OISE: OBJECTIVE-INDUCED SPLITS IN ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

What a frozen vision encoder splits images on first depends on whether its training objective matches a whole-image target or predicts hidden content. We examine this relationship in 21 frozen vision encoders, comprising 15 core models spanning objectives and architectures and six language-supervised variants spanning scale, pretraining data, a convolutional backbone, and a VLM tower. A fixed clustering procedure induces a binary partition within each class without group labels. We compare assignments on the same images using the Adjusted Rand Index (ARI), whose chance expectation is zero. On UrbanCars, 19 of the 21 encoders have self-consistency of 0.56–0.93 across discovery draws. Eight self-supervised cross-view encoders produce similar partitions (ARI 0.56–0.81). Label- and language-supervised models occupy an intermediate position, agreeing with the cross-view models at 0.37 on average. MAE, data2vec, I-JEPA, and the hybrid DINOv2 have low agreement with whole-image models (0.07) and with one another (0.06). The exception is MAE–data2vec (0.32 on UrbanCars and at most 0.08 elsewhere). Within the original four-dataset analysis, the three latent-prediction models (data2vec, I-JEPA, and DINOv2) are jointly isolated only on UrbanCars, where partitions are most stable. Matched training reproduces the ResNet-18 ordering with the same images, epochs, and batch size. Controlled interventions link partition selection to cross-view augmentation and show that relative variance can shift the partition among stably expressed factors. Finally, label-free partition agreement predicts cross-encoder group-transfer gains, although the gains are limited to a few worst-group accuracy points.[1]

1 INTRODUCTION

What a frozen vision encoder splits images on first depends on whether its training objective matches a whole-image target or predicts hidden content.

Frozen vision encoders can organize the same images around different factors, such as background, a co-occurring object, or brightness. Conversely, independently trained encoders may prioritize the same factor despite having unaligned feature spaces. This organization is relevant when representations are compared, used for group discovery, or substituted as feature extractors. We study which properties of pretraining predict the dominant partition induced by a frozen encoder.

Language supervision is one possible organizing principle, given the differences between language-supervised and vision-only encoders (Radford et al., 2021; Santurkar et al., 2023). An alternative is convergence across objectives as data and model scale increase, as proposed by the Platonic Representation Hypothesis (Huh et al., 2024). Evaluating these possibilities requires comparisons that do not assume aligned feature coordinates (Kornblith et al., 2019; Raghu et al., 2021). Standard approaches compare representational geometry. We instead compare a discrete property induced by that geometry, namely how different encoders partition the *same images*.

For each encoder, a fixed unsupervised pipeline fits a class-conditioned binary partitioning rule on a sample of images and applies it to a fixed pool of images shared by every encoder. The Adjusted Rand Index (ARI) compares the resulting sample assignments without aligning feature coordinates. It is also unchanged by exchanging the two labels of either partition. We measure self-consistency across resampled image samples to distinguish low cross-encoder agreement from instability of the partitions themselves. The partitioner is a fixed diagnostic procedure, not a new clustering method.

[1] Code: https://anonymous.4open.science/r/OISE-ED68/

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.

On UrbanCars, agreement is associated with target locality (whole-image versus hidden-content targets) rather than a simple language-supervised versus vision-only distinction. Objectives that match whole-image targets, including labels, captions, and augmented views, produce related partitions. Agreement is highest among the eight self-supervised cross-view encoders, with externally supervised (label- and language-supervised) models occupying an intermediate position. Hidden-content objectives, by contrast, have low agreement both with whole-image models and with one another, so they do not form a coherent alternative family. DINOv2 is treated as a hybrid because it combines whole-image and masked-token terms while also differing in data, scale, and regularization.

We evaluate this association on new random samples drawn after the grouping was fixed, on further datasets, with matched training, and with controlled interventions. These experiments separate the empirical pattern from its proposed explanation and identify conditions under which it does not replicate.

Contributions.

- **The finding.** We identify an association between target locality and partition agreement. Cross-view encoders have the highest mutual agreement on UrbanCars, followed by externally supervised models. New samples drawn after the grouping was fixed confirm this ordering, while cross-dataset tests establish its limits (Secs. 4.1–4.3).
- **Controlled training.** Matched training reproduces the separation between cross-view and masked-reconstruction ResNet-18 runs with the same images, epochs, and batch size. This supports an objective-plus-recipe explanation, not an isolated causal effect of the loss, because each objective retains its prescribed augmentation. One ViT-S run (MoCo-v3) does not follow this ordering (Sec. 4.4).
- **The mechanism.** Controlled interventions show that cross-view augmentation can suppress a factor in the root partition. Among stably expressed factors, increasing relative variance can shift the partition towards that factor, with an objective-dependent transition (Sec. 4.5).
- **The use.** Partition agreement predicts cross-encoder pseudo-group transfer after accounting separately for self-consistency and linear CKA. The modest worst-group gains provide downstream validation of the analysis rather than a standalone robust-training method (Sec. 4.6).

2 Related Work

Representation comparison. Representation-comparison methods address differences in feature coordinates between independently trained encoders. CKA (Kornblith et al., 2019) is one such approach, and Raghu et al. (2021) use it to compare the internal representations of ViTs and CNNs. The closest large-scale comparison is Ciernik et al. (2025), who analyse 64 vision models with CKA and RSA. They find that relative pairwise similarities involving self-supervised models transfer more consistently across datasets than those involving classification or image-text models. We distinguish self-supervised objectives by target locality and study agreement between induced partitions within a dataset. This partition-level analysis also supports the transfer and axis-removal tests in Section 4.6.

Model stitching. Model stitching tests whether a learned adapter between two networks' layers preserves downstream performance (Lenc & Vedaldi, 2015; Bansal et al., 2021). Recent work extends this question to frozen foundation models (Mai et al., 2026), including adapters trained by feature matching. Stitching evaluates compatibility after learning an adapter. Our comparison uses independently fitted partition rules without a cross-model adapter or encoder fine-tuning and measures agreement without group labels.

Label-free group discovery. GEORGE (Sohoni et al., 2020), EIIL (Creager et al., 2021), and Domino (Eyuboglu et al., 2022) identify latent or spurious groups without group annotations, and JTT (Liu et al., 2021a) upweights examples that ERM misclassifies. Their focus is on groups within a model or representation. We ask whether independently trained encoders induce similar partitions and whether this agreement predicts the transfer of pseudo-groups between encoders.

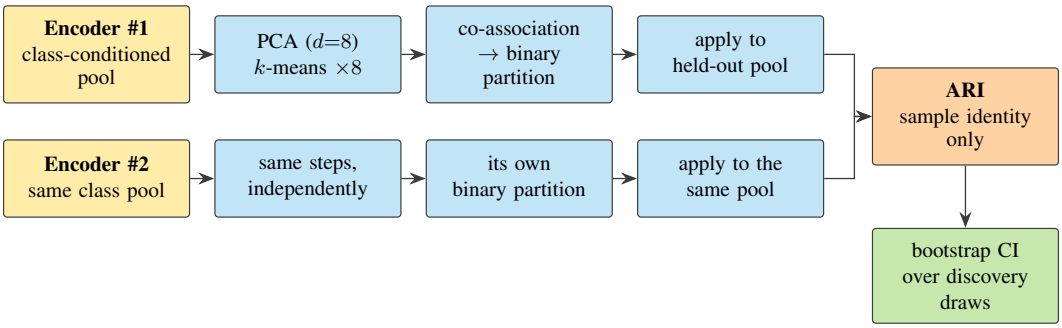


Figure 1: Comparison pipeline. Each encoder discovers a binary root partition in its own feature space and applies the fitted rule to the same held-out images. Cross-encoder comparison uses only sample assignments.

Objective families in visual pretraining. Joint-embedding methods match representations of augmented views through contrast, assignment, distillation, or redundancy reduction (Chen et al., 2021; Caron et al., 2020; Asano et al., 2020; Assran et al., 2022; Caron et al., 2021; Grill et al., 2020; Chen & He, 2021; Bardes et al., 2022; Zbontar et al., 2021). We refer to these as *cross-view whole-image* objectives. Label- and language-supervised objectives also define a whole-image target, but use external supervision. Hidden-content objectives reconstruct masked pixels or predict hidden tokens or latents (He et al., 2022; Bao et al., 2022; Baevski et al., 2022; Assran et al., 2023); DINOv2 combines a masked-token term with a whole-image distillation target (Oquab et al., 2024). Prior work finds differences in the global and local information encoded by contrastive and masked-image ViTs (Park et al., 2023). We examine whether the whole-image versus hidden-content distinction also predicts agreement between the dominant partitions of independently trained encoders, evaluate this association under matched training, and test its relevance to group transfer.

3 Method

We compare frozen encoders in two stages. Each encoder independently discovers a class-conditioned binary partition. We then apply all fitted partitioning rules to a shared held-out pool and compare only the resulting sample assignments, never raw feature coordinates (Figure 1).

3.1 Discovering and applying a root partition

Let $Z \in \mathbb{R}^{n \times d}$ contain the L2-normalized embeddings of a class-conditioned discovery pool ($n = 300$). Class labels define the pool, but no group labels are used to fit the partition. We project onto the top eight principal components, $f_i = (z_i - \bar{z})B$, and run $M = 20$-cluster k-means $N = 8$ times. If $c_i^{(t)}$ is the assignment in run t, the co-association matrix is

$$C_{ij} = \frac{1}{N} \sum_{t=1}^{N} \mathbb{1}[c_i^{(t)} = c_j^{(t)}]. \tag{1}$$

Average-linkage clustering of distance $1 - C$ produces a binary *root partition* $b_i \in \{0, 1\}$, our operational measure of an encoder's dominant split. The constants were inherited from an earlier group-discovery pipeline and were not selected for the present comparison. The primary contrast remains positive across the tested alternatives (Appendix A.5).

Let μ_0 and μ_1 be the two centroids in projected space. We store $(B, \bar{z}, \mu_0, \mu_1)$ and assign a reference image by

$$\hat{b}(z) = \mathbb{1}[\|(z - \bar{z})B - \mu_1\| < \|(z - \bar{z})B - \mu_0\|]\,. \tag{2}$$

The *held-out pool* is fixed by image identity and shared across encoders. No additional parameters are fitted on it. It shares images with discovery draws (60% on UrbanCars on average), so it is not a disjoint test set. Removing all overlapping images leaves the sign and interpretation of the main comparisons unchanged (Appendix A.16).

3.2 Partition agreement, stability, and inference

For encoders A and B on discovery draw r, let $a_r(A, B)$ be the ARI between their held-out binary assignments. ARI is invariant to swapping the two sides of either partition and has expectation zero for independent random partitions (Hubert & Arabie, 1985). It is not used here as a clustering-quality score because there is no ground-truth partition. Instead, it measures whether two encoders make the same sample-level distinction.

We interpret cross-encoder agreement in relation to partition stability. We define an encoder's *self-consistency* as its mean ARI across pairs of resampled discovery draws. On UrbanCars, 19 of the 21 encoders have self-consistency of 0.56–0.93. The two exceptions are DINOv2 (0.24) and ViT-L/16 (supervised) (0.39). We report raw ARI throughout and use agreement normalized by the geometric mean of the two self-consistencies only as a secondary diagnostic (Appendix A.17).

Our primary contrast compares within-set and between-set agreement. For a set S hypothesized to share structure and a comparison set S', define

$$\Delta = \frac{1}{R} \sum_{r} \left[\text{mean}_{\{A,B\} \subset S} \, a_r(A, B) - \text{mean}_{A \in S, B \in S'} \, a_r(A, B) \right]. \tag{3}$$

Within-S' agreement is a secondary diagnostic of whether the comparison set forms a coherent group. We compute one statistic per draw and obtain 95% percentile-bootstrap intervals from 5,000 resamples. Because discovery pools overlap across draws, these intervals quantify the stability of the procedure under resampling, not uncertainty over a population of independent images. Exact validity criteria, seed aggregation, and draw ranges appear in Appendix A.1–A.2.

3.3 Encoders and objective categories

We distinguish objectives by the locality of their target. *Whole-image targets* include class labels, captions, and representations of another augmented view. We further separate externally supervised whole-image models from self-supervised cross-view models. *Hidden-content objectives* predict masked pixels, tokens, or latent targets. DINOv2 is explicitly marked as a hybrid because it combines DINO's whole-image self-distillation with an iBOT masked-token term.

The initial preregistered analysis used eight encoders: six label- or language-supervised models, DINOv2-L, and MAE-L. The expanded core roster contains 15 objective- and architecture-diverse encoders: three externally supervised whole-image models (CLIP-L/14, SigLIP-L/16 and ViT-L/16 (supervised)), eight self-supervised cross-view models, and four models containing hidden-content prediction. Six additional language-supervised lineage variants test scale, pretraining data, a convolutional backbone, and a VLM tower in the appendix. Table 3 lists every model and target. The expanded taxonomy was developed after inspecting extension results. We therefore separate its exploratory analysis from a subsequent test on untouched draws in Section 4.3. When we describe the initial eight-encoder analysis, *supervised group* keeps its original meaning: the six label- or language-supervised encoders.

4 Experiments

We first evaluate partition agreement on UrbanCars (Li et al., 2023), where controlled background and co-object factors support interpretation of the discovered partitions. Subsequent experiments test fresh draws, additional datasets, matched training, controlled interventions, and group transfer. We interpret agreement using whole-image mean self-consistency and a regime threshold of 0.3. This threshold was fixed before EuroSAT's self-consistency was computed, although its contrast was already known. Draws without a valid two-sided partition are excluded by a fixed rule.

4.1 Finding I: agreement among externally supervised encoders

The initial eight-encoder protocol was fixed before its evaluation draws were analysed. On those draws, every pair among the six externally supervised models has positive ARI (0.31–0.78), whereas their agreement with DINOv2 is only 0.10–0.19 and five of their six pairs with MAE have non-positive ARI. The prespecified within-supervised minus supervised-to-DINOv2/MAE contrast summarizes this gap: $\Delta = +0.497$ $[0.449, 0.544]$ over 37 valid draws.

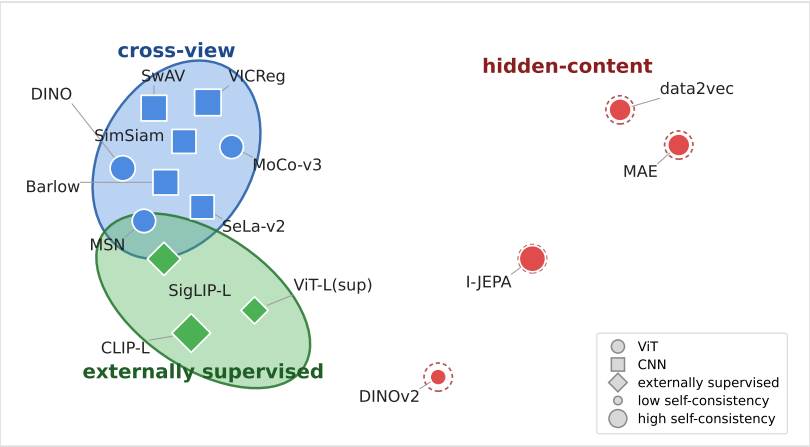


Figure 2: The 15 core encoders positioned by mutual root-partition agreement on UrbanCars using metric MDS of $1 - \text{ARI}$. The layout is illustrative (values: Appendix A.4; all 21 encoders: Figure 11). The whole-image models form two neighbouring shaded clusters, cross-view (blue) and externally supervised (diamonds). The hidden-content models are separated from these clusters and do not form a compact cluster of their own. Their highest pairwise agreement is between data2vec and MAE (0.32). Marker size denotes self-consistency.

	ext. supervised	cross-view	hidden-content
ext. supervised (3)	**0.48** ↑	**0.37** ↑	0.04 ↓
cross-view (8)	**0.37** ↑	**0.65** ↑	0.08 ↓
hidden-content (4)	0.04 ↓	0.08 ↓	0.06 ↓

Table 1: Mean root-partition ARI between objective categories on UrbanCars (draws 180–219); ↑/↓ mark high/low agreement. Diagonal entries average distinct encoder pairs within a category. Hidden-content includes the hybrid DINOv2 model; its low within-category agreement indicates that it does not form a coherent second family.

