the target recognizer's correct-class probability change, the deployed evidence F_t is label-free: it contains neither Y nor $p_t(Y)$, and the decoder and scaler are fitted only on source data. Under the fixed source decoder, nothing forces F_t to rank H_t better than matched responses F_m from another recognizer. Indeed, donor contexts provide cases in which the frozen source interpretation fails even though target-supervised holdouts recover more target-benefit information from F_t. The empirical question is therefore whether source-learned *coordinate semantics* align with the target recognizer, not whether a model trivially predicts a loss quantity that was given to it.

The experiments give three separations (Fig. 1). First, a matched Direct-NLL router transfers substantially worse benefit ordering than explicit benefit supervision, including on supervised CNNs unseen during source-router training: **source task optimization does not imply portable benefit ordering**.

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.

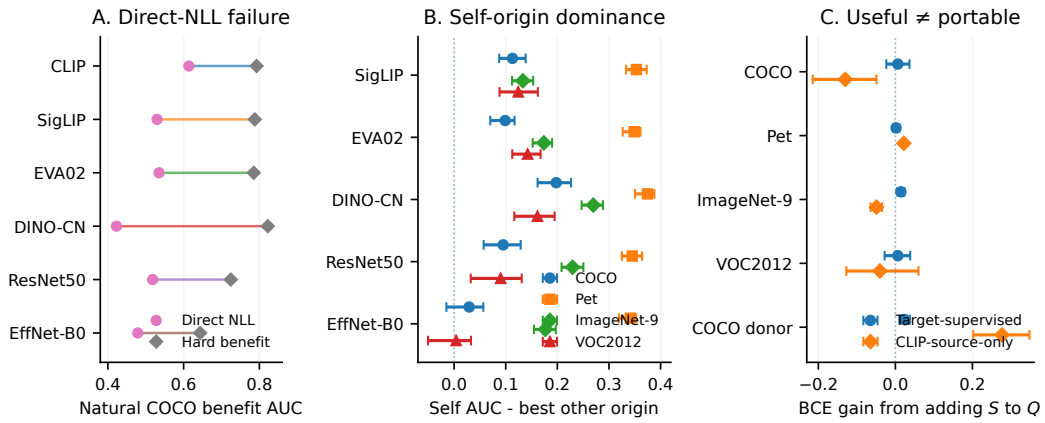

Figure 1: **Three empirical separations define the benefit interface.** *Left:* on natural COCO, explicit Hard-benefit supervision transfers a substantially better benefit ordering than a matched CLIP-source router trained directly on routed NLL; points are descriptive AUCs under the same source protocol. *Middle:* with one unchanged CLIP benefit decoder, each non-source target's own response interface beats its strongest alternative evidence origin across COCO, Pet, precision-corrected ImageNet-9, and VOC2012; bars are paired selected-best 95% intervals. *Right:* adding structural relations S to confidence family Q can help a target-supervised learner yet hurt the same test set when the (Q, S) interpretation is learned only on CLIP; intervals are prespecified 99% family-adjusted contrasts. Thus direct task optimization, evidence origin, and conditional usefulness are distinct transfer questions.

Second, with one frozen CLIP decoder, the target recognizer's own evidence is the highest-AUC origin among six recognizers in all 20 non-source target–dataset cases across COCO, Pet, ImageNet-9, and VOC2012: **the reusable object is a recognizer-relative response interface, not a source model's scene score**. Third, structural relations can improve held-out target-supervised prediction beyond a broad confidence family while the same source-learned augmentation moves in the opposite direction on ImageNet-9: **conditional usefulness does not imply portable interpretation**. A compatibility analysis formalizes sufficient conditions for shared ranking, while failed multi-source and confidence-combination repairs delineate the boundary. The contribution is this decomposition of model-relative value transfer—which response coordinates transfer, which interpretation does not, and why these properties must be tested separately—rather than a new universally better gate.

2 Relation to Information Value and Routing

System risk versus transfer of a model-relative value target. Learning-to-defer studies selection among predictors under a specified system risk, including multiple-expert consistency guarantees (Mozannar & Sontag, 2020; Verma et al., 2023; Mao et al., 2025). RouteLLM already shows that a preference-trained query router can generalize to unseen strong/weak model pairs without retraining (Ong et al., 2025). Other recent new-model routers construct explicit target-model capability evidence: Identity-Free Deferral builds a few-shot competence profile; ICL-Router profiles candidates on a query set; ZeroRouter embeds a model from anchor queries; ContextualRouter estimates candidate performance from historical query outcomes; and SCOPE retrieves prior model behavior to predict performance and cost (Strong et al., 2026; Wang et al., 2026; Yan et al., 2026; Varangot-Reille et al., 2026; Cao et al., 2026). Our question is different: the target label itself is recognizer-relative incremental context benefit, not generic response quality or a fixed strong/weak choice. The unchanged deployment map receives no target correctness history or anchor evaluation and reads only that recognizer's own core/context/full response relations. Value-of-information and dynamic feature acquisition motivate utility-based selection (Bilgic & Getoor, 2007; Covert et al., 2023; Gadgil et al., 2024), but do not by themselves establish decoder portability. We compute all three views and do not claim acquisition-cost savings.

Value and confidence are established ingredients. The utility and limits of visual context have been studied directly, including upper bounds on contextual improvement for object detection (Barnea & Ben-Shahar, 2019). Dynamic multimodal fusion and Learning to Route learn instance-dependent information use (Han et al., 2022; Ajirak et al., 2025); DyMo uses before/after cross-entropy decrease as a modality-value proxy (Du et al., 2026). Neither context utility nor loss decrease is our novelty. ConfidNet learns confidence using label-aware targets (Corbière et al., 2019). Risk Advisor is model-agnostic across black-box classifier families but fits a failure-risk meta-learner from the deployed classifier's own training records; confidence differences can estimate aggregate accuracy under shift (Lahoti et al., 2023; Guillory et al., 2021). Our deployment decoder instead receives no target-recognizer fitting records and predicts that recognizer's per-example probability change between views. Confidence and pseudo-label gains are therefore necessary controls.

Conditional information and ranking. Residual-based conditional-independence methods explicitly address interaction and modeling assumptions (Sudarshan et al., 2023). Our residual-only portability experiment is not such a test; our nested-feature comparison measures finite-learner held-out risk, not unrestricted conditional mutual information. Proper-loss ranking guarantees and multi-source adaptation also predate this work (Agarwal, 2014; Gao & Zhou, 2015; Mansour et al., 2008). We specialize these foundations to recognizer-relative benefit ordering and test the resulting distinctions, without claiming new general ranking or conditional-independence theory.

3 Benefit Information and Its Decoder

Model-relative observations. For core X, context C, and recognizer m, define

$$A_m = \log p_{m,f}(Y \mid X, C) - \log p_{m,c}(Y \mid X), \qquad H_m = \mathbf{1}[A_m > 0]. \tag{1}$$

Positive benefit need not change top-1 correctness. For Bayes-optimal nested posteriors, $\mathbb{E}[A] = I(Y; C \mid X) \geq 0$, although realized benefit may be negative; finite recognizers can also have negative mean benefit. Label-free evidence F_m comprises 14 scalars: three pairwise JS divergences, three top-two margins, three entropies, three prediction-agreement indicators, and two within-model feature cosines. None uses Y or the correct-class probability, and no common embedding basis is assumed across recognizers.

A source MLP b_s minimizes benefit BCE to estimate $P(H_s = 1 \mid F_s)$. With source-training standardization $\mathcal{N}_s$, deployment uses $b_s(\mathcal{N}_s F_t)$ without changing parameters, scaler, or source-selected checkpoint. Benefit AUC is evaluated against H_t; reported seed averages average metrics, not predictions.

Benefit interface and distinct transfer questions. We call the common coordinate construction $m \mapsto F_m$ a *benefit interface*: every recognizer instantiates the same label-free response relations from its own core/context/full outputs, without a shared embedding basis. Five questions must be kept separate: *training objective* (what source risk supervises the router), *evidence alignment* (whose F_m best orders H_t under a fixed decoder), *decoder portability* (whether source-fitted b_s remains useful on F_t), *conditional usefulness* (whether structural relations improve prediction beyond a specified confidence family), and *decision value* (whether a score improves the downstream routing criterion). The Direct-NLL comparison isolates the first; the origin swap isolates the second; the donor and (Q, S) experiments separate the third and fourth. No one property is treated as implying the others.

Evidence-origin control. For matched objects and contexts, set

$$M_{t,m} = \mathrm{AUC}(H_t, b_s(\mathcal{N}_s F_m)), \qquad \Delta_{st} = M_{t,t} - M_{t,s}. \tag{2}$$

The decoder and target event are fixed while the evidence-producing recognizer changes. Original crops, masks, and donor assignments are matched; native normalization and neutral fills remain recognizer-specific. A positive Δ_{st} excludes the narrow explanation that this source map merely reuses its own scene-benefit score. It does not exclude every possible source-only predictor. We call $M_{t,t} > \max_{m \neq t} M_{t,m}$ *self-origin dominance*. This finite-model property neither identifies exclusive information nor guarantees a universal diagonal. Source realized advantage, which knows Y, is an additional diagnostic rather than a deployable baseline or an AUC upper bound.

Available information versus a fixed boundary. We also fit $g_s(F_s)$, $g_t(F_t)$, and $g_{st}(F_s, F_t)$ to the same target event using matched held-out folds and budgets. Better target or joint performance demonstrates additional predictive information accessible to those learners, not an unrestricted information-theoretic identification. These fits use target labels and are never deployment routers.

Residual-only failure is not conditional sufficiency. Let Q denote the tested confidence family and S the remaining structural relations. For any fixed residualizer a, (Q, S) and $(Q, S - a(Q))$ determine each other. Dropping Q can nevertheless erase interactions: independent uniform signs Q, S with $H = \mathbf{1}[QS > 0]$ have uninformative individual coordinates but a perfectly informative pair. Thus weak residual-only transfer cannot establish $H \perp S \mid Q$. For binary log loss, a nested-feature comparison satisfies

$$L(q_Q) - L(q_{Q,S}) = I(H; S \mid Q) + \epsilon_Q - \epsilon_{Q,S}, \tag{3}$$

where each ϵ is the fitted predictor's population excess log loss over its Bayes optimum. Appendix B.6 proves the identities and gives the counterexample. The familiar Bayes-risk identity motivates our supervision comparison; finite fitted risk differences are not CMI estimates.

A sufficient condition for portable ordering. On finite source supports, suppose benefit-conditioned densities satisfy

$$f_{m,y}(z) = a_m(z) r_y(z) / Z_{m,y}, \qquad a_m(z), r_y(z) > 0. \tag{4}$$

The recognizer-specific nuisance tilt is shared between benefit classes, giving log-likelihood ratio $h(z) + c_m$ with $h = \log(r_1/r_0)$. Within-recognizer pairwise logistic risk then has minimizers $h + c$ on connected comparison support. A target whose likelihood ratio preserves this ordering has Bayes-optimal ranking. Appendix C states support conditions and a drift bound. Pooled BCE can reverse an otherwise compatible order because source membership varies with z. This motivates, but does not guarantee success of, the within-recognizer intervention. None of these population assumptions follows from high observed AUC.

4 Experimental Design

Frozen sources and numerical consistency. Source records come from CLIP ViT-B/16, SigLIP ViT-B/16, EVA02 ViT-B/16, and DINOv3 ConvNeXt-L (Radford et al., 2021; Zhai et al., 2023; Sun et al., 2023; Siméoni et al., 2025). Each uses the same 5,000 ImageNet training and 1,000 disjoint validation images, with genuine and different-class donor boundaries around a 192×192 core. The 14–32–32–1 GELU MLP uses 100 epochs, AdamW .001 learning rate, .0001 decay, and seeds 17/23/41. Hard checkpoints minimize source-validation benefit BCE. The matched Direct-NLL baseline instead backpropagates the routed classification loss of the posterior mixture $(1-r)p_c + rp_f$ through the same gate and selects by source-validation routed NLL; it receives no benefit-event target. Simpler controls use their declared source-only criteria.

Current events use cancellation-resistant float64 log-domain arithmetic. An earlier FP32 correction changed ConvNeXt-source Hard labels while leaving the VLM Hard fits bit-identical. A subsequent independent audit found that subtractive *superclass* log probabilities also censored tiny ImageNet-9 benefits: 468 DINO events changed, all with $|A| < 8 \times 10^{-15}$. Complement-based evaluation of the same probability sum corrects this without changing any frozen expert or router. All affected ImageNet-9 tables are refreshed; both numerical histories and non-negligible-benefit sensitivities are retained (Appendix A.5).

