WHERE MODELS DIVERGE: LINEAR PREDICTABILITY TRACES THE DIRECTION OF REPRESENTATIONAL CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study what directional differences in cross-model predictability can reveal about how two models organize their representations on the same data. To do this, we fit linear maps between a pair of models' representations in both directions and use their *predictive asymmetry* – the difference in held-out linear predictability between models – as a diagnostic tool. This signal reveals which distinctions one model preserves relative to another, localizes which examples drive these differences, and can identify where relative performance gains between two models lie, without requiring any labels, concept/dictionary extraction, or clustering. Validating this diagnostic through controlled experiments, we show that it can recover differences in supervision granularity, training data coverage, and training progression between model pairs. Changing only the supervision objective can even reverse the direction of linear recoverability between a fine-tuned model and its parent. We apply predictive asymmetry to twelve domain-specialized foundation models and trace how their representations differ from their parent models after fine-tuning. Our experiments demonstrate that *predictive asymmetry gives a label-free account of specialization*: it tracks relative training exposure between specialist and parent and localizes where specialist–parent probe advantages hold. Predictive asymmetry therefore provides a simple, label-free view of where and in which direction model representations diverge.

1 INTRODUCTION

Learned representations are commonly used to understand how architecture, data, and training objectives shape the behavior of a model (Nguyen et al., 2020; Raghu et al., 2022; Neyshabur et al., 2021; Kumar et al., 2022). Many analyses compare representations across models with metrics that measure how *similarly* two models organize the same data (Hotelling, 1992; Raghu et al., 2017; Kriegeskorte et al., 2008; Kornblith et al., 2019; Klabunde et al., 2025). These model comparisons have been conducted extensively across architectures, objectives, and even modalities, helping us better understand models (Neyshabur et al., 2021; Kumar et al., 2022; Huh et al., 2024). Most such comparisons reduce the relationship to a single global similarity score, which says how much two representations align overall, but not how differences are structured across the data.

The structure of representational disagreement can be informative: it can be used to reason about what is preserved or changed through training and adaptation (Neyshabur et al., 2021; Kumar et al., 2022), how differences in what models encode relate to differences in downstream behavior (Liu et al., 2025; Tong et al., 2024), and whether models converge to similar internal organizations (Huh et al., 2024). In particular, systematic differences concentrated on specific inputs or distinctions can reveal model differences that a global similarity score cannot capture.

Several recent works study these differences directly. Some localize differences to concepts, neighborhoods, or observations (Vielhaben et al., 2026; Kondapaneni et al., 2025; 2026; Jiralerspong & Bricken, 2026; Thasarathan et al., 2025) while others compare representations through learned mappings between their spaces (Muzellec & Kar, 2026; Kostousov et al., 2026; Balogh & Jelasity, 2025). Prior work has shown that cross-representation predictability can be asymmetric (Muzellec & Kar, 2026), but has largely treated that asymmetry as a global property of a model pair. We ask whether the directional mismatch itself has structure: can it tell us which distinctions differ, where those differences occur, and whether they reflect training history or downstream specialization?

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.

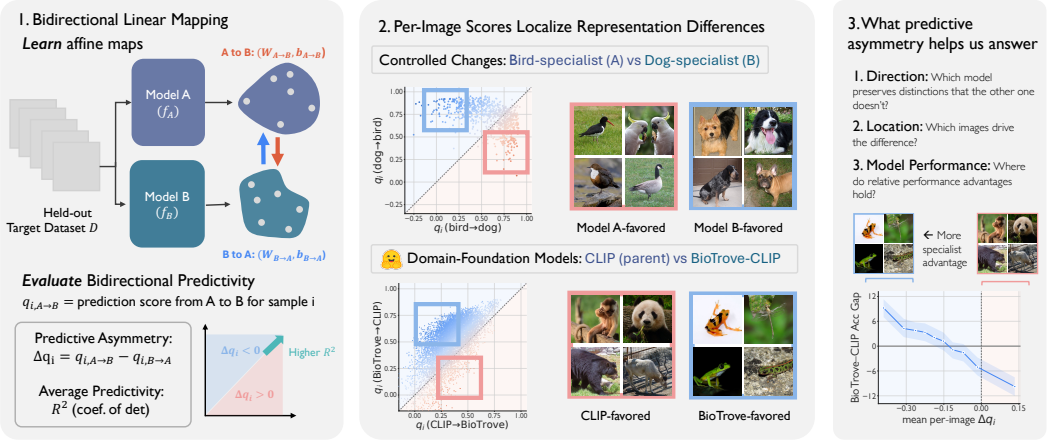

Figure 1: **Predictive asymmetry reveals where and how model representations differ.** (1) We fit affine maps between two models' representations in both directions; the per-image difference in prediction scores is the predictive asymmetry Δq_i. (2) Images off the diagonal are predicted better in one direction, separating images favored by each model, both for controlled specialists (top) and a released domain specialist (bottom). (3) Asymmetry indicates how two models organize information differently, on which images, and where relative performance shifts.

We directly study this *predictive asymmetry*, which we define as the difference between held-out linear predictability in two directions. We ask:

What does predictive asymmetry reveal about where and how two models differ?

Our central thesis is that bidirectional linear prediction provides three complementary kinds of information that symmetric similarity discards. The difference between directional prediction scores gives *Direction*: which representation preserves distinctions that are relatively less accessible from the other. Its per-image decomposition gives *Location*: which observations contribute to that mismatch. Finally, relating this label-free geometry to independently measured downstream probe performance gives *Downstream Performance*: where representational direction coincides with relative behavioral advantage. This allows us to understand differences between representations without extracting concepts, graphs, or dictionaries.

We first validate that predictive asymmetry recovers known differences in how two models were trained. Through controlled experiments, we show that predictive asymmetry identifies which model received finer-grained supervision for particular classes, which classes were seen by only one model during training, and which model was trained for longer, better than other methods. We then show that changing only the supervision target can reverse the direction of asymmetry.

Next, we use predictive asymmetry to study twelve domain-specific foundation models to understand how their fine-tuning changed their representations. In practice, it is useful to know what changed after fine-tuning, on which inputs, and where those changes translate into a performance advantage. Our analysis shows that predictive asymmetry can provide a label-free audit of specialization, with local patterns robust to changes in evaluation-population composition and rank perturbations.

We make three contributions:

1. We introduce a per-example decomposition of bidirectional linear predictability that localizes directional representational mismatch without requiring concepts, labels, or learned dictionaries.
2. We validate this asymmetry through controlled interventions, showing that its sign identifies differences in supervision granularity, training exposure, and training progression, and remains stable under population controls.
3. Across twelve domain specialists, we find that specialization can preserve or suppress different distinctions relative to the parent, reorganizes similar classes across same-domain specialists, and tracks relative training exposure and specialist–parent probe advantage.

2 Related Works

Studying model representations. Metrics such as CCA (Hotelling, 1992), SVCCA (Raghu et al., 2017), RSA (Kriegeskorte et al., 2008), CKA (Kornblith et al., 2019), and neighborhood-similarity measures like mKNN (Klabunde et al., 2025) have been used extensively to study model architectures (Nguyen et al., 2020; Raghu et al., 2022), training objectives (Xie et al., 2022; Park et al., 2024), adaptation and fine-tuning (Neyshabur et al., 2021; Kumar et al., 2022), as well as general convergence across models (Huh et al., 2024). Other works study properties of individual representations, such as intrinsic dimensionality, and link them to generalization (Ansuini et al., 2019) and downstream performance (Rao et al., 2026). Both families provide global symmetric summaries of representational similarity, but do not show how differences are distributed across inputs or whether they are directional.

Mapping between model representations. An adjacent line of work studies representations by relating them through learned mappings. Linear or orthogonal maps can align independently learned spaces across languages, models, and modalities (Mikolov et al., 2013; Merullo et al., 2022; Jha et al., 2026), and model stitching tests whether one representation can functionally replace another via a simple map between model layers (Bansal et al., 2021; Hernandez et al., 2023). More broadly, measuring how well one representation can predict another (often linearly) serves as a similarity metric, both between networks (Li et al., 2015; Kornblith et al., 2019) and between models and the brain (Muzellec & Kar, 2026). Following prior work, we also focus on *linear* maps, which have been shown to be a simple yet effective approach that can transfer substantial structure across learned representations (Merullo et al., 2022; Bansal et al., 2021).

A recurring finding in this literature is that such mappings are asymmetric: stitching can succeed better in one direction than the other (Bansal et al., 2021), and forward versus reverse model–brain predictivity can vary (Muzellec & Kar, 2026), demonstrating that mapping direction is itself informative. RSVC uses cross-model prediction in both directions over an extracted concept representation to identify concepts that are shared, missing, or differently represented across models (Kondapaneni et al., 2025). We decompose the same directional signal to individual observations in the full representation space, without requiring a concept decomposition, and show that its local structure recovers functional differences between models.

Model Diff & Explainable AI. A related line of work asks whether the structure of learned representations can be made interpretable, and whether differences between representations can be made explicit. Classical interpretability methods explain individual models through attribution (Selvaraju et al., 2017; Sundararajan et al., 2017; Zhang et al., 2025) or interpretable concepts (Kim et al., 2018; Ghorbani et al., 2019; Fel et al., 2023). More recent approaches compare models directly: RDX identifies and visualizes groups of inputs that are organized differently across representation spaces (Kondapaneni et al., 2026), shared sparse representations expose common and model-specific features (Thasarathan et al., 2025; Jiralerspong & Bricken, 2026), and ModelDiff relates functional differences to dependence on training examples (Shah et al., 2022). We show that without learned dictionaries, neighborhood graphs, or access to training data, analyzing the asymmetry of a simple linear map between two representations is sufficient to recover where and how models diverge.

3 Method

We fit linear maps between two representations and measure how much the two representations share, which one holds more unpredicted variation, and which observations produce that difference.

Cross-validated affine prediction. Let models A and B encode the same n images as $x_i \in \mathbb{R}^{d_A}$ and $y_i \in \mathbb{R}^{d_B}$, each normalized to unit Euclidean norm. For direction $A \to B$ and fold f we fit an affine map (W_f, b_f) by ridge regression with a variance-scaled penalty selected on the training fold only (Appendix A.2), and record the held-out residual $r_i^{A\to B} = y_i - W_f x_i - b_f$ for every i in the validation fold. Repeating over all folds yields a residual for every image.

Predictive asymmetry. Residuals are not comparable across directions because they live in different target spaces, so we normalize each by the pooled held-out error of predicting the target's training-fold mean, a single constant E_B^{mean} per direction. The per-image score and its asymmetry are

$$q_i^{A\to B} = 1 - \frac{\|r_i^{A\to B}\|_2^2}{E_B^{\text{mean}}}, \qquad \Delta q_i = q_i^{A\to B} - q_i^{B\to A}, \tag{1}$$

and their means over the dataset are the out-of-sample coefficient of determination $R^2_{A\to B} = \frac{1}{n}\sum_i q_i^{A\to B}$ and the global predictive asymmetry $\Delta R^2 = \frac{1}{n}\sum_i \Delta q_i$; we also report mean predictivity $\overline{R}^2 = \frac{1}{2}(R^2_{A\to B} + R^2_{B\to A})$. R^2 equals one for perfect reconstruction and zero for the mean predictor. $\overline{R}^2$ measures how much linear structure the two representations share; $\Delta R^2 > 0$ means A holds variation that B cannot linearly predict, and $\Delta R^2 < 0$ the reverse. Because the normalizing constant is fixed over the population, Δq_i localizes the global asymmetry to images, classes, or any subset by averaging, without refitting; it is not itself a coefficient of determination for one image.

Geometric interpretation. In the infinite-data, unregularized limit the optimal map decomposes into canonical components with correlations ρ_k, and $R^{2,*}_{A\to B} = \sum_k w_k^B \rho_k^2$ where w_k^B is component k's share of B's variance (Kornblith et al., 2019). Reversing the direction keeps the same ρ_k but reweights by A's variance shares, so asymmetry arises only when the two models allocate variance differently across components that differ in predictability; whitening both representations removes it for equal dimension, which is why we do not whiten. Derivations are in Appendix B.5.

4 Experimental Design

Although prior work has shown that mapping direction can be informative, it has not established what directional differences between learned models reveal about how their representations differ. We address this gap by evaluating this diagnostic in controlled settings where the source and location of change are known. We then test how supervision determines its direction, and finally apply the same diagnostic to released specialists.

Mapping and Linear Probing Setup. For every model, we extract the final pooled encoder output before any projection head, normalized to unit Euclidean norm. We fit ridge-regularized affine maps in both directions with five-fold cross-validation over images, selecting the variance-scaled penalty per fold (Section 3). For some experiments, we compare the representational analysis against downstream classification performance. For this, we fit ℓ_2-regularized logistic regression (L-BFGS) on the same held-out folds used for the linear maps; full details are in Appendix A. Since predictive asymmetry depends on the model pair and evaluation population, its magnitude is not calibrated across pairs; we therefore interpret ΔR^2 within a given pair rather than comparing absolute values across different model pairs.