Self-supervision is not one coherent family here, since DINOv2 and MAE do not agree with each other, and architecture alone does not explain the result: of three ViT-L models without text supervision, only the label-supervised one aligns with the supervised group.

4.2 Finding II: Asymmetric agreement across objective categories

To test whether this pattern reflects external supervision or a broader property of the training target, we expand the roster. The 15-encoder roster distinguishes cross-view self-supervision from hidden-content prediction. DINO, VICReg, Barlow Twins, and SimSiam agree with one another at 0.57–0.81 and with the other cross-view models at per-encoder means of 0.60–0.65. Their agreement with the three externally supervised models is lower (0.32–0.39), and agreement with hidden-content models is at most 0.09. The hidden-content models, in turn, have low mutual agreement, except for data2vec–MAE (0.32). The resulting structure is therefore asymmetric: whole-image objectives produce related partitions, with the highest agreement among cross-view models, whereas hidden-content objectives do not form a coherent second family (Figure 2 and Table 1). Figure 3 illustrates one discovery draw. CLIP-L/14 and VICReg both separate images by background, whereas MAE separates a small set of dark images. The displayed images are selected by proximity to each side's mean and illustrate the partitions rather than establish their agreement.

Lineage and architecture checks preserve this separation, and SimMIM and BEiT-B agree most with MAE (Appendices A.9 and A.10).

4.3 Finding III: Fresh-draw confirmation and dataset dependence

The expanded grouping was developed through two revisions, so analyses of the development draws are exploratory with respect to that taxonomy. We therefore fixed five numerical predictions and

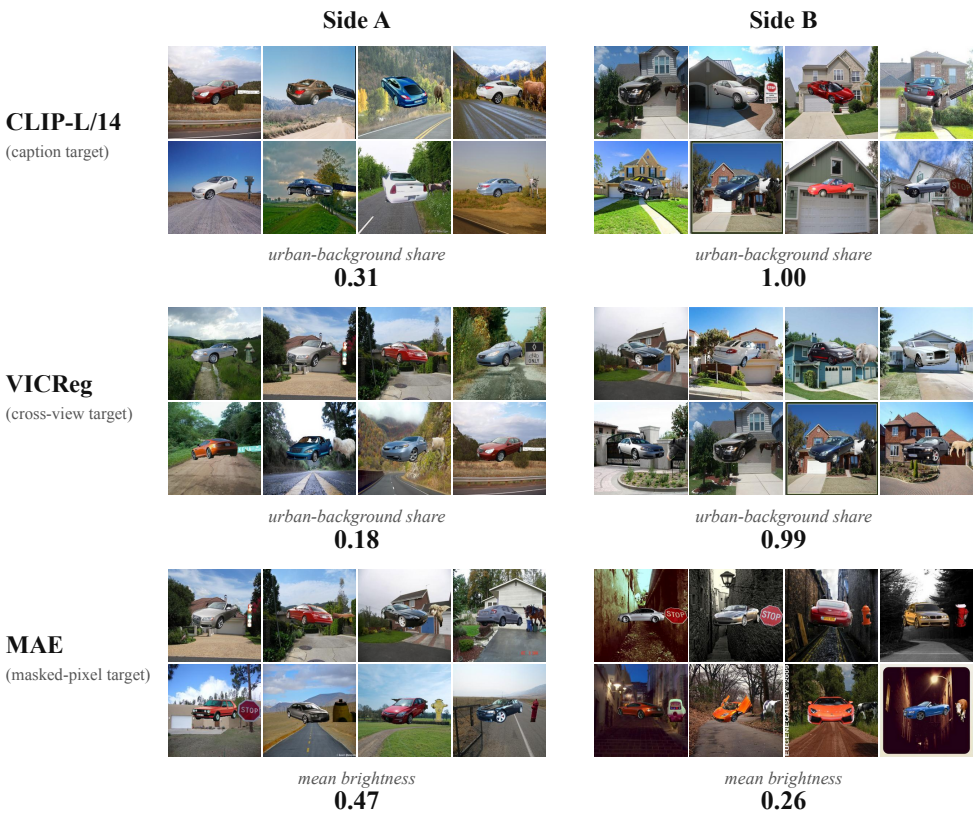
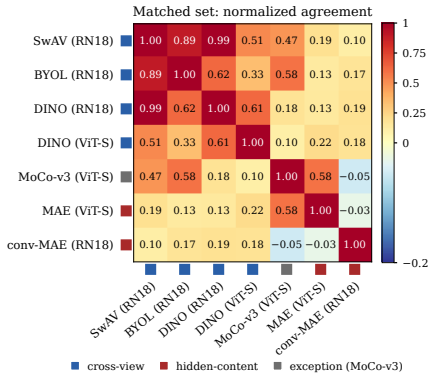


Figure 3: Three encoders' root partitions fitted on discovery draw 180 and applied to the same 300 held-out UrbanCars images (class 1). Each cell shows the eight images nearest its side's mean in the encoder's feature space. Side labels are oriented to agree with CLIP-L/14. Values are computed over all images on each side: urban-background share (CLIP, VICReg) or mean greyscale thumbnail brightness in $[0, 1]$ (MAE). Mean ARI with CLIP-L/14 over draws 180–219 is 0.32 for VICReg and -0.06 for MAE. Further montages appear in Appendix A.21.

evaluated them on previously unused draws 700–739. All five were satisfied, including positive whole-image versus hidden-content contrasts on UrbanCars (+0.428 $[0.395, 0.458]$) and CelebA (+0.152 $[0.134, 0.172]$), an intermediate externally supervised tier, and bounds on MAE and DINOv2 agreement (Appendix A.14).

Across datasets, the original eight-encoder contrast remains positive on CelebA (Liu et al., 2015) (+0.283 $[0.221, 0.351]$), ISIC (Codella et al., 2019; Bissoto et al., 2020) (+0.125 $[0.091, 0.163]$), and MetaShift (Liang & Zou, 2022) (+0.145 $[0.106, 0.187]$). However, the intermediate supervised tier reverses on ISIC, and latent-prediction models align with other groups on some datasets. MetaShift's whole-image mean self-consistency (0.17) is below the regime threshold, limiting interpretation of its positive contrast. EuroSAT (Helber et al., 2019) passes the threshold but does not support a positive primary contrast (-0.005 $[-0.017, 0.007]$).

Draw-based replications on Waterbirds and SpuCo Dogs (Joshi et al., 2023) (four class pools, all within the regime) reproduce the whole-image versus hidden-content contrast ($+0.047$ to $+0.105$). Within-cross-view agreement also exceeds cross-view-to-hidden-content agreement by 0.133–0.264, all with intervals above zero. The intermediate-tier criterion is met on three of the four pools, but the externally supervised models have low self-consistency there, limiting interpretation of their position (Appendix A.11). On the Spawrious corgi and bulldog pools, confidence intervals exclude zero for both the original eight-encoder contrast ($+0.298$, $+0.228$) and the 15-encoder whole-image versus hidden-content contrast ($+0.150$, $+0.198$). Both pools pass the stability criterion, but the intermediate-tier criterion is met only on bulldog (Appendix A.12). These results establish that self-consistency aids interpretation of weak agreement but does not determine whether objective-related structure emerges (Appendix A.13).



Run	Compared with	ARI
Cross-view RN18	each other	0.23–0.35
	each other (norm.)	0.62–0.99
MAE ViT-S	cross-view RN18	≤ 0.06
	pretrained MAE	0.37
conv-MAE RN18	cross-view RN18	0.04–0.11
	itself (SC)	0.69
DINO ViT-S	cross-view RN18	0.16–0.37

Figure 4: The matched set. Left: self-consistency-normalized agreement. Cross-view ResNet-18 runs have higher mutual agreement than agreement with masked-reconstruction runs. DINO-ViT-S also agrees with them, whereas MoCo-v3-ViT-S does not follow this ordering. Right: ARI for the comparisons cited in the text, including pretrained MAE, which is absent from the heatmap. Values are raw unless marked (SC: self-consistency; Appendix A.18).

Positive contrasts also occur one level below the root, with smaller values (+0.035 to +0.100 across four datasets). We treat this depth-2 result as supporting evidence. Full results and alignment corrections appear in Appendix A.6.

4.4 Matched training supports an objective-plus-recipe association

We compare training recipes on the same COCO images (Lin et al., 2014) (118k) for 200 epochs with batch size 256, holding data exposure and training duration fixed. SwAV, BYOL, and DINO use ResNet-18 with their prescribed cross-view augmentations, while MAE uses ViT-S with masked reconstruction. Additional DINO-ViT-S, MoCo-v3-ViT-S, and conv-MAE-ResNet-18 runs compare objectives within and across architectures. Each configuration is trained once with its objective-specific augmentation, so this experiment tests objective-plus-recipe differences, not the loss in isolation or seed variability.

The three cross-view ResNet-18 runs have mutual raw ARI of 0.23–0.35, or 0.62–0.99 after self-consistency normalization (Figure 4). They also agree more with five pretrained cross-view models than with the initial supervised six. MAE's agreement with these ResNet-18 runs is at most 0.06, whereas its agreement with pretrained MAE is 0.37 $[0.30, 0.45]$. Conv-MAE provides the within-architecture comparison. It agrees with the cross-view ResNet-18 runs at 0.04–0.11 despite self-consistency of 0.69. Conversely, DINO-ViT-S agrees with those runs at 0.16–0.37. Thus, architecture alone does not account for the ResNet-18 grouping.

The preregistered raw ARI threshold of 0.5 was not met, and the normalized analysis was introduced post hoc. Conv-MAE does not agree with other hidden-content models. MoCo-v3-ViT-S also departs from the proposed ordering, agreeing with MAE-ViT-S about as much as with cross-view runs. We retain the raw endpoint and report these exceptions with the normalized results (Appendix A.18).

4.5 Controlled interventions link partition selection to augmentation

An ingredient analysis over all 210 encoder pairs identifies associations with UrbanCars agreement. The two largest correlations in magnitude involve a shared whole-image target (Spearman ρ=+0.75) and hidden-content prediction in exactly one model (ρ=−0.76, backbone-identity permutation $p<0.001$). Shared text targets, shared pretraining data, and colour augmentation also correlate with agreement, so the observational roster does not identify an isolated causal ingredient.

Because the matched runs retain objective-specific augmentations, we test augmentation directly in a coloured-MNIST (LeCun et al., 1998) testbed with illumination, digit shape, and corner-patch factors. Architecture, data, and optimization budget are fixed across seven objectives. With brightness augmentation, cross-view root partitions have no measured alignment with illumination (ARI 0.00). Removing that augmentation makes SimCLR and BYOL partition by illumination (0.96)

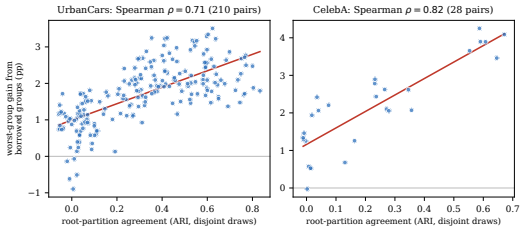


Figure 5: Root-partition agreement predicts the worst-group-accuracy gain when one encoder uses another's discovered groups. Each point is an unordered encoder pair, with transfer gain averaged over both directions. Predictor and outcome draws are disjoint; their image pools can overlap.

and agree with the masked-latent data2vec variant (0.89). Masked objectives retain illumination- or patch-aligned partitions (0.52–0.93). The intervention therefore changes partition alignment without changing the cross-view loss.

A second intervention increases corner-patch size while holding the illumination setting fixed. As patch variance increases relative to illumination variance, the root partition can shift towards the patch. The data2vec-lite model exhibits the clearest transition near the empirical variance reversal. BYOL becomes stably patch-aligned near that reversal, although its weak top-aligned factor changes earlier, and MAE-lite transitions earlier still. These results support an objective-dependent response to relative variance rather than a universal transition threshold (Appendix A.20).

4.6 Partition agreement predicts cross-encoder group transfer

We next test whether structural agreement predicts a downstream outcome. For each ordered pair $A \to B$, we discover class-conditioned pseudo-groups with A and train a group-reweighted linear head on standardized features from B. Transfer gain is the target's worst-group accuracy improvement over ERM. Pairwise outcomes average both transfer directions. The predictor is ARI from draw indices disjoint from those used to measure transfer (Figure 5).

Across 210 UrbanCars pairs, transfer gain correlates with partition agreement at ρ=0.71 (p<0.0002 by backbone-identity permutation). The corresponding CelebA correlation is 0.82 (p=0.0028, 28 pairs). Controlling for linear CKA leaves partial ρ of +0.57 and +0.45, respectively. In the 15-encoder UrbanCars and eight-encoder CelebA analyses, the association also persists among cross-lineage pairs (ρ=0.72, 0.78) and after controlling for self-consistency (partial ρ +0.51, +0.83). Partition agreement therefore predicts transfer beyond these controls in the tested settings.

Practical gains are modest. For the 15-encoder UrbanCars roster at an induced spurious-correlation strength of $c = 0.8$ and the eight-encoder CelebA roster at natural skew, true-group reweighting improves worst-group accuracy by 6.9 and 5.2 points, respectively. Selecting the highest-agreement source yields 2.3 and 3.2 points, compared with 1.9 and 2.2 for a random source, but it does not consistently outperform the target encoder's own groups. These results support source ranking rather than a standalone group-selection algorithm. A complementary intervention removes each encoder's root axis, the difference between the mean standardized features of its two root groups, before refitting the head. Worst-group accuracy improves for whole-image encoders, whereas MAE's confidence intervals include zero. Random directions and the first principal component produce smaller mean gains for the whole-image encoders (Appendix A.8).

5 Discussion

5.1 Evidence for objective-dependent partition structure

The results support an association between objective targets and the dominant partitions of frozen representations. On UrbanCars, contrastive learning, assignment, distillation, and redundancy reduction yield related partitions despite differences in loss formulation. Whole-image targets may favour distinctions that are stable across views or informative about an image-level label or caption. This interpretation is consistent with the observed agreement, but the pretrained roster alone does not identify its cause.

The controlled testbed provides a narrower mechanistic result. Changing brightness augmentation changes whether a cross-view root partition aligns with illumination. Among stably expressed factors, the variance intervention can shift alignment towards a factor whose relative variance increases, and the transition differs across objectives. These experiments support augmentation-dependent suppression and objective-dependent selection, not a universal rule for the dominant factor.