Recognizers and supervised heads. Targets add supervised ResNet50 and EfficientNet-B0 (He et al., 2016; Tan & Le, 2019), neither used in source router training. DINOv3 ConvNeXt is a convolutional inference architecture with distilled pretraining, not independent supervised pretraining. On COCO/Pet/VOC, CNNs share normalized-feature/class-mean/normalized-prototype head construction with cosine scale 50; VLMs use fixed text templates. Prototype fitting uses training labels only: 1,600 COCO objects (20/class), 740 Pet images (20/breed), and 220 VOC crops (11/class). Evaluation labels fit neither heads nor deployment routers.

Natural and intervened contexts. The controlled COCO cohort contains 1,457 objects from 1,116 val2017 images across 80 categories (Lin et al., 2014). Object-centered crops map once to 224×224.

Cores preserve target-mask pixels; donor images replace only their complement. Same-class, target-absent high-PMI, and target-absent low-PMI donors are drawn using three fixed train2017 manifests. Genuine views retain the original crop. A separate natural cohort excludes all controlled-cohort images, giving 1,346 objects in 961 photographs and 63 represented categories; heads remain 80-way.

Oxford-IIIT Pet uses its official trainval/test lists and foreground/background trimaps (Parkhi et al., 2012). Empty-support masks are excluded by annotation validity before completed evaluation, leaving 3,673 trainval and 3,662 test photographs. All six recognizers share the test set. This changes mask, crop, category, and head statistics jointly; it is not an isolated dataset effect.

Separate external confirmations. The internally prespecified ImageNet-9 confirmation, recorded before its benefit-transfer outcomes, uses all 4,050 aligned validation foreground IDs (Xiao et al., 2021). Supplied `only_fg`, `no_fg`, and `original` images provide core, context-only, and full views. Evidence remains 1,000-way, but benefit compares probability summed over the correct nine-superclass group without renormalizing away unmapped mass. Supervised CNNs retain native pretrained 1,000-way heads. Five non-source contrasts use paired 99% marginal intervals, adjusted to nominal 95% family coverage.

VOC2012 was separately specified before its recognizer outcomes (Everingham et al., 2010). An annotation-only feasibility check failed the initial 20/class prototype plan. Before inference, a retained v2 amendment changed only head fitting to disjoint Main/train boxes with the mechanically feasible balanced cap 11/class. All 952 eligible Segmentation/val instances from 453 images are evaluated. Its original stricter interval/success rules remain unchanged; this is a pre-outcome amended protocol, not an external preregistration.

Diagnostics, holdouts, and uncertainty. Source/target/joint diagnostics use fixed-capacity histogram boosting and inner-validated logistic replication. Original-image folds keep repeated objects and donor views together. Strict donor folds additionally separate donor identities. A corrected COCO class-held-out control retains the original held-out classes but purges all overlapping training-image identities; the earlier class-only folds were not image-disjoint. Bootstrap intervals condition on fixed fits and manifests, not unknown pretraining populations or full CV retraining.

The post-P7 conditional-increment diagnostic fixes Q to 13 confidence coordinates: three margins, three entropies, their full–core changes, three agreements, and core/full maximum probabilities. S is three JS divergences and two cosines. Both Q and (Q, S) use 150 boosting iterations, 15 leaves, learning rate .05, minimum leaf size 20, $L_2 = 1$, and no early stopping. A fixed three-fold image assignment reserves fold zero as test; donor data requires both endpoints in the test fold and neither in fitting. Two additional CLIP-training-only learners are evaluated on those same test rows. Five condition-mean target-supervised BCE contrasts form one family with 5,000 paired image-cluster draws and 99% marginal intervals; donor uncertainty also has a two-endpoint resampling sensitivity. No new target result selects features, hyperparameters, or a deployment method.

Matched-object regime diagnostics compare the same 1,457 objects with genuine/donor context, including a regime-aware decoder versus a blind decoder with separate regime intercepts. Source-only interventions include within-recognizer pairing WR4, global pairing GR4, pooled BCE4, balanced BBCE4, and predefined WR3. All candidates and unsuccessful source-only selection are retained; protocol details are in the supplement.

5 Results

5.1 Direct task optimization does not recover the portable benefit ordering

The strongest "nothing special is needed" baseline trains the same CLIP-source MLP directly on routed NLL. It sees the same frozen experts and 14 response features as Hard-benefit supervision, uses the same source images, optimization budget, initialization seeds, and matched minibatch order, and selects its checkpoint by source routed NLL. Yet on natural COCO its benefit AUC is only .423–.614 across the six recognizers, versus .644–.822 for Hard (Table 1; Fig. 1A). The gap is therefore not caused by target fitting, model capacity, or a richer observation available to Hard.

Table 1: **Direct task-loss optimization does not recover the portable benefit ordering.** All learned rules use the fixed CLIP source. Direct-NLL and Hard use the same 14-D MLP, source records, initialization protocol, and optimization budget, but are trained for routed NLL and benefit BCE respectively. Margin/MSP use target-model outputs; logistic regression uses the same 14 response features.

Target	Direct NLL	MSP	Margin	Logistic	Hard
CLIP	.614	.809	**.826**	.784	.793
SigLIP	.530	.797	**.853**	.780	.788
EVA02	.535	.811	**.853**	.785	.785
DINO-CN	.423	.651	.785	.787	**.822**
ResNet50	.518	.666	**.764**	.686	.725
EffNet-B0	.479	.632	**.724**	.626	.644

Table 2: **Same-image, same-target-label evidence-origin control.** The identical CLIP router and source scaler receive target evidence F_t or source evidence F_s. Every score predicts H_t. Source advantage A_s sees the label and is diagnostic only. Intervals use 2,000 paired image-cluster draws.

Dataset	Target	$b_s(F_t)$	$b_s(F_s)$	A_s	Paired difference [95% CI]
COCO	DINO-CN	0.822	0.624	0.734	$+0.198$ [0.163,0.235]
COCO	ResNet50	0.725	0.603	0.656	$+0.121$ [0.084,0.161]
COCO	EffNet-B0	0.644	0.576	0.639	$+0.067$ [0.028,0.108]
Pet	DINO-CN	0.912	0.520	0.469	$+0.392$ [0.369,0.415]
Pet	ResNet50	0.915	0.509	0.536	$+0.406$ [0.386,0.427]
Pet	EffNet-B0	0.898	0.503	0.505	$+0.395$ [0.375,0.415]

This is not only an AUC/objective mismatch. On the two supervised CNNs that never participate in source-router training, Hard also attains lower routed NLL than Direct-NLL: 2.1693 versus 2.5285 on ResNet50 and 2.6351 versus 3.0543 on EfficientNet. We do *not* infer that Hard is universally task-optimal: calibrated margin, logistic, or fixed mixtures can be stronger in particular regimes, and Direct-NLL can be competitive on its source distribution. The narrower conclusion is the one needed here: directly optimizing source routing loss does not recover the reusable target-benefit ordering.

5.2 The target recognizer's own responses instantiate the reusable interface

The complete natural COCO/Pet source–target matrices are reported in Appendix Table 9. Three VLM sources transfer to natural COCO CNNs with AUC ranges .821–.844 on ConvNeXt, .706–.739 on ResNet50, and .640–.682 on EfficientNet; the corresponding Pet ranges are .912–.947, .907–.923, and .876–.901. Corrected ConvNeXt-source transfer is weak on COCO but stronger on Pet, so architecture alone does not explain portability.

The sharper control fixes both the decoder and the target event and changes only which recognizer instantiates the response interface. The frozen CLIP decoder gains approximately $+.198/+.121/+.067$ AUC from target rather than matched CLIP responses on the three COCO CNNs, and $+.392/+.406/+.395$ on Pet (Table 2). Under the full 14-D interface, self evidence is the best of six origins in all 20 non-source target–dataset cases across COCO, Pet, ImageNet-9, and VOC2012 (Fig. 1B). Nominal self-minus-selected-best intervals are positive in 18; EfficientNet on COCO and VOC remains unresolved. This is a descriptive multi-model regularity, not 20 independent replications or a simultaneous 20-case significance claim.

The separately frozen confirmation families preserve the same direction. Precision-refreshed ImageNet-9 has five positive adjusted target-over-source contrasts, mean $+.238$ with 99% interval [.223,.252] (Table 3); VOC v2 gives mean $+.126$ [.091,.164], with four of five positive prespecified family intervals. Margin nevertheless wins on several COCO/Pet/ImageNet-9 targets. Self-origin dominance therefore identifies the response *origin* that best matches target-relative benefit under a fixed decoder; it is not a claim of universal superiority over confidence.

Table 3: **Internally prespecified ImageNet-9 confirmation on 4,050 held-out foreground IDs.** Router, scaler, heads and observations are unchanged. The nine-superclass event now uses cancellation-resistant complement arithmetic (Appendix A.5). Paired 99% marginal intervals across five targets give nominal 95% Bonferroni family coverage. ResNet50 and EfficientNet retain their native 1,000-way heads.

Target	$b_s(F_t)$	$b_s(F_s)$	Target margin	Δ_{t-s} [99% CI]
SigLIP	.727	.574	.728	$+.154$ [.126,.181]
EVA02	.737	.563	.742	$+.174$ [.146,.200]
DINO-CN	.812	.501	.843	$+.311$ [.282,.341]
ResNet50	.770	.475	.677	$+.295$ [.269,.321]
EffNet-B0	.746	.490	.757	$+.255$ [.228,.284]

Table 4: **Information availability, not deployment performance.** Capacity-matched nonlinear predictors are fitted using target labels under identical five-fold original-image splits. Columns are out-of-fold benefit AUC. $L_t - L_s$ is benefit BCE, not classification NLL; intervals condition on fitted folds.

Data	Target	$g(F_s)$	$g(F_t)$	$g(F_s, F_t)$	$L_t - L_s$ [95% CI]
COCO natural	DINO-CN	0.696	0.847	0.865	-0.168 [-0.215,-0.124]
COCO natural	ResNet50	0.657	0.765	0.787	-0.117 [-0.157,-0.077]
COCO natural	EffNet-B0	0.635	0.746	0.761	-0.088 [-0.125,-0.053]
COCO donor	DINO-CN	0.777	0.893	0.909	-0.118 [-0.133,-0.103]
COCO donor	ResNet50	0.764	0.842	0.861	-0.066 [-0.080,-0.053]
COCO donor	EffNet-B0	0.727	0.788	0.819	-0.054 [-0.069,-0.041]
Pet	DINO-CN	0.554	0.959	0.957	-0.352 [-0.368,-0.336]
Pet	ResNet50	0.589	0.931	0.926	-0.310 [-0.327,-0.293]
Pet	EffNet-B0	0.517	0.926	0.922	-0.350 [-0.365,-0.334]

5.3 Conditional usefulness does not imply portable interpretation

Decoder portability is stricter than evidence alignment. On donor composites, the frozen target-versus-source origin contrast falls to about $+.012/-.033/-.065$ on the CNNs; the latter two intervals are negative. Yet capacity-matched target-supervised diagnostics achieve .893/.842/.788 AUC from target evidence versus .777/.764/.727 from source evidence, with joint evidence improving further (Table 4). Thus the target responses remain predictively useful where the source boundary fails.

Tighter holdouts preserve that distinction. Target+donor-disjoint fitting retains target-minus-source AUC $+.125/+.066/+.071$, all with positive adjusted intervals; corrected COCO class-and-image-disjoint fitting retains $+.167/+.134/+.120$ [.120,.215]/[.088,.180]/[.074,.167]. Pet breed-held-out increments are $+.430/+.335/+.419$. These controls weaken image, class, and donor memorization explanations without identifying unrestricted information.

Confidence is a major carrier of the interface: deleting all margins reduces self-origin maxima from 15/15 to 11/15 on the original three-dataset grid, and a confidence-residualized structural decoder is self-first in only 3/15 cases. But residual-only failure cannot establish conditional sufficiency because the conditioning variables are absent. The matched (Q, S) versus Q comparison addresses the actual conditional question. Average held-out benefit BCE improves by .0142 [.0056,.0234] on ImageNet-9 and .0238 [.0104,.0377] on donor COCO; the three other condition intervals include zero (Table 5). These are finite-learner predictive gains, not CMI estimates or causal mediation.