4.1 Controlled experiments

4.1.1 ImageNet controls: Recovering Known Training Changes

We train ResNet-50 models from scratch on subsets of ImageNet (Deng et al., 2009), varying supervision granularity, training-set coverage, and training duration independently. These interventions give us known answers to the questions our diagnostic is meant to recover: which held-out evaluation samples should reflect the training difference, and which of the two models received the supervision, data exposure, or additional training responsible for it. Full details are in Appendix A.2.2.

1) Supervision granularity. We draw 50 classes from five WordNet superclasses—dogs, birds, insects, fish, and reptiles—with ten classes per superclass. We train a *fine* model (50-way), a *coarse* model (5-way), and five *mixed-granularity* specialists. Each specialist distinguishes the ten classes within one superclass but collapses each of the other four to a single label, resulting in matched 14-way tasks that differ only in which data subset receives fine-grained supervision. We fit the linear maps on held-out images from all 50 classes. In the fine–coarse comparison, the fine model is explicitly trained to preserve within-superclass distinctions that the coarse model can ignore. If predictive asymmetry captures this difference, with the coarse model as A and the fine model as B we expect $\Delta R^2 < 0$. The mixed specialists provide a stronger localization test: when comparing, for example, a bird specialist (A) with a dog specialist (B), we expect bird images to have $\Delta q_i > 0$, dog images to have $\Delta q_i < 0$, and the remaining classes to show little directional asymmetry. Since the fitted maps are affected by the evaluation population, we additionally test whether localization persists when fitting only on the two affected superclasses, rather than the full held-out eval set.

2) Training-set coverage. We train models on overlapping ImageNet-100 classes, with a reference model A and variants B sharing 0%, 25%, 50%, 75%, or 100% of their classes. The maps are fit on held-out images from the union of both models' training classes, so the analysis population contains A-private, B-private, and shared classes. If predictive asymmetry traces differential training

exposure, we expect $\Delta q_i > 0$ for images from A-only classes, $\Delta q_i < 0$ for images from B-only classes, and little directional asymmetry for shared classes.

3) Training progression. We compare checkpoints from the same training run, holding architecture, objective, and training data fixed while varying only training duration. This setting tests whether predictive asymmetry orders representations by training progress. If additional training produces relatively unrecovered structure, an earlier Model A and later Model B should yield $\Delta R^2 < 0$.

Alternative Metrics. We compare predictive asymmetry against five per-observation alternatives. Two are directional and therefore comparable to Δq_i: *information imbalance* (Acevedo et al., 2025), the asymmetry in how well nearest neighbors in one representation remain neighbors in the other, and *local participation ratio*, the difference in effective dimensionality of each image's neighborhood measured separately in each representation (Amsaleg et al., 2022), which tests whether asymmetry is explained by local dimensionality alone. Three are symmetric and can localize a change but cannot attribute it to a model: *symmetric predictivity*, an ablation of our own score that averages the two directions instead of differencing them; *mutual-kNN disagreement* (Klabunde et al., 2025), the inverse shared-neighbor fraction, and *PNKA* (Kolling et al., 2023), cosine disagreement between kernel rows. Full details are in Appendix A.3.

Evaluating recovery of known training changes. Because each controlled intervention specifies the expected source of the representational difference, we can directly test whether each metric recovers it. For supervision granularity and training-set coverage, each held-out sample has two ground-truth labels: whether it belongs to a population treated differently by the two models and, if so, which model received the finer supervision or class exposure. We evaluate *detection* by asking whether score magnitude separates affected from unaffected samples (AUROC), and *attribution* by asking whether score sign points to the model that received the intervention (sign accuracy). For training progression, we report the fraction of checkpoint pairs for which the global score identifies the later checkpoint. Symmetric metrics are evaluated only on detection.

4.1.2 CLIP FINE-TUNING: HOW ADAPTATION AFFECTS PREDICTIVE ASYMMETRY

The ImageNet controlled experiments test whether predictive asymmetry recovers known training differences between independent models, but not between a foundation model and its fine-tuned counterpart. We isolate this by starting from the same CLIP DataComp ViT-L/14 parent (Radford et al., 2021; Gadre et al., 2023) and fine-tuning on the same iNat21 bird images with the same optimization, varying only the supervision: **1) SPECIES** uses 427 fine-grained species labels, **2) ORDER** uses 26 coarser taxonomic "order" level labels, and **3) RANDOM** uses a random 26-way partition.

Species supervision preserves species-level distinctions, whereas Order groups species according to an existing taxonomic hierarchy. Random uses the same 26-class size profile as Order but groups species randomly. Thus, if parent-favored asymmetry follows simply from using a coarse 26-way objective, Order and Random should behave similarly; if instead it reflects collapsing distinctions already organized by the pretrained model, the two may diverge. We compare each checkpoint to the shared parent throughout training.

4.2 APPLICATION TO RELEASED DOMAIN SPECIALISTS

We next apply predictive asymmetry where the source of representational change is not known in advance. We study twelve publicly released domain-adapted vision–language models spanning six domains: pathology (PLIP (Huang et al., 2023), QuiltNet (Ikezogwo et al., 2025)), biodiversity (BioCLIP (Stevens et al., 2024), BioTrove (Yang et al., 2025), TaxaBind (Sastry et al., 2024)), remote sensing (GeoRSCLIP (Zhang et al., 2024), RemoteCLIP (Liu et al., 2024), RSDiX (Terlizzi et al., 2025), SkyCLIP (Wang et al., 2023)), camera traps (WildCLIP (Gabeff et al., 2024)), fashion (FashionCLIP (Chia et al., 2022)), and street imagery (StreetCLIP (Haas et al., 2023)). We focus on models intended to provide broad domain representations rather than models fine-tuned for a single downstream task. Each specialist is paired with its foundation-model parent; where available, multiple independently trained specialists within a domain let us compare which representational patterns recur across adaptation recipes. We also note that TaxaBind inherits BioCLIP initialization, making their classwise agreement non-independent and we recognize this in Section 6.2; RSDiX is not fully independent of RESISC45, so its gains reflect in-distribution specialization rather than held-out transfer.

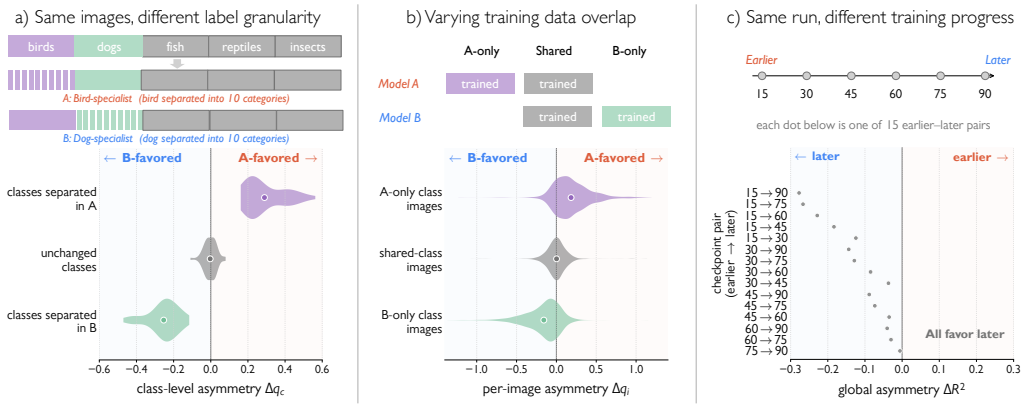

Figure 2: **Predictive asymmetry recovers the source of representational change across three ImageNet interventions.** Blue/left indicates Model B holds unpredicted variation; red/right indicates Model A does. (a) **Granularity:** Class-level asymmetry (Δq_c) correctly attributes specific classes to the model that received fine-grained supervision for them. (b) **Coverage:** Per-image asymmetry (Δq_i) correctly attributes private images to the model that observed them during training, without using class labels. (c) **Progression:** Global asymmetry (ΔR^2) correctly identifies the later checkpoint as the harder-to-predict representation across all 15 evaluated pairs.

Method	Supervision granularity		Training-set coverage		Progression
	Detect changed class (AUROC)	Identify fine-labeled model (bal. sign acc.)	Detect private image (AUROC)	Identify trained model (bal. sign acc.)	Identify later checkpoint (acc.)
Directional Methods					
Pred. asymmetry (ours)	**.997**$\pm$.003	**.996**$\pm$.006	**.722**$\pm$.033	**.787**$\pm$.028	**15/15**
Info. imbalance	.924$\pm$.023	.924$\pm$.030	.618$\pm$.036	.491$\pm$.011	14/15
Local PR	.548$\pm$.051	.503$\pm$.051	.566$\pm$.042	.335$\pm$.054	**15/15**
Symmetric Methods					
Sym. predictivity	.989$\pm$.006	N/A	.718$\pm$.052	N/A	N/A
Mutual-kNN	.771$\pm$.048	N/A	.654$\pm$.036	N/A	N/A
PNKA	.816$\pm$.068	N/A	.643$\pm$.062	N/A	N/A

Table 1: **Baseline comparison on ImageNet controlled interventions.** This table evaluates whether representation metrics can successfully identify *where* models differ and *which* model received additional training. **Granularity:** "Detect" isolates fine-grained classes; "Identify" attributes the change to the correct model. **Coverage:** "Detect" isolates images seen by only one model; "Identify" attributes them to the model that saw them during training. **Progression:** Accuracy identifying the later checkpoint across 15 pairs. Detection uses AUROC. Attribution uses balanced sign accuracy at the fixed zero threshold, with exact zeros counted as errors. Symmetric methods lack directional scores and cannot assign direction or order checkpoints. $\pm$ is 95% bootstrap half-width. **Bold** = best per task.

Because predictive asymmetry is population-dependent, we evaluate each parent–specialist pair on an in-domain benchmark where the distinctions targeted by specialization are present. Class labels are not used to fit the cross-predictive maps or compute asymmetry, but are introduced afterward for classwise and downstream analyses. To assess robustness, we additionally test whether the results persist under class-balanced evaluation populations and matched-rank PCA controls. Where training-corpus metadata are available, we additionally relate local asymmetry to training exposure. Table 4 in Appendix A.4 gives the parent and evaluation-set details for each model.

5 Controlled Experiments

5.1 ImageNet controls: predictive asymmetry recovers training differences

Supervision granularity. The fine–coarse comparison shows the expected global asymmetry: the fine model contains substantial variation that is not linearly recoverable from the coarse model ($\Delta R^2 \approx -.49$). More importantly, the mixed specialists localize this difference. Across pairwise comparisons, classes finely supervised only in A are consistently A-favored, classes finely super-

vised only in B are B-favored, and identically supervised classes remain near zero (Figure 2a). Because the paired models match in architecture, training images, and total task granularity, the sign isolates *which population* received the finer labels rather than a global difference in task difficulty. Table 1 quantifies the corresponding localization and attribution performance. This localization is also stable to the fitting population: restricting the maps from all 50 classes to only the 20 classes whose supervision differs preserves the detection and localization (99.3%; Appendix B.4).

Training-set coverage. Predictive asymmetry also recovers differential data exposure (Figure 2b). Across the overlap sweep, held-out images from A-private classes are A-favored, images from B-private classes are B-favored, and shared classes remain near zero. Thus, local asymmetry identifies both which images correspond to differential training exposure and which model observed their class. Quantitative detection and attribution results are reported in Table 1.

Training progression. Across all 15 earlier–later checkpoint pairs from the same training run, predictive asymmetry correctly identifies the later checkpoint (Figure 2c).

Comparison with alternative metrics. Table 1 shows that predictive asymmetry is the only alternative that performs strongly across all five controlled tasks. Symmetric metrics can detect that a change occurred but cannot assign it to a model, and symmetric predictivity nearly matches ours on detection alone. Among directional baselines, Information Imbalance weakens on coverage attribution, while local PR orders checkpoints well but is near chance on localization and attribution. Dimensionality alone therefore does not explain the signal.

5.2 CLIP FINE-TUNING: SUPERVISION DETERMINES REPRESENTATIONAL ASYMMETRY

Controlled CLIP fine-tuning shows that changing only the supervision target is sufficient to reverse the direction of predictive asymmetry. Species supervision becomes specialist-favored ($\Delta R^2 < 0$), while order supervision becomes parent-favored ($\Delta R^2 > 0$; Figure 3 (left)). This is not simply an effect of the number of training classes: RANDOM matches ORDER in label count and class-size profile yet is strongly specialist-favored. Thus, direction depends on which distinctions the label partition preserves or suppresses relative to those already accessible in the parent.

Downstream probes support this interpretation. From a 66.8% species / 84.2% order parent, species fine-tuning improves both to 77.1% / 89.8%; order fine-tuning raises order performance to 91.5% but reduces species performance to 62.7%, below the parent; and RANDOM improves both to 73.1% / 87.4%. Asymmetry tracks the preservation or compression of fine-grained distinctions.

6 STUDYING DOMAIN-SPECIALIZED REPRESENTATIONS

We now apply the same analysis as a diagnostic tool to the published domain foundation models described in Section 4.2. We explore 1) which direction does predictive asymmetry take for parent-specialist pairs across domains? 2) what factors are associated with these local asymmetry patterns? and 3) do these asymmetries correlate with relative downstream performance?