Hidden-content models show a less uniform pattern. Across the original four datasets, MAE has low agreement with whole-image models. The three latent-prediction models are jointly isolated on UrbanCars, but some align with supervised partitions on CelebA or ISIC. Different pixel, token, and latent targets may favour different statistics, but the present evidence does not isolate these effects. It supports whole-image convergence more directly than a common hidden-content mechanism.

DINOv2 differs from DINO in its masked-token term, data, scale, and regularization. An ImageNet-1k iBOT model agrees with DINO at 0.70, so a masked-token term alone does not explain DINOv2's separation. DINOv2's low self-consistency (0.24) further limits interpretation of its disagreement.

5.2 Implications for representational convergence

Our findings qualify an objective-independent reading of the Platonic Representation Hypothesis (Huh et al., 2024): eight ImageNet-1k models share a dominant UrbanCars partition (ARI 0.56–0.81), whereas DINOv2, trained on 142M images, agrees with no core encoder above 0.25. This concerns one dataset-conditional partition and is limited by DINOv2's low self-consistency, so it does not refute convergence; it shows that scale alone does not ensure agreement of the dominant partition within this roster.

5.3 Scope and limitations

The empirical pattern is dataset-dependent. CelebA and ISIC exhibit weaker or reordered agreement, MetaShift has low partition stability, and EuroSAT passes the stability criterion without a positive primary contrast. Spawrious supports the whole-image versus hidden-content contrast on both pools, but the intermediate-tier criterion is met only on bulldog. On Waterbirds and SpuCo Dogs, positive contrasts coexist with low self-consistency among the externally supervised models, so these results support cross-view agreement more directly than coherence across all whole-image models. Self-consistency is therefore an interpretive diagnostic, not a sufficient criterion for the proposed structure.

Matched runs fix images, epochs, and batch size but not augmentation, and the controlled interventions use a small synthetic testbed. The taxonomy was developed after the extension results; the fresh-draw test confirms its predictions without making that development confirmatory.

Failed predictions further delimit the findings. Whole-image agreement is not uniform on ISIC, and latent-prediction isolation does not generalize across datasets. In the matched set, the raw ARI threshold fails, conv-MAE does not join other hidden-content runs, and MoCo-v3-ViT-S does not follow the proposed ordering. Cross-view models in the testbed do not converge on shape. Downstream transfer provides a ranking signal, not a standalone practical method.

We claim structural agreement, not quality. We measure whether encoders induce similar dominant partitions. This comparison does not establish that the shared distinction is semantic, optimal, or useful.

6 Conclusion

What a frozen vision encoder splits images on first depends on whether its objective matches a whole-image target or predicts hidden content. On UrbanCars, whole-image objectives induce related partitions, whereas hidden-content objectives neither share those partitions nor form a single alternative family. Matched training supports an objective-plus-recipe association, and controlled interventions link partition selection to augmentation and relative factor variance. Partition agreement also predicts cross-encoder group transfer, with modest worst-group accuracy gains. These findings concern structural agreement rather than partition quality, and their cross-dataset failures define the limits of the proposed account.

AI USE STATEMENT

We used generative AI tools (Claude Code and OpenAI ChatGPT/Codex) to assist with experiment and analysis code, debugging, manuscript drafting and editing, claim–result consistency checks, and related-work citation checks. Experimental design, stopping decisions, interpretation, and claim scope were determined by the authors. The authors reviewed all AI-assisted text and code, checked reported numbers against source scripts and result files, and take responsibility for the final content. AI-assisted audits identified five errors in earlier analyses. All were corrected and are disclosed in Appendices A.8, A.6, and A.15. Superseded results are not used in the paper.

ETHICS STATEMENT

CelebA and MetaShift are used as released, without new data collection. ISIC2018 is used through its public bias-annotated subset. The observation that a depth-2 partition on CelebA aligns with apparent gender is reported as evidence about representation structure, not as a method for inferring a protected attribute. We do not propose or endorse inferring gender or other sensitive attributes from images without consent. The intended use of the group-discovery analysis is defensive: diagnosing and reducing reliance on spurious features rather than identifying individuals or attributes.

REPRODUCIBILITY STATEMENT

Appendix A.2 reports every protocol constant, seed, and draw range, and Appendix A.1 defines every statistic. One deterministic script generates headline numbers and tables from cached per-encoder partitions and experiment outputs. The linear CKA ablation and matched-set partitions are recomputed from extracted feature files. Each preregistered prediction is stated with the experiment that evaluates it, and corrections and failed predictions are retained in the appendix. Code, configurations, cached partitions, and result files are available at `https://anonymous.4open.science/r/OISE-ED68/`.

REFERENCES

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision (ECCV)*, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, 2022.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? Not so fast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *International Conference on Computer Vision (ICCV)*, 2021.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *International Conference on Machine Learning (ICML)*, 2025.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)*, 2021.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations (ICLR)*, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark and a more realistic dataset. *arXiv preprint arXiv:2306.11957*, 2023.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations (ICLR)*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Aengus Lynch, Gbètondji J-S Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

Zheda Mai, Ke Zhang, Fu-En Wang, Zixiao Ken Wang, Albert Y. C. Chen, Lu Xia, Min Sun, Wei-Lun Chao, and Cheng-Hao Kuo. Revisiting model stitching in the foundation model era. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations (ICLR)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

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

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

A Appendix

A.1 Formal definitions

Throughout, an encoder E is frozen. We denote by $z = E(x) \in \mathbb{R}^d$ its L2-normalized embedding of image x, extracted once and then held fixed.

Discovery draw and held-out pool. For class y and draw index r, the discovery pool D_r is 300 images of class y sampled without replacement from the validation split with $\texttt{RandomState}(910000 + 7919r)$. The held-out pool H is the pool of draw $r{=}999$, fixed by image identity and shared by every encoder.

Root partition (Eq. 1). With $Z \in \mathbb{R}^{n\times d}$ the embeddings of D_r, $\bar{z}$ their mean and $B \in \mathbb{R}^{d\times 8}$ the top-8 PCA basis, $f_i = (z_i - \bar{z})B$. For $t = 1, \ldots, N$ ($N{=}8$), k-means with $M{=}20$ clusters

gives labels $c_i^{(t)}$. We form $C_{ij} = \frac{1}{N}\sum_t \mathbb{1}[c_i^{(t)} = c_j^{(t)}]$. Average-linkage agglomerative clustering on distance $1 - C$ with two clusters gives $b_i \in \{0, 1\}$. A split is valid only if both sides have ≥ 5 images. The stored rule is $(B, \bar{z}, \mu_0, \mu_1)$ with $\mu_s = \text{mean}\{f_i : b_i = s\}$.

Held-out application (Eq. 2). On H, $\hat{b}(z) = \mathbb{1}[\|(z - \bar{z})B - \mu_1\| < \|(z - \bar{z})B - \mu_0\|]$. No parameters are refit on H. Because H and D_r are drawn from the same validation images, they share images (on UrbanCars 60% of H on average). Thus H is a fixed reference pool shared by every encoder, not a disjoint test set.

Seeds and the per-draw partition. Each draw is run with S=5 seeds, giving $\hat{b}^{(s)}$ on H. Seeds are sign-aligned to seed 1 (flip $\hat{b}^{(s)} \mapsto 1 - \hat{b}^{(s)}$ if it agrees with $\hat{b}^{(1)}$ on fewer than half of H) and combined by majority vote, $\hat{b}_r = \mathbb{1}[\frac{1}{S}\sum_s \hat{b}^{(s)} \geq \frac{1}{2}]$, on the intersection of the seeds' valid masks.

Cross-encoder agreement. For encoders A, B and draw r, $a_r(A, B) = \text{ARI}(\hat{b}_r^A, \hat{b}_r^B)$ on the images of H valid for both, provided at least 8 such images and two sides on each. ARI is the Hubert–Arabie index, $(\text{RI} - \mathbb{E}[\text{RI}])/(\max \text{RI} - \mathbb{E}[\text{RI}])$, 0 in expectation for independent partitions. The reported value is $\bar{a}(A, B) = \frac{1}{R}\sum_r a_r(A, B)$ over the R draws valid for the pair, with a percentile bootstrap over draws (5000 resamples).

Primary contrast (Eq. 3) and secondary diagnostic. Δ is the draw-level mean of within-S minus S-to-S' mean ARI. The secondary diagnostic is the mean a_r within S'.

Self-consistency and normalized agreement. For an encoder A, $\text{SC}(A) = \text{mean}_{r<r'} \text{ARI}(\hat{b}_r^A, \hat{b}_{r'}^A)$ over draws on which A produced a two-sided partition. The normalized agreement of a pair is $\bar{a}(A, B)/\sqrt{\text{SC}(A)\,\text{SC}(B)}$, clipped to $[-1, 1]$.

Depth-2 conditional contrast. Each root side is split again by the same rule, giving a depth-2 label $\ell \in \{0, 1, 2, 3\}$ with root side $\lfloor \ell/2 \rfloor$ and sub-side $\ell \bmod 2$. After aligning the two encoders' root orientations, for a pair on draw r and root side $s \in \{0, 1\}$, let I_s be the images of H both encoders assign to root side s. If $|I_s| \geq 8$ with two sub-sides each, $a_{r,s}^{(2)}(A, B) = \text{ARI}(\ell^A \bmod 2,\ \ell^B \bmod 2)$ on I_s, and $a_r^{(2)} = \text{mean}_s\, a_{r,s}^{(2)}$. The depth-2 contrast applies Eq. 3 to $a_r^{(2)}$.

Group transfer. For an ordered pair $A \to B$, pseudo-groups on B's head-training set are $(y, \hat{b}^A)$. A logistic head on per-dimension standardized features of B is fit with inverse-group-frequency weights, and $\text{WGA}_B(\cdot)$ is its worst-group accuracy over the true cells of the test set. The transfer gain is $G(A \to B) = \text{WGA}_B(\text{groups}_A) - \text{WGA}_B(\text{ERM})$, the pair value is $\frac{1}{2}[G(A \to B) + G(B \to A)]$, and the predictor is $\bar{a}(A, B)$ computed on disjoint draws. Association is Spearman ρ over pairs, with p from 5,000 permutations of backbone identities applied to the ARI matrix.

Axis removal. On a draw's head-training pool, root sides are discovered on the raw embeddings and features are then standardized (fit on that pool). The encoder's own axis is $a = (\mu_1 - \mu_0)/\|\mu_1 - \mu_0\|$ with μ_s the mean standardized feature of root side s of the class-1 images. We project standardized train and test features as $z' = z - (z \cdot a)\, a$ and refit the same ERM head. We report Δ WGA with a draw-level bootstrap.

Ingredient analysis. Each of the 21 objectives is coded on nine binary ingredients $g_k(E)$. For pair (A, B) the features are $\mathbb{1}[g_k(A) = g_k(B) = 1]$ (both) and $\mathbb{1}[g_k(A) \neq g_k(B)]$ (exactly one). The outcome is $\bar{a}(A, B)$. Each feature's Spearman ρ over the 210 pairs is tested by permuting backbone identities.

A.2 Full protocol constants and draw-range ledger

A.3 Chronology of the roster extensions and grouping revisions

The self-supervised extensions were evaluated sequentially as the grouping hypothesis was revised. MSN and SwAV were preregistered. SeLa-v2 and MoCo-v3 used the same fixed protocol on previously unused draws (Appendix A.4). MSN (Assran et al., 2022), trained without labels or text, agreed with the initial supervised six (0.28–0.51). This observation motivated a hypothesis based on discrete, balanced categorical assignment. SwAV and SeLa-v2 (Caron et al., 2020; Asano et al., 2020), both using Sinkhorn–Knopp balanced codes, were consistent with that hypothesis. They

Constant	Value
Sampling	
Discovery draw r	300 images per class, drawn without replacement from the validation split with `RandomState`$(910000{+}7919\,r)$.
Held-out pool	Class-conditioned pool of draw $r{=}999$, fixed by image identity, identical for every encoder; no rule is fit on it. Shares images with the discovery pools (UrbanCars 60% on average).
Seeds per draw	5; sign-aligned majority vote over the seeds that produce a root partition. A draw enters a pair's ARI only if the two votes share ≥ 8 held-out images with two sides each.
Recurrence between discovery pools	UrbanCars 60% (300 of 500 class-1 validation images per draw); CelebA and ISIC 15–26%. CIs therefore measure procedure stability, not image-sampling variation.
Partition pipeline	
Root partition	PCA $d_{\text{PCA}}{=}8 \to k$-means $M{=}20$, $N_{\text{runs}}{=}8 \to$ co-association $\to$ average-linkage 2-cut; nearest-centroid application to the held-out pool.
Depth-2 variant	Fixed pipeline as at the root (specified before the adaptive variant). Branch-size-adaptive variant ($M{=}\max(5, \lfloor n/10 \rfloor)$, $N_{\text{runs}}{=}16$ below 150 images) reported second (Appendix A.6).
Uncertainty	Draw-level paired percentile bootstrap, 5000 resamples, 95%.
Draw ranges	
Development	160–169 (protocol frozen after these).
Primary contrast (8-encoder)	UrbanCars 220–259 ($n{=}37$; Figure 6), CelebA 40–79 (39), ISIC 50–89 (40), MetaShift 50–89 (39), EuroSAT 50–89 (40), UrbanCars+RN50x64 380–419 (38); each opened once, after the protocol was frozen.
Roster / dataset / matched set	180–219 (40 attempted draws) on every dataset (Tables 25, 29, Figure 7). Eight-backbone contrast recomputed here: +0.476 $[0.429, 0.525]$ (UrbanCars), +0.334, +0.111, +0.115, all inside the primary ranges' CIs.
Depth-2 draws	UrbanCars 1000–1039, others 200–239.
Confirmation of the final grouping	700–739 on UrbanCars and CelebA, opened once after the five predictions were fixed (Appendix A.14).
Pipeline checks	N_{runs} sensitivity 800–819; design ablation 900–919 (Appendix A.5).
Waterbirds, SpuCo Dogs	180–219 within each class pool of the test split, five seeds per draw (Appendix A.11).
Spawrious	180–219 within each breed pool; the 15-encoder analysis uses one seed per draw (Appendix A.12).
Transfer predictor	640–659, disjoint from the outcome draws 180–219 (Section 4.6); axis removal 180–199.

Table 2: Protocol constants, grouped by role. None was chosen on a confirmatory result. Appendix A.5 evaluates alternative pipeline settings.

agreed with MSN and the supervised models and with each other at 0.82 on the draws then available.

On the common draws 180–219, the highest agreement is CLIP-L/14–CLIP-L/14 (LAION) (0.82), two instances of the same objective trained on different data. The highest self-supervised agreement is VICReg–Barlow Twins (0.81). MoCo-v3 (Chen et al., 2021) then provided a test without balanced assignment. Its mean agreement was 0.32 with the initial supervised six and ≤ 0.10 with DINOv2 and MAE. Its largest agreements were instead with MSN, SwAV, and SeLa-v2 (0.56–0.66). This result motivated revising the original grouping.

A.4 Full pairwise tables

Per-pair numbers for the eight-encoder roster and for every extension in Section 4 follow. Figure 7 summarizes the sequence of roster extensions before the pairwise tables.