ImageNet-9 gives the decisive reversal. Adding the same structural coordinates to a CLIP-training-only learner *increases* target BCE by .0491 [.0347,.0645] on exactly the observations where target-supervised augmentation reduces BCE by .0142 [.0056,.0234] (Fig. 1C). COCO also gives a negative source-only increment; Pet and donors are positive and VOC is unresolved. The interpretation of useful structure is therefore not invariant across recognizers: conditional usefulness can be real while its source-trained decoder is wrong.

Table 5: **Conditional usefulness need not transfer.** Positive entries are benefit BCE reductions $L(Q) - L(Q, S)$ averaged over the three CNNs, where Q is the complete 13-D confidence family and S comprises three JS divergences and two feature cosines. Target-supervised columns use fixed-budget learners fitted only on disjoint target training images; the secondary source-only column uses two learners fitted on CLIP training records. Both compare identical test examples. Paired 99% marginal intervals use 5,000 image-cluster draws; the five primary target-supervised contrasts have nominal 95% Bonferroni family coverage. Intervals are conditional on the fixed fits. They are not estimates of unrestricted conditional mutual information. Donor fit/test sets also share no donor identity.

Data	Test rows / images	Target-supervised	Source-only (secondary)
COCO natural	452 / 321	.0063 [-.0236,.0366]	−.1300 [-.2147,-.0491]
Oxford-IIIT Pet	1221 / 1221	.0018 [-.0063,.0103]	.0220 [.0173,.0265]
ImageNet-9	1350 / 1350	.0142 [.0056,.0234]	−.0491 [-.0645,-.0347]
VOC2012	299 / 151	.0065 [-.0280,.0384]	−.0402 [-.1272,.0600]
COCO donor	1454 / 367	.0238 [.0104,.0377]	.2770 [.2021,.3479]

Table 6: **Matched objects reveal a change in decoding.** The same 1,457 objects receive genuine or nine donor contexts. P_+ is benefit probability conditional on an increased margin. The final comparison uses target-supervised, image-held-out predictors: regime-specific decoding versus a blind decoder with per-regime intercept correction.

Target	P_+(genuine)	P_+(donor)	$\Delta_G - \Delta_D$ [95% CI]	Shape–prior BCE [95% CI]
DINO-CN	0.703	0.295	+0.188 [0.155,0.218]	−0.005 [-0.008,-0.001]
ResNet50	0.768	0.289	+0.166 [0.127,0.200]	−0.005 [-0.009,0.000]
EffNet-B0	0.781	0.381	+0.121 [0.081,0.161]	−0.013 [-0.019,-0.007]

5.4 Boundary evidence and failed universal repairs

The donor reversal is not explained only by comparing different objects. On the same 1,457 instances, genuine minus donor context preserves a positive change in origin contrast of $+.188/+.166/+.121$. An increased full-view margin accompanies benefit in .703/.768/.781 of genuine cases but only .295/.289/.381 of donor cases. Relative to a blind predictor with regime-specific intercepts, a regime-aware finite learner improves BCE by .0047 on ConvNeXt and .0132 on EfficientNet; ResNet50 is unresolved (Table 6). These controls support a change in finite-learner decoding, not failure of every calibration family.

High benefit AUC also need not improve classification decisions. Pet core/full disagreements contain only 86/203/280 CNN examples; the CLIP map's AUC there is .516/.529/.547 versus .912/.915/.898 on complete cohorts, and margin is stronger on those disagreement subsets. Likewise, the attempted source-only repairs do not produce a universal gate: WR4 reaches .601/.540 on the two unseen natural COCO CNNs versus .603/.542 for matched BCE4 and .764/.724 for margin; source-only leave-one-recognizer-out selection chooses BBCE4 and does not solve transfer. The retained Hard+margin combiner beats margin at the adjusted dataset mean only on VOC and is below Hard alone at every dataset mean. These negative results bound the central claim rather than compete with it: shared coordinates can transfer without a universally portable decoder or decision rule.

6 Discussion and Conclusion

The results reduce to three empirical statements. *First*, direct source task-loss optimization does not recover the portable target-benefit ordering, even under matched architecture, evidence, data, and optimization. *Second*, when a frozen decoder does transfer, the strongest evidence is consistently produced by the target recognizer's own response interface rather than by carrying the source model's scene score. *Third*, information and interpretation separate: structural relations can add predictive value beyond confidence on a target while the same source-learned augmentation moves in the wrong direction. Together these statements identify the reusable object more precisely than "a

universal router": shared response coordinates can expose recognizer-relative benefit, but their decision boundary is only conditionally portable.

This framing also explains why the negative results matter. Donor contexts can break a frozen decoder without eliminating target-relative information; multi-source pooling, within-recognizer ranking, residual-only transfer, and Hard+margin combination do not supply a universal repair. The scientific distinction is therefore among *benefit supervision, evidence origin, conditional predictive usefulness, portable interpretation, and decision value*. The same distinction may arise whenever the realized value of an optional observation or intervention changes with the base model, although our empirical evidence is limited to visual context and does not establish cross-task transfer.

The scope is finite. Masks are supplied annotations; COCO/Pet/VOC CNN heads are prototypes; DINOv3 is distilled, and supervised CNNs share ImageNet pretraining. VOC has a disclosed pre-inference feasibility amendment and small rare classes. Protocols were internally fixed at different stages, not externally preregistered. Numerical and class-split corrections are disclosed with old artifacts retained. New held-out intervals condition on fixed fits and manifests, and finite-learner gains do not identify CMI. No cross-task, confidence-independent, causal-mediation, or universal routing claim follows. The compatibility theory states sufficient conditions; the experiments test distinctions and failure modes rather than verify every population assumption.

Reproducibility Statement

The supplement records protocol chronology, source/head/evaluation separation, both numerical corrections, the COCO class/image purge, and all unsuccessful interventions. Frozen caches, source states, split identities, per-example scores, and versioned evidence hashes support reconstruction. Independent calculations refit the new conditional learners, rebuild their features and events, reproduce all 5,000-draw primary intervals, and check corrected class-disjoint intervals and ImageNet-9 primary results. These computational checks do not replace human source, theorem, anonymity, and release-permission review.

AI Use Statement

Generative AI tools were used only to provide limited assistance with method implementation and result interpretation. Experimental design and mathematical derivations and proofs were conducted independently by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014.

Marzieh Ajirak, Oded Bein, Ellen Rose Bowen, Dora Kanellopoulos, Avital Falk, Faith Gunning, Nili Solomonov, and Logan Grosenick. Learning to route: Per-sample adaptive routing for multimodal multitask prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2304.

Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7412–7420, 2019. doi: 10.1109/CVPR.2019.00759.

Mustafa Bilgic and Lise Getoor. VOILA: Efficient feature-value acquisition for classification. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pp. 1225–1230, 2007.

Qi Cao, Shuhao Zhang, Ruizhe Zhou, Ruiyi Zhang, Peijia Qin, and Pengtao Xie. Models under SCOPE: Scalable and controllable routing via pre-hoc reasoning. *arXiv preprint arXiv:2601.22323*, 2026. doi: 10.48550/arXiv.2601.22323.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6424–6447. PMLR, 2023.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, 2026.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations*, 2024.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 939–945, 2015.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20707–20717, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Preethi Lahoti, Krishna Gummadi, and Gerhard Weikum. Responsible model deployment via model-agnostic uncertainty learning. *Machine Learning*, 112:939–970, 2023. doi: 10.1007/s10994-022-06248-y.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43035–43066. PMLR, 2025.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Joshua Strong, Pramit Saha, Yasin Ibrahim, Cheng Ouyang, and J. Alison Noble. Identity-free deferral for unseen experts. In *International Conference on Learning Representations*, 2026.

Mukund Sudarshan, Aahlad Puli, Wesley Tansey, and Rajesh Ranganath. DIET: Conditional independence testing with marginal dependence measures of residual information. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10343–10367. PMLR, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Clovis Varangot-Reille, Christophe Bouvard, and Antoine Gourru. Generalising LLM routing using past performance retrieval: A few-shot router is sufficient. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 304–319. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-srw.22.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11415–11434. PMLR, 2023.

Chenxu Wang, Hao Li, Yiqun Zhang, Linyao Chen, Jianhao Chen, Ping Jian, Qiaosheng Zhang, and Shuyue Hu. ICL-Router: In-context learned model representations for LLM routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33413–33421, 2026. doi: 10.1609/aaai.v40i39.40628.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Cheng Yan, Wuyang Zhang, Zhiyuan Ning, Fan Xu, Ziyang Tao, Lu Zhang, Bing Yin, and Yanyong Zhang. Breaking model lock-in: Cost-efficient zero-shot LLM routing via a universal latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36483–36490, 2026. doi: 10.1609/aaai.v40i43.40970.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12985–12994, 2020.

A Protocol, Data Separation, and Numerical Audit

This supplement distinguishes original-protocol single-source fits, new source-only ranking interventions, and target-supervised diagnostics. They answer different questions and cannot substitute for one another in a table. The controlled COCO experiments informed the research design before the new interventions. The second-dataset Pet outcomes were not used to choose a router or ranking algorithm. Internal timestamped plans are retained; they are not external preregistrations.

A.1 Source records and original-protocol fitting

All source datasets contain the same 5,000 ImageNet training and 1,000 disjoint validation images balanced across 1,000 classes. Genuine and different-class donor views share a fixed 192×192 core inside a 224×224 canvas. CLIP, SigLIP, and EVA use frozen text heads. DINOv3 ConvNeXt-L uses a 1,000-way normalized-prototype head fitted on five separate training images per class. Head-fitting, router-training, and router-validation identities are disjoint. These are frozen experts; routing experiments do not train or fine-tune their backbone parameters.

The evidence vector is identical in every architecture: three JS divergences between predictive distributions, three logit top-two margins, three entropies, three top-1 agreement indicators, and two within-model feature cosines. The feature order is recorded in the code and cached arrays. The MLP is Linear(14,32)–GELU–Linear(32,32)–GELU–Linear(32,1), with a sigmoid only when a score is used as a probability or mixture. All feature standardization comes from source training; target data never changes it.

Original single-source fits use seeds 17/23/41, 100 epochs, AdamW learning rate .001, weight decay .0001, and batches of 512 original images with both contexts. Initialization and batch order are matched across Hard-benefit and direct-NLL objectives. H_{BCE} minimizes source-validation benefit BCE over all 1,000 validation images. A second checkpoint from the same Hard trajectory minimizes routed NLL, as does the direct-NLL checkpoint. Linear neural controls use the same budget. A separate logistic classifier with all 14 features selects $C \in \{.01, .1, 1, 10, 100\}$ by source benefit BCE. Scalar margin/MSP/entropy controls use source-only logistic calibration when a probability is required. Raw-score AUC is not conflated with calibrated-gate loss.

A.2 Stable realized benefit

For fixed cached logits, correct-class log probability is evaluated using

$$\log p_y = -\operatorname{logaddexp}\big(0, \operatorname{logsumexp}_{k \neq y}(z_k - z_y)\big)\,. \tag{5}$$

The log-domain arithmetic uses float64 and avoids subtracting nearly equal large numbers when the correct-class probability is close to one. Original float32 log-softmax censored positive ConvNeXt benefits near zero. On its 10,000 source training views, the corrected sign changes for 1,473 genuine and 177 wrong-context records. On its 2,000 validation views, 280 genuine and 23 wrong records change. No VLM-source Hard labels change.

The original all-validation source experiment was therefore rerun with the corrected arithmetic, without changing features, images, budgets, optimization, or checkpoint criteria. All three VLM Hard networks and selected checkpoints are bit-identical to the old fits. Direct-NLL weights differ only at small floating-point scale with unchanged selected epochs. ConvNeXt Hard weights and some checkpoints change substantively. The original artifacts are retained, but every current single-source matrix uses the corrected fits and stable target events. The weak old ConvNeXt-source row is not reused as evidence for a numerical-invariant architecture law.

Stable log-probabilities do not recover precision absent from the backbone's float32 logits. Moreover, strict positive benefit is not the same as a practically meaningful loss reduction. The fixed $|A_t| > 10^{-6}$ sensitivity and prediction-disagreement analyses make those distinctions explicit. Identical experts have exactly zero advantage, and a neutral condition with no positive benefit events has undefined AUC rather than chance AUC.