6.1 RELEASED SPECIALISTS SHOW HETEROGENEOUS PREDICTIVE ASYMMETRY

We observe that there is no single consistent "direction of predictive asymmetry" with respect to their parent across domain foundation models, as shown in Figure 3 (right). For the three biodiversity foundation models, the parents recover less of the specialist embeddings than vice versa ($\Delta R^2 = -0.179$ to -0.160), which we term "specialist-favored", while two of the remote sensing foundation models show the opposite trend ("parent-favored") ($\Delta R^2 = 0.094$). While remaining models show globally symmetric trends between parent and specialist (within the 95th percentile of cross-seed same objective variations), some (RemoteCLIP and WildCLIP) still show substantial image-level asymmetries in both directions, while others (FashionCLIP, StreetCLIP, and SkyCLIP) instead have high mutual recoverability. Therefore this shows that ΔR^2 and $\bar{R}^2$ provide complementary views of representational change: direction versus overall recoverability. Because these patterns could depend on the evaluation population or representation capacity, we test both directly: classwise patterns remain stable under population rebalancing (median $\rho = .998$) and after reducing both representations to 256 PCA dimensions ($\geq 50\%$ reduction; median $\rho = .974$; Appendix B.4).

The direction of specialist structure reflects the distinctions emphasized during training, consistent with our controlled CLIP experiments (Section 5.2). Fine-grained supervision yields specialist-

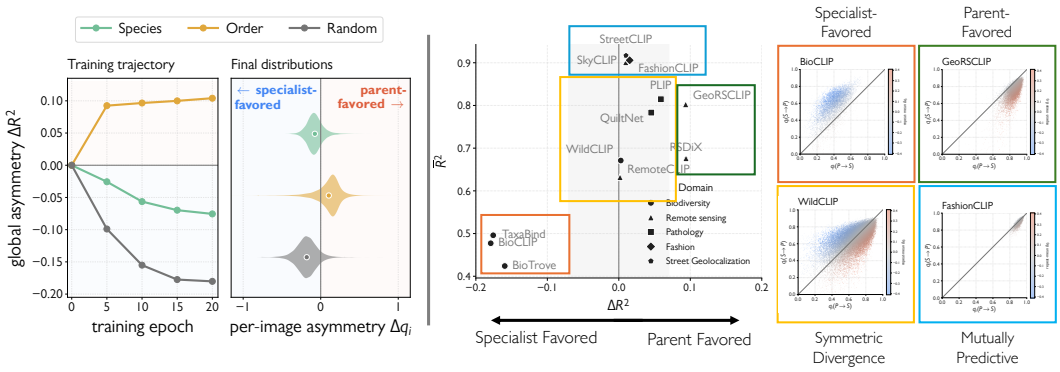

Figure 3: **(Left) Predictive asymmetry reflects which distinctions fine-tuning preserves or compresses.** ORDER collapses species distinctions, yielding parent-favored asymmetry, while others yield specialist-favored asymmetry. **(Right) Domain specialization varies in both reciprocal predictability and direction.** Each point denotes a released parent–specialist pair. The vertical axis shows mean bidirectional recoverability $\bar{R}^2$, while the horizontal axis shows directional asymmetry ΔR^2. Released specialists occupy qualitatively different regions of this space.

favored asymmetry, while collapsing within-group distinctions yields parent-favored asymmetry. Accordingly, specialist-favored biological models use highly fine-grained taxonomic supervision (tens to hundreds of thousands of species) (Stevens et al., 2024; Yang et al., 2025; Sastry et al., 2024), whereas parent-favored remote-sensing models use much coarser semantic groupings (tens of scene categories across millions of images) or objectives that encourage invariance among similar examples (Zhang et al., 2024; Terlizzi et al., 2025). Thus, specialization is consistent which distinctions training preserves versus collapses, rather than simply adding linearly accessible structure.

6.2 Local asymmetry reveals semantic patterns across domain specialists

We find that independently trained specialists within the same domain largely agree on which classes are relatively parent- or specialist-favored, even when their global asymmetry differs. Classwise asymmetry is strongly correlated across biological ($\rho = 0.80$–0.98), pathology ($\rho = 0.87$), and remote-sensing ($\rho = 0.57$–0.85) specialists. The highest agreement, BioCLIP–TaxaBind ($\rho = .98$), is expected to be less independent because TaxaBind inherits BioCLIP initialization. Even though GeoRSCLIP is globally parent-favored while RemoteCLIP is near symmetric, their relative class ordering is similar ($\rho = 0.726$). These shared class orderings are also semantically structured: insects and amphibians tend to be more specialist-favored than mammals in biological models; diagnostically specific tissue compartments more specialist-favored than background, adipose, or mucus in pathology; and anthropogenic structures more specialist-favored than natural scenes in remote sensing. Therefore, local asymmetry reveals a reproducible domain-level organization even when global asymmetry differs across specialists.

For six specialists with available dataset statistics, we find that differences within this shared organization track relative training exposure between CLIP and the specialist. Classes that are more prevalent in the specialist's training data relative to a DataComp-1B proxy for parent exposure are more specialist-favored ($\rho = -0.30$ to -0.63; Figure 4), and in every model relative exposure tracks classwise asymmetry more than specialist or parent training frequency alone. Relative exposure also tracks differences between specialists: compared with GeoRSCLIP's RS5M training data, basketball courts, overpasses, and storage tanks are $5.5\times$, $3.2\times$, and $2.9\times$ more prevalent in RemoteCLIP's training mixture, and each shifts toward the specialist-favored end of the class ranking. Thus, specialists share a broad domain-level semantic organization, while differences in training exposure corresponding with which concepts each emphasizes.

6.3 Local representational direction tracks relative probe advantage

The preceding analyses use only representation geometry. We next ask whether this label-free geometric quantity identifies where the representations differ in downstream linear probing performance. We bin held-out images by specialist-favoring $-\Delta q_i$ and independently measure the specialist–parent linear-probe accuracy gap within each bin and find that image-level asymmetry local-

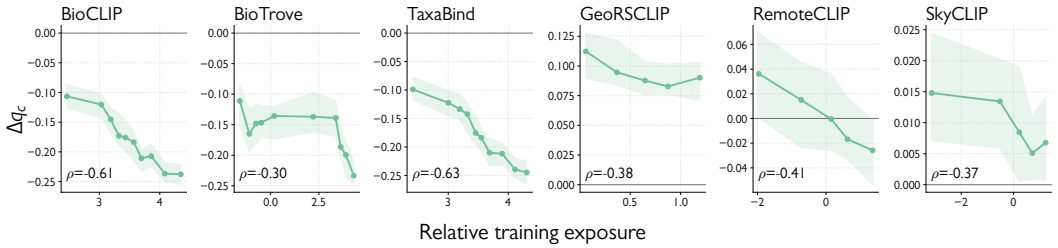

Figure 4: Δq_c **aggregated over classes tracks relative training exposure between specialist and parent.** Δq_c shows stronger correlation with *relative* exposure than training data frequency or CLIP prior/DataComp-1B frequency alone.

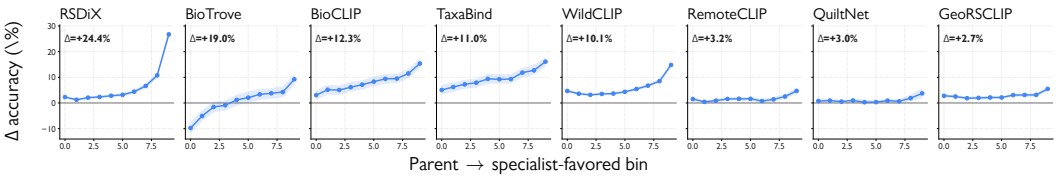

Figure 5: q_i **stratifies where the specialist's relative advantage concentrates.** Images are binned by specialist-favoring asymmetry; the vertical axis shows the specialist-minus-parent probe accuracy gap. Panels show the eight pairs whose gap widens toward the specialist-favoring end. Remaining four pairs show little change in relative probe gap across asymmetry. Full results are in Figure 12

izes where a specialist's performance advantage over its parent is concentrated. For the eight pairs with substantial directional asymmetry, the specialist–parent accuracy gap is larger at the specialist-favoring end (Figure 5). The strongest effect is RSDiX, where the gap grows from approximately $+2$ to $+26$ percentage points, and for BioTrove reverses from a 10-point parent advantage to a 9-point specialist advantage.

Predictive asymmetry distinguishes cases where the parent and specialist perform similarly across the whole dataset from cases where similar overall accuracy masks opposing local strengths. BioTrove, for example, is only 0.7 pp better than its parent overall, yet shifts $\sim$ 19-point in relative advantage across the asymmetry axis. By contrast, near-symmetric pairs with high mutual recoverability such as FashionCLIP, StreetCLIP, and SkyCLIP show little change in relative accuracy across this axis, confirming that the absence of representational asymmetry corresponds to the absence of localized behavioral divergence.

Mutual predictivity, on the other hand, identifies data regimes where the two models are more likely to disagree, tracking it in 11 of 12 pairs. Together, the two signals separate *where behavior changes* from *which model is relatively stronger*, enabling a more informative understanding of specialization. Full classwise and accuracy-decomposition analyses are reported in Appendix B.2.

7 Limitations and Conclusions

Limitations. Predictive asymmetry measures linear accessibility, not information or usefulness Xu et al. (2020): nonlinearly encoded structure may appears unpredictable, and relatively added or removed variation may not always be useful for downstream performance (i.e., Random 26-class model in Sec. 5.2 produced the same specialist-favored sign as species supervision). Also, the exposure analysis covers only 6 of 12 specialists and relies on DataComp-1B captions as a proxy for parent exposure. Finally, released specialists differ from their parents in data, scale, and objective at once, so those results are correlational and the causal claims rest on the controlled experiments.

Conclusion. Where one representation fails to linearly predict another, the failure carries information about how the two models differ. Predictive asymmetry stores structured evidence of difference between two models' representations. We first validate the diagnostic through controlled experiments and then apply it to domain specialized foundation models to analyze how it can recover the effects of fine-tuning on the representations. Bidirectional linear prediction thus offers a simple, label-free way to audit what adaptation changed in a representation, and where. Future work can extend the study of predictive asymmetry across modalities and across models of different scales.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing in the main manuscript and appendix, including improving clarity, grammar, and presentation, as well as assisting with LaTeX formatting. We also used generative AI tools for retrieval and discovery, including identifying potentially relevant prior work and datasets. In addition, generative AI tools were used for research execution, including assistance with code generation for implementing and evaluating research ideas.

We reviewed all AI-assisted work. Suggested related work, datasets, and other resources were independently checked by the authors before inclusion or use. AI-generated or AI-assisted code was reviewed, verified, and tested for correctness by the authors. AI-assisted writing were also manually reviewed and revised by all the authors.

We take responsibility for the final content of this work, including all text, claims, code, results, and other artifacts produced with the aid of generative AI.

REFERENCES

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images. *arXiv preprint arXiv:2505.17101*, 2025.

Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. Intrinsic dimensionality estimation within tight localities: A theoretical and experimental analysis, 2022. URL https://arxiv.org/abs/2209.14475.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Ahmad Taher Azar, Mohamed Tounsi, Suliman Mohamed Fati, Yasir Javed, Syed Umar Amin, Zafar Iqbal Khan, Shrooq Alsenan, and Jothi Ganesan. Automated system for colon cancer detection and segmentation based on deep learning techniques. *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, 15(1):1–28, 2023.

András Balogh and Márk Jelasity. How not to stitch representations to measure similarity: Task loss matching versus direct matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15472–15480, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi: 10.1109/jproc.2017.2675998. URL http://dx.doi.org/10.1109/JPROC.2017.2675998.

Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, Nov 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-23052-9. URL https://doi.org/10.1038/s41598-022-23052-9.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023.

Reuben Fernandes. Streetview global: annotated street-level imagery with scene descriptions and vqa, sourced from mapillary. https://huggingface.co/datasets/Reubencf/streetview-global, 2026.

Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9):3770–3786, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. URL https://arxiv.org/abs/2304.14108.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023. URL https://arxiv.org/abs/2302.00275.

Adriano Hernandez, Rumen Dangovski, Peter Y. Lu, and Marin Soljacic. Model stitching: Looking for functional similarity between representations, 2023. URL https://arxiv.org/abs/2303.11277.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology, 2025. URL https://arxiv.org/abs/2306.11207.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2026. URL https://arxiv.org/abs/2505.12540.

Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms. *arXiv preprint arXiv:2602.11729*, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Camila Kolling, Till Speicher, Vedant Nanda, Mariya Toneva, and Krishna P Gummadi. Investigating the effects of fairness interventions using pointwise representational similarity. *arXiv preprint arXiv:2305.19294*, 2023.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. In *International Conference on Learning Representations*, volume 2025, pp. 12850–12881, 2025.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational difference explanations. *Advances in Neural Information Processing Systems*, 38:71478–71527, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Samuel Kostousov, Abhinn Kaushik, and Brokoslaw Laschowski. Bidirectional representational alignment between biological and artificial neural networks. *arXiv preprint arXiv:2608.18244*, 2026.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. URL https://arxiv.org/abs/2202.10054.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Yiming Liu, Yuhui Zhang, Dhruba Ghosh, Ludwig Schmidt, and Serena Yeung-Levy. Data or language supervision: What makes clip better than dino?, 2025. URL https://arxiv.org/abs/2510.11835.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Mapillary contributors. Mapillary: Street-level imagery, powered by collaboration and computer vision. https://www.mapillary.com/. Images licensed under CC BY-SA 4.0.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Sabine Muzellec and Kohitij Kar. Reverse predictivity for bidirectional comparison of neural networks and biological brains. *Nature Machine Intelligence*, 8(3):474–488, 2026.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning?, 2021. URL https://arxiv.org/abs/2008.11687.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things. *Uncovering How Neural Network Representations Vary with Width and Depth*, pp. 1–25, 2020.

Young-Jin Park, Hao Wang, Shervin Ardeshir, and Navid Azizan. Quantifying representation reliability in self-supervised learning models, 2024. URL https://arxiv.org/abs/2306.00206.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022. URL https://arxiv.org/abs/2108.08810.

Arjun Rao, Marc Rußwurm, Konstantin Klemmer, and Esther Rolf. Measuring the intrinsic dimension of earth representations. In *International Conference on Learning Representations*, volume 2026, pp. 113200–113224, 2026.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications, 2024. URL https://arxiv.org/abs/2411.00683.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms, 2022. URL https://arxiv.org/abs/2211.12491.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life, 2024. URL https://arxiv.org/abs/2311.18803.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Andrea Terlizzi, Angelo Nazzaro, Lorenzo Bernardi, Francesco Bardozzo, and Roberto Tagliaferri. Rsdix: Lightweight and data-efficient vlms for remote sensing through self-distillation. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE, 2025.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Johanna Vielhaben, Dilyara Bareeva, Jim Berend, Wojciech Samek, and Nils Strodthoff. Beyond scalars: Concept-based alignment analysis in vision transformers. *Advances in Neural Information Processing Systems*, 38:71379–71414, 2026.

Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing, 2023. URL https://arxiv.org/abs/2312.12856.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling, 2022. URL https://arxiv.org/abs/2205.13543.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints, 2020. URL https://arxiv.org/abs/2002.10689.

Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K. Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, Arti Singh, Soumik Sarkar, Nirav Merchant, Chinmay Hegde, and Baskar Ganapathysubramanian. Biotrove: A large curated image dataset enabling ai for biodiversity, 2025. URL https://arxiv.org/abs/2406.17720.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9611–9620. IEEE, 2025.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. ISSN 1558-0644. doi: 10.1109/tgrs.2024.3449154. URL http://dx.doi.org/10.1109/TGRS.2024.3449154.

A Implementation Details

A.1 Preprocessing and feature stage

For all experiments, we use the final pooled representation from the visual encoder before the classifier head or CLIP image-projection layer. For the controlled torchvision ResNet-50 models, we use the 2,048-dimensional vector after global average pooling and before the classification head. For OpenCLIP models, including the Section 5.2 models, we use the pooled vision-tower output before `visual.proj`. For Hugging Face CLIP models, we use `vision_model.pooler_output` before `visual_projection`. SkyCLIP is the only model in these sections loaded with timm; we use its native class-token-pooled pre-logits representation, returned by `forward_head(..., pre_logits=True)`, before the checkpoint classifier.

In Section 6, BioCLIP, BioTrove, TaxaBind, RemoteCLIP, GeoRSCLIP, and WildCLIP use OpenCLIP loaders; PLIP, QuiltNet, FashionCLIP, RSDiX, and StreetCLIP use Hugging Face CLIP loaders; and SkyCLIP uses timm. Before fitting the maps, we scale each feature vector to unit Euclidean norm.

A.2 Ridge estimator and cross-validation

We follow the column-vector convention of Section 3: for direction $A \to B$, $W_f x_i + b_f$ predicts y_i, with $W_f \in \mathbb{R}^{d_B \times d_A}$.

Preprocessing. Each feature vector is normalized to unit Euclidean norm. The training-set mean is subtracted after normalization and recomputed for every candidate fit and final refit.

Objective. For an outer fold f with training set $\mathcal{T}_f$, define the source covariance $C_f = |\mathcal{T}_f|^{-1} \sum_{i \in \mathcal{T}_f} (x_i - \bar{x}_f)(x_i - \bar{x}_f)^\top$ and variance factor $\tau_f = \mathrm{tr}(C_f)/d_A$. The ridge map minimizes

$$\min_{W,b} \frac{1}{n_f} \sum_{i \in \mathcal{T}_f} \|y_i - W x_i - b\|_2^2 + \lambda_f \tau_f \|W\|_F^2, \qquad n_f = |\mathcal{T}_f|, \tag{2}$$

where τ_f puts the regularization penalty on a comparable scale across source representations. We solve the closed-form normal equations in float64, with intercept $\hat{b}_f = \bar{y}_f - \hat{W}_f \bar{x}_f$.

Cross-validation. We use five-fold cross-validation. For each outer fold f, one of the remaining four folds is held out as an inner validation set for selecting λ_f, and the other three are used to fit candidate models. We choose λ_f from $\{10^k : k = -8, \ldots, 6\}$ by mean squared Euclidean prediction error on the inner validation fold. We break ties in favor of the smaller value and allow values at the edge of the search grid. We then refit (W_f, b_f) on all of $\mathcal{T}_f$ and record residuals on the held-out fold $\mathcal{V}_f$ without renormalizing the mapped predictions. The reverse direction is fit independently using the same fold splits. Every reported residual is computed out of fold; held-out data are never used to fit parameters or select hyperparameters.

A.2.1 Linear probing implementation

We fit ℓ_2-regularized multinomial logistic regression (scikit-learn LBFGS) using the same five-fold structure as the affine maps: three folds fit the model, one selects regularization strength $C \in \{10^{-3}, \ldots, 10^2\}$ by macro-averaged NLL, and the held-out fold provides the predictions we report. Features are row-ℓ_2-normalized with balanced class weights. We evaluate four variants per fold: the native A and B representations and the two mapped representations, after re-normalizing the cross-validated predictions to unit norm.

A.2.2 Controlled experiment setup

Shared ImageNet training configuration. All three controls use torchvision ResNet-50 models trained from scratch for 90 epochs with SGD, batch size 256, initial learning rate 0.1, momentum 0.9, weight decay 10^{-4}, label smoothing 0.1, and cosine learning-rate decay. We save checkpoints at epochs $15, 30, 45, 60, 75, 90$; unless stated otherwise, analyses use epoch 90.

Superclass (WordNet root; eligible leaves)	Selected ImageNet-1K leaf classes (WordNet ID)
Dogs (n02084071; 118)	Pekinese (n02086079), bluetick (n02088632), black-and-tan coonhound (n02089078), redbone (n02090379), Staffordshire bull terrier (n02093256), Border terrier (n02093754), Norwich terrier (n02094258), Border collie (n02106166), miniature pinscher (n02107312), French bulldog (n02108915).
Birds (n01503061; 59)	Hen (n01514859), ostrich (n01518878), house finch (n01532829), water ouzel (n01601694), sulphur-crested cockatoo (n01819313), toucan (n01843383), goose (n01855672), limpkin (n02013706), oystercatcher (n02037110), king penguin (n02056570).
Fish (n02512053; 16)	Tench (n01440764), tiger shark (n01491361), stingray (n01498041), barracouta (n02514041), eel (n02526121), coho salmon (n02536864), rock beauty (n02606052), anemone fish (n02607072), lionfish (n02643566), pufferfish (n02655020).
Reptiles (n01661091; 36)	Loggerhead turtle (n01664065), box turtle (n01669191), common iguana (n01677366), whiptail lizard (n01685808), frilled lizard (n01688243), American alligator (n01698640), hognose snake (n01729322), garter snake (n01735189), night snake (n01740131), horned viper (n01753488).
Insects (n02159955; 27)	Ladybug (n02165456), ground beetle (n02167151), long-horned beetle (n02168699), fly (n02190166), grasshopper (n02226429), cricket (n02229544), cockroach (n02233338), dragonfly (n02268443), ringlet butterfly (n02277742), lycaenid butterfly (n02281787).

Table 2: **Superclass composition of the supervision-granularity control.** "Eligible leaves" is the number of ImageNet-1K leaves below the specified WordNet root before sampling; every experiment uses the same 10 selected leaves per superclass.

Supervision granularity. Before training, we select 10 ImageNet-1K leaf classes from each of five WordNet superclasses (Table 2). We selected these superclasses because each has at least 10 eligible leaf classes and forms a clean semantic grouping. All objectives therefore use the same 50 classes and images.

The fine objective has 50 labels, the coarse objective has five superclass labels, and each of the five mixed objectives retains the 10 fine labels for one superclass while collapsing each of the other four superclasses to one label, for 14 labels total.

We trained six fine and six coarse runs (seeds 0–5) and five runs for each mixed objective (seeds 0–4). The mixed-specialist benchmark uses all 10 unordered objective pairs at each of the five matched seeds, yielding 50 cells. Each run uses 63,609 ImageNet training images. We also have a held-out 2,000-image evaluation set (40 per class; five folds of 400), which all paper results use.

Training-set coverage. We use the Hugging Face dataset `ilee0022/ImageNet100`, whose train and validation splits contain 117,000 and 13,000 images over 100 classes. We used commit `c55b2f2967c034db17be30f7d430e41c80fd4281` of the Hugging Face dataset. Because the dataset was not loaded with an explicit revision, we additionally identify samples by source split/index and image SHA-256 in the stored manifests. From the 117,000-image training split, we construct deterministic per-class optimization/validation splits with seed 20260825. Using seed 42, we assign model A a fixed random set of 50 classes. Each model B is also assigned 50 classes, with overlap levels of $0, 25, 50, 75$, or 100% with A. Since 25 and 75 percent of 50 are non-integers, these conditions use 13 and 37 shared classes, respectively (Table 3). All primary overlap models use training seed 0 and the epoch-90 checkpoint.

The evaluation population is drawn only from the dataset's 13,000-image validation split. We use 80% of it (10,442 images), with the other 20% held-out for any other additional analysis. For each overlap condition, map fitting and evaluation use the development images whose classes lie in the union of the two training datasets; so the analyzed population shrinks as overlap increases. Training support labels (shared, A-private, or B-private) are used only for evaluation and never for fitting the maps.

Training overlap	Shared	A-private	B-private	A train/val	B train/val	Evaluation
0%	0	50	50	52,630/5,827	52,713/5,830	10,442
25%	13	37	37	52,630/5,827	52,725/5,834	9,096
50%	25	25	25	52,630/5,827	52,737/5,838	7,835
75%	37	13	13	52,630/5,827	52,696/5,833	6,585
100%	50	0	0	52,630/5,827	52,630/5,827	5,255

Table 3: **ImageNet-100 class-overlap design.** Train/val counts are the optimization and internal-validation images for each model. The last column gives the evaluation images used by the representation analysis.

Training progression. All 15 reported comparisons are the $\binom{6}{2}$ earlier–later pairs from one run, not 15 independent training runs. Specifically, they use the seed-0 `model_A` ResNet-50 from the ImageNet-100 coverage experiment, trained on its fixed 50-class roster with 52,630 optimization images and 5,827 internal-validation images. We compare its exact checkpoints at epochs $15, 30, 45, 60, 75$, and 90 while holding architecture, class roster, image set, labels, seed, and optimizer trajectory fixed. Every pair is evaluated on the same 2,000-image, five-fold evaluation set used by the granularity control; the remaining 500 images are held out for additional analyses. Thus, this experiment tests whether the metric correctly orders checkpoints from the same training run, and the 15 pairwise outcomes should not be read as independent replicates.

A.3 Baseline details

We compare predictive asymmetry against five per-observation alternatives. All methods use the same aligned evaluation images and the same orientation of models A and B. For directional metrics, we use score magnitude for detection and score sign for attribution. Symmetric scores are evaluated only on detection. Except for PNKA, the geometry-based baselines use row-ℓ_2-normalized representations.

Information imbalance (II). Following Acevedo et al. (2025), let $n_A(i)$ be the nearest non-self neighbor of image i in representation A, and let $r^B_{i,j}$ be the rank of image j among the non-self neighbors of i in representation B. We define

$$I_i(A\to B) = \frac{2}{N-1}\, r^B_{i,n_A(i)}, \qquad \Delta I_i = I_i(B\to A) - I_i(A\to B).$$

Positive ΔI_i favors model A, matching the sign convention of Δq_i. We use $|\Delta I_i|$ for detection.