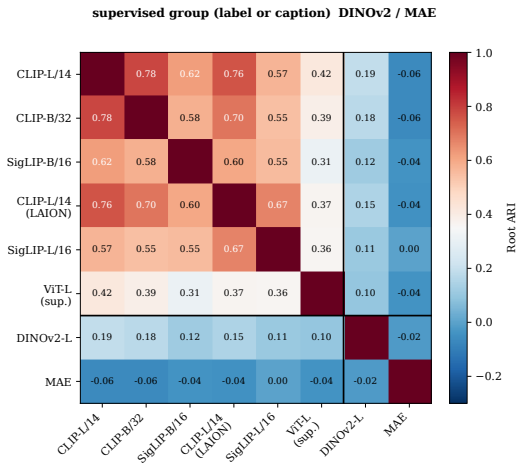


Figure 6: Full pairwise root ARI, UrbanCars confirmatory sample (numbers in Table 5). The supervised block (top-left) is positive throughout. Agreement with DINOv2 and MAE (right) is lower, including for the DINOv2–MAE pair.

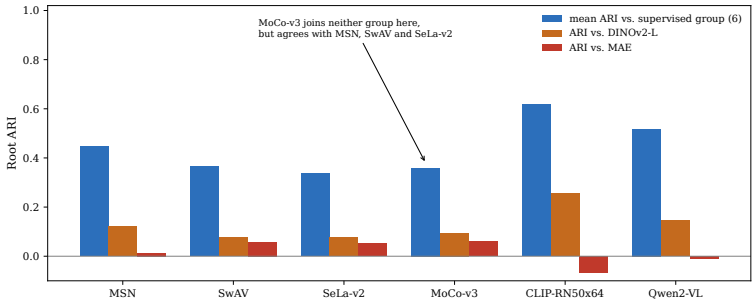


Figure 7: The six roster additions in the order they were tested. All used the same frozen protocol on unopened draws. MSN, SwAV, and Qwen2-VL were preregistered. Each addition has higher mean agreement with the initial supervised six than with DINOv2 or MAE. MoCo-v3, the addition that first broke the balanced-assignment account, joins neither group here but agrees with MSN, SwAV and SeLa-v2 (Section 4.2). Values are on the common draws 180–219.

Backbone	Objective family	Target	Arch.
CLIP-L/14	language (image–text)	whole-image	ViT
CLIP-B/32	language (image–text)	whole-image	ViT
CLIP-B/16	language (image–text)	whole-image	ViT
CLIP-L/14 (LAION)	language (image–text)	whole-image	ViT
CLIP RN50x64	language (image–text)	whole-image	CNN
SigLIP-B/16	language (image–text)	whole-image	ViT
SigLIP-L/16	language (image–text)	whole-image	ViT
Qwen2-VL	language (VLM tower)	whole-image	ViT
ViT-L (sup.)	supervised (labels)	whole-image	ViT
MSN-L	self-sup. (masked siamese)	whole-image	ViT
SwAV	self-sup. (clustering)	whole-image	CNN
SeLa-v2	self-sup. (clustering)	whole-image	CNN
MoCo-v3	self-sup. (contrastive)	whole-image	ViT
DINO	self-sup. (self-distill.)	whole-image	ViT
VICReg	self-sup. (redundancy red.)	whole-image	CNN
Barlow Twins	self-sup. (redundancy red.)	whole-image	CNN
SimSiam	self-sup. (self-distill.)	whole-image	CNN
MAE-L	masked reconstruction	hidden content	ViT
data2vec	masked latent regression	hidden content	ViT
I-JEPA	masked latent prediction	hidden content	ViT
DINOv2-L	self-distill. + masked token	whole+hidden	ViT

Table 3: The full 21-backbone roster by objective family, target locality, and architecture. The distinction in Section 4 follows the *target* column, separating whole-image objectives (language-, label-, and self-supervised) from hidden-content prediction. DINOv2 carries both a whole-image and a masked-token term.

Group	Backbone	Training paradigm	Reference
Supervised (label or caption)	CLIP-L/14	image-text contrastive	(Radford et al., 2021)
Supervised (label or caption)	CLIP-B/32	image-text contrastive	(Radford et al., 2021)
Supervised (label or caption)	CLIP-L/14 (LAION)	image-text contrastive	(Radford et al., 2021)
Supervised (label or caption)	SigLIP-B/16	image-text sigmoid-contrastive	(Zhai et al., 2023)
Supervised (label or caption)	SigLIP-L/16	image-text sigmoid-contrastive	(Zhai et al., 2023)
Supervised (label or caption)	ViT-L/16 (supervised)	class-label supervised	(Dosovitskiy et al., 2021)
Self-supervised (first grouping)	DINOv2-L	self-distillation + masked tokens	(Oquab et al., 2024)
Self-supervised (first grouping)	MAE (ViT-L)	masked-pixel reconstruction	(He et al., 2022)

Table 4: The eight-encoder confirmatory roster and grouping, frozen before the confirmatory sample (Section 4) was analysed.

Qwen2-VL as a vision-language tower. Qwen2-VL (Wang et al., 2024) extends the comparison to a vision tower trained with a multimodal language model. Its training includes ∼600B-token image–text alignment, followed by joint training on ∼800B multimodal tokens. Unlike LLaVA-1.5 (Liu et al., 2024), it does not use a pretrained CLIP tower (Appendix A.7). In the preregistered evaluation (n=40), agreement with the six externally supervised encoders is 0.42–0.66, compared with 0.20 for DINOv2 and −0.02 for MAE. The group contrast is $+0.026$ $[-0.020,\ 0.072]$. These results place the final vision tower near the externally supervised models, but do not isolate the effect of its autoregressive training stage.

A.5 Full pipeline-design ablation

Every confirmatory number in the main text also uses a fixed N_{runs}=8 k-means repeats per co-association matrix (Section 3.1). Re-running the primary contrast on a fresh, previously unopened draw range (n=16–18) at $N_{\text{runs}} \in \{4, 8, 16\}$ gives 0.466 $[0.391, 0.538]$, 0.497 $[0.436, 0.554]$, and 0.499 $[0.437, 0.560]$ respectively (Figure 8). The paper's own value reproduces on this independent sample, and all three intervals overlap.

We extended this check to three further pipeline design choices: PCA dimensionality D_{PCA}=8, consensus cluster count M=20, and Adjusted Rand Index as the agreement statistic. None of these was tuned against the primary contrast. The constants were inherited from an earlier label-free

Pair	Root ARI (draw-level, 95% CI)
CLIP-L/14–CLIP-B/32	0.784 [0.739, 0.822]
CLIP-L/14–SigLIP-B/16	0.621 [0.563, 0.671]
CLIP-L/14–CLIP-L/14(LAION)	0.757 [0.693, 0.818]
CLIP-L/14–SigLIP-L/16	0.572 [0.517, 0.626]
CLIP-L/14–ViT-L(supervised)	0.418 [0.308, 0.526]
CLIP-B/32–SigLIP-B/16	0.582 [0.522, 0.632]
CLIP-B/32–CLIP-L/14(LAION)	0.699 [0.637, 0.759]
CLIP-B/32–SigLIP-L/16	0.547 [0.486, 0.607]
CLIP-B/32–ViT-L(supervised)	0.390 [0.288, 0.491]
SigLIP-B/16–CLIP-L/14(LAION)	0.600 [0.545, 0.649]
SigLIP-B/16–SigLIP-L/16	0.554 [0.494, 0.608]
SigLIP-B/16–ViT-L(supervised)	0.314 [0.226, 0.396]
CLIP-L/14(LAION)–SigLIP-L/16	0.674 [0.606, 0.733]
CLIP-L/14(LAION)–ViT-L(supervised)	0.369 [0.268, 0.469]
SigLIP-L/16–ViT-L(supervised)	0.361 [0.273, 0.447]
CLIP-L/14–DINOv2-L	0.188 [0.122, 0.255]
CLIP-B/32–DINOv2-L	0.177 [0.115, 0.238]
SigLIP-B/16–DINOv2-L	0.119 [0.073, 0.165]
CLIP-L/14(LAION)–DINOv2-L	0.148 [0.095, 0.203]
SigLIP-L/16–DINOv2-L	0.110 [0.068, 0.154]
ViT-L(supervised)–DINOv2-L	0.096 [0.041, 0.153]
CLIP-L/14–MAE	-0.063 [-0.070, -0.056]
CLIP-B/32–MAE	-0.063 [-0.073, -0.052]
SigLIP-B/16–MAE	-0.038 [-0.045, -0.031]
CLIP-L/14(LAION)–MAE	-0.042 [-0.055, -0.029]
SigLIP-L/16–MAE	0.002 [-0.017, 0.030]
ViT-L(supervised)–MAE	-0.036 [-0.050, -0.021]
DINOv2-L–MAE	-0.019 **[**-0.031**,** -0.007**]**

Table 5: Full pairwise root ARI, UrbanCars confirmatory sample (Section 4.1).

Pair	Root ARI (95% CI)
RN50x64–CLIP-L/14	0.771 [0.691, 0.836]
RN50x64–CLIP-B/32	0.748 [0.663, 0.823]
RN50x64–CLIP-L/14(LAION)	0.692 [0.597, 0.780]
RN50x64–SigLIP-B/16	0.566 [0.497, 0.631]
RN50x64–SigLIP-L/16	0.481 [0.391, 0.565]
RN50x64–ViT-L(supervised)	0.384 [0.289, 0.473]
RN50x64–DINOv2-L	0.229 [0.148, 0.310]
RN50x64–MAE	-0.064 [-0.076, -0.050]

Table 6: RN50x64 vs. roster, n=38 (draws 380–419, Table 2).

group-discovery pipeline built for a different purpose. In that pipeline, d_{PCA} had been chosen on CelebA with true group labels, without any encoder comparison. We had not checked whether the headline result depended on these choices. We re-ran the primary contrast for the frozen eight-encoder roster, class 1, on a second fresh, previously unopened draw range (n=20, draws 900–919). We varied one choice at a time from the paper's baseline. The usual minimum-cluster-size floor excluded some draws in each configuration. The retained count is noted in each row.

Figure 9 visualizes the same sweep.

A.6 Depth-2 analysis, implementation corrections, and sensitivity

The conditional contrast in Section 4.3 required an audit of the extraction code. Depth-2 child nodes are keyed by each seed's *raw*, unaligned root branch. The routine combining them into one per-draw label instead selected a child by the *aligned* root label. Because k-means cluster indices carry no canonical sign, a seed's orientation is arbitrarily flipped relative to the reference roughly 50% of the time. For these seeds, the mismatch silently zeroed their contribution to the whole draw. We confirmed the error by instrumenting synthetic and real cases. Among 10 consecutive draws, 3 with

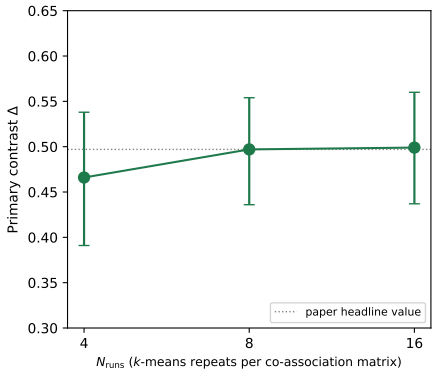


Figure 8: Primary contrast as a function of N_{runs}, on a fresh draw range never used for any other confirmatory number in this paper.

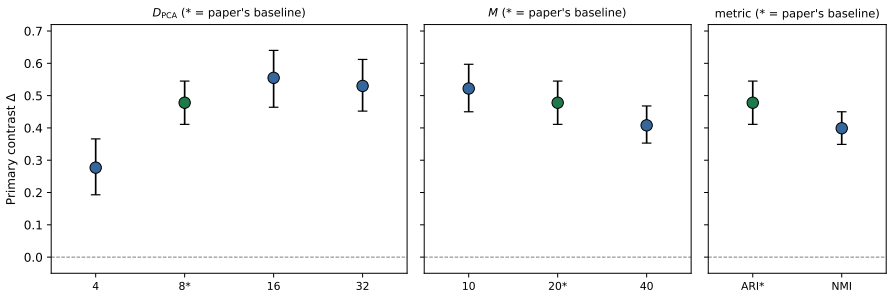


Figure 9: The same ablation as Table 12, plotted with one panel per design choice. The paper's baseline (starred, green) is repeated in each panel. All seven distinct configurations exclude zero, and the baseline is not an outlier in either direction.

Pair	Root ARI (95% CI)
MSN–CLIP-L/14	0.455 [0.393, 0.524]
MSN–CLIP-B/32	0.430 [0.366, 0.498]
MSN–SigLIP-B/16	0.463 [0.423, 0.505]
MSN–CLIP-L/14(LAION)	0.505 [0.443, 0.573]
MSN–SigLIP-L/16	0.488 [0.425, 0.550]
MSN–ViT-L(supervised)	0.275 [0.199, 0.355]
MSN–DINOv2-L	0.116 [0.073, 0.165]
MSN–MAE	0.003 [−0.011, 0.016]

Table 7: MSN vs. roster, confirmatory scale (Appendix A.3).

Pair	Root ARI (95% CI)
SwAV–CLIP-L/14	0.361 [0.316, 0.397]
SwAV–CLIP-B/32	0.355 [0.302, 0.411]
SwAV–SigLIP-B/16	0.425 [0.367, 0.475]
SwAV–CLIP-L/14(LAION)	0.448 [0.386, 0.512]
SwAV–SigLIP-L/16	0.605 [0.532, 0.669]
SwAV–ViT-L(supervised)	0.247 [0.185, 0.310]
SwAV–DINOv2-L	0.074 [0.044, 0.106]
SwAV–MAE	0.043 [0.026, 0.067]
SwAV–MSN	**0.663 [0.598, 0.723]**

Table 8: SwAV vs. roster and vs. MSN, confirmatory scale (Appendix A.3).

a flip collapsed to zero retained points. We fixed the error by re-associating each child with its seed's actual orientation before combining. After the fix, 0 draws collapsed. The fix alone raised confirmatory retention from an average of 4.2–4.9 of 15 possible supervised–supervised pairs per draw to 12–13. It does not affect any root-level result, which never calls this code path.

A second finding from the audit concerned branch size rather than a bug. Root-branch imbalance is associated with depth-2 formation failure (Spearman ρ=0.735, $p<10^{-4}$, n=24 backbone×dataset cells). A smaller branch gives the M=20-cluster consensus step too few points per cluster to be stable. We address this with a branch-size-scaled variant. For $n < 150$, it uses $M_{\text{cluster}} = \max(5, \lfloor n/10 \rfloor)$ and doubles the consensus repeats. It is identical to the fixed pipeline at the n=300 root scale and changes only smaller depth-2/depth-3 branches. This variant was introduced after the fixed pipeline's confirmatory results were seen, so the main text reports the fixed pipeline first (Table 13). The conditional contrast compares two encoders' depth-2 splits only on the samples they place on the same root side. Because root labels are arbitrary per encoder, we align the second encoder's root orientation to the first before conditioning. An earlier version omitted this cross-encoder alignment. Applying it shifts the four fixed-pipeline contrasts by at most 0.02. On ISIC, it moves the lower confidence bound from just below zero to just above it. Table 13 reports the aligned values. Under the fixed pipeline, the depth-2 contrast excludes zero on all four datasets, although only marginally on ISIC (+0.035 $[0.006, 0.062]$). The adaptive variant raises the contrasts, lifting ISIC to +0.040 $[0.016, 0.065]$, with its interval farther above zero.