Table 7: **Analysis chronology and evidential status.** "Frozen" means the listed analysis choices were recorded before the corresponding new outcomes were inspected; none of these internal plans is presented as an external preregistration. Negative and boundary results remain in the manuscript.

Analysis	What was already known when specified	Role in the paper
COCO/Pet mainline	COCO informed research development; Pet was added after the initial COCO program	Discovery/replication evidence; not confirmatory preregistration
ImageNet-9 family	COCO/Pet results and source router were fixed; ImageNet-9 benefit-transfer outcomes were unseen	Internally prospective confirmation with a prespecified five-target success rule; later strict-sign arithmetic correction retained the rule and frozen model outputs
P1–P5 robustness	COCO/Pet/ImageNet-9 transfer and U3/U4 results were known	Outcome-frozen attacks on donor reuse, class prevalence, and evidence-origin alternatives
P6 feature deletion	P1–P5 outcomes were known	Outcome-frozen test of direct confidence-family dependence
VOC2012 v2	Earlier datasets/P1–P6 were known; no VOC recognizer output had been generated. V1 failed at annotation-only feasibility	Non-ImageNet-derived pre-outcome external confirmation; only CNN-head source/cap was amended before inference
Hard+margin	P6 and VOC outcomes were known	Prespecified secondary operational-complementarity test; failed and retained
P7 residualization	P1–P6 outcomes were known	Outcome-frozen source-only residual portability test; failure retained
P8 conditional increment	All prior transfer, VOC, P6/P7, and Hard+margin outcomes were known	Post-P7 diagnostic fixed before its new fits; distinguishes residual-only failure from conditional predictive usefulness

A.3 COCO CROPS AND DONOR CONSTRUCTION

The controlled cohort comprises 1,457 val2017 annotation IDs from 1,116 original images in all 80 categories. At most 20 eligible targets per class are selected by the fixed manifest. The object-centered crop is mapped once to 224 pixels. Masks are rasterized from instance annotations and conservatively preserve pixels touched by foreground coverage under resizing. Boolean selection preserves target and donor pixels exactly; it is not alpha compositing.

For source crop P, donor image canvas D, binary target support M, and recognizer-specific neutral fill μ, the views are

$$X = M \odot P + (1-M) \odot \mu, \qquad V_D = M \odot P + (1-M) \odot D, \qquad C_D = M \odot \mu + (1-M) \odot D.$$

A per-sample check verifies object-pixel equality, donor-pixel equality, and invariant recovered core. Source fill and input normalization follow the recognizer's declared protocol. Same-class donors may contain another instance of the category. High/low-PMI donors exclude the target category and use co-occurrence statistics from train2017; PMI is an operational proxy, not a human semantic judgment. Three manifests share target identities and are not counted as three independent collections of objects.

The original-photo COCO cohort excludes every controlled-cohort source image before fixed hash selection. It contains 1,346 instances in 961 images and 63 represented categories; the classification head remains 80-way. Visible co-object PMI strata use train2017 reference distributions and exclude the target category. Stratum counts are 138 atypical, 358 middle, 140 typical, and 710 with no scored co-object. Full views remain original crop pixels; the core and context-only views remain artificial ablations. This is natural scene variation, not a measurement of longitudinal deployment drift.

A.4 OXFORD-IIIT PET CONFIRMATION AND EXCLUSIONS

The official images and annotations archives were downloaded from the Oxford distribution and validated against the published resource checksums. We use the official trainval/test annotation lists. All photographs are decoded and all trimaps are checked before evaluation. The original lists contain 3,680 trainval and 3,669 test images. Seven masks in each split have only background label 2, so no

Table 8: Complete corrected donor transfer: pooled AUC / equal-condition macro AUC. All rows use original all-validation single-source fits, not the new rank-objective fits.

Source / target	CLIP	SigLIP	EVA02	DINO-CN	ResNet50	EffNet-B0
CLIP	0.823/0.791	0.827/0.788	0.831/0.782	0.710/0.668	0.656/0.621	0.591/0.562
SigLIP	0.799/0.759	0.830/0.785	0.837/0.779	0.815/0.779	0.737/0.669	0.659/0.584
EVA02	0.799/0.767	0.846/0.810	0.840/0.797	0.771/0.738	0.720/0.677	0.655/0.605
DINO-CN	0.550/0.556	0.637/0.639	0.573/0.566	0.767/0.740	0.644/0.653	0.568/0.552

Table 9: **Complete natural-image transfer on two datasets.** Rows are fixed ImageNet-trained benefit routers; columns are target recognizers. Entries average per-fit AUC over seeds 17/23/41. All single-source rows use the original all-validation protocol with stable benefit arithmetic. No router is fitted on either target dataset; CNN heads use training-split prototypes.

Source / target	CLIP	SigLIP	EVA02	DINO-CN	ResNet50	EffNet-B0
COCO original photographs						
CLIP	0.793	0.788	0.785	0.822	0.725	0.644
SigLIP	0.780	0.783	0.782	0.821	0.706	0.640
EVA02	0.799	0.837	0.810	0.844	0.739	0.682
DINO-CN	0.459	0.583	0.490	0.673	0.577	0.512
Oxford-IIIT Pet original photographs						
CLIP	0.911	0.943	0.920	0.912	0.915	0.898
SigLIP	0.900	0.938	0.910	0.947	0.923	0.901
EVA02	0.915	0.945	0.923	0.917	0.907	0.876
DINO-CN	0.586	0.790	0.772	0.963	0.913	0.895

object-preserving core can be formed; those records are excluded by annotation validity, not model performance. The final sets contain 3,673 trainval and 3,662 test images.

The excluded training images are saint_bernard_108, Egyptian_Mau_162, Egyptian_Mau_165, Egyptian_Mau_196, leonberger_18, miniature_pinscher_14, and saint_bernard_15. The excluded test images are Egyptian_Mau_20, japanese_chin_199, keeshond_7, Persian_259, saint_bernard_60, saint_bernard_78, and wheaten_terrier_195. The pre-exclusion manifest and the failed initial cache attempt are retained; no result on a smaller completed subset is substituted for the final cohort.

The full photograph is resized to 224×224 once. Foreground label 1 and unknown-boundary label 3 are retained; positive bilinear support after mask resizing defines the core. Core and context-only are complementary ablations using model-specific neutral fill. All six recognizers use the same photographs and 37-class ordering. VLM breed prompts follow the existing fixed templates. CNN heads use the same hash-selected 20 valid trainval images per breed, 740 in total. Features are normalized before class averaging, and class prototypes are normalized again, with logit scale 50. No test image or label fits a head, router, calibrator, or source-selection rule. Pet provides fine-grained classification generality, not a new detection or grounding task.

Complete single-source transfer matrices. For completeness, Table 9 reports every corrected source–target direction on natural COCO and Pet. These matrices support the headline evidence-origin analysis but are not used to select the source decoder or any target-specific rule.

A.5 Internally prespecified ImageNet-9 / Backgrounds Challenge confirmation

This confirmation was specified in a dated internal extension plan before its ImageNet-9 benefit-transfer outcomes were inspected. The source router is the unchanged CLIP Hard-benefit model used in the COCO evidence-origin study; no ImageNet-9 record changes its weights, source standardization, checkpoint, or temperature. The source router was trained and validated on ImageNet training images, whereas the 4,050 Backgrounds Challenge foreground identities are validation images.

We align the benchmark's `original`, `only_fg`, and `no_fg` conditions by foreground ImageNet ID. Alignment and the official ImageNet-to-ImageNet-9 map are checked for every record: all 4,050 IDs are unique, mapped, and have identical nine-class targets in all three conditions. We define the

core from `only_fg`, context-only evidence from `no_fg`, and the full view from the uncomposited `original` image. Unlike the central-square source protocol, these views come directly from the benchmark's supplied foreground/background interventions.

The router still observes the same 14 label-free relations among the recognizer's 1,000-way core, context-only, and full outputs. The benefit event is instead evaluated at the nine-superclass task level. For ImageNet-to-9 mapping M and superclass y_9,

$$A_t^{(9)} = \log \sum_{k:M(k)=y_9} p_{t,f}(k) - \log \sum_{k:M(k)=y_9} p_{t,c}(k), \qquad H_t^{(9)} = \mathbf{1}[A_t^{(9)} > 0].$$

Unmapped ImageNet probability mass remains outside the nine superclass sums and is not renormalized. This is deliberate: the benefit event asks whether the recognizer moves probability mass toward the correct superclass, while the evidence remains independent of y_9.

CLIP, SigLIP, and EVA retain the same 1,000-way prompt-ensemble heads used in source experiments. DINOv3 ConvNeXt uses the already fitted 1,000-way normalized prototype head. ResNet50 and EfficientNet-B0 deliberately retain their native pretrained ImageNet classifiers in this benchmark rather than the COCO/Pet prototype construction. Thus the confirmation changes not only image cohort and target label aggregation but also the supervised CNN head protocol. It is not an architecture-only causal comparison.

The confirmatory family contains the five non-source targets. Each paired contrast compares the same fixed CLIP map on target evidence versus CLIP evidence from the same foreground ID and evaluates both against $H_t^{(9)}$. Two thousand foreground-ID bootstrap draws use identical multiplicities for both evidence origins. The five 99% marginal intervals provide nominal 95% Bonferroni family coverage; they are not 99% simultaneous intervals. Before outcomes, success required a positive mean contrast and at least three positive adjusted intervals. After the arithmetic correction below, all five remain positive: SigLIP $+.154$ [.126,.181], EVA02 $+.174$ [.146,.200], DINOv3 $+.311$ [.282,.341], ResNet50 $+.295$ [.269,.321], and EfficientNet $+.255$ [.228,.284]. The mean is $+.238$ with a separate 99% mean interval [.223,.252]. CLIP-as-target is an exact zero evidence-origin control. The result strengthens evidence for target-response portability; it does not imply that the transferred MLP beats target margin on every target.

Superclass numerical correction. A subsequent independent long-double audit found cancellation in the original subtraction of superclass and total log normalizers. For correct superclass B, the current implementation evaluates

$$\log P(B) = -\operatorname{logaddexp}(0, \operatorname{logsumexp}_{k \notin B} z_k - \operatorname{logsumexp}_{k \in B} z_k)\,.$$

This is the same prespecified, non-renormalized probability sum, with unmapped ImageNet classes included in the complement. It changes 468 of 4,050 DINO benefit signs, including 452 records previously equal to zero; all changed advantages have magnitude below 8×10^{-15}. DINO positive events increase from 2,286 to 2,736. The other five recognizers have zero sign changes. Current float64 events agree exactly with the separate long-double reconstruction. Frozen logits, router weights, scaler, folds and hyperparameters are unchanged. The correction refreshes all affected primary, all-origin, within-superclass, information, P6/P7, Hard+margin and P8 results in a new versioned directory; old results are retained. The DINO primary origin contrast changes from $+.304$ to $+.311$, and the five-target mean from $+.236$ to $+.238$. This repairs strict-sign arithmetic, not missing backbone precision or the practical relevance of vanishingly small benefits. The fixed $|A| > 10^{-6}$ sensitivity is retained with counts in the arithmetic audit. The original confirmation chronology and success rule remain unchanged.

The diagnostic table uses the same target event but permits target labels inside five-fold foreground-ID training. Its fixed-capacity nonlinear learner and inner-validated logistic replication both find much stronger target-evidence than source-evidence prediction on every non-source recognizer. Joint evidence does not improve over target-only evidence under the nonlinear learner in this benchmark; the corrected DINO logistic difference is small and descriptive. This differs from the COCO donor diagnostic and is therefore evidence against a generic claim that source and target evidence should always be concatenated. These fits diagnose accessible information only and are never used to select or calibrate the source-only deployment router.

Table 10: **ImageNet-9 target-supervised information diagnostic.** Entries are five-fold foreground-ID held-out AUC / benefit BCE. All current rows use the corrected stable-superclass event; folds, learners and budgets are unchanged. These fits use target labels and are not deployment routers.