Local participation ratio (PR). For each image and representation, we select its 20 nearest non-self neighbors under cosine distance from within the same ground-truth class. We combine the query with these neighbors, center the resulting neighborhood at its own mean, and compute the covariance eigenvalues $\{\lambda_{ij}\}$. The local participation ratio is

$$d^{\mathrm{PR}}_i = \frac{\left(\sum_j \lambda_{ij}\right)^2}{\sum_j \lambda^2_{ij}}, \qquad \Delta d^{\mathrm{PR}}_i = d^{\mathrm{PR}}_A(i) - d^{\mathrm{PR}}_B(i).$$

Positive Δd^{PR}_i favors model A, and we use $|\Delta d^{\mathrm{PR}}_i|$ for detection. Because this baseline is computed separately in each representation, it does not measure correspondence between them.

Symmetric predictivity. This ablation uses the same cross-validated maps as predictive asymmetry but removes direction:

$$s^{\mathrm{pred}}_i = 1 - \tfrac{1}{2}\left(q^{A\to B}_i + q^{B\to A}_i\right).$$

Mutual-kNN. Using the exact $k = 10$ nearest non-self neighbors under cosine distance, we define

$$s^{\mathrm{mKNN}}_i = 1 - \frac{|\mathcal{N}^A_{10}(i) \cap \mathcal{N}^B_{10}(i)|}{10}.$$

PNKA. Following Kolling et al. (2023), let $X_{A,c}$ and $X_{B,c}$ denote the raw encoder representations after centering each feature column over the evaluation population. We form the linear kernels

$$K_A = X_{A,c}X_{A,c}^\top, \qquad K_B = X_{B,c}X_{B,c}^\top,$$

and define pointwise kernel disagreement as

$$s_i^{\mathrm{PNKA}} = 1 - \frac{\langle K_A[i,:], K_B[i,:]\rangle}{\|K_A[i,:]\|_2 \, \|K_B[i,:]\|_2}.$$

PNKA is symmetric and is therefore used only for detection.

A.3.1 CLIP FINE-TUNING: INTERPRETING THE DIRECTION OF PREDICTIVE ASYMMETRY

Here we test what determines the *direction* of asymmetry during fine-tuning, asking whether different supervision targets can push the same pretrained representation in opposite directions.

We select 427 bird species from 26 taxonomic orders in the iNaturalist 2021 training split. For each species, 50 images are withheld before fine-tuning, giving 101,064 training images and 21,350 held-out images. We start from the same OpenCLIP 3.0.0 DataComp ViT-L/14 parent (`datacomp_xl_s13b_b90k`) and fine-tune three models on the same training population with the same objective and optimization schedule, changing only the label partition:

- **Species** (427 classes): fine-grained supervision that requires the model to distinguish individual species. If fine-tuning introduces species-level representational variation that cannot be well predicted from the parent representation, we expect specialist-favored asymmetry.
- **Order** (26 classes): coarse taxonomic supervision that does not require distinctions between species within the same order. If fine-tuning suppresses some of those distinctions while the parent retains them, we expect parent-favored asymmetry.
- **Random26** (26 classes): a fixed seed-0 permutation of species into 26 pseudo-classes. It matches Order in label cardinality, order-name caption vocabulary, and the number of species assigned to each class, while removing the taxonomic relationship between species and labels. This tests whether the direction depends on which distinctions the labels encourage the model to preserve, rather than simply on using a 26-class objective.

All arms use captions of the form `a photo of {taxon}, a type of bird.` and a symmetric multi-positive CLIP loss: within the distributed global batch, every image–text pair sharing a class caption is treated as a positive, with probability mass distributed uniformly across those positives in both directions. We freeze the text tower and update the image tower. To prevent taxonomic imbalance, each model uses inverse-frequency class-balanced sampling with replacement.

We train in bfloat16 mixed precision for 20 epochs on four GPUs with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$), a global batch size of 256, learning rate 10^{-5}, weight decay 0.1, 500 linear-warmup steps, and cosine learning-rate decay. We save checkpoints at epochs 5, 10, 15, and 20. The training curves reported in the paper use seed 0 and compare each checkpoint to the unchanged parent.

The evaluation population comprises the 21,350 images withheld from the 427 trained species. We assign 17,080 of these images to the fixed evaluation population used in the reported five-fold out-of-fold maps and native probes, while reserving 4,270 images as a held-out partition for further analysis. Species and order probes use the same five image folds as the affine maps. These probes test whether asymmetry relative to the parent tracks changes in how linearly separable fine- and coarse-grained distinctions are.

A.4 RELEASED SPECIALIST DETAILS

Table 4 lists the twelve pairs. Parents are `ViT-B-16-quickgelu` and `ViT-B-32-quickgelu` (OpenAI CLIP), `openai/clip-vit-large-patch14-336`, and two LAION ViT-B/32 checkpoints; we use the publicly released specialist checkpoints.

Three domains contain multiple independently trained specialists that share a parent and an evaluation set–biodiversity (three), remote sensing (four, of which three share a parent), and pathology (two).

Domain	Evaluation set	Reference	Specialists
Pathology	CRC histology (Azar et al., 2023)	CLIP B/32	PLIP, QuiltNet
Biodiversity	Rare Species (Stevens et al., 2024)	CLIP B/16	BioCLIP, BioTrove, TaxaBind
Remote sensing	RESISC45 (Cheng et al., 2017)	CLIP B/32 variants	GeoRSCLIP, RemoteCLIP, RS-DiX, SkyCLIP
Camera traps	Snapshot Serengeti (Gabeff et al., 2024)	CLIP B/16	WildCLIP
Fashion	DeepFashion (Liu et al., 2016)	LAION CLIP B/32	FashionCLIP
Street imagery	Mapillary (Mapillary contributors; Fernandes, 2026)	CLIP L/14@336	StreetCLIP

Table 4: **Released specialist comparisons.** Each specialist is evaluated against its corresponding foundation reference on one in-domain population; "variants" distinguishes the OpenAI and LAION B/32 references used by the remote-sensing releases. TaxaBind is initialized from BioCLIP rather than independently from the listed CLIP reference and is marked as shared-init in Table 10.

Estimating the seed-variation baseline. We estimate the reference band in Figure 3 by comparing models trained under the same condition with different random seeds. We form all seed pairs *within* these conditions, giving $6\binom{4}{2} + 7\binom{5}{2} = 106$ pairwise comparisons. We pool the 106 values of $|\Delta R^2|$ and use their empirical 95th percentile, 0.0698 (displayed as 0.070), as the seed-variation reference band.

A.5 Training-exposure analysis

This section describes the corpus counts and statistical analysis used for Figure 4. The corpus counts are computed independently of the representation analysis: they do not use R^2, evaluation images, or downstream predictions. Comparable class-level counts are available for six specialists— BioCLIP, TaxaBind, BioTrove, RemoteCLIP, GeoRSCLIP, and SkyCLIP—so the confirmatory analysis is restricted to those models.

A.5.1 Exposure score

For specialist s and evaluation class c, let $n_{s,c}$ be the number of specialist-training items matched to the class concept, N_s the size of the corresponding specialist pool, $n_{W,c}$ the number of matching captions in the parent-side web proxy, and N_W the size of that proxy pool. The reported relative exposure is the difference of smoothed log shares,

$$E_{s,c} = \log_{10}\left(\frac{n_{s,c} + 0.5}{N_s}\right) - \log_{10}\left(\frac{n_{W,c} + 0.5}{N_W}\right). \tag{3}$$

The 0.5 pseudo-count is applied to both numerators, including zero-count classes. Denominators are not incremented. Thus $E_{s,c} = 0$ means equal corpus share, rather than equal raw count; positive values mean that the concept is over-represented in specialist training relative to the web proxy.

Biodiversity Exposure Calculation. For the biodiversity specialists, the unit of analysis is biological family. BioCLIP uses the TreeOfLife-10M training split. TaxaBind uses a proxy for cumulative training exposure: TreeOfLife-10M inherited through BioCLIP initialization, the 2,552,501-example iSatNat training population derived from geolocated iNaturalist-2021 observations, and the independently collected 74,910-example iSoundNat training split of iNaturalist observations containing both an image and audio. Each training record is counted once under its assigned biological family. We treat this as a proxy for exposure rather than an exact training mixture because the TaxaBind image encoder was sequentially patched across modality-specific fits. BioTrove uses the reconstructed 39.86M-image BioTrove-Train subset described in the released-corpus analysis. For all three specialists, the parent-side quantity is the share of DataComp-1B captions that explicitly name the same biological family.

Remote-sensing pools and counting unit. RemoteCLIP uses all RET-3, DET-10, and SEG-4 training records released with the model. We count an image once for class c if any associated caption contains a primary phrase for c after the matching procedure below. The component totals are 13,717, 110,799, and 41,171 images, respectively, giving $N_s = 165,687$; caption counts and RET-3 filename-derived scene labels are used only for validation and do not enter Eq. 3.

GeoRSCLIP uses RS5M. For the 2,869,256 web-derived `pub11` pairs, we count captions. For the curated `rs3` portion—990,848 MillionAID, 344,385 BigEarthNet, and 727,144 fMoW images—we count each image once if any of its five caption fields matches. These are combined with $N_s = 4,931,633$. SkyCLIP uses the 2,554,817-image, top-50%-by-CLIP SkyScript training list. We scan its `title` field, containing the central OpenStreetMap object's tags, because that is the text field used to train the released top-50%-by-CLIP checkpoint. The broader `title_multi_objects` count, which also includes surrounding objects, is retained for validation but is not used in the reported analysis.

The parent proxy for all three remote-sensing specialists is the same DataComp-1B scan. Before counting classes, we restrict it to captions containing at least one overhead-imagery context phrase listed below. This retains 2,046,169 of 1,387,173,656 captions (0.148%), so $N_W = 2,046,169$. A retained caption is counted once for every class whose primary vocabulary it matches; the denominator remains the number of retained captions, not the sum of class matches.

Text normalization and match precedence. For RemoteCLIP, RS5M, and DataComp, text is HTML-unescaped, NFKC-normalized, case-folded, and reduced to space-separated word characters. Matches require whole-token or whole-phrase boundaries. We also match the literal suffix variants `s` and `es`. For multiword phrases, we also match a concatenated spelling (for example, `storage tank` also matches `storagetank`) because the RemoteCLIP detection and segmentation captions use concatenated category names; the Aho–Corasick DataComp scanner adds `s` to this concatenated form, while the regular-expression corpus counters admit `s` and `es`. If an exclusion phrase appears, we do not count that class for the caption. Otherwise a primary match takes precedence over a broad match. The reported exposure axes use *only primary* matches; broader near-neighbor matches are recorded for validation but are not added to $n_{s,c}$ or $n_{W,c}$. SkyScript is processed with the same normalization and precedence, but uses the separate OpenStreetMap-rendered primary vocabulary listed below.

Exact overhead-context vocabulary. The DataComp remote-sensing prefilter is the union of the following 65 phrases: `aerial view`, `aerial image`, `aerial photo`, `aerial photograph`, `aerial photography`, `aerial shot`, `aerial imagery`, `aerial picture`, `aerial drone`, `satellite image`, `satellite images`, `satellite imagery`, `satellite view`, `satellite photo`, `satellite photograph`, `satellite picture`, `satellite map`, `satellite data`, `from above`, `seen from above`, `viewed from above`, `view from above`, `bird s eye view`, `birds eye view`, `bird eye view`, `birdseye view`, `top view`, `top down view`, `overhead view`, `overhead shot`, `overhead image`, `overhead photo`, `drone view`, `drone shot`, `drone photo`, `drone photograph`, `drone photography`, `drone image`, `drone footage`, `uav image`, `uav imagery`, `remote sensing`, `remotely sensed`, `orthophoto`, `orthoimage`, `orthomosaic`, `google earth`, `google maps satellite`, `bing maps`, `sentinel 2`, `sentinel-2`, `landsat`, `planet labs`, `maxar`, `worldview`, `spot 6`, `spot 7`, `pleiades`, `digitalglobe`, `digital globe`, `earth observation`, `earth from space`, `from space`, `nasa earth observatory`, and `copernicus`.

B Additional Results

This section reports the full controlled-experiment results and the released-specialist results referenced in Sections 5–6. Unless noted otherwise, all values are from the final analysis set. Confidence intervals resample the appropriate independent unit for each experiment while holding the trained checkpoints fixed.

B.1 Controlled experiments: full results

B.1.1 Complete controlled-granularity evaluation

Table 5 reports both global recoverability and evaluation metrics for every unordered pair of mixed-supervision objectives and all five matched training seeds.

Figures 6, 7, and 8 show the corresponding per-image predictivity patterns: the canonical fine–coarse comparison, each matched birds–dogs seed, and all ten mixed-objective pairs, respectively.

B.1.2 Complete controlled-coverage evaluation

Table 6 reports global recoverability and results separately for shared and private classes over the full ImageNet-100 overlap sweep. A-private and B-private images have asymmetry of opposite signs, while shared-class images remain near zero; the 100% overlap condition is retained as a negative control.

B.1.3 Controlled CLIP endpoints and probes

Table 7 reports model recoverability, predictive asymmetry, and native species/order macro-F1 for the three matched-pool CLIP fine-tuning arms at epoch 20. Results for all checkpoints are provided in the machine-readable artifact.