For leave-one-backbone-out sensitivity, we drop each backbone in turn and recompute the fixed-pipeline contrast from the conditional pairwise agreements. The contrast stays positive for every removal except MAE, the most isolated backbone. Removing MAE shrinks the whole-image/hidden-content gap. The contrast stays positive on UrbanCars, ISIC, and MetaShift but falls to near zero on CelebA (-0.01). The depth-2 gap therefore rests in part on MAE's separation from the other encoders. Figure 10 shows one representative CelebA draw.

Number and content of depth-2 splits. At matched confirmatory scale (n=40, stability-screened against a geometry-randomized null, minimum-size floor of 10 applied post hoc) on UrbanCars, we ran this screen on the eight-encoder roster plus MSN. BEiT-L was tried and excluded because it produced no genuine bipartition. Only CLIP-L/14 and CLIP-B/32 expose any confirmed depth-2 occurrence (five each). The other seven, DINOv2 and MAE among them, expose confirmed structure only at the root. CLIP-B/32's depth-2 occurrences re-split the depth-1 background factor (BA_{bg} 0.759, $BA_{\text{co-obj}}$ 0.567). CLIP-L/14's track the co-object factor above chance and above its own BA_{bg} (0.733 vs. 0.649). A follow-up asking whether this was driven by CLIP's training data

Pair	Root ARI (95% CI)
SeLa-v2–CLIP-L/14	0.352 [0.322, 0.378]
SeLa-v2–CLIP-B/32	0.357 [0.315, 0.403]
SeLa-v2–SigLIP-B/16	0.415 [0.380, 0.447]
SeLa-v2–CLIP-L/14(LAION)	0.405 [0.364, 0.452]
SeLa-v2–SigLIP-L/16	0.552 [0.484, 0.617]
SeLa-v2–ViT-L(supervised)	0.254 [0.195, 0.315]
SeLa-v2–DINOv2-L	0.067 [0.042, 0.093]
SeLa-v2–MAE	0.026 [0.019, 0.032]
SeLa-v2–MSN	0.682 [0.607, 0.748]
SeLa-v2–SwAV	**0.817 [0.736, 0.883]**

Table 9: SeLa-v2 vs. roster and vs. MSN/SwAV, confirmatory scale (Appendix A.3).

Pair	Root ARI (95% CI)
MoCo-v3–CLIP-L/14	0.282 [0.222, 0.338]
MoCo-v3–CLIP-B/32	0.259 [0.203, 0.312]
MoCo-v3–SigLIP-B/16	0.343 [0.269, 0.412]
MoCo-v3–CLIP-L/14(LAION)	0.348 [0.268, 0.424]
MoCo-v3–SigLIP-L/16	0.461 [0.377, 0.536]
MoCo-v3–ViT-L(supervised)	0.198 [0.128, 0.269]
MoCo-v3–DINOv2-L	0.065 [0.033, 0.097]
MoCo-v3–MAE	0.100 [0.052, 0.155]

Table 10: MoCo-v3 vs. roster, confirmatory scale (Appendix A.3).

(OpenAI vs. LAION-2B) or by patch granularity (ViT-B/32 vs. B/16) did not meet the prespecified criterion on a second independent draw range. LAION's interval, 0.618 $[0.582, 0.657]$, overlaps OpenAI's 0.644 $[0.590, 0.703]$. B/16 vs. B/32 missed the preregistered non-overlapping-CI condition by a narrow margin ($+0.047$ $[0.007, 0.087]$). This result does not satisfy the registered criterion despite the positive mean difference.

A.7 Qwen2-VL extraction detail

Qwen2-VL-2B-Instruct's vision tower has no single CLS-token summary. It outputs one 1536-dimensional token per spatially merged patch region (a 2×2 merge of the raw 14×14 patch grid, via the model's own PatchMerger). This is the representation the model feeds to its language model as visual tokens, not an earlier or intermediate layer. Since different images can in principle produce different token counts under Qwen2-VL's dynamic-resolution preprocessing, we recover each image's token count from the model's `image_grid_thw` output. We mean-pool only that image's tokens to obtain a 1536-dimensional vector and then L2-normalize it, as for every other backbone. Extraction uses fp32 throughout the roster. We did not use fp16/bf16, despite their common use for large VLMs, to keep numerical precision consistent. Extraction also required



Figure 10: CLIP-L/14 on CelebA, one representative draw. Each depth-2 cell shows the held-out images nearest to its centroid. The two children of one root branch (left pair) separate by apparent age. The two children of the other (right pair) separate by apparent gender. *Male* faces concentrate in the second panel of the right pair, although gender is neither provided to nor targeted by the discovery procedure. Across draws for CLIP-L/14 and DINOv2-L, the root captures gender in 3–8% of draws, depth-2 in 58–62%.

Pair	Root ARI (95% CI)
Qwen2-VL–CLIP-L/14	0.556 [0.491, 0.617]
Qwen2-VL–CLIP-B/32	0.573 [0.517, 0.628]
Qwen2-VL–SigLIP-B/16	0.553 [0.485, 0.617]
Qwen2-VL–CLIP-L/14(LAION)	0.627 [0.573, 0.683]
Qwen2-VL–SigLIP-L/16	0.659 [0.600, 0.709]
Qwen2-VL–ViT-L(supervised)	0.416 [0.344, 0.484]
Qwen2-VL–DINOv2-L	0.203 [0.139, 0.271]
Qwen2-VL–MAE	−0.020 [−0.031, −0.010]

Table 11: Qwen2-VL vs. roster, confirmatory scale (Appendix A.4).

Configuration	Primary contrast	95% CI	n draws
$D_{\text{PCA}} = 4$	0.277	[0.193, 0.366]	16
$D_{\text{PCA}} = 8$ (baseline)	0.478	[0.411, 0.545]	18
$D_{\text{PCA}} = 16$	0.555	[0.464, 0.640]	15
$D_{\text{PCA}} = 32$	0.530	[0.452, 0.612]	17
$M = 10$	0.522	[0.450, 0.597]	20
$M = 20$ (baseline)	0.478	[0.411, 0.545]	18
$M = 40$	0.408	[0.353, 0.468]	15
metric = ARI (baseline)	0.478	[0.411, 0.545]	18
metric = NMI	0.399	[0.349, 0.450]	18

Table 12: Pipeline-design ablation, draws 900–919 (n=20), all other settings held at the paper's baseline. Every row excludes zero. The baseline appears three times, so the table covers seven distinct configurations. The baseline D_{PCA}=8 does not maximize the estimated contrast. The estimated contrast decreases as cluster count M increases but remains positive throughout the tested range. Replacing ARI with Normalized Mutual Information on the same discovered partitions also yields a positive contrast. Effect magnitudes across these two metrics are not directly comparable.

`torch.backends.cudnn.enabled=False`, the environment-specific fix already used for SwAV and in the multibackbone extraction script. Qwen2-VL's patch embedding uses a `Conv3d` that otherwise triggers `CUDNN_STATUS_NOT_INITIALIZED` in this environment.

A.8 Cross-encoder group transfer: protocol, ablations, and both datasets

Protocol. For each discovery draw r in the confirmatory range 180–219, the paper's discovery sample (300 images per class, identical for all backbones since it depends only on labels and r) is the head-training set and the remaining 400 UrbanCars images (50 per true cell) the test set. On the head-training set we induce a spurious correlation of strength c by keeping every image whose background and co-occurring object both agree with its class, a fraction $(1-c)/c$ of singly misaligned and $((1-c)/c)^2$ of doubly misaligned images (c=0.8: $n \approx 235$, smallest cell ≈ 4). For a source encoder A we fit the class-conditioned root partition on A's discovery draw as in Section 3 and apply it to the head-training images, giving pseudo-groups (y, branch). For a target B, we fit a logistic head on per-dimension standardized features with inverse-group-frequency sample weights (the RWG recipe). We report worst-group accuracy over the eight true cells on the test set. An earlier alternative used the group-balancing subsample mechanism of Kirichenko et al. (2023). We omitted their separate held-out retraining split because there is no upstream ERM-trained backbone here to correct. This alternative gave the same sign with a smaller effect and was noisier, as the smallest pseudo-group contains only a handful of images. For each target, the reference conditions are ERM (no groups), SELF (its own pseudo-groups), and ORACLE (true cells). The transfer gain of an ordered pair is $\text{WGA}_B(\text{groups}_A) - \text{WGA}_B(\text{ERM})$. The pair value averages both directions. The predictor is the root ARI of the pair computed on dev draws 640–659 (Spearman 0.971 with the confirmatory-draw ARI). To account for dependence between pairs, p-values come from 5,000 backbone-identity permutations of the ARI matrix. CelebA uses the same protocol with its natural blond/gender skew (no induced correlation), four true cells, and the eight backbones with CelebA features.

	draws	fixed pipeline (prespecified)	n	branch-size adaptive	n
UrbanCars	1000–1039	+0.053 $[0.032, 0.073]$	40	+0.062 $[0.042, 0.082]$	40
CelebA	200–239	+0.066 $[0.030, 0.110]$	37	+0.101 $[0.064, 0.138]$	39
ISIC	200–239	+0.035 $[0.006, 0.062]$	35	+0.040 $[0.016, 0.065]$	37
MetaShift	200–239	+0.100 $[0.080, 0.124]$	40	+0.146 $[0.110, 0.190]$	40

Table 13: Depth-2 conditional contrast (mean over draws, 95% draw-level bootstrap CI) under the fixed pipeline and the branch-size-adaptive variant.

Feature-standardization correction. An earlier version of this experiment fitted the head on unstandardized features. MSN's features have per-dimension standard deviation 3×10^{-4}, so its head was degenerate (accuracy 0.50), and every other head was under-fit (mean ERM worst-group accuracy 0.22 on UrbanCars, 0.26 on CelebA, against 0.50 and 0.59 once standardized). The under-fit heads produced transfer gains of 8–33 worst-group points and the same rank correlations as below. All reported transfer estimates use the standardized head. The corrected effect sizes are an order of magnitude smaller, with the correlation structure retained. Superseded estimates are disclosed here but are not used as evidence.

	UrbanCars, 15 backbones (105 pairs)				UrbanCars, 21 (210)	CelebA, 8 (28)
	c=0.7	c=0.8	c=0.85	c=0.9	c=0.8	natural skew
Spearman ρ (perm. p)	+0.34 (0.0554)	+0.65 (<.0002)	+0.72 (<.0002)	+0.57 (0.0022)	+0.71 (<.0002)	+0.82 (0.0028)
Partial ρ (ARI | CKA) / (CKA | ARI)	+0.29 / −0.15	+0.49 / −0.10	+0.61 / −0.22	+0.41 / +0.04	+0.57 / −0.20	+0.45 / −0.07
Partial ρ (ARI | self-consistency)	+0.15	+0.51	+0.57	+0.41	–	+0.83
Tercile gains, low/mid/high (pp)	0.5/0.9/0.9	1.3/2.1/2.4	1.2/2.3/3.0	1.6/2.4/3.2	0.9/2.0/2.2	1.0/2.1/3.3
Max-ARI source / random / self (pp)	1.0/0.7/0.8	2.3/1.9/2.1	2.8/2.2/2.5	2.9/2.4/2.8	2.1/1.7/1.9	3.2/2.2/2.4
Max-ARI beats random / self (targets)	9/8 of 15	10/7 of 15	11/10 of 15	11/9 of 15	14/10 of 21	6/7 of 8
ERM WGA / oracle gain (pp)	0.65 / 3.3	0.50 / 6.9	0.40 / 7.8	0.29 / 8.4	0.45 / 6.4	0.59 / 5.2

Table 14: Group transfer with the standardized head on UrbanCars at four skew levels (15 backbones), the full UrbanCars roster (21, c=0.8), and CelebA at natural skew. Permutation p-values use 5,000 backbone-identity permutations. CKA uses the same class-1 features. Oracle gain is measured over ERM. Partial-ρ and CKA details appear in the text and Appendix A.17.

Ablations (standardized head, recomputed after the correction). (i) *Lineage.* Restricting to cross-lineage pairs (no CLIP–CLIP or SigLIP–SigLIP) leaves ρ=0.72 (94 pairs, 15 backbones) and 0.78 (24 pairs, CelebA). Controlling for same-lineage and same-family indicators leaves partial ρ +0.70 and +0.59. Controlling instead for pair self-consistency leaves partial ρ +0.51 and +0.83. Partial correlations here and in Section 4.6 are rank-based. An earlier version of this ablation used raw residuals.

(ii) *Predictor cost.* ARI from 5/10/20 draws: ρ=0.57/0.53/0.65 (UrbanCars), 0.79/0.82/0.82 (CelebA).

(iii) *Direction.* Directed gains correlate with the target's oracle headroom (ρ=0.52, 0.18), so the statistic should be read together with the target's headroom.

(iv) *Predictor draws.* ARI from the outcome draws themselves gives ρ=0.69 against 0.65 from disjoint draws (UrbanCars). Two earlier ablations were not rerun with the standardized head and are not reported: a DFR-subsample head, and CelebA depth-2 pseudo-groups.

Removing the encoder's own axis. Because each partition is inspectable, we first asked whether its alignment with a known shortcut (balanced accuracy of the root partition against UrbanCars' background or CelebA's gender label) predicts shortcut vulnerability. It does on UrbanCars (15 encoders: Spearman −0.54 with ERM worst-group accuracy, p=0.04) and not on CelebA (8 encoders: +0.38, p=0.35). Partition alignment and classifier reliance are distinct quantities. A preregistered follow-up therefore tested the effect of removing the discovered axis. On each draw, we take the difference between the two root sides' standardized feature means on the class-1 pool. We project this axis out of train and test features and refit the same standardized ERM head under the same skew. The prediction that average accuracy would *drop* failed (it changes by at most 0.010), and the worst-group gains are small: +0.009 to +0.031 on UrbanCars (CI excludes zero for 13/13 whole-image encoders) and +0.018 to +0.048 on CelebA (6/6), against +0.001/−0.001 for MAE (CIs including zero on both datasets) and +0.016/+0.002 for DINOv2 (excluding zero on UrbanCars only; Table 16). Every whole-image encoder has a positive gain, whereas MAE's confidence intervals include zero. DINOv2's gain excludes zero only on UrbanCars. The standardized head leaves an ERM-to-oracle

gap of 8 (UrbanCars) and 6 (CelebA) worst-group points for the whole-image encoders in Table 15. The gains recover 25% and 55% of that gap (Table 15). We report the fraction for completeness, not as evidence of practical effectiveness.

	Encoders	ERM WGA	after	Δ WGA	avg. acc. before→after	oracle WGA	gap recovered
UrbanCars	whole-image (13)	0.49	0.51	+0.009 to +0.031	0.78→0.78	0.57	25% (12–36%)
	DINOv2-L	0.60	0.62	+0.016 $[0.004, 0.029]$	0.86→0.86	0.69	19%
	MAE-L	0.31	0.31	+0.001 $[-0.004, 0.006]$	0.67→0.67	0.39	2%
CelebA	whole-image (6)	0.63	0.66	+0.018 to +0.048	0.89→0.89	0.69	55% (30–70%)
	DINOv2-L	0.40	0.40	+0.002 $[-0.003, 0.008]$	0.86→0.86	0.44	5%
	MAE-L	0.53	0.53	-0.001 $[-0.008, 0.005]$	0.89→0.89	0.57	−3%

Table 15: Projecting out each encoder's own root axis, then refitting the standardized ERM head on draws 180–199. The oracle uses true-group reweighting. The roster comprises the 15 UrbanCars and 8 CelebA backbones of the four-skew transfer run, not the core roster. Per-encoder Δ and CIs appear in Table 16.