Target	$g(F_s)$	$g(F_t)$	$g(F_s, F_t)$
Fixed-capacity nonlinear learner			
SigLIP	.619 / .581	.842 / .426	.838 / .432
EVA02	.642 / .602	.864 / .421	.858 / .429
DINO-CN	.553 / .638	.849 / .442	.845 / .449
ResNet50	.570 / .680	.966 / .232	.965 / .235
EffNet-B0	.595 / .630	.954 / .255	.952 / .259
Inner-validated logistic regression			
SigLIP	.655 / .563	.802 / .478	.802 / .476
EVA02	.662 / .591	.823 / .479	.823 / .477
DINO-CN	.563 / .625	.838 / .464	.838 / .464
ResNet50	.593 / .665	.958 / .290	.958 / .290
EffNet-B0	.608 / .614	.931 / .359	.929 / .362

The independent audit reconstructs the nine-class benefit events and all source/target fixed-map AUCs from retained logits, validates all 4,050 mapping identities, and recomputes the five paired intervals plus their mean using a separate scikit-learn implementation. Corrected retained evidence and hashes are in `experiments/results/in9_precision_refresh_20260916/`; the original `experiments/results/strong_accept_in9_20260915/` record is historical. Both identify the unchanged inference caches by SHA-256. Benchmark accuracy values used for this benefit study come from summed nine-superclass probabilities, not the official top-1-ImageNet-then-map Backgrounds Challenge accuracy convention, and are not compared to published benchmark leaderboards.

A.6 VOC2012 NON-IMAGENET-DERIVED EXTERNAL CONFIRMATION

After the robustness analyses above were fixed, we recorded a separate PASCAL VOC2012 confirmation before generating any VOC recognizer output (Everingham et al., 2010). The first protocol deliberately failed closed at an annotation-only feasibility screen: under the frozen mask-purity and expanded-crop-inside-image rules, several Segmentation/train classes had fewer than the requested 20 prototype examples. This failure and the original protocol hash are retained. Before any model inference, a v2 feasibility amendment changed *only* the CNN-head training source/cap: official VOC2012 Main/train detection boxes, which have zero image-ID overlap with Segmentation/val, and the mechanically determined balanced cap $K = \min(20, \min_c n_c) = 11$. Evaluation endpoints, the five-target family, source CLIP decoder, 14-D evidence, view construction, and success criterion were unchanged. Thus v2 is prospectively frozen with respect to model outcomes but is not presented as an unamended preregistration.

The evaluation cohort contains all 952 valid Segmentation/val target instances from 453 images after the frozen annotation/geometric rules. Full is the unmodified natural object-centered crop; core preserves the target-instance mask with recognizer-specific neutral fill outside; context-only is its complement. VLMs use fixed 20-class prompt heads. The three CNNs use 220 Main/train crops, 11 per class, with the same normalized-feature/class-mean/normalized-prototype construction and cosine scale 50. No validation object fits a head, router, scaler, temperature, or feature subset.

The primary family uses the unchanged precision-refreshed CLIP Hard-BCE decoder and compares $b(F_t)$ with matched $b(F_{\mathrm{CLIP}})$ against one target event. Two thousand image-cluster draws use seed 9262. Before outcomes, success required a positive five-target mean and at least three of five positive prespecified Bonferroni family-99% intervals. The result passes: the mean target-minus-CLIP evidence AUC is $+.126$ with 99% interval [.091,.164], and four of five family intervals are positive. EfficientNet is retained as the boundary ($+.013$ [$-.066, .100$]).

The secondary all-origin control again gives self evidence the largest point AUC for all five non-source targets. Selected-best 95% intervals are positive for SigLIP, EVA02, DINO-CN, and ResNet50; EfficientNet is again unresolved ($+.004$ [$-.051, .033$]). Equal-class macro target-minus-CLIP gaps

Table 11: **Prospectively frozen-with-respect-to-outcomes VOC2012 external confirmation.** The unchanged CLIP source decoder is evaluated on 952 valid Segmentation/val instances from 453 images. A blind annotation-only v1 feasibility failure (20 prototype examples/class unavailable) was retained before any VOC model output; v2 uses disjoint official Main/train boxes with the mechanically determined balanced cap $K = 11$ for the three CNN heads. Primary intervals are the prespecified Bonferroni family-99% intervals. Within-class gaps are equal-class macro target-minus-CLIP evidence AUC with nominal 95% class-bootstrap intervals.

Target	$b(F_t)$	$b(F_{\mathrm{CLIP}})$	Margin	$\Delta_{t-\mathrm{CLIP}}$ [family 99%]	Within-class Δ [95%]
SigLIP	.861	.683	.841	+.177 [.114,.232]	+.240 [.164,.324]
EVA02	.893	.737	.837	+.157 [.108,.207]	+.212 [.152,.274]
DINO-CN	.794	.606	.664	+.187 [.107,.269]	+.310 [.209,.393]
ResNet50	.727	.630	.719	+.097 [.022,.183]	+.177 [.053,.293]
EffNet-B0	.623	.609	.628	+.013 [-.066,.100]	+.097 [-.013,.208]

are positive in point estimate for all five and have positive 95% intervals for the same four targets. The transferred full decoder also exceeds raw target margin in point AUC on SigLIP/EVA/DINO/ResNet (.861/.893/.794/.727 versus .841/.837/.664/.719), while EfficientNet margin is slightly higher (.628 versus .623). Frozen P6 transfer shows the same confidence-family boundary: NoMargin is self-first on three of five non-source targets, and Structural evidence on two of five.

A separate audit reconstructs all three CNN prototypes from retained training features, rebuilds stable benefit events from logits, applies the source decoder directly from its state dictionaries/scaler, recomputes every primary AUC and family interval using sklearn weighted ROC AUC, and checks the all-origin/within-class/P6 point statistics. It exactly recovers the primary mean $+.126$, 99% interval [.091,.164], four positive family intervals, and the prespecified success decision. This is an implementation cross-check, not independent data collection.

A.7 Sampling units and evidence checks

Main transfer AUCs average metrics from the three source fits. Absolute scores are never ensembled across seeds before AUC unless an ensemble is explicitly named. Original-image-cluster bootstrap uses 2,000 draws, seed 915. All objects and all their repeated donor views share the same bootstrap multiplicity. Pet has one target object per image, so this becomes image bootstrap. Intervals are conditional on fixed source fits and manifests; they are not uncertainty intervals over all possible backbones or pretraining datasets. Three primary CNN origin contrasts within a natural dataset have Bonferroni 98.333% intervals; other large-grid comparisons are descriptive unless separately stated.

The extension audit reconstructs all transfer AUCs with scikit-learn from retained scores and events. An independent long-double log-probability formula checks stable event arithmetic. All six Pet identities and labels match the manifest. Three CNN prototype heads are recomputed from saved training features using NumPy. Each CNN is reloaded for fresh inference on 16 Pet images; under the same chunk-size contract, both core and full logits match exactly. The three primary Pet origin intervals are independently regenerated using 2,000 scikit-learn weighted-AUC draws. These are arithmetic and implementation checks, not independent data collection or independent laboratory replication.

B Information-Origin and Context-Regime Diagnostics

B.1 Target supervision is confined to diagnostic fits

For every diagnostic, CLIP provides source evidence and the selected target provides target evidence. Their ground-truth category, object, context, and target benefit event are aligned. Five outer folds partition original-image identities; every object and repeated context belonging to an image stays in one fold. Source, target, and joint predictors use identical train/test membership. Neither predictions nor feature normalization see held-out images during fitting.

Table 12: Target-supervised logistic replication, reporting AUC / benefit BCE. Regularization is chosen in an inner image-disjoint split. This is not the deployment algorithm.

Target	Data	F_s	F_t	(F_s, F_t)
DINO-CN	COCO natural	0.694/0.623	0.815/0.514	0.824/0.505
DINO-CN	COCO donor	0.751/0.464	0.822/0.405	0.846/0.387
DINO-CN	Pet	0.559/0.590	0.966/0.238	0.964/0.243
ResNet50	COCO natural	0.678/0.635	0.771/0.551	0.780/0.546
ResNet50	COCO donor	0.745/0.476	0.778/0.456	0.809/0.430
ResNet50	Pet	0.626/0.611	0.927/0.340	0.926/0.341
EffNet-B0	COCO natural	0.645/0.645	0.754/0.573	0.764/0.564
EffNet-B0	COCO donor	0.720/0.565	0.719/0.566	0.762/0.533
EffNet-B0	Pet	0.519/0.694	0.920/0.395	0.919/0.394

The nonlinear learner is histogram gradient boosting with 150 iterations, 15 maximum leaves per tree, learning rate .05, minimum leaf size 20, L_2 penalty 1, no early stopping, and fixed random seed. Equal model settings prevent one feature set from receiving a larger tree budget. Joint features have twice the observation dimension, so this is not a claim of equal VC dimension or equal total information access. The experiment deliberately asks whether those extra observations help at a fixed nonlinear budget.

The logistic replication selects regularization on an inner original-image split, then refits using outer-training data only. Its standardization is fitted within the corresponding training split. Reported AUC is computed from out-of-fold scores, and benefit BCE is $\log(1 + e^z) - Hz$. Paired log-loss intervals resample original-image clusters conditional on those fitted folds. Because training sets overlap between outer folds, this bootstrap omits full retraining uncertainty; it is not an exact confidence interval for conditional mutual information.

The Pet diagnostic strengthens the target-over-source conclusion, but joint inputs do not improve on target inputs alone. Additional source measurements can be redundant or harmful to a finite learner. The linear donor EfficientNet comparison is also a useful boundary: target-only linear performance is not clearly better than source-only, while joint evidence is stronger. Both facts are retained rather than selecting whichever learner gives the broadest claim.

B.2 Outcome-frozen robustness to donor identity and semantic class

After the initial image-grouped diagnostics were known, we recorded `paper/STRONG_ACCEPT_ROBUSTNESS_PROTOCOL_2026-09-15.md` before inspecting the additional outcomes below. It is an internal outcome-frozen extension, not an external preregistration. The purpose is to test two specific residual alternatives: memorization through reused donor images and prediction from class-correlated benefit prevalence.

For the controlled COCO diagnostic, target-image IDs and actual donor-image IDs are independently assigned to five deterministic hash folds using distinct fixed seeds. In outer fold f, a test observation is used only when *both* endpoints have fold f; a training observation is used only when *neither* endpoint has fold f. Rows with exactly one endpoint in f are unused for that fold. Consequently no target image and no donor image can appear in both the fit and test sets. Across the five folds this gives 2,617 held-out views from 997 target images, drawn from 13,113 total donor views and 11,328 distinct donor images. The same logistic and histogram-gradient-boosting settings are used for source, target, and joint evidence. Nonlinear target-minus-source AUC is $+.125$ [.097,.155], $+.066$ [.035,.098], and $+.071$ [.038,.100] for DINO-CN, ResNet50, and EfficientNet, where intervals are the three-target 98.333% family intervals. Logistic differences are $+.076$, $+.030$, and $+.014$; the EfficientNet logistic interval includes zero. The finite-capacity nonlinear result therefore survives strict endpoint separation, while the weakest linear-style case is retained as a boundary.

For natural COCO and Pet, complete ground-truth classes are assigned to outer folds. The original COCO class-only implementation was not simultaneously image-disjoint: 28/73/40/42/47 original images occurred in both fit and test because one image can contain multiple categories. The corrected analysis retains exactly the original class assignments (seed 91543) and all 1,346 test observations, but removes every training row whose original image occurs in that fold's test set. This purges

Table 13: **Target-supervised robustness diagnostics.** Nonlinear target-evidence minus source-evidence held-out AUC, with 98.333% marginal intervals adjusted over three CNNs. Corrected COCO class splits retain all old test rows while purging every overlapping training image. Historical class-only results are retained separately.

Diagnostic	Target	Δ AUC [98.333% CI]
COCO donor: both endpoints disjoint	DINO-CN	$+.125$ [.097,.155]
	ResNet50	$+.066$ [.035,.098]
	EffNet-B0	$+.071$ [.038,.100]
COCO natural: class + image disjoint	DINO-CN	$+.167$ [.120,.215]
	ResNet50	$+.134$ [.088,.180]
	EffNet-B0	$+.120$ [.074,.167]
Pet: breed held out	DINO-CN	$+.430$ [.404,.456]
	ResNet50	$+.335$ [.309,.361]
	EffNet-B0	$+.419$ [.394,.446]

44/100/63/59/65 rows; zero class and image overlap remains. With unchanged learners, corrected nonlinear target-minus-source AUC is $+.167/+.134/+.120$, with three-target 98.333% intervals [.120,.215]/[.088,.180]/[.074,.167]. Historical class-only values $+.176/+.141/+.138$ are preserved but not presented as image-disjoint evidence. Pet has one example per image, so its breed-held-out increments $+.430/+.335/+.419$ are unchanged. These controls support held-out-class predictive information without relying on a known test photograph.