B.2 Domain-specialized representations: full results

B.2.1 Released-specialist recoverability and native probes

For the released models, we report the quantities used in the Section 6 analyses: both directional recoverabilities, their mean and signed difference, the complete twelve-model cloud grid, and native parent-versus-specialist accuracy and macro-F1. Rows use the same domain-grouped order throughout.

B.3 Class agreement between independently trained models

Table 10 shows that specialists trained for the same domain tend to rank classes similarly by asymmetry. The two independent biodiversity comparisons both have $\rho \approx .80$, while the six remote-sensing comparisons range from .57 to .85. Pathology also has high agreement ($\rho = .87$), although it contains only nine classes. The BioCLIP–TaxaBind correlation is highest ($\rho = .98$), although these models are not fully independent because TaxaBind was initialized from BioCLIP. Thus, specialists tend to agree on which classes are relatively parent- or specialist-favored. This does not imply that they have the same global asymmetry or that the agreement has the same causal origin.

To test whether shared evaluation data alone explain this agreement, we also evaluated specialists from other domains on the same images and classes. On Rare Species, the mean correlation among the three biodiversity specialists was .856, compared with .194 between a biodiversity specialist and one of eight other-domain specialists (24 comparisons; $\Delta\rho = .662$, class-bootstrap 95% CI $[.612, .712]$). The two independently trained biodiversity comparisons were each approximately .80, so this contrast is not driven only by the shared BioCLIP–TaxaBind initialization. On RESISC45, the mean across the six remote-sensing pairs was .722, compared with $-.197$ across 28 remote-versus-other-domain comparisons ($\Delta\rho = .919$ $[.752, 1.037]$). Both same-domain gaps had exact model-label permutation $p = .0061$. On the nine-class CRC population, PLIP–QuiltNet had $\rho = .867$, compared with a mean of .436 across 16 pathology-versus-other-domain comparisons; we treat this smaller contrast as descriptive because only nine classes are available. Shared evaluation data

Table 5: Complete controlled-granularity evaluation. Endpoint geometry and the population-resolved evaluation are shown together for all ten unordered mixed-supervision pairs. Values are means ± standard deviations over five independently trained matched seeds. Detection distinguishes the two differently supervised superclasses from the other three; attribution tests whether the sign identifies which endpoint received fine supervision.

Fine-supervision pair	$\overline{R}^2$	ΔR^2	Detection AUROC	Attribution accuracy
birds / dogs	0.688±0.015	-0.050±0.010	0.999±0.001	1.000±0.000
birds / fish	0.697±0.013	-0.019±0.034	0.994±0.011	0.990±0.022
birds / insects	0.693±0.025	-0.020±0.020	0.999±0.001	1.000±0.000
birds / reptiles	0.706±0.030	-0.005±0.025	0.994±0.012	1.000±0.000
dogs / fish	0.680±0.005	+0.026±0.031	1.000±0.000	1.000±0.000
dogs / insects	0.664±0.016	+0.029±0.017	0.999±0.001	1.000±0.000
dogs / reptiles	0.686±0.025	+0.048±0.020	0.998±0.004	0.990±0.022
fish / insects	0.681±0.022	+0.002±0.035	0.998±0.004	1.000±0.000
fish / reptiles	0.698±0.014	+0.021±0.039	0.998±0.005	0.990±0.022
insects / reptiles	0.688±0.015	+0.021±0.027	0.995±0.005	0.990±0.022

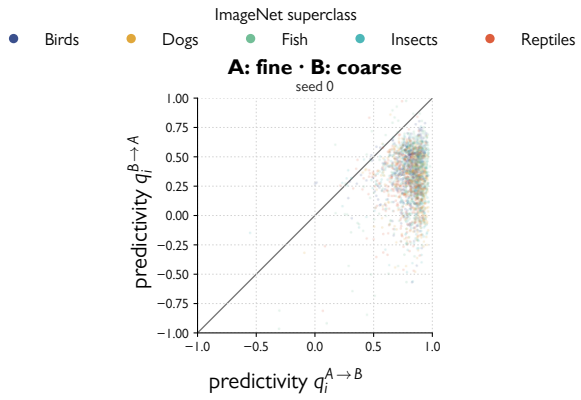

Figure 6: Per-image predictivity for the canonical seed-0 fine–coarse comparison. Model A uses 50 fine labels and model B uses five superclass labels; color denotes superclass.

alone therefore do not explain the within-domain agreement in Table 10. This auxiliary cross-domain sweep used the canonical fold assignment and an earlier post-map-normalized row-ℓ_2 affine protocol, so we report it as a sensitivity analysis rather than pooling it with the repeated-CV estimates in the table.

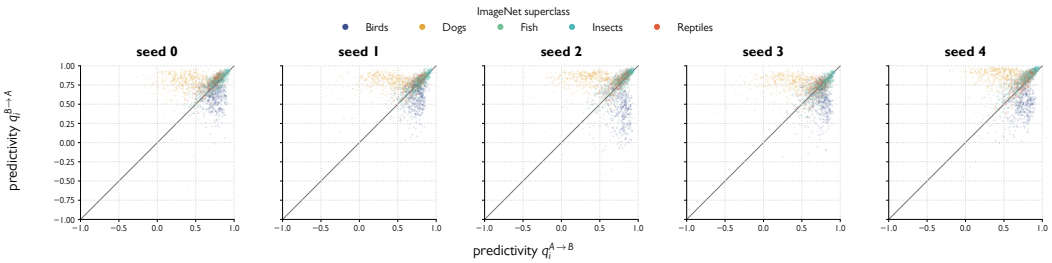

Figure 7: Birds-fine (A) versus dogs-fine (B), separately for all five matched training seeds. Each panel contains the same 2,000 held-out images.

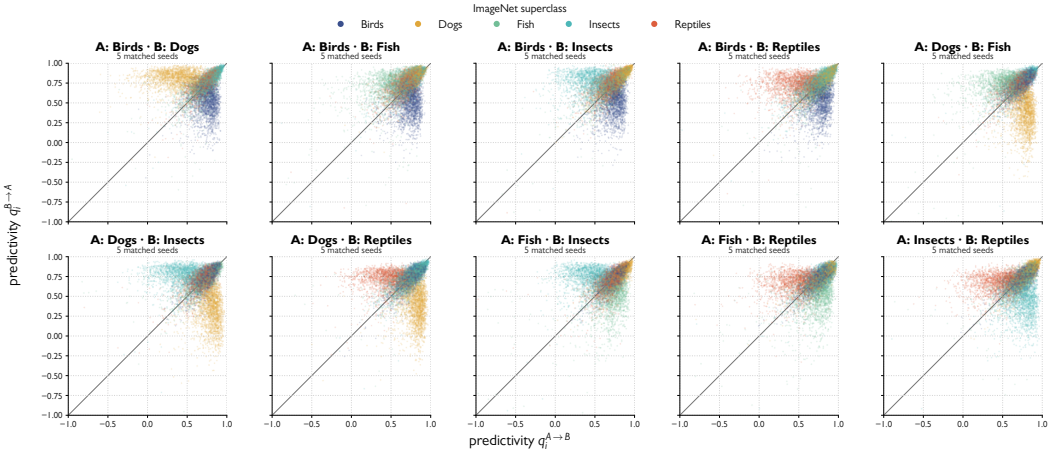

Figure 8: Per-image predictivity for all ten unordered pairs of the five mixed-supervision objectives. Each panel pools the five matched seeds; colors denote superclasses and panel titles define endpoints A and B.

Table 6: Combined global and population-resolved results for the ImageNet-100 training-coverage control. Global recoverability columns use symmetric 95% leakage-group bootstrap half-widths; population means and evaluation metrics use symmetric 95% class-bootstrap half-widths. Detection distinguishes private from shared images, while attribution asks whether the sign identifies the endpoint trained on the image's class. Dashes mark undefined quantities.

Overlap (%)	$\overline{R}^2$	ΔR^2	A-private Δq_i	Shared Δq_i	B-private Δq_i	Detection AUROC	Attribution accuracy
0	0.413±0.003	+0.046±0.006	0.216±0.022	—	-0.126±0.023	—	0.759±0.021
25	0.452±0.003	+0.014±0.006	0.186±0.028	0.013±0.016	-0.161±0.025	0.700±0.030	0.774±0.025
50	0.495±0.003	+0.012±0.005	0.184±0.029	0.002±0.012	-0.158±0.029	0.714±0.027	0.797±0.024
75	0.565±0.004	+0.003±0.005	0.195±0.040	-0.001±0.010	-0.184±0.023	0.752±0.030	0.819±0.027
100	0.710±0.004	+0.029±0.002	—	0.029±0.004	—	—	—

Table 7: Combined representation and probe results for the matched-pool CLIP control at epoch 20 (R0). P is the frozen parent and A is the trained arm. The first three columns report bidirectional recoverability and predictive asymmetry; species and order columns report native macro-F1 on the common held-out population.

Supervision	$R^2_{P\to A}$	$R^2_{A\to P}$	ΔR^2	Species P	Species A	Order P	Order A
species	0.587	0.663	-0.076	0.668	0.773	0.842	0.898
order	0.763	0.658	+0.104	0.668	0.626	0.842	0.915
random26	0.448	0.628	-0.180	0.668	0.734	0.842	0.874

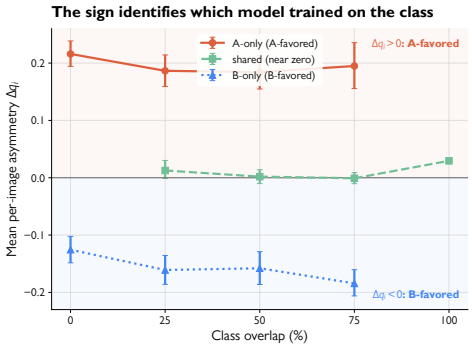

Figure 9: Mean held-out asymmetry for each support population across the ImageNet-100 overlap sweep, with 95% class-bootstrap intervals.

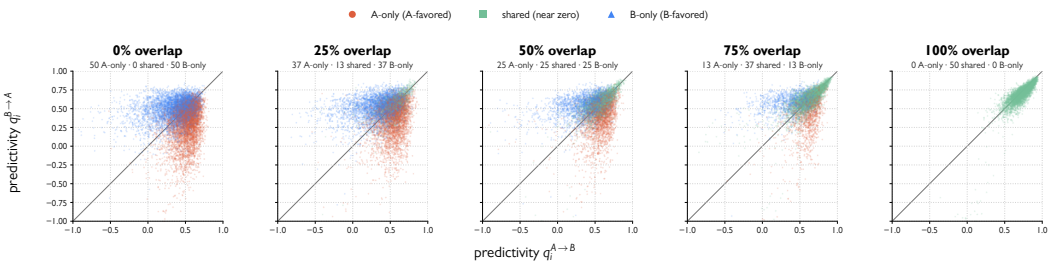

Figure 10: Bidirectional predictivity across the complete ImageNet-100 class- overlap sweep. Points are held-out images colored by whether their class was private to A, shared, or private to B.

Table 8: Global released-specialist recoverability, grouped by domain. Values are point estimates ± symmetric 95% leakage-group bootstrap half-widths, conditional on the released checkpoints and repeat-mean scores.