UrbanCars (c=0.8)	ERM WGA	after (Δ, 95% CI)	acc.	CelebA	ERM WGA	after (Δ, 95% CI)	acc.
CLIP-L/14	0.64	0.65 (+0.017 [0.006,0.027])	0.87→0.87	CLIP-L/14	0.61	0.63 (+0.018 [0.009,0.027])	0.89→0.88
CLIP-B/32	0.40	0.43 (+0.028 [0.014,0.042])	0.72→0.73	CLIP-B/32	0.67	0.71 (+0.037 [0.020,0.055])	0.90→0.90
CLIP-L/14 (LAION)	0.68	0.70 (+0.024 [0.010,0.038])	0.88→0.89	CLIP-L/14 (LAION)	0.64	0.69 (+0.048 [0.033,0.062])	0.90→0.90
SigLIP-B/16	0.68	0.70 (+0.018 [0.005,0.034])	0.88→0.89	SigLIP-B/16	0.69	0.73 (+0.043 [0.031,0.055])	0.90→0.90
SigLIP-L/16	0.71	0.73 (+0.020 [0.009,0.029])	0.90→0.91	SigLIP-L/16	0.70	0.74 (+0.039 [0.029,0.051])	0.90→0.90
ViT-L (sup.)	0.38	0.40 (+0.013 [0.005,0.020])	0.70→0.71	ViT-L (sup.)	0.45	0.47 (+0.022 [0.012,0.033])	0.87→0.87
DINOv2-L	0.60	0.62 (+0.016 [0.004,0.029])	0.86→0.86	DINOv2-L	0.40	0.40 (+0.002 $[-0.003, 0.008]$)	0.86→0.86
MAE-L	0.31	0.31 (+0.001 $[-0.004, 0.006]$)	0.67→0.67	MAE-L	0.53	0.53 (-0.001 $[-0.008, 0.005]$)	0.89→0.89
MSN-L	0.25	0.27 (+0.017 [0.004,0.030])	0.64→0.65				
SwAV	0.22	0.24 (+0.021 [0.010,0.032])	0.65→0.66				
SeLa-v2	0.23	0.25 (+0.019 [0.009,0.030])	0.62→0.63				
MoCo-v3	0.32	0.35 (+0.031 [0.016,0.047])	0.67→0.68				
CLIP RN50x64	0.69	0.72 (+0.027 [0.012,0.043])	0.89→0.90				
Qwen2-VL	0.64	0.65 (+0.009 [0.001,0.019])	0.86→0.87				
CLIP-B/16	0.55	0.56 (+0.016 [0.004,0.027])	0.81→0.81				

Table 16: Projecting out each encoder's own label-free root axis and refitting the standardized ERM head: worst-group accuracy before and after (Δ with draw-level bootstrap CI) and average accuracy before→after, mean over draws 180–199. Every UrbanCars and CelebA encoder with transfer features is shown. MAE's axis is brightness, unrelated to either task.

Controls. We fixed the control rule before running the experiment. Using the same draws, skews, and head, we tested three alternative directions per draw and encoder. These were a random unit direction inside the top-8 PCA subspace of the class-1 training features, the first principal component, and a random direction in the full feature space. The own-axis condition reproduces the run above (maximum difference below 10^{-15} on both datasets). For the whole-image encoders, mean ΔWGA is +0.020 $[0.015, 0.025]$ for the own axis against +0.003, +0.007 and +0.001 on UrbanCars, and +0.034 $[0.029, 0.040]$ against +0.004, +0.000 and +0.001 on CelebA. The prespecified paired difference against the random in-subspace direction is +0.017 $[0.013, 0.020]$ and +0.031 $[0.025, 0.037]$. The gain is specific to the discovered axis on both datasets, including CelebA, where that axis does not track gender.

A.9 Roster extensions: scale, pretraining data, architecture, and a VLM tower

The core analysis uses one encoder per objective and architecture family (15 in total). Six additional language-supervised variants test the effects of lineage, scale, pretraining data, and architecture. CLIP-B/32, CLIP-B/16, and CLIP-L/14 (LAION, Schuhmann et al., 2022) vary patch size, scale, and data. CLIP-RN50x64 uses a convolutional backbone, SigLIP-B/16 varies scale within the SigLIP lineage, and Qwen2-VL provides a VLM vision tower. All have their closest core neighbour among the externally supervised encoders and low agreement with hidden-content models (Table 17). They are treated as extension checks, not separate findings. CLIP-B/32, CLIP-L/14 (LAION), and SigLIP-B/16 also belong to the original eight-encoder confirmatory roster (Section 4.1), whose grouping is retained as registered.

Restricting the roster to one CLIP and one SigLIP preserves whole-image agreement and separation from hidden-content models (Table 18). The observed structure is therefore not explained solely by repeated language-supervised lineages.

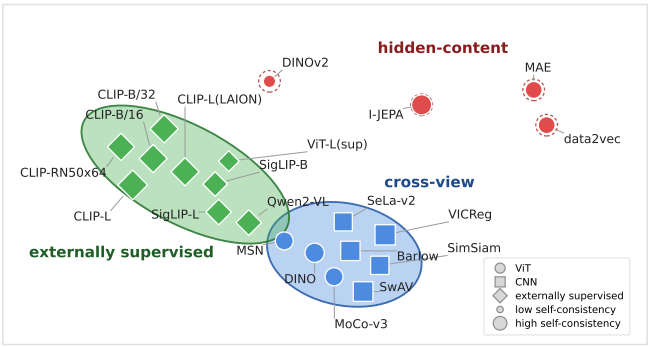


Figure 11: All twenty-one encoders, as Figure 2. The six extensions fall inside the externally supervised cluster.

An additional label-supervised ConvNeXt-Base (Liu et al., 2022), trained on ImageNet-1k, has self-consistency 0.63. Its agreement is 0.43 with the core externally supervised trio and 0.57 with the cross-view models. This result extends the whole-image association to another convolutional architecture.

Extension	self-cons.	vs. ext. sup. 3	vs. cross-view 8	vs. hidden 4	closest core encoder
CLIP-B/32	0.83	0.59	0.31	0.04	CLIP-L/14 0.79
CLIP-L/14 (LAION)	0.82	0.64	0.36	0.04	CLIP-L/14 0.82
CLIP-B/16	0.82	0.59	0.32	0.04	CLIP-L/14 0.77
CLIP RN50x64	0.77	0.57	0.28	0.04	CLIP-L/14 0.78
SigLIP-B/16	0.59	0.46	0.37	0.04	CLIP-L/14 0.54
Qwen2-VL	0.66	0.51	0.51	0.05	SigLIP-L/16 0.62

Table 17: The six roster extensions on UrbanCars (draws 180–219), against the core externally supervised trio, the eight cross-view models, and the four hidden-content models.

Roster	within whole-image	within ext. sup.	ext. sup. × cross-view 8	whole × hidden
full roster (17 whole-image)	0.48	0.57	0.36	0.06
core roster (11 whole-image)	0.52	0.48	0.37	0.07

Table 18: Mean pairwise ARI with and without the CLIP- and SigLIP-lineage duplicates (UrbanCars, draws 180–219).

A.10 Two more hidden-content backbones

The core roster contains four hidden-content encoders and eight cross-view encoders. We evaluated two further hidden-content backbones under the same frozen UrbanCars protocol to widen the hidden-content set. SimMIM (Xie et al., 2022) on Swin-B (Liu et al., 2021b) is a masked-pixel objective on a windowed architecture that is neither a plain ViT nor a CNN, and BEiT-B is a masked visual-token objective distinct from MAE, data2vec and I-JEPA. Both were extracted once from the same image manifest as every other backbone. Unlike an earlier BEiT-L extraction, which produced no dominant root bipartition on UrbanCars and is absent from the roster, both yield valid partitions. SimMIM has self-consistency 0.46 over 24 valid draws, and BEiT-B has 0.27 over 34. BEiT-B falls below the 0.3 stability threshold, limiting interpretation of its cross-encoder agreement.

Both models have low agreement with the whole-image models (SimMIM -0.01, BEiT-B 0.00) and with the externally supervised trio (-0.04, -0.01). Their highest agreement is with MAE, at 0.43 for SimMIM and 0.24 for BEiT-B (Table 19). The pixel-reconstruction pair MAE–SimMIM has higher agreement than MAE–BEiT. SimMIM also extends the observed separation to a windowed Swin backbone. These models are robustness checks and are excluded from the headline contrasts, which retain the frozen roster.

Hidden-content backbone	self-cons.	n	vs. whole-image	vs. hidden 4	closest core encoder
SimMIM (Swin-B, pixel)	0.46	24	-0.01	0.17	MAE 0.43
BEiT-B (ViT-B, token)	0.27	34	0.00	0.09	MAE 0.24

Table 19: Two further hidden-content backbones on UrbanCars (draws 180–219), evaluated under the same frozen protocol, against the eleven whole-image models and the four core hidden-content models. Both have low agreement with the whole-image block and their highest agreement with MAE. They provide a robustness check on the hidden-content set and are not added to the headline roster.

A.11 Waterbirds and SpuCo Dogs: draw-based replication

We evaluate the 15-encoder core roster on two further natural-image datasets with a background shortcut: Waterbirds (Sagawa et al., 2020), CUB birds composited on Places backgrounds, and SpuCo Dogs (Joshi et al., 2023), large versus small dog breeds photographed indoors or outdoors. We use the test splits, whose backgrounds are balanced within each class: Waterbirds landbirds (n=4,510) and waterbirds (1,284), and SpuCo small and large dogs (1,000 each, 500 per background). Each class pool is analysed separately with the protocol of Appendix A.1: draws 180–219 of 300 images, the held-out pool of draw 999, and a five-seed majority vote. The tested predictions were fixed before the features were analysed. Every encoder yields a valid partition on all 40 draws in every pool.

All four pools pass the regime criterion (Table 20), the waterbird pool narrowly. In all four, the whole-image versus hidden-content contrast and the difference between within-cross-view and cross-view-to-hidden-content agreement have intervals above zero, and the four hidden-content models agree with one another less than the cross-view models do. The cross-view root partitions follow background (balanced accuracy mostly 0.8–0.97).

Two aspects differ from UrbanCars. First, the three externally supervised models are unstable in these pools (self-consistency 0.10–0.28, except SigLIP-L/16 on large dogs, 0.51) and agree with one another at only 0.05–0.10. The whole-image agreement is therefore carried mainly by the cross-view models. Both tier differences have intervals above zero on three of the four pools (all but waterbirds), so the prespecified intermediate-tier criterion is met there, but the low self-consistency of the supervised models limits interpretation of their position. Second, hidden-content isolation is weaker: data2vec and I-JEPA also follow background on the landbird pool (balanced accuracy 0.85 and 0.87). This agrees with the scope stated in Section 4.3, where latent-prediction isolation is established only on UrbanCars. An earlier single-pool check on the Waterbirds validation split (five PCA seeds) gave the same picture: cross-view agreement 0.37 against 0.19 with hidden-content models on landbirds, and no coherence among the supervised models (0.00).

Pool	n	whole-image SC	Δ whole vs. hidden	within cv $-$ cv$\times$hidden	within cv	within hidden	within sup. 3	A	B
Waterbirds, landbird	4510	0.59	$+0.105$ $[0.087, 0.124]$	$+0.264$ $[0.233, 0.297]$	0.56	0.15	0.06	$+0.435$ $[0.394, 0.477]$	$+0.054$ $[0.043, 0.065]$
Waterbirds, waterbird	1284	0.33	$+0.047$ $[0.033, 0.061]$	$+0.133$ $[0.104, 0.162]$	0.25	0.03	0.05	$+0.222$ $[0.187, 0.259]$	-0.008 $[-0.016, 0.000]$
SpuCo, small dogs	1000	0.37	$+0.071$ $[0.054, 0.089]$	$+0.136$ $[0.109, 0.165]$	0.29	0.08	0.08	$+0.184$ $[0.141, 0.228]$	$+0.048$ $[0.033, 0.064]$
SpuCo, large dogs	1000	0.40	$+0.090$ $[0.066, 0.114]$	$+0.141$ $[0.111, 0.171]$	0.30	0.05	0.10	$+0.116$ $[0.085, 0.148]$	$+0.065$ $[0.039, 0.090]$

Table 20: Draw-based replication on Waterbirds and SpuCo Dogs (test splits, 15 core encoders, draws 180–219, five-seed majority vote, 95% draw-level CIs). Whole-image SC is the mean self-consistency of the eleven whole-image encoders; cv denotes the eight cross-view models; A and B are the tier differences of Appendix A.16. The supervised trio has low self-consistency in these pools, which limits interpretation of A and B.

A.12 Spawrious: replication on controlled backgrounds

Spawrious (Lynch et al., 2023) provides a further test with controlled whole-scene backgrounds. It contains four synthetic dog breeds against six background settings: beach, desert, dirt, jungle, mountain, and snow. We evaluate the eight-encoder confirmatory roster within each breed, fixing the classification target while background varies. The protocol uses 300-image discovery draws 180–219 and applies each fitted partitioning rule to a fixed held-out pool.

The corgi pool (n=2,400, 400 per background) passes the regime criterion with mean self-consistency 0.50 across the six externally supervised encoders, the whole-image subset of this roster. Mean agreement among the six externally supervised models is 0.43 $[0.36, 0.50]$, compared

Breed pool	SC (sup. 6)	SC ≥ 0.3	within sup. 6	sup. $\times$ {DINOv2,MAE}	contrast Δ
corgi	0.50	yes	0.43	0.13	$+0.298\ [0.232, 0.364]$
bulldog	0.39	yes	0.37	0.14	$+0.228\ [0.164, 0.301]$

Table 21: Spawrious primary contrast on two breed pools, using draws 180–219 and the eight-encoder confirmatory roster. Both pools pass the regime criterion and have positive within-supervised minus supervised-to-{DINOv2,MAE} contrasts. On corgi, the five language-supervised encoders achieve 0.95–0.98 one-versus-rest background balanced accuracy, compared with 0.73 for label-supervised ViT-L and 0.59 for DINOv2.

Breed pool	whole-image SC	within cross-view	ext. sup. $\times$ cross-view	within hidden	whole vs. hidden Δ	A	B
corgi	0.50	0.30	0.27	0.08	$+0.150\ [0.121, 0.180]$	$+0.028\ [-0.015, 0.073]$	$+0.161\ [0.121, 0.202]$
bulldog	0.49	0.39	0.28	0.06	$+0.198\ [0.162, 0.235]$	$+0.109\ [0.046, 0.179]$	$+0.171\ [0.130, 0.213]$

Table 22: Spawrious, 15-encoder core roster (draws 180–219). Whole-image SC is the mean self-consistency of the eleven whole-image encoders. Δ is mean ARI within the whole-image models minus their mean ARI with the hidden-content models. A and B are the three-tier differences of Appendix A.16. Both pools have positive Δ with CIs above zero. The intermediate-tier criterion is met on bulldog but remains unconfirmed on corgi, where the CI for A includes zero.

with 0.13 between these models and DINOv2/MAE. The contrast is $+0.298\ [0.232, 0.364]$, with a draw-level bootstrap interval excluding zero. The five language-supervised encoders align with one-versus-rest background partitions at 0.95–0.98 balanced accuracy, selecting the best background per draw. The label-supervised ViT-L control reaches 0.73, with lower self-consistency on this pool. DINOv2 has weaker background alignment, with best one-versus-rest balanced accuracy 0.59.