B.3 Fixed-map all-origin and within-class controls

The deployment-side robustness check does not fit any target-supervised learner. For fixed source decoder b_s, target t, and evidence-producing recognizer m, it evaluates the matrix $M_{t,m}$ from Eq. 2. Every row uses one fixed target event and identical observations. On natural COCO, Pet, and ImageNet-9, every non-source target has its own evidence as the pointwise best of the six origins (15/15 cases). To avoid treating the strongest non-self origin as prespecified, each paired bootstrap draw subtracts the *maximum* non-self origin AUC from the self-origin AUC. Four of five COCO non-source intervals, all five Pet intervals, and all five ImageNet-9 intervals are positive. The sole boundary is natural-COCO EfficientNet: point gap $+.029$, selected-best 95% interval $[-.015, .057]$.

A separate equal-class analysis removes all between-class ranking. Classes with only one target benefit state have undefined within-class AUC and are excluded rather than assigned chance performance. The evaluable-class counts for DINO-CN/ResNet50/EfficientNet are 57/57/55 on COCO and 37/37/37 on Pet. Their macro target-minus-CLIP evidence gaps are $+.184$ [.114,.257], $+.125$ [.048,.196], $+.112$ [.030,.199] on COCO and $+.396$ [.356,.432], $+.403$ [.365,.441], $+.396$ [.356,.436] on Pet. On ImageNet-9, all nine superclasses are evaluable for every target; the five non-source macro gaps range from $+.156$ to $+.316$, with every superclass-bootstrap interval positive. These controls use target labels only to define the evaluation event and within-class strata; they do not fit, normalize, calibrate, or select the frozen decoder.

B.4 Source-only confidence-family ablation

A final outcome-frozen control asks whether the all-origin diagonal is simply inherited from direct confidence magnitudes. After P1–P5 were known but before inspecting this ablation, we fixed two CLIP-source Hard-BCE fits with the same source data, optimization budget, validation criterion, and seeds as the main family. *NoMargin* removes all three single-view margins. *Structural* additionally removes all three single-view entropies, leaving only the three pairwise JS divergences, three prediction-agreement indicators, and two feature cosines. No target record enters training, scaling, checkpointing, or feature-subset choice.

The full 14-D interface gives self-origin rank one in 15/15 non-source target–dataset cases, with positive selected-best intervals in 14/15. Removing all margins reduces these counts to 11/15 and 8/15; the Structural variant gives 10/15 and 9/15 after the superclass arithmetic correction. The failure pattern is informative rather than discarded. On natural COCO, NoMargin is self-first for only EVA02

Table 14: **Self-origin dominance under one unchanged CLIP source decoder.** For every non-source target, we apply the same source router to evidence produced by each of the six recognizers on the identical observation and rank all origins against the same target benefit event. "Self best" counts targets whose own evidence has the largest AUC; "CI>0" counts paired 95% selected-best intervals whose lower endpoint is positive. The last column removes between-class ranking by macro-averaging the target-minus-CLIP origin contrast within ground-truth classes/superclasses. VOC2012 is a separately specified pre-outcome v2 external extension after the recorded annotation-only head-feasibility amendment; it is not presented as an unamended preregistration. No target label fits or selects a router in this table.

Dataset	Self best	CI>0	Min. self–best gap	Within-class origin gap
COCO natural	5/5	4/5	+.029	+.112–+.184
Oxford-IIIT Pet	5/5	5/5	+.341	+.396–+.403
ImageNet-9	5/5	5/5	+.133	+.156–+.316
VOC2012	5/5	4/5	+.004	+.097–+.310

Table 15: **Source-only confidence-family deletion on the three-dataset grid.** Cells give self-origin rank-one / positive nominal selected-best 95% interval counts. NoMargin removes three margins; Structural also removes three entropies. Source fits are unchanged; ImageNet-9 uses corrected benefit events. VOC transfer is reported separately.

Ablation	COCO	Pet	ImageNet-9	Total
NoMargin	2/5 / 0/5	4/5 / 4/5	5/5 / 4/5	11/15 / 8/15
Structural	1/5 / 1/5	4/5 / 4/5	5/5 / 4/5	10/15 / 9/15

and DINO-CN, and Structural only for DINO-CN. On Pet, both ablations fail for DINO-CN but remain self-first with positive selected-best intervals on SigLIP, EVA02, ResNet50, and EfficientNet. On corrected ImageNet-9, Structural evidence is self-first in all five targets but has positive selected-best intervals in four; DINO-CN remains the interval boundary. Thus a direct-margin explanation is not sufficient in every regime, but confidence-family coordinates are an important part of the transferable interface; we do not claim a confidence-free universal geometry.

B.5 Standalone residual transfer is weak (P7)

After P6 was known, we froze a stricter source-only conditional test before inspecting its outcomes. The nuisance vector contains all three margins and entropies, their full–core changes, all three top-1 agreements, and core/full maximum posterior probability. Five-fold residualizers are fit only on the 5,000 CLIP ImageNet source-training identities; genuine/wrong views of one image remain together. Source-training structural residuals are strictly out-of-fold. The same residualizer ensemble is then used unchanged on COCO, Pet, and ImageNet-9. No target fitting, target scaling, target sign selection, or target binning is allowed.

The nuisance family is not weak: on held-out CLIP source validation it explains about 91.5–92.3% of the frozen Hard-logit variance and 73–78% of the three JS-coordinate variance; cosine R^2 is lower (.29/.34). Residual standard deviations remain nonzero, and a new source-only decoder trained on the five structural residuals still reaches .659–.662 source-validation benefit AUC. Thus the test removes substantial confidence-predictable structure without producing a constant score.

For the residual-structural endpoint, self evidence is rank one in 3/15 non-source cases, with a positive nominal selected-best interval in 2/15; the mean gap is $-.020$ AUC. The two resolved positives are DINO-CN on COCO ($+.072$ [.033,.094]) and corrected ImageNet-9 ($+.039$ [.015,.054]). Pet has no self-rank-one residual-structural target. Residualizing the frozen Hard score gives 5/15 rank-one and 5/15 resolved positives, mean gap $-.024$. These results establish weak standalone residual portability. They do *not* establish conditional independence or confidence mediation: the residual decoder lacks the confidence variables needed to exploit possible interactions. The nested-feature analysis below directly tests that different question.

Table 16: **Source-only residual-only transfer (P7).** Structural features or the frozen Hard score are residualized against the 13-D confidence family using CLIP-source cross-fitting. Entries give self-origin rank / self-minus-best AUC [nominal 95% CI]. Confidence is absent from the residual decoder: this is not a conditional-independence test. ImageNet-9 labels are precision-corrected; every source fit is unchanged.

Data	Target	Structural residual	Score residual
COCO natural	SigLIP	6 / −.087 [-.122,-.056]	3 / −.048 [-.086,-.015]
	EVA02	3 / −.055 [-.089,-.025]	4 / −.076 [-.114,-.041]
	DINO-CN	1 / .072 [.033,.094]	1 / .154 [.115,.193]
	ResNet50	6 / −.084 [-.119,-.052]	2 / −.022 [-.066,.018]
	EffNet-B0	5 / −.054 [-.096,-.027]	5 / −.073 [-.116,-.034]
Oxford-IIIT Pet	SigLIP	5 / −.042 [-.068,-.019]	1 / .036 [.013,.054]
	EVA02	2 / −.015 [-.040,.009]	6 / −.054 [-.076,-.039]
	DINO-CN	4 / −.013 [-.037,.002]	1 / .207 [.181,.225]
	ResNet50	4 / −.017 [-.046,.002]	1 / .105 [.083,.125]
	EffNet-B0	4 / −.025 [-.053,-.007]	3 / −.010 [-.035,.007]
ImageNet-9	SigLIP	3 / −.004 [-.030,.015]	5 / −.053 [-.075,-.033]
	EVA02	1 / .022 [-.006,.036]	4 / −.038 [-.061,-.016]
	DINO-CN	1 / .039 [.015,.054]	1 / .201 [.178,.219]
	ResNet50	6 / −.026 [-.048,-.018]	6 / −.321 [-.340,-.304]
	EffNet-B0	2 / −.015 [-.035,.003]	6 / −.366 [-.386,-.346]

A separate implementation reconstructs the nuisance vector, residualizer ensemble predictions, source and residual decoder scores, complete AUC matrices, self ranks, and point gaps directly from retained caches/state dictionaries. A 200-draw independent bootstrap smoke audit reproduces the interval directions. Support diagnostics also argue against a trivial "all residuals vanished" explanation: target structural-residual standard deviations are generally comparable to or larger than source validation. Some DINO nuisance coordinates lie outside CLIP-source ranges, so especially large DINO residual effects are treated as portability under source-only extrapolation rather than a target-conditioned estimate.

B.6 Residual-only portability versus conditional predictive information

The source-only residual analysis P7 asks whether a decoder can transfer using the residuals *alone*. Its negative result cannot establish that the removed confidence coordinates are sufficient. The following elementary consequence of conditional entropy makes the distinction explicit. It is a diagnostic-identification boundary, not a claim of a new general information identity or conditional-independence test. Related residual-based tests require their own assumptions and treatment of interactions (Sudarshan et al., 2023).

Let $H \in \{0, 1\}$ be the benefit event, Q the observed confidence family, and S the structural features. Fix a measurable residualizer $a(Q)$, which may have been estimated on an independent source dataset. Define $R = S - a(Q)$. Write $L(q_U) = \mathbb{E}[-H \log q_U(U) - (1 - H) \log(1 - q_U(U))]$ and $L^*(U) = \inf_q L(q_U)$ over measurable probability predictors. Assume fitted predictors have finite expected log loss; Bayes boundary probabilities use the usual $0 \log 0 = 0$ convention.

Proposition 1 (Residualization does not identify conditional sufficiency)**.** *For a fixed residualizer,*

$$I(H; R \mid Q) = I(H; S \mid Q), \qquad L^*(Q) - L^*(Q, S) = I(H; S \mid Q).$$

Nevertheless, $I(H; R)$ can be zero while $I(H; S \mid Q) > 0$. For fitted predictors q_Q and $q_{Q,S}$,

$$L(q_Q) - L(q_{Q,S}) = I(H; S \mid Q) + \epsilon_Q - \epsilon_{Q,S},$$

where $\epsilon_U = \mathbb{E}\ \mathrm{KL}(\mathrm{Ber}(P(H = 1 \mid U)) \| \mathrm{Ber}(q_U(U))) \geq 0$.

Proof. The maps $(Q, S) \mapsto (Q, S - a(Q))$ and $(Q, R) \mapsto (Q, R + a(Q))$ are inverse measurable transformations. They generate identical information about H conditional on Q, proving the first identity. Binary log loss decomposes into conditional entropy plus the displayed Bernoulli KL term, so $L^*(U) = \mathsf{H}(H \mid U)$. Taking the difference gives both risk identities. For the counterexample, take independent uniform signs $Q, S \in \{-1, +1\}$ and $H = \mathbf{1}[QS > 0]$. Then $a(Q) = \mathbb{E}[S \mid Q] = 0$ is an exact conditional-mean residualizer and $R = S$. Individually Q and R leave $P(H = 1) = 1/2$,

whereas their pair determines H. Therefore $I(H;R)=0$, $I(H;S\mid Q)=\log 2$, individual Bayes BCE is $\log 2$, and joint Bayes BCE is zero. □

Even a perfect conditional-mean residualizer can therefore leave interactions that a residual-only predictor cannot use. Finite-learner improvement after adding S need not imply strictly positive population CMI: a redundant transformation can reduce approximation error for a restricted learner. Conversely, a zero or negative fitted improvement does not establish zero CMI, because estimation or approximation error can dominate a positive conditional-information term. Equation 3 contains both excess risks. Our experiment estimates held-out predictive improvement for a fixed learner, not the information quantities or a causal mediation effect. The exact four-state counterexample is checked numerically; it is not presented as synthetic validation of the vision experiment.

B.7 POST-P7 NESTED-FEATURE DIAGNOSTIC (P8)

The internal plan `paper/CONDITIONAL_INCREMENT_PROTOCOL_2026-09-16.md` was recorded after earlier COCO/Pet/ImageNet-9/VOC, P6, P7, and Hard+margin outcomes were known but before the new fits below. This is a post-P7 analysis on retained benchmarks, not a new independently collected or externally preregistered confirmation. No split or hyperparameter was chosen from its new outcomes.