Specialist	n	classes	$R^2_{P\to S}$	$R^2_{S\to P}$	$\overline{R}^2$	ΔR^2
Biodiversity						
BioCLIP	9,586	400	0.388±0.003	0.567±0.003	0.477±0.003	-0.179±0.003
BioTrove	9,586	400	0.344±0.003	0.504±0.004	0.424±0.003	-0.160±0.003
TaxaBind	9,586	400	0.408±0.003	0.584±0.003	0.496±0.003	-0.176±0.003
Remote sensing						
GeoRSCLIP	25,200	45	0.849±0.001	0.756±0.001	0.803±0.001	+0.094±0.001
RemoteCLIP	25,200	45	0.633±0.002	0.631±0.002	0.632±0.001	+0.002±0.002
RSDiX	25,200	45	0.723±0.002	0.629±0.002	0.676±0.001	+0.094±0.002
SkyCLIP	25,200	45	0.906±0.001	0.896±0.001	0.901±0.001	+0.010±0.000
Pathology						
PLIP	5,740	9	0.844±0.002	0.785±0.003	0.815±0.002	+0.059±0.003
QuiltNet	5,740	9	0.806±0.002	0.761±0.004	0.783±0.003	+0.045±0.003
Camera traps						
WildCLIP	68,647	38	0.672±0.001	0.670±0.001	0.671±0.001	+0.003±0.001
Fashion						
FashionCLIP	8,000	44	0.913±0.001	0.898±0.001	0.906±0.001	+0.015±0.000
Street imagery						
StreetCLIP	4,000	135	0.922±0.001	0.912±0.001	0.917±0.001	+0.010±0.000

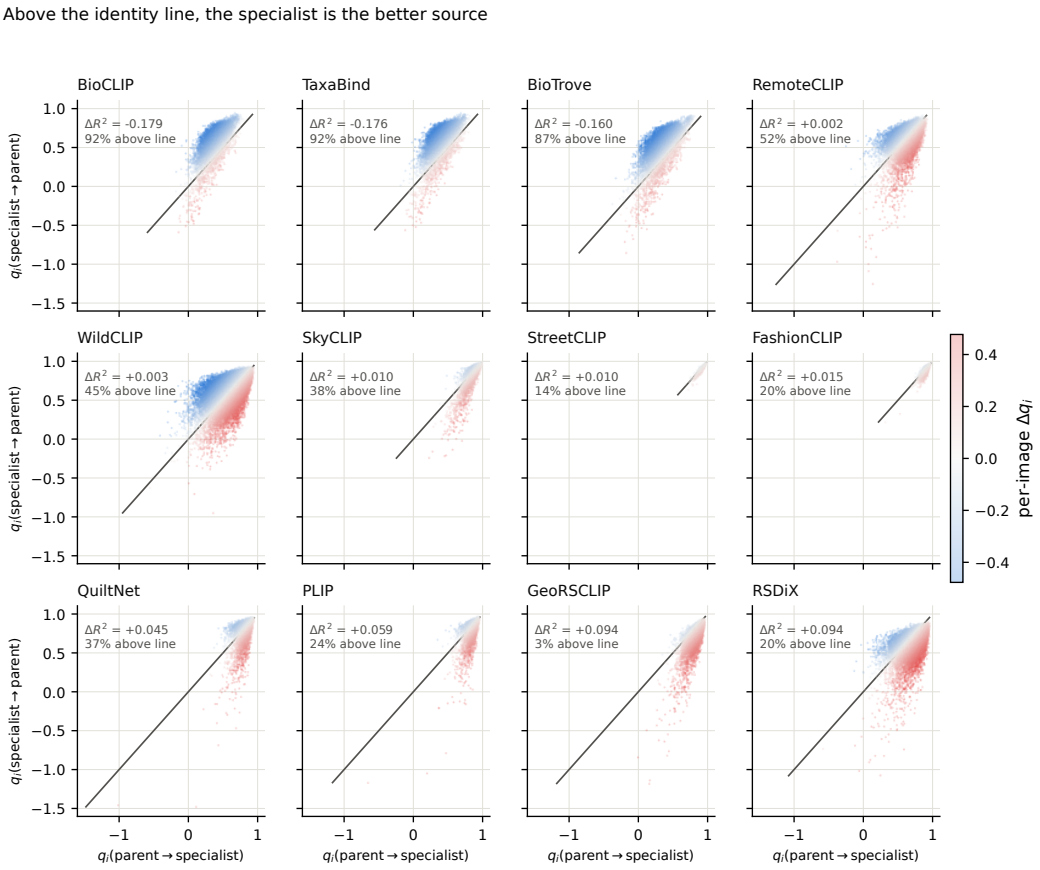

Figure 11: Held-out per-image bidirectional recoverability for all twelve released parent–specialist pairs. Each panel uses its own displayed range; color encodes repeat-mean per-image asymmetry on the common color scale.

Table 9: Native downstream linear probes, grouped by domain. Parent (P) and specialist (S) columns are mean percentages over ten matched cross-fitting assignments. Δ is specialist minus parent in percentage points; its uncertainty is the standard deviation of the ten paired differences.

Specialist	Acc. P	Acc. S	Δ Acc.	Macro-F1 P	Macro-F1 S	Δ Macro-F1
Biodiversity						
BioCLIP	68.8	76.9	+8.1$\pm$0.4	68.6	76.9	+8.3$\pm$0.4
BioTrove	68.8	69.5	+0.7$\pm$0.4	68.6	69.6	+1.1$\pm$0.4
TaxaBind	68.8	78.3	+9.5$\pm$0.4	68.6	78.3	+9.8$\pm$0.4
Remote sensing						
GeoRSCLIP	92.4	95.2	+2.8$\pm$0.2	92.4	95.2	+2.8$\pm$0.2
RemoteCLIP	92.4	94.1	+1.7$\pm$0.1	92.4	94.1	+1.7$\pm$0.1
RSDiX	92.4	98.7	+6.2$\pm$0.1	92.4	98.7	+6.3$\pm$0.1
SkyCLIP	93.4	93.7	+0.3$\pm$0.1	93.4	93.7	+0.3$\pm$0.1
Pathology						
PLIP	98.3	99.3	+1.0$\pm$0.1	98.0	99.1	+1.1$\pm$0.1
QuiltNet	98.3	99.4	+1.1$\pm$0.1	98.0	99.3	+1.3$\pm$0.2
Camera traps						
WildCLIP	85.2	91.1	+5.8$\pm$0.2	72.8	81.6	+8.8$\pm$0.5
Fashion						
FashionCLIP	60.4	61.3	+0.9$\pm$1.7	32.4	33.6	+1.1$\pm$1.6
Street imagery						
StreetCLIP	78.1	76.4	-1.7$\pm$0.5	34.0	34.4	+0.3$\pm$0.9

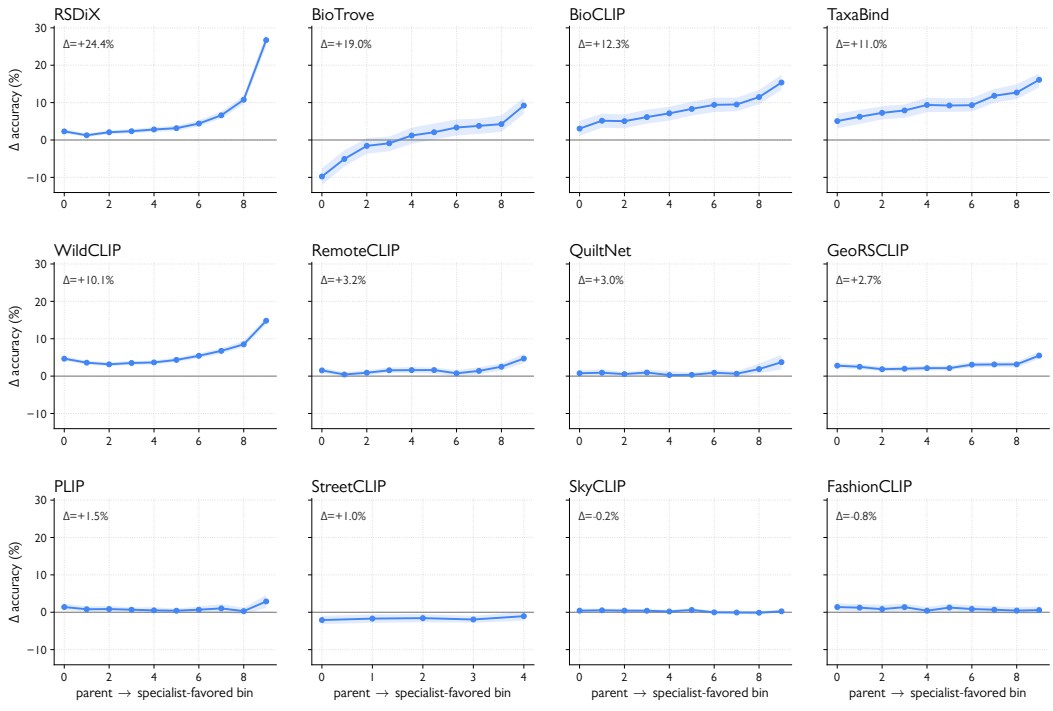

Figure 12: q_i **stratifies where the specialist's relative advantage concentrates.** Full result ordered by observed total probe gap.

Pair	Evaluation set	n_{cls}	ρ [95% CI]	BH q	Relationship
PLIP–QuiltNet	crc val he 7k	9	0.87 [0.34, 1.00]	0.0046	small-n descriptive
BioCLIP–TaxaBind	rare species	400	0.98 [0.97, 0.98]	0.000125	shared init
BioTrove–TaxaBind	rare species	400	0.80 [0.75, 0.84]	0.000125	independent
BioCLIP–BioTrove	rare species	400	0.80 [0.75, 0.84]	0.000125	independent
RemoteCLIP–RSDiX	resisc45	45	0.85 [0.72, 0.91]	0.000125	independent
GeoRSCLIP–RSDiX	resisc45	45	0.84 [0.71, 0.91]	0.000125	independent
GeoRSCLIP–RemoteCLIP	resisc45	45	0.73 [0.57, 0.83]	0.000125	independent
GeoRSCLIP–SkyCLIP	resisc45	45	0.70 [0.55, 0.79]	0.000125	independent
RemoteCLIP–SkyCLIP	resisc45	45	0.62 [0.48, 0.70]	0.000125	independent
RSDiX–SkyCLIP	resisc45	45	0.57 [0.39, 0.68]	0.000222	independent

Table 10: **Separately trained domain specialists have high agreement on relative bidirectional predictability of classes.** Spearman correlations compare repeat-averaged class-level Δq_c fingerprints on a shared evaluation set. Confidence intervals bootstrap classes, and q values apply Benjamini–Hochberg correction to the ten prespecified within-domain permutation tests. "Shared init" marks the BioCLIP–TaxaBind comparison; all other pairs were trained independently of one another. The nine-class pathology result is reported as descriptive despite its corrected significance.

Specialist	Rows retained	Median-cap ρ	Same-N ρ
Biodiversity			
BioCLIP	96.3%	$.9978 \pm .0005$	$.9976 \pm .0003$
BioTrove	96.3%	$.9977 \pm .0003$	$.9976 \pm .0003$
TaxaBind	96.3%	$.9980 \pm .0003$	$.9975 \pm .0003$
Remote sensing			
GeoRSCLIP	99.2%	$.9992 \pm .0003$	$.9994 \pm .0002$
RemoteCLIP	99.2%	$.9995 \pm .0004$	$.9993 \pm .0004$
RSDiX	99.2%	$.9990 \pm .0003$	$.9989 \pm .0004$
SkyCLIP	99.2%	$.9997 \pm .0002$	$.9997 \pm .0002$
Pathology			
PLIP	79.2%	$1.0000 \pm .0000$	$.9950 \pm .0081$
QuiltNet	79.2%	$.9833 \pm .0000$	$.9933 \pm .0086$
Camera traps			
WildCLIP	17.4%	$.8732 \pm .0110$	$.9672 \pm .0211$
Fashion			
FashionCLIP	9.8%	$.6265 \pm .0480$	$.3581 \pm .2114$
Street imagery			
StreetCLIP	11.2%	$.5802 \pm .0243$	$.3701 \pm .0649$

Table 11: Evaluation-population controls for all released specialists. Entries are mean $\pm$ standard deviation over ten refitted subsamples of the Spearman correlation with the original classwise asymmetry fingerprint.

Matched rank	Median class ρ	Range across pairs
64	.699	.281–.933
128	.875	.476–.964
256	.974	.677–.983

Table 12: Sensitivity to matched low-rank representation capacity. Class ρ compares each rank-controlled classwise fingerprint with its original full-width fingerprint after averaging scores over ten cross-fitting assignments.

B.4 Evaluating robustness of predictive asymmetry

B.4.1 Evaluation-population controls

In the controlled granularity experiment, restricting the fit from all 50 classes to the 20 classes with different supervision preserves the classwise ordering ($\rho = .987$) and model attribution (sign accuracy $= .993$; 1,000 paired class observations). For this analysis, we recompute centering, ridge selection, both affine maps, and normalization from scratch.

For each released specialist we cap every evaluation class at that dataset's median class count, repeat ten deterministic subsamples, and refit both map directions. We also use a random subsample of the same size to distinguish the effect of class balancing from the effect of using fewer fitting images.

The median median-cap correlation is .998. The lower correlations for WildCLIP, FashionCLIP, and StreetCLIP coincide with the strongest reductions in fitting data; the same-N control shows that much of this drop is due to the smaller fitting set. Also, FashionCLIP and StreetCLIP are two specialist models with high mutual predictability and low asymmetry to begin with, with little cross-class structure.

B.4.2 Matched-capacity controls

The matched-capacity control fits centered, unwhitened PCA independently for each model within each cross-fitting training scope, projects both models to the same rank, and refits both maps. Native widths are already matched within every released pair, so PCA reduces representation capacity rather than correcting a dimension mismatch.

Table 13: Controlled-benchmark robustness to the cross-predictive map family. Detection entries are AUROC, attribution entries are balanced accuracy at the fixed zero threshold, and intervals are 95% bootstrap half-widths. Later checkpoint reports correctly ordered pairs out of 15. Orthogonal Procrustes preserves detection but, because its two directional residual norms are equal up to numerical precision, cannot provide signed attribution.

Map	Granularity detect.	Granularity attrib.	Coverage detect.	Coverage attrib.	Later checkpoint
Affine ridge	$0.997_{\pm 0.003}$	$0.996_{\pm 0.006}$	$0.722_{\pm 0.031}$	$0.787_{\pm 0.028}$	15/15
Orthogonal Procrustes	$0.999_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.826_{\pm 0.044}$	$0.500_{\pm 0.000}$	13/15
Residual MLP	$0.994_{\pm 0.006}$	$0.996_{\pm 0.006}$	$0.663_{\pm 0.033}$	$0.720_{\pm 0.021}$	15/15

Table 14: Released-specialist sensitivity to the map family. Agreement columns compare each method with the paper's affine-ridge analysis across all twelve pairs; image-level quantities are averaged across pairs. The ridge row is the reference, not an independent replication.