The separately evaluated bulldog pool also passes the criterion (mean self-consistency 0.39 for the same six encoders) and yields a positive contrast, $+0.228\ [0.164, 0.301]$ (Table 21). These results replicate the original contrast in two breed pools on a dataset distinct from UrbanCars.

We then extracted the remaining ten core encoders (the eight cross-view models, data2vec, and I-JEPA). Together with the core supervised trio, DINOv2, and MAE from the original roster, these form the 15-encoder core set. We applied the protocol of Section 4.2 on both breed pools, with one seed per draw rather than the five-seed majority vote of Appendix A.1. Both pass the regime criterion when it is recomputed over all eleven whole-image encoders, with mean self-consistency 0.50 on corgi and 0.49 on bulldog.

Both pools support the whole-image versus hidden-content contrast. It is $+0.150\ [0.121, 0.180]$ on corgi and $+0.198\ [0.162, 0.235]$ on bulldog. Cross-view models have mean mutual agreement 0.30 and 0.39, compared with 0.08 and 0.06 among the four hidden-content models (Table 22). The tier analysis provides more qualified support. On both pools, externally supervised models agree more with cross-view models than with hidden-content models ($B = +0.161$ and $+0.171$, CIs excluding zero). On corgi, the difference between within-cross-view and supervised-to-cross-view agreement is $A = +0.028\ [-0.015, 0.073]$. This interval does not establish a positive difference, but it also does not establish equivalence between the groups. On bulldog, both tier differences have intervals above zero ($A = +0.109\ [0.046, 0.179]$).

Within-cross-view and supervised-to-cross-view agreement are lower than on UrbanCars. Four encoders on corgi and five on bulldog have self-consistency below 0.3, including data2vec and I-JEPA on both pools. This instability limits interpretation of low cross-encoder agreement.

A.13 Dataset-dependent stability and counterexamples

We distinguish stable partition structure from low-stability results using the whole-image encoders' mean self-consistency on draws 180–219, with a threshold of 0.3. We fixed this threshold before computing EuroSAT's self-consistency, although its contrast was already known (Table 23). MetaShift falls below it. Seven of its fifteen encoders have self-consistency under 0.2, MSN produces a valid partition on only 18 of 40 draws, and no annotated factor dominates (indoor/outdoor balanced accuracy 0.56–0.64). Its primary contrast, run with the frozen protocol on unopened draws, is nevertheless positive (+0.145 $[0.106, 0.187]$), and its six-backbone extension (Table 24) shows the same ordering as UrbanCars at a lower level. EuroSAT passes the criterion. Its whole-image en-

coders are the six supervised models, with mean self-consistency 0.48. The range across all eight encoders is 0.22–0.75. Yet the supervised pairs agree at only 0.12, and the contrast is -0.005 $[-0.017, 0.007]$. The dataset mean also obscures variation across encoders. In particular, three of the eight encoders (CLIP-B/32, CLIP-L/14 (LAION), MAE) fall below 0.3, and among the stable ones the strongest pair crosses the family line (SigLIP-L/16–DINOv2 0.39, against SigLIP-L/16–ViT-L 0.31). Stability is therefore not sufficient for shared partitions, and MetaShift's weaker but positive contrast below the threshold shows it is not strictly necessary either. We do not have a label-free criterion that predicts EuroSAT.

Dataset	encoders	whole-image SC	SC ≥ 0.3	primary contrast Δ	within cross-view	whole $\times$ hidden
UrbanCars	21	0.72	yes	$+0.497$ $[0.449, 0.544]$	0.65	0.06
CelebA	21	0.41	yes	$+0.283$ $[0.221, 0.351]$	0.36	0.08
ISIC	14	0.38	yes	$+0.125$ $[0.091, 0.163]$	0.29	0.12
MetaShift	15	0.17	no	$+0.145$ $[0.106, 0.187]$	0.21	0.03
EuroSAT	8	0.48	yes	-0.005 $[-0.017, 0.007]$	—	0.13

Table 23: Dataset stability and agreement. The encoder count and descriptive agreement columns refer to the available roster on draws 180–219. The primary-contrast column instead uses the original eight-encoder roster and the dataset-specific confirmatory ranges in Table 2, including 50–89 for EuroSAT. Within cross-view uses each dataset's available cross-view encoders. Whole $\times$ hidden denotes mean ARI between whole-image and hidden-content models.

		MetaShift		
Added backbone	sup. 6	SSL 4	DINOv2	MAE
DINO ViT-B/16	0.11	0.12	0.00	-0.01
VICReg RN50	0.11	0.06	-0.01	-0.01
Barlow Twins RN50	0.10	0.10	0.01	-0.02
SimSiam RN50	0.12	0.11	-0.01	-0.02
data2vec ViT-B	0.05	0.04	0.00	0.00
I-JEPA ViT-H/14	0.01	-0.01	-0.01	0.05

Table 24: The six added backbones on MetaShift, as in Table 25. SSL 4 denotes MSN, SwAV, SeLa-v2, and MoCo-v3. MetaShift is outside the regime.

A.14 Confirming the final grouping on previously unused draws

The grouping in Section 4.2 was revised twice using draws 180–219 (Appendix A.3). Analyses on those draws are therefore exploratory with respect to the grouping. We fixed the following five predictions before evaluating the unchanged grouping on previously unused draws 700–739.

(P1) The UrbanCars whole-image–hidden-content contrast has a CI above zero. (P2) The same holds for CelebA. (P3) Both three-tier differences (Appendix A.16) have CIs above zero. (P4) MAE agrees with no core encoder above 0.10 except data2vec, and data2vec–MAE agreement is at least 0.20. (P5) DINOv2 agrees with no whole-image encoder above 0.35.

All five criteria were satisfied. The contrasts were +0.428 $[0.395, 0.458]$ on UrbanCars, +0.152 $[0.134, 0.172]$ on CelebA, +0.314 $[0.27, 0.35]$ for tier A, and +0.292 $[0.26, 0.33]$ for tier B. Excluding data2vec, MAE's largest agreement was 0.08, compared with 0.25 for data2vec. DINOv2's largest whole-image agreement was 0.23. This evaluation is confirmatory for the revised predictions, not for the preceding development of the grouping.

A.15 Self-supervised extensions and checkpoint-cache correction

Table 25 uses public checkpoints: DINO ViT-B/16 (`facebook/dino-vitb16`), data2vec-vision base (`facebook/data2vec-vision-base`), I-JEPA ViT-H/14 (`facebook/ijepa_vith14_1k`, mean-pooled patch tokens since the model has no CLS token), VICReg, Barlow Twins and SimSiam ResNet-50 from the authors' releases, with the same manifest, L2-normalization and class-1 root-partition pipeline as every other backbone. Predictions (no new backbone joins the initial supervised six and none converges with DINOv2 or MAE) were written before extraction. One artefact was caught and fixed before analysis. The VICReg and

Added backbone	UrbanCars sup. 3	SSL 3	DINOv2	MAE	MoCo-v3	CelebA sup. 3	SSL 3	DINOv2	MAE	MoCo-v3
DINO ViT-B/16	0.39	0.65	0.08	0.02	0.62	0.15	0.37	0.26	0.01	0.47
VICReg RN50	0.37	0.63	0.09	0.05	0.71	0.11	0.31	0.19	0.01	0.30
Barlow Twins RN50	0.37	0.63	0.08	0.05	0.66	0.12	0.28	0.18	0.01	0.44
SimSiam RN50	0.32	0.59	0.06	0.06	0.62	0.10	0.35	0.25	0.02	0.44
data2vec ViT-B	−0.01	0.11	−0.03	0.32	0.15	0.01	0.00	0.00	0.01	0.01
I-JEPA ViT-H/14	0.03	0.06	0.02	0.01	0.07	0.20	0.10	0.04	0.00	0.16

Table 25: Six added self-supervised backbones, class-1 root partitions, draws 180–219. Both datasets use the same 21-backbone roster and columns. SSL 3 denotes MSN/SwAV/SeLa-v2. MoCo-v3 has its own column because its UrbanCars pattern is intermediate. Results for ISIC and MetaShift appear in Tables 26 and 24, respectively.

Barlow Twins `torch.hub` entry points both save their weights under the same file name in the shared checkpoint cache, so loading one after the other silently reused the first model's weights. The two feature matrices were bit-identical, and their partitions had ARI 1.00. Loading each from its own URL into a distinct file gives a mean feature cosine of 0.26 and a pairwise ARI of 0.81.

Cross-dataset detail. The cross-dataset results provide partial support for the exploratory grouping. On CelebA, ISIC, and MetaShift, MAE's agreement with every other model is ≤ 0.08. Within this core-roster comparison across the four main datasets, its highest agreement is with data2vec on UrbanCars (0.32). Whole-image models agree with each other on CelebA (0.36–0.48) and among the three ResNets on ISIC (0.48–0.52). However, I-JEPA agrees with the initial supervised six on CelebA/ISIC (0.28/0.35), and data2vec agrees with the ResNet trio on ISIC. DINOv2 has the least self-consistent partition on UrbanCars (0.24) but is self-consistent on CelebA (0.62, Appendix A.17). On CelebA, it reaches 0.18–0.26 agreement with the whole-image models.

Cross-dataset test (preregistered). Before extracting the six backbones on CelebA and ISIC, we registered two predictions. (M1) The data2vec and I-JEPA models agree with the supervised core, DINOv2, MAE, and each other at ARI < 0.15. (W1) DINO, VICReg, Barlow Twins, and SimSiam agree with each other at mean ARI ≥ 0.30. Each also agrees with the supervised core more than with the hidden-content models. On CelebA, W1 passed and M1 failed. On ISIC, both failed (Table 26).

Dataset	Backbone	vs. sup. 6	vs. DINOv2	vs. MAE	vs. data2vec	vs. I-JEPA	vs. whole-image 4
CelebA	DINO	0.16	0.26	0.01	0.00	0.17	0.38
CelebA	VICReg	0.12	0.19	0.01	0.01	0.13	0.44
CelebA	Barlow Twins	0.14	0.18	0.01	0.00	0.16	0.42
CelebA	SimSiam	0.09	0.25	0.02	0.00	0.08	0.44
CelebA	data2vec	0.00	0.00	0.01	—	0.07	0.01
CelebA	I-JEPA	0.28	0.04	0.00	0.07	—	0.13
ISIC	DINO	0.26	0.25	0.02	−0.01	0.35	0.08
ISIC	VICReg	0.06	0.05	0.00	0.23	0.09	0.36
ISIC	Barlow Twins	0.06	0.11	−0.01	0.18	0.05	0.36
ISIC	SimSiam	0.04	0.06	0.00	0.23	0.05	0.36
ISIC	data2vec	0.02	−0.02	0.06	—	0.01	0.16
ISIC	I-JEPA	0.35	0.11	0.06	0.01	—	0.13

Table 26: Cross-dataset replication of Table 25 (root partitions, draws 180–219). Whole-image pairwise agreement: CelebA 0.36–0.48; ISIC 0.48–0.52 among VICReg/Barlow Twins/SimSiam and 0.04–0.09 between DINO and each of them.

Figure 12 summarizes the cross-dataset contrasts.

A.16 Supervised tier, pool overlap, and stability sensitivity

Three tiers. We fixed the interpretation rule before running the test. Per draw, A is mean ARI within the cross-view models minus mean ARI between them and the three supervised encoders of the core roster, and B is that between-group mean minus mean ARI between those encoders and the hidden-content models. Using all six supervised models of the confirmatory roster instead gives the same readings (A = +0.29, B = +0.32 on UrbanCars; +0.24, +0.03 on CelebA; +0.19, −0.03 on ISIC; +0.15, +0.08 on MetaShift). We prespecified an intermediate tier when both confidence intervals lie above zero. This criterion is met on UrbanCars, CelebA, and MetaShift. On ISIC, B

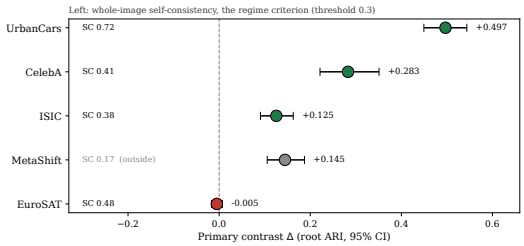


Figure 12: Primary contrasts across datasets with 95% draw-level confidence intervals. Dataset-specific draw ranges and retained sample sizes appear in Table 2.

is negative, and the supervised models agree less with cross-view models than with hidden-content models (Table 27).

Dataset	A	B	Reading
UrbanCars	+0.273 $[0.235, 0.309]$	+0.330 $[0.297, 0.361]$	intermediate
CelebA	+0.244 $[0.196, 0.295]$	+0.038 $[0.017, 0.060]$	intermediate
ISIC	+0.187 $[0.131, 0.244]$	-0.033 $[-0.054, -0.010]$	not intermediate
MetaShift	+0.162 $[0.119, 0.208]$	+0.068 $[0.042, 0.095]$	intermediate

Table 27: Intermediate-tier test for externally supervised models, with 95% draw-level CIs over draws 180–219. A compares within-cross-view agreement with supervised-to-cross-view agreement. B compares supervised-to-cross-view agreement with supervised-to-hidden-content agreement. The intermediate-tier criterion requires both intervals to be above zero.

Overlap-free held-out pool. H shares on average 60% of its images with each discovery pool on UrbanCars (10%, 14% and 27% on CelebA, ISIC and MetaShift). Recomputing every pair ARI on $H \setminus D_r$, and self-consistency on $H \setminus (D_r \cup D_{r'})$, changes UrbanCars within-cross-view agreement from 0.65 to 0.63. The eight-encoder contrast changes from +0.476 to +0.475 $[0.428, 0.523]$. Whole-image-to-hidden-content agreement and mean whole-image self-consistency remain 0.06 and 0.72, respectively, at the reported precision. On the other datasets, every change is below 0.01. Images the rule was fit on do not drive the agreement. This check does not remove the recurrence of images between discovery pools themselves.

Normalized agreement without unstable encoders. We exclude encoders with self-consistency below 0.2. This removes 0, 2, 0, and 7 encoders on UrbanCars, CelebA, ISIC, and MetaShift, respectively. Without clipping, within-whole-image normalized agreement is 0.67, 0.51, 0.44 and 0.68, and whole-image-to-hidden-content agreement is 0.12, 0.20, 0.26 and 0.10. The clipping in the main analysis had touched 2, 18, 5 and 11 of 204, 204, 85 and 99 pairs (whole-image pairs and whole-image–hidden-content pairs; pairs among hidden-content models are not normalized).

A.17 Partition self-consistency and normalized agreement

Partition stability provides an empirical reference for interpreting cross-encoder ARI, rather than a formal upper bound. We define an encoder's *self-consistency* as the mean ARI between its root partitions on different draws, applied to the same held-out pool (Table 28), and the *normalized agreement* of a pair as its ARI divided by the geometric mean of the two self-consistencies. Three observations are relevant.