The fixed 13-D confidence family Q is identical to P7: core/context/full top-two logit margins; full-minus-core margin; three entropies; full-minus-core entropy; three top-1 agreements; and core/full maximum softmax probability. The five structural coordinates S are exactly the three JS divergences and two within-model feature cosines. We compare Q (13-D) against (Q,S) (18-D); unlike P7, the conditioning coordinates remain available. Both fit histogram gradient boosting with 150 iterations, 15 maximum leaves, learning rate .05, minimum leaf size 20, $L_2=1$, fixed random state 9153, and no early stopping or target hyperparameter search. Equal tree budgets do not imply equal statistical complexity across different input dimensions.

For each condition, complete image IDs receive one of three deterministic approximately balanced fold assignments (seed 9291). Fold zero is a single test set; the other folds supply training. Source and target feature comparisons use exactly the same test rows. On COCO donors, donor IDs receive a separate assignment (seed 9293): test requires both endpoints in fold zero, training requires neither. Crossing rows are not used. This gives 5,829 training and 1,454 test donor views, with 367 test original images. Unlike the earlier 2,617-view P1 OOF diagnostic, P8 is an independent fit/test split with no overlapping outer-training fits. All held-out IDs and masks are saved.

The three target recognizers are fixed to DINO-CN, ResNet50, and EfficientNet-B0. Their mean BCE reduction $L(Q)-L(Q,S)$ is the primary condition-level endpoint. The five conditions are COCO natural, Pet, ImageNet-9, VOC, and COCO graded donors. Five thousand paired image-cluster bootstrap draws (seed 9297) preserve all target scores and labels associated with an image. Each 99% marginal interval has Bonferroni adjustment over these five primary means for nominal 95% family coverage. Individual target intervals are descriptive 95% intervals. All intervals condition on fixed fitted models and benchmark manifests. They do not include backbone-training or model-population uncertainty.

The primary conditional BCE reductions are positive in point estimate in all five conditions, but resolved only on ImageNet-9 (.0142 [.0056,.0234]) and donors (.0238 [.0104,.0377]); COCO, Pet, and VOC remain uncertain. The prespecified donor sensitivity independently resamples target and donor endpoints and multiplies their row multiplicities (seed 9299). Its 99% interval is [.0027,.0449], retaining the donor direction. Individual target gains are heterogeneous: neither DINO interval excludes zero in the corrected ImageNet-9 and donor conditions. The mean effect is not asserted for every recognizer.

For the secondary portability comparison, two additional models with the same budgets are trained on CLIP ImageNet *training* records only, from Q and (Q,S) respectively. Source validation is recorded descriptively; it selects neither model nor checkpoint. Applied unchanged to the same target test rows, structural augmentation improves mean BCE on Pet and donors, worsens it on COCO and ImageNet-9, and is unresolved on VOC. In particular, ImageNet-9 augmentation has source-only reduction $-.0491$ $[-.0645,-.0347]$, opposite the positive target-supervised reduction. These secondary intervals are not pooled with the five primary tests to imply simultaneous coverage

Table 17: **Post-P6 source-only Hard+margin complementarity.** The source-fitted affine combiner is unchanged. Entries are five-target mean AUC differences with 98.75% marginal intervals adjusted over four datasets (nominal 95% family coverage). ImageNet-9 uses stable-superclass events. The prespecified rule requires three positive Hard+Margin-minus-Margin intervals; only VOC passes.

Data	Hard+Margin − Margin	Hard+Margin − Hard
COCO natural	−.059 [-.081,-.038]	−.016 [-.019,-.013]
Oxford-IIIT Pet	−.018 [-.022,-.014]	−.004 [-.004,-.004]
ImageNet-9	−.001 [-.010,.009]	−.010 [-.011,-.008]
VOC2012	.035 [.004,.067]	−.006 [-.010,-.003]

over ten claims. The comparison supports a separation between conditional predictive usefulness and the portability of these particular fitted decoders, not a generally superior source-trained algorithm.

Independent reconstruction and retained corrections. A separate implementation rebuilds all confidence and structural coordinates from raw caches, reconstructs benefit signs using long-double complement arithmetic, refits both source and target conditional learners from the declared training rows, and checks saved test logits. It independently reproduces all primary image-bootstrap arrays and the donor two-endpoint arrays, not merely their rounded interval endpoints. The same audit verifies every corrected COCO class/image split, all OOF AUC/BCE values, and 2,000-draw three-target adjusted AUC intervals. P2 OOF refitting uncertainty is not supplied by that audit. Its long-double event check discovered the superclass-cancellation issue described in Appendix A.5; the original P8 ImageNet-9 output is retained and corrected results use the same fixed split and budgets. Neither P7 source fitting nor P8 source fitting was changed by the correction.

B.8 Source-only Hard + margin is not broadly additive

After P6 and the VOC outcomes were known, we recorded a secondary source-only complementarity test before fitting its new combiner. The frozen CLIP Hard logit and raw full-minus-core margin are the only two inputs. For each Hard seed, a $C = 1$ two-feature logistic regression is fitted on a deterministic half of the 1,000 CLIP source-validation identities; paired genuine/wrong rows move together. Scaling and coefficients are source-only and are never changed on COCO, Pet, ImageNet-9, or VOC. Standardized Hard coefficients are about $+1.56$, while the margin coefficients are small and negative ($-.13$ to $-.15$), so the source fit does not collapse to a margin-only rule.

The prespecified broad-complementarity criterion fails: only one of four dataset-mean Hard+Margin-minus-Margin family intervals is positive, versus a required three. Only 6/20 individual target–dataset point differences exceed raw margin. More importantly, the combined score is below the original Hard score in all four dataset means, with negative family intervals throughout. The sole individual Hard+Margin-minus-Hard positive point is VOC DINO-CN ($+.0049$). The initial independent reconstruction rebuilds events and scores and repeats all dataset intervals. The precision-corrected ImageNet-9 recomputation retains the failed decision: its mean Hard+Margin-minus-Margin gap is now $-.001$ $[-.010, .009]$, and its combined score remains below Hard alone. Historical pre-correction intervals are retained separately. Thus the transferable benefit interface is not established as an operationally additive correction to target margin.

B.9 Same-object genuine versus donor changes

A mechanism comparison based only on the separate natural and controlled cohorts would confound context construction with object selection. We instead deduplicate the genuine view from the 1,457-object mask cohort and compare it with all nine donor views per object. Source and target response arrays align by annotation, original image, true category, and donor identity. Bootstrap draws use the same original-image multiplicities for genuine and donor contrasts. The reported genuine-minus-donor change therefore compares effects on the same objects.

Two-dimensional surface axes are fixed by pooled source-training quintiles, never by target outcomes: JS(core,full) versus full-minus-core margin, and JS(context,full) versus that margin difference. Cells with fewer than 20 observations in either regime are excluded from common-support decomposition

and their excluded mass is reported. Display smoothing is Jeffreys smoothing, not invented target evidence. Coarse cells do not control all 14-dimensional variation and cannot establish a sufficient statistic.

For regime j let $p_j(k)$ be its mass on a supported cell and $r_j(k)$ the observed benefit probability. On renormalized common support,

$$\sum_k p_1(k)r_1(k) - \sum_k p_0(k)r_0(k) = \frac{1}{2}\sum_k [p_1(k) - p_0(k)][r_1(k) + r_0(k)] + \frac{1}{2}\sum_k [p_1(k) + p_0(k)][r_1(k) - r_0(k)].$$

The first term describes occupancy change and the second within-cell response change. It is an algebraic decomposition, not a causal attribution. For the JS(core,full) projection, the within-cell terms are about $-.208/-.324/-.305$ for ConvNeXt/ResNet/EfficientNet, compared with total changes $-.302/-.349/-.344$. The other projection and all cell counts remain in the result record, including cases that do not support a universal stable surface.

B.10 Prior-only control

A genuine/donor indicator can improve benefit prediction simply because the regimes have different positive-label frequencies. The follow-up diagnostic therefore compares a blind decoder with two regime-specific intercept adjustments against a regime-aware decoder with global intercept adjustment. Fit, calibration, and test images are disjoint and identical across the competitors. Both use the fixed nonlinear capacity. Genuine and donor data have equal aggregate weight, so nine donor repeats do not dominate one genuine observation.

The regime-aware-minus-prior-only benefit BCE differences are $-.00471$ $[-.00848, -.00065]$ for ConvNeXt, $-.00451$ $[-.00917, .00024]$ for ResNet50, and $-.01321$ $[-.01946, -.00687]$ for EfficientNet. The ResNet interval is not an equivalence result and does not support a definite shape gain. These are post-initial-diagnostic controls, not an unlabeled policy and not proof that every prior-only calibration family fails.

C Transfer Conditions and Their Failure Modes

C.1 A compatible likelihood-ratio coordinate

We state a finite-support proposition to make identifiability explicit. Let $\mathcal{Z}$ be finite, and optimize over real-valued scores f on the union of source supports. For each source m, the benefit-conditioned distributions have the same support S_m for both labels, and satisfy

$$f_{m,y}(z) = a_m(z)r_y(z)/Z_{m,y}, \qquad z \in S_m,$$

with positive a_m and r_y. Construct a graph with vertices $\bigcup_m S_m$ and an edge between two points if some source includes both. Assume it is connected. Define $h(z) = \log[r_1(z)/r_0(z)]$, and positive source weights w_m.

Proposition 2 (Compatible within-recognizer ordering)**.** *The population objective*

$$\mathcal{L}(f) = \sum_m w_m \mathbb{E}_{Z^+\sim f_{m,1}, Z^-\sim f_{m,0}} \log\left(1 + \exp(f(Z^-) - f(Z^+))\right) \tag{6}$$

has minimizers $f = h + c$ on the connected union of supports. If a target has both benefit classes, support contained in that union, and class-conditional likelihood ratio strictly increasing in h, every such minimizer has Bayes-optimal target ranking. Source benefit priors can differ arbitrarily in $(0, 1)$.

Proof. For a source-supported unordered pair z, z', set $d = f(z) - f(z')$, $\alpha = f_{m,1}(z)f_{m,0}(z')$, and $\beta = f_{m,1}(z')f_{m,0}(z)$. The pair's objective is $\alpha\log(1 + e^{-d}) + \beta\log(1 + e^{d})$. Its strictly convex optimum is

$$d = \log(\alpha/\beta) = \log\frac{f_{m,1}(z)}{f_{m,0}(z)} - \log\frac{f_{m,1}(z')}{f_{m,0}(z')} = h(z) - h(z').$$

The nuisance tilt and normalization constants cancel. Diagonal $z = z'$ contributions are constant, and all off-diagonal pair weights are positive. Hence h simultaneously minimizes every supported pair term, so it minimizes their positively weighted sum. Every global minimizer must minimize each nonconstant pair term; connectedness propagates equality of score differences to give $f = h + c$. A target's posterior odds are its likelihood ratio multiplied by a positive prior-odds constant. Its Bayes ordering is consequently the ordering of h. The pair-sampling objective conditions on benefit class and contains no source prior. □

This proposition is an application of established pairwise-logistic ranking theory (Gao & Zhou, 2015), not a claim to have discovered logistic consistency. The useful specialization is the recognizer-dependent nuisance tilt and connected comparison-support condition. A disconnected support allows separate additive constants whose cross-component ordering is unconstrained. Unseen target support also cannot be ordered from this proposition. Real-valued neural features with finite data require estimation and approximation assumptions; our implementation's finite minibatch sampler is not a certificate of population optimality.