Map	Mean $\lvert\Delta R^2\rvert$	Global sign agree.	Image ρ	Image sign agree.
Affine ridge	0.071	1.000	1.000	1.000
Orthogonal Procrustes	0.187	0.417	-0.060	0.426
Residual MLP	0.081	1.000	0.860	0.883

Agreement rises monotonically with retained rank. At the reported rank $k = 128$, the median classwise correlation is .875; at $k = 256$ it is .974. These results show robustness to these particular perturbations; they do not imply invariance to arbitrary changes in the evaluation population.

B.4.3 Map-family robustness

We repeat the analysis with the canonical affine ridge map, centered orthogonal Procrustes, and an overparameterized nonlinear residual MLP while holding the evaluation populations and split assignments fixed.

The controlled localization results do not depend on affine ridge: Procrustes and the residual MLP retain high mixed-supervision detection, and both provide meaningful coverage localization. Procrustes serves as a useful control rather than a directional alternative: its orthogonality makes the two residual norms directionally symmetric, so its attribution accuracy is exactly chance. Across released specialists, the residual MLP gives the same global asymmetry sign as ridge for all 12 pairs and has mean imagewise rank correlation .860 with ridge; Procrustes does not preserve the signed specialist pattern.

B.4.4 Why use the bidirectional difference?

Table 15 compares Δq_i with each constituent direction on the controlled granularity, coverage, and checkpoint-progression benchmarks. For detection and model ranking, Δq_i matches or improves upon the better one-way score in every task. More fundamentally, a one-way score depends on which model is chosen as source and which as target, and it has no orientation-independent zero point. By contrast, Δq_i changes sign exactly when the endpoints are swapped.

Table 15: Bidirectional difference versus either direction alone on the controlled benchmarks. All first-block entries are AUROC. Detection uses negative directional recoverability because low recoverability marks a changed class; the B-owned rows orient each score so larger values favor endpoint B. The gain is paired Δq_i minus the better one-way score. Progression reports all 15 earlier–later checkpoint pairs and the mean error under swapping endpoint order. A one-way score can rank endpoints after choosing an orientation, but only Δq_i is antisymmetric and has an orientation-independent zero threshold.

Controlled task	Δq_i	$A \to B$ only	$B \to A$ only	Δq_i gain [95% CI]
Granularity: changed-class detection	0.997	0.758	0.838	+0.159 [+0.126, +0.194]
Granularity: B-owned class	1.000	0.997	0.993	+0.002 [+0.000, +0.005]
Coverage: private-class detection	0.722	0.655	0.679	+0.043 [+0.038, +0.046]
Coverage: B-private class	0.883	0.757	0.731	+0.126 [+0.113, +0.138]
Progression: later checkpoint identified	15/15	0/15	15/15	—
Progression: mean swap error	0.000	1.240	1.240	—

B.5 Relating predictive asymmetry to CCA

Here we derive the connection between directional recoverability and CCA and state the assumptions under which it holds. We assume a fixed evaluation population and preprocessing pipeline throughout. We use $R^{2,*}$ to distinguish population-optimal, unregularized predictivity from the empirical ridge score R^2.

B.5.1 Linear recoverability as explained variance

Let a and b denote centered representations from encoders A and B. Centering is applied after the row normalization described in the main text. Define

$$\Sigma_A = \mathbb{E}[aa^\top], \qquad \Sigma_B = \mathbb{E}[bb^\top], \qquad \Sigma_{AB} = \mathbb{E}[ab^\top], \tag{4}$$

with $\Sigma_{BA} = \Sigma_{AB}^\top$. We assume positive-definite covariances and nonzero total variance. If a covariance is singular, the same derivation applies on its nonzero-variance subspace; d_A and d_B then denote effective ranks.

Expanding the expected prediction error gives

$$\begin{aligned}\mathbb{E}\|b - Wa\|_2^2 = \ & \operatorname{tr}(\Sigma_B) - 2\operatorname{tr}(W\Sigma_{AB}) \\ & + \operatorname{tr}(W\Sigma_A W^\top).\end{aligned} \tag{5}$$

The minimizer is

$$W^*_{A\to B} = \Sigma_{BA}\Sigma_A^{-1}. \tag{6}$$

Its residual covariance and expected squared error are

$$\Sigma_{B|A} = \Sigma_B - \Sigma_{BA}\Sigma_A^{-1}\Sigma_{AB}, \tag{7}$$

$$E^*_{A\to B} = \operatorname{tr}(\Sigma_B) - \operatorname{tr}\big(\Sigma_{BA}\Sigma_A^{-1}\Sigma_{AB}\big). \tag{8}$$

The second term is the amount of target variance explained by the optimal linear predictor.

Under independent source–target correspondence, $\Sigma_{AB} = 0$. The optimal predictor is then the target mean, with error

$$E^{*,\text{null}}_{A\to B} = \operatorname{tr}(\Sigma_B). \tag{9}$$

Consequently,

$$\boxed{R^{2,*}_{A\to B} = \frac{\operatorname{tr}\big(\Sigma_{BA}\Sigma_A^{-1}\Sigma_{AB}\big)}{\operatorname{tr}(\Sigma_B)}.} \tag{10}$$

This quantity lies in $[0, 1]$ and measures the fraction of target variance explained by the optimal linear predictor.

B.5.2 Canonical decomposition

CCA decomposes the whitened cross-covariance:

$= \Sigma_A^{-1/2}\Sigma_{AB}\Sigma_B^{-1/2} = URV^\top$, (11) where U and V are complete orthonormal bases and R is rectangular diagonal. Its diagonal entries are the canonical correlations $\rho_i \in [0, 1]$.

For each paired component, the canonical variates are

$$z_i^A = u_i^\top \Sigma_A^{-1/2} a, \qquad z_i^B = v_i^\top \Sigma_B^{-1/2} b. \tag{12}$$

These scalar scores have unit variance and satisfy

$$\operatorname{Cov}(z_i^A, z_j^A) = \operatorname{Cov}(z_i^B, z_j^B) = \delta_{ij}, \qquad \operatorname{Cov}(z_i^A, z_j^B) = \rho_i \delta_{ij}. \tag{13}$$

Thus, the optimal linear prediction of z_i^B from all source canonical variates is $\rho_i z_i^A$, with residual variance $1-\rho_i^2$. The predicted quantity is the canonical variate across observations, not the canonical direction itself.

Using

$$\Sigma_{AB} = \Sigma_A^{1/2} U R V^\top \Sigma_B^{1/2}, \tag{14}$$

the numerator of equation 10 becomes

$$\begin{aligned} &\mathrm{tr}\big(\Sigma_{BA}\Sigma_A^{-1}\Sigma_{AB}\big) \\ &\quad = \mathrm{tr}\big(\Sigma_B^{1/2} V R^\top R V^\top \Sigma_B^{1/2}\big) \\ &\quad = \sum_{i=1}^{d_B} \rho_i^2 \, v_i^\top \Sigma_B v_i, \end{aligned} \tag{15}$$

where unmatched target directions have $\rho_i = 0$. Since V is complete and orthonormal,

$$\sum_{i=1}^{d_B} v_i^\top \Sigma_B v_i = \mathrm{tr}(\Sigma_B). \tag{16}$$

Therefore,

$$\boxed{R^{2,*}_{A\to B} = \sum_{i=1}^{d_B} w_i^B \rho_i^2, \qquad w_i^B = \frac{v_i^\top \Sigma_B v_i}{\mathrm{tr}(\Sigma_B)}.} \tag{17}$$

This is the weighted squared-canonical-correlation expression for linear-regression explained variance (Kornblith et al., 2019).

Interpreting the weights. The target can be reconstructed from its canonical scores as

$$b = \sum_i \ell_i^B z_i^B, \qquad \ell_i^B = \Sigma_B^{1/2} v_i. \tag{18}$$

Because these scores have unit variance and are mutually uncorrelated, component i contributes $\|\ell_i^B\|_2^2$ to total target variance:

$$w_i^B = \frac{\|\ell_i^B\|_2^2}{\mathrm{tr}(\Sigma_B)}. \tag{19}$$

The reconstruction vectors ℓ_i^B need not be orthogonal.

Note that w_i^B is not the variance of the canonical variate; that variance is one by construction. Specifically, the extraction vector in the original feature coordinates is $c_i^B = \Sigma_B^{-1/2} v_i$, so

$$(c_i^B)^\top \Sigma_B c_i^B = 1, \qquad w_i^B = \frac{(c_i^B)^\top \Sigma_B^2 c_i^B}{\mathrm{tr}(\Sigma_B)}. \tag{20}$$

Relation to PCA. Canonical directions need not coincide with principal components. If the target covariance has eigendecomposition $\Sigma_B = \sum_j \lambda_j e_j e_j^\top$, then

$$w_i^B = \frac{\sum_j \lambda_j (e_j^\top v_i)^2}{\sum_j \lambda_j}. \tag{21}$$

Thus, the variance contribution $\|\ell_i\|_2^2$ is a weighted average of the target's PCA eigenvalues, with weights given by squared overlaps between the whitened canonical direction and the covariance eigenvectors. A large contribution indicates stronger alignment with high-variance PCA directions. Unlike PCA, however, CCA selects components according to their cross-representation correlations rather than their variance. Thus, w_i^B measures how much target variance a canonical component contributes, while ρ_i^2 measures how well that component is linearly predicted from the source.

B.5.3 Directional asymmetry

For equal effective dimensions $d_A = d_B = d$, both directions use the same canonical correlation spectrum:

$$R^{2,*}_{A\to B} = \sum_i w_i^B \rho_i^2, \qquad R^{2,*}_{B\to A} = \sum_i w_i^A \rho_i^2. \tag{22}$$

Their difference is

$$\boxed{\Delta R^{2,*}(A,B) = \sum_i (w_i^B - w_i^A)\rho_i^2.} \tag{23}$$

Asymmetry vanishes if the variance weights agree or if all canonical correlations are equal. Asymmetry therefore requires both unequal variance allocation and nonuniform predictability across components.

Equivalently, let

$$\varepsilon_B = 1 - R^{2,*}_{A\to B}, \qquad \varepsilon_A = 1 - R^{2,*}_{B\to A}. \tag{24}$$

Then

$$\Delta R^{2,*}(A,B) = \varepsilon_A - \varepsilon_B. \tag{25}$$

Positive asymmetry means that a larger fraction of A's variance remains linearly unexplained by B than vice versa. The difference alone does not identify either unexplained fraction, motivating the separate reporting of both directional scores.

For unequal effective dimensions, equation 17 applies separately to each target, including its unmatched zero-correlation components.

B.6 Whitening

Consider population-whitened representations

$$a_{\mathrm{w}} = \Sigma_A^{-1/2} a, \qquad b_{\mathrm{w}} = \Sigma_B^{-1/2} b. \tag{26}$$

Their covariance matrices are identities. Consequently, every target component has equal variance weight:

$$R^{2,*}_{A_{\mathrm{w}}\to B_{\mathrm{w}}} = \frac{\sum_i \rho_i^2}{d_B}, \qquad R^{2,*}_{B_{\mathrm{w}}\to A_{\mathrm{w}}} = \frac{\sum_i \rho_i^2}{d_A}. \tag{27}$$

For equal effective dimensions, both scores equal the mean squared canonical correlation and their asymmetry is zero. For unequal effective dimensions, whitening does not generally remove asymmetry.

A low whitened score means that canonical components are, on average, poorly predicted linearly. It does not imply that every component is poorly predicted, nor does it identify which components matter for a downstream task. These identities assume no additional row normalization after whitening, which would change the covariance geometry. Under these assumptions, recoverability reduces to the standard CCA summary: mean squared canonical correlation.

B.6.1 Scope and empirical estimation

Finite-sample ridge prediction. The identities above apply to population-optimal, unregularized maps. Empirical ridge errors additionally reflect regularization, sample size, and estimation uncertainty. Empirical estimates approach the population result as estimation error and effective regularization vanish. For the mean-loss convention in Eq. equation 2, effective regularization is $\lambda_f \tau_f$. Exact CCA decompositions and whitening symmetry are therefore not guaranteed for finite-sample ridge scores.

Centering and population-specific fitting. The covariance derivation requires a fitted mean, implemented through an unpenalized intercept or training-fold centering. Row normalization alone does not center the embeddings. The derivation also assumes that the predictor is fit on the same population on which it is evaluated. If a global map is evaluated on a subset, its error then also reflects how well the global map transfers to that subset.

Limits of interpretation. Recoverability measures linear accessibility in the chosen embedding geometry. Low recoverability can arise from nonlinear recoding without information loss. Conversely, an invertible linear transformation of a representation remains perfectly recoverable in both directions in the population limit, even if its covariance changes substantially. Variance weighting reflects reconstruction of the embedding, not the task relevance of the reconstructed variation. We therefore interpret directional scores together with residual analyses and downstream experiments when studying specialization.