C.2 Quantitative sensitivity to support and conditional drift

Let P be a source distribution of evidence and benefit, Q a target, $\eta_P = P(H = 1 \mid F)$, and $\eta_Q = Q(H = 1 \mid F)$. AUC gives half credit to score ties, and AUC_Q^* denotes a Bayes-optimal target ordering by η_Q. Let $q(F) \in (0, 1)$ be a source probability predictor. Suppose $Q_F \ll P_F$ and $dQ_F/dP_F \le \kappa$, the target benefit prevalence π_Q lies in $(0, 1)$, and $g : [0, 1] \to [0, 1]$ is strictly increasing and L-Lipschitz. Define

$$\epsilon = \mathbb{E}_{P_F} \mathrm{KL}(\mathrm{Ber}(\eta_P) \| \mathrm{Ber}(q)), \qquad \delta = \mathbb{E}_{Q_F} |\eta_Q - g(\eta_P)|.$$

Then

$$\mathrm{AUC}_Q^* - \mathrm{AUC}_Q(q) \le \min\left\{1, \frac{\delta + L\sqrt{\kappa\epsilon/2}}{\pi_Q(1 - \pi_Q)}\right\}. \tag{7}$$

To see this, the symmetric pairwise ranking expression bounds regret for a score $r \in [0, 1]$ by $\mathbb{E}_Q|\eta_Q - r|/[\pi_Q(1 - \pi_Q)]$: on every reversed pair the posterior difference is at most the sum of the two estimation errors, with half credit for ties. Choose $r = g(q)$; strict monotonicity preserves q's AUC. The triangle inequality gives $\mathbb{E}_Q|\eta_Q - g(q)| \le \delta + L\mathbb{E}_Q|\eta_P - q|$. Cauchy–Schwarz, density-ratio control, and Bernoulli Pinsker yield $\mathbb{E}_Q|\eta_P - q| \le \sqrt{\kappa\epsilon/2}$.

The source excess risk here is a *population* quantity, not measured training BCE. The bound is often loose, particularly for rare benefit events or poor support overlap. It uses known proper-loss/ranking arguments (Agarwal, 2014). Neither the bins nor the grouped-CV results estimate all its terms or verify its assumptions. Its role is to distinguish conditional compatibility from marginal overlap and source fitting error.

C.3 Why pooling can reverse an ordering

Consider $z \in \{-1, 1\}$ and source posteriors $\eta_1(z) = \sigma(z + a)$, $\eta_2(z) = \sigma(z - a)$ with $a = 1 + \log 4$. Both increase with z. Let their marginals be $(.9, .1)$ and $(.1, .9)$ and pool the sources equally. Their posteriors are approximately $(.8, .9673)$ and $(.0327, .2)$, but the pooled posterior is $(.7233, .2767)$ and decreases with z. A target with uniform marginal and posterior $(.2, .8)$ gives AUC .8 to either increasing source score, versus .2 to the pooled posterior. The source-conditioned log-odds differ only by intercepts, yet evidence-dependent source membership changes the pooled relationship.

More generally, write a pooled posterior as $\bar\eta(z) = \sum_m \rho_m(z)\eta_m(z)$ where $\rho_m(z)$ is the conditional source membership. For two points z_1, z_2, adding and subtracting $\sum_m \rho_m(z_2)\eta_m(z_1)$ yields

$$\bar\eta(z_2) - \bar\eta(z_1) \ge \min_m[\eta_m(z_2) - \eta_m(z_1)] - \mathrm{range}_m\, \eta_m(z_1)\, \mathrm{TV}(\rho(z_2), \rho(z_1)).$$

This is a sufficient order-stability certificate, not a necessary condition. It complements existing multiple-source adaptation results (Mansour et al., 2008). Our empirical failure cannot uniquely be attributed to this mechanism because finite approximation and optimization also matter. Numerical counterexamples and random finite-distribution bound checks are regression tests, not substitutes for the proof.

Table 18: **Prespecified source-only ranking intervention.** All candidates use matched features, budgets and source-AUC checkpointing. WR constructs within-recognizer pairs; GR ignores recognizer membership; BBCE balances classes within each recognizer. ResNet/EfficientNet never enter training. The source-only selector chose BBCE4, not a target-best rule.

Data	Target	BCE4	BBCE4	GR4	WR4	WR3	Margin
COCO	DINO-CN	0.697	0.696	0.691	0.693	0.846	0.785
COCO	ResNet50	0.603	0.597	0.593	0.601	0.737	0.764
COCO	EffNet-B0	0.542	0.541	0.534	0.540	0.667	0.724
Pet	DINO-CN	0.963	0.962	0.961	0.961	0.943	0.958
Pet	ResNet50	0.929	0.927	0.929	0.927	0.927	0.924
Pet	EffNet-B0	0.914	0.910	0.913	0.911	0.908	0.910

D Source-Only Ranking Intervention and Negative Results

The primary candidate WR4 pools four recognizers but forms every positive–negative training pair within one recognizer. The source ID is pairing metadata, never an input or a target-specific embedding. WR3 uses only the three VLMs. GR4 forms pairs across the pooled data, ignoring identity; BBCE4 averages equal positive and negative BCE contributions within each recognizer; BCE4 is the ordinary pooled baseline. These controls separate pair origin, class balancing, and a mere change of objective.

All candidates use the same 14–32–32–1 network, 100 epochs, AdamW $.001/.0001$, original-image batch size 512, and seeds 17/23/41. Every included recognizer contributes the same original-image batch and both context views. WR samples 512 positive–negative pairs with replacement per source; GR uses the corresponding total number of pooled pairs. Pair sampling uses a separate deterministic generator. The implementation records skipped single-label batch groups. Gradients do not receive target labels or recognizer identity as a feature.

Source validation is split by deterministic original-image hash. The checkpoint half selects the earliest maximum macro source AUC for *every* candidate; the remaining images are reserved for source-only calibration/selection roles. This differs from the original all-validation benefit-BCE-selected single-source experiment and is not used to replace those rows. Each full candidate and each of four leave-one-recognizer-out variants of WR4/GR4/BCE4/BBCE4 is trained with all three seeds: 21 configurations and 63 fits in total.

The source-only selector averages omitted-recognizer AUC across the four leave-one-out tasks on 150 original validation images per omitted recognizer that are not used for checkpointing. Its scores are .7371 for raw margin, .8225 for WR4, .8209 for GR4, .8239 for BCE4, and .8256 for BBCE4; it selects BBCE4 before target outcomes. The frozen candidate hashes and source-only scores are retained. A pre-target serialization mismatch in one fit prompted recomputation of all selection scores; they were exactly identical, the chosen rule was unchanged, and both manifests were retained. No target result justified the reconciliation.

On natural COCO, the new pairwise primary does not beat the matched four-source BCE baseline on the unseen CNNs, and both are substantially below margin. WR3 is better there but is the declared ablation. The unchanged candidates behave differently on Pet, where four-source AUC can be high without a clear ranking-loss advantage. These are recorded failed attempts to turn compatibility into a generally successful source-only method, not a universal-router contribution. Four primary natural COCO comparisons (WR4 versus BCE4/margin on the two unseen targets) use 98.75% Bonferroni intervals in addition to nominal paired intervals. All other broad-grid contrasts remain descriptive.

An earlier independent recheck of BCE-selected U4 used ten new seeds and three 300-epoch extensions per pool. U4 was below U3 on both unseen natural CNNs for every new seed. That prior negative result remains archived; this ranking experiment changes the training objective and checkpoint criterion and does not replace it. Results that favor U4 in some low-PMI donor conditions are also retained. Neither study asserts failure in every regime or for every possible source mixture.

Table 19: **Pet disagreement and non-negligible-benefit checks.** All scores are the fixed CLIP map unless labelled margin. Disagreement is label-free; the $|A_t| > 10^{-6}$ subset is a label-aware numerical diagnostic, not a deployment rule. Point estimates are not new significance claims.

Target	$n_{\neq}$	Hard$_{\neq}$	Margin$_{\neq}$	$n_{\lvert A\rvert>10^{-6}}$	Hard	Source evidence
CLIP	430	0.626	0.642	3662	0.911	0.911
SigLIP	182	0.632	0.697	3660	0.943	0.559
EVA02	217	0.608	0.652	3660	0.920	0.570
DINO-CN	86	0.516	0.728	1277	0.881	0.536
ResNet50	203	0.529	0.634	2374	0.883	0.493
EffNet-B0	280	0.547	0.603	2757	0.882	0.497

E Benefit Ranking Versus Decision Value

AUC predicts the sign of a target-probability change. It neither weights its magnitude nor states whether either expert predicts the correct class. In Pet, many CNN predictions are saturated and most core/full top-1 predictions agree. The disagreement counts and their much weaker AUC therefore qualify the high whole-cohort ranking values. The non-negligible-benefit subset is fixed at 10^{-6} as a numerical sensitivity check, not selected to maximize a result; it uses evaluation labels and is not a deployment filter.

We additionally evaluate two label-free pseudo-label gains after observing the high aggregate AUC. One substitutes the full prediction for Y in Eq. 1; the other substitutes the core prediction. They are simple controls, not new acquisition policies or reproductions of DyMo's complete calibration procedure (Du et al., 2026). On Pet CNNs, full-pseudo-label AUC is .878/.842/.843 and core-pseudo-label AUC is .884/.816/.822, versus fixed CLIP-map AUC .912/.915/.898. Raw target margin is nevertheless stronger than that map on all three CNNs. On examples where core and full predict the same correct class, pseudo-label advantage equals realized advantage exactly; that algebraic fact does not make its uncalibrated magnitude a globally optimal benefit-probability ordering.

For a posterior mixture $p_r = (1-r)p_c + rp_f$,

$$L_r = L_c - \log(1 - r + re^A).$$

The optimal coefficient depends on the conditional advantage magnitudes, not only $P(A > 0 \mid F)$. For example, fix $p_c = (.80, .02, .18)$ and $p_f = (.88, .10, .02)$. Two conditional label laws $P = (.8, 0, .2)$ and $Q = (0, .8, .2)$ have the same benefit probability .8, but require optimal mixture weights 0 and .85. The first improvement is $.8\log(1 + .1r) + .2\log(1 - 8r/9)$, whose concave maximum occurs at zero. The second replaces .1 by 4 in the first logarithm and has its concave maximum at .85. This is a decision-identification boundary, not the main ranking-transfer theorem.

Fixed-quota tests at .25/.50/.75 full-view use are also retained for every source-only candidate. The largest scores use the full prediction and the remainder use core, with deterministic label-free tie handling. These scores are not automatically calibrated posterior-mixture weights, and evaluating both experts does not save their inference cost. An earlier controlled CLIP-source ConvNeXt test at 50% use improved NLL relative to margin without a distinguishable accuracy improvement; the expanded Pet results do not turn that observation into general task superiority. Full endpoint metrics are stored with the per-image results for every candidate rather than choosing a favorable operating point from target labels.

F Provenance and Reproduction

The current working manuscript is built from its main source, this supplement, generated portability tables, and the information-versus-decoder figure. Older manuscript versions and data-generating code remain archived and are not silently overwritten. Source and target cache hashes identify each experiment. The refreshed original source protocol, new ranking protocol, and target-supervised diagnostic protocol have separate result directories and output names.

From the project root, the main stages are:

```
python3 -m experiments.submission_precision_refresh train
python3 -m experiments.submission_precision_refresh \
  evaluate-parallel
python3 -m experiments.submission_precision_refresh summary
python3 -m experiments.recognizer_rank_transfer train
python3 -m experiments.recognizer_rank_transfer loro
python3 -m experiments.recognizer_rank_transfer select
python3 -m experiments.recognizer_rank_transfer evaluate
python3 -m tools.prepare_pets_verified run
python3 -m experiments.pets_rank_confirmation cache-all
python3 -m experiments.pets_rank_confirmation evaluate
python3 -m experiments.pets_rank_confirmation diagnostic
python3 -m experiments.audit_rank_extension --device cuda:0
python3 -m experiments.pet_prediction_gain_control
python3 -m experiments.in9_benefit_interface cache \\
  --output-root ./results/strong_accept_in9
python3 -m experiments.in9_benefit_interface evaluate \\
  --output-root ./results/strong_accept_in9
python3 -m experiments.audit_in9_benefit_interface \\
  --root ./results/strong_accept_in9
python3 -m experiments.strong_accept_robustness p1 \\
  --target dinov3_convnext_l
python3 -m experiments.strong_accept_robustness p2-coco \\
  --target dinov3_convnext_l
python3 -m experiments.strong_accept_robustness p2-pet \\
  --target dinov3_convnext_l
python3 -m experiments.strong_accept_robustness p3-coco \\
  --target dinov3_convnext_l
python3 -m experiments.strong_accept_robustness p4-coco
python3 -m experiments.strong_accept_robustness p4-pet
python3 -m experiments.strong_accept_robustness p5-in9
python3 -m experiments.portability_paper_assets
bash paper/compile_iclr2027.sh
```

The default diagnostic target list may be narrowed to the three CNNs for the primary replication. Dataset acquisition requires the official archive availability; inference requires local pretrained checkpoints. The cached-data evaluation and figure regeneration do not require backbone retraining.