(i) Self-consistency is dataset-dependent. Most encoders have a stable root axis on UrbanCars (0.56–0.93) but far less so on CelebA and ISIC, so the smaller raw contrasts there partly reflect instability rather than disagreement.

(ii) DINOv2's isolation on UrbanCars coincides with the lowest self-consistency in the roster (0.24). Its root partition is less stable under our procedure than those of MAE (0.56) and I-JEPA (0.85), which also have low agreement with other encoders. On CelebA, where DINOv2 is self-consistent (0.62), it also agrees more with the whole-image models (0.18–0.26).

(iii) Normalization has a large effect for small models trained from scratch. The SwAV and BYOL ResNet-18 runs in Appendix A.18 have self-consistency 0.26 and 0.29, so their raw mutual ARI of

0.24 becomes 0.89 after normalization. This is comparable to SwAV–SeLa-v2 (0.97) or VICReg–SimSiam (1.00), and higher than DINOv2–MAE (−0.05).

Self-consistency and family agreement across the four datasets. Across the whole-image encoders available on each dataset (17, 17, 10, 11), an encoder's self-consistency tracks its mean agreement with its own family with Spearman +0.25 (UrbanCars), +0.62 (CelebA), +0.04 (ISIC) and +0.58 (MetaShift). The whole-image family's mean self-consistency is 0.72, 0.41, 0.38 and 0.17. Raw within-family agreement normalizes to 0.67, 0.50, 0.43 and 0.71, and whole-image-to-hidden-content agreement to 0.12, 0.20, 0.26 and 0.12.

Two cautions apply. First, self-consistency counts only draws on which the encoder produced a two-sided partition, as does pairwise ARI. Second, the normalized ratio is clipped to $[-1, 1]$ per pair. Before clipping, division by self-consistency near zero can produce large ratios. Several MetaShift encoders have low self-consistency (0.09–0.10). The MetaShift normalized values are therefore the least reliable. MSN produced a valid MetaShift partition on only 18 of 40 draws, and its MetaShift pairs are on those draws.

What the root axes track (balanced accuracy of the root partition against an annotated factor). On MetaShift, indoor/outdoor alignment is weak for every whole-image encoder (0.56–0.64). The values for data2vec, DINOv2, and MAE are 0.67, 0.52, and 0.57, respectively. No single labelled factor dominates on this dataset.

ISIC has seven annotated dermoscopy artefacts. The supervised, CLIP, and DINO models split on patches, ruler, or gel border (0.70–0.77). The self-supervised ResNet-50 models (VICReg, Barlow Twins, and SimSiam) split on hair (0.64–0.65), accounting for low within-family agreement. DINOv2 splits on patches (0.79). I-JEPA tracks gel border (0.85), ink (0.81), and ruler (0.79). This overlap in annotated factors is consistent with I-JEPA's agreement of 0.35 with supervised models. MAE tracks no factor above 0.60.

In the matched set on UrbanCars, the whole-image ResNet-18 runs track background at 0.66–0.73. DINO-ViT-S(v2) tracks background at 0.58 and brightness at 0.65. Conv-MAE tracks no factor above 0.56.

iBOT comparison. DINOv2 differs from DINO in its masked-token term, data, scale, and regularization (Section 5.1). We examine whether a masked-token term is sufficient to produce the observed separation using the official iBOT ViT-B/16 checkpoint (Zhou et al., 2022). Like DINO, this model uses ImageNet-1k at ViT-B scale, without KoLeo or Sinkhorn centring. It agrees with DINO at 0.70, its largest pairwise agreement in the roster, compared with 0.08 for DINOv2–DINO. Its mean agreement is also higher with whole-image models (0.46) than with hidden-content models (0.06). Thus, a masked-token term does not necessarily produce separation. The comparison does not identify which other changes explain DINOv2's result.

A refuted prediction. Before measuring, we predicted that whole-image root partitions would align with image brightness more on CelebA and ISIC (no dominant scene factor) than on UrbanCars. They do not. Brightness balanced accuracy is 0.51–0.59 for every whole-image encoder on every dataset. Only MAE's split is a brightness split (0.65 UrbanCars, 0.84 CelebA, 0.60 ISIC, 0.62 MetaShift), and data2vec's on UrbanCars only (0.60). Which axis the latent-prediction models (DINOv2, data2vec, I-JEPA) split on, and why it coincides with the supervised axis on CelebA and ISIC but not on UrbanCars, remains open.

A.18 Matched training protocols and additional results

Section 4.4 gives the headline result and Figure 4. Table 29 reports each run against the frozen roster.

Within-set agreement. We report raw ARI followed by agreement normalized by self-consistency. The cross-view ResNet-18 pairs are SwAV–BYOL $0.24/0.89$, SwAV–DINO $0.35/0.99$, and BYOL–DINO $0.23/0.62$. DINO-ViT-S(v2) agrees with DINO, SwAV, and BYOL at $0.37/0.61$, $0.23/0.51$, and $0.16/0.33$, respectively. MAE-ViT-S agrees with the cross-view ResNet-18 runs at 0.04–0.06 raw and 0.13–0.19 normalized. Conv-MAE-RN18 agrees with the cross-view ResNet-18 runs at 0.04–0.11 raw and 0.10–0.19 normalized, and with MAE-ViT-S at −0.01.

Encoder	UrbanCars	CelebA	ISIC
CLIP-L/14	0.93	0.27	0.24
CLIP-B/32	0.83	0.46	0.32
CLIP-L/14 (LAION)	0.82	0.63	0.22
SigLIP-B/16	0.59	0.60	0.33
SigLIP-L/16	0.65	0.55	0.46
ViT-L/16 (sup.)	0.39	0.20	0.48
MSN	0.64	0.30	—
SwAV	0.80	0.65	—
SeLa-v2	0.66	0.68	—
MoCo-v3	0.65	0.37	—
CLIP-RN50x64	0.77	0.14	—
Qwen2-VL	0.66	0.28	—
CLIP-B/16	0.82	0.35	—
DINO	0.76	0.45	0.30
VICReg	0.86	0.31	0.47
Barlow Twins	0.78	0.30	0.40
SimSiam	0.68	0.40	0.54
DINOv2-L	0.24	0.62	0.35
MAE	0.56	0.46	0.33
data2vec	0.56	0.58	0.50
I-JEPA	0.85	0.38	0.94
SwAV (ResNet-18, COCO)	0.26	—	—
BYOL (ResNet-18, COCO)	0.29	—	—

Table 28: Self-consistency (mean ARI between an encoder's own root partitions on different draws, applied to the shared held-out pool). We use draws 180–219 and 5 seeds. Dashes mark backbones not extracted for that dataset, not missing values.

Run	self-cons.	ext. sup. 6	cross-view 5	hidden 3	same objective, pretrained
SwAV (RN18)	0.26	0.05	0.18	0.07	SwAV-RN50: 0.18
BYOL (RN18)	0.29	0.11	0.28	0.06	—
DINO (RN18)	0.47	0.03	0.11	0.05	DINO: 0.12
MAE (ViT-S)	0.33	-0.03	0.04	0.21	MAE: 0.37 $[0.30, 0.45]$
DINO (ViT-S, v2)	0.78	0.00	0.02	0.05	DINO: 0.01
conv-MAE (RN18)	0.69	0.01	0.00	-0.02	MAE: -0.03
MoCo-v3 (ViT-S)	0.23	0.05	0.18	0.16	MoCo-v3: 0.19
DINO (ViT-S, v1)	0.18	-0.01	0.00	0.01	collapsed (cos. 0.997)

Table 29: Matched set against the frozen roster, draws 180–219 (within-set pairs, raw and normalized below).

Conv-MAE therefore supports the preregistered low-agreement prediction for the cross-view ResNet-18 runs (P1). It *fails* the prediction that hidden-content runs agree with each other and with pretrained hidden-content models (P2). Its agreement is -0.01 with MAE-ViT-S, -0.03 with pretrained MAE, and -0.02 with the three pretrained hidden-content models. Its root partition is stable (self-consistency 0.69) but tracks none of the labelled factors.

Training protocol. We preregistered that whole-image runs would converge with each other and with pretrained whole-image models more than with the initial supervised six. We also predicted that MAE would agree with every whole-image run and pretrained whole-image model below 0.15, and with pretrained hidden-content models above that threshold. All runs share the images (COCO train2017, 118,287, resized to 256) and budget (200 epochs, batch 256). They share an augmentation family within each objective category, as detailed below. All are trained from scratch with `solo-learn` (da Costa et al., 2022). We evaluate them against the frozen roster using the paper's pipeline (UrbanCars class-1 root partitions, draws 180–219). SwAV, BYOL, and DINO use ResNet-18 with whole-image objectives. MAE uses ViT-S/16 with a masked objective and a single crop by construction.

Each run uses its objective's recipe in `solo-learn`. The three whole-image runs share LARS, a warm-up-cosine schedule, checkpointing every epoch, and two-view `symmetric` augmentation.

The augmentation uses two 224-px crops with colour jitter, greyscale, and blur. Each method uses its published base learning rate. MAE and the conv-MAE run below use AdamW with random-resized-crop and flip only, as prescribed for masked reconstruction. The objective therefore varies together with its augmentation. The testbed's P4 examines augmentation of a reconstruction objective separately (Appendix A.19), but does not isolate a loss-only effect in the matched set.

Multi-crop simplification. SwAV and DINO's original recipes also use multi-crop (additional small local views), an ingredient not coded in the ingredient analysis of Section 4.5 and not varied across the matched set's own runs. The two-view SwAV and DINO runs test the clustering and self-distillation ingredients, respectively, without multi-crop. They are valid but partial instances of those objectives, not full reproductions. BYOL's original recipe has no multi-crop, so its run is unaffected. Root partitions are extracted from the final checkpoint and evaluated with the identical script used for every other backbone in the paper, with no backbone-specific tuning.

ViT-S runs and failed predictions. The two ViT-S whole-image runs used solo-learn's DINO and MoCo-v3 recipes at the same budget. The first DINO-ViT-S run (v1) used $65{,}536$ prototypes with no teacher temperature warm-up and collapsed on the 118k-image set (self-consistency 0.18, feature cosine 0.997). The re-run (v2) used $4{,}096$ prototypes, a 30-epoch temperature warm-up, gradient clipping at 3 and learning rate 10^{-3}, and produced a more stable partition (self-consistency 0.78). Four further preregistered predictions failed and are reported here. The first preregistration, written under a discrete-assignment account, predicted that the SwAV run would agree with the initial supervised six more than the DINO and BYOL runs. This failed (0.05 against 0.03 and 0.11). It also predicted that DINO–BYOL would agree less than SwAV agrees with the pretrained MSN/SwAV/SeLa-v2 trio. This prediction also failed (0.23 against 0.18). The revised preregistration, frozen before the ViT-S runs, predicted that the DINO-ViT-S and MoCo-v3-ViT-S runs would agree at ≥ 0.5. This failed (0.04 raw, 0.10 normalized). It also predicted agreement with pretrained cross-view models at ≥ 0.4, which failed (0.02 and 0.18).

Convolutional masked reconstruction. The conv-MAE run is a ResNet-18 trained from scratch on the same images and budget with 16×16-patch 75% pixel masking, a 7×7 decoder and per-patch-normalized MSE (MAE defaults transposed to a CNN). Its root partition tracks neither background, co-object nor brightness (mean side differences ≤ 0.09).

A.19 Objective ingredients and controlled factor interventions

Predictions for the ingredient analysis and initial testbed were recorded before evaluation. The testbed settings were subsequently revised after the pilot described below. We report the predictions and this change alongside the results.

Ingredient analysis. For the 21 UrbanCars backbones, we code nine binary ingredients from the original papers. These are a whole-image target, a hidden-content target, input masking, colour/brightness augmentation, EMA teacher, prototypes, Sinkhorn–Knopp, text targets, and class labels. Architecture family and pretraining data are coded separately. The indicator *whole-image target* covers matching to a label, caption, or augmented view, and is therefore broader than the self-supervised *cross-view* category (it was recorded under the name "cross-view target" in the preregistration).

For each of the 210 pairs, the outcome is root ARI on draws 180–219. Pair-level features indicate whether both models share ingredient k or exactly one has it. We use 3000 backbone-identity permutations to account for pair dependence. The prespecified signs were positive for a shared whole-image target, negative when exactly one model has a hidden-content target, and non-positive when both do. Null effects were predicted for EMA, prototypes, Sinkhorn–Knopp, and input masking. The first three associations were not statistically significant, whereas input masking had correlation -0.565 (p=0.0013). The nonsignificant results do not establish that the corresponding effects are zero.

Input masking co-occurs with hidden-content prediction in this roster, limiting their separation. Shared text targets and pretraining data are additional confounds addressed by the matched set (Appendix A.18). In joint OLS (R^2=0.83), the cross-view and one-hidden indicators are correlated at -0.74. With 21 models, their conditional effects are difficult to distinguish. The regression identifies associations, while the controlled testbed below tests the effect of augmentation directly.

Indicator (pair-level)	Spearman ρ	perm. p	Frozen prediction
both: whole-image target	$+0.749$	0.0003	> 0 ✓
one: hidden-content target	-0.765	< 0.0005	< 0 ✓
both: hidden-content target	-0.206	0.019	≤ 0 ✓
one: input masking	-0.565	0.0013	(confounded with hidden content)
one: colour augmentation	-0.245	0.0007	—
both: text target	$+0.455$	0.0003	— (lineage)
same pretraining data	$+0.352$	0.0003	— (confound)
both: EMA teacher	-0.155	0.24	null
both: prototypes	$+0.099$	0.45	null
both: Sinkhorn–Knopp	$+0.035$	0.84	null
same architecture family	-0.075	0.53	null

Table 30: Ingredient analysis over 210 backbone pairs (UrbanCars root ARI, draws 180–219). p from 3000 backbone-identity permutations. Checkmarks indicate that the directional predictions were satisfied. Null denotes a prespecified expectation, not a conclusion of equivalence.

Controlled testbed. The coloured-MNIST testbed ($28\times28\times3$) contains three binary factors. F1 is illumination, set by global brightness multipliers $\times0.6$ and $\times1.0$, and has the largest pixel variance. F2 is shape, defined within the even class by digits $\{0, 6, 8\}$ versus $\{2, 4\}$ and preserved by the augmentation. F3 is the colour of a 6×6 corner patch. The classification target is parity. Factor labels are Bernoulli($\frac{1}{2}$) and independent of class.

The rendering procedure introduces two qualifications. Illumination is applied after the patch is drawn, so F3's raw magnitude is coupled to F1 even though their labels are independent. Shape variance is inherited from MNIST rather than set by the experimenter and is smaller than the designed patch variance in the training tensors. F2 is therefore an available residual factor, not a controlled-variance intervention equivalent to F1 or F3. Its smaller variance is consistent with the failed shape prediction below, but that association does not establish its cause.

A three-convolution encoder (64-d) is trained from scratch for 3000 steps with batch size 256 and five seeds. The seven objectives are supervised learning, SimCLR (Chen et al., 2020), BYOL, SwAV-lite (Sinkhorn–Knopp, $K=32$), DINO-lite (EMA teacher, centring, $K=64$), MAE-lite (4×4 patches, 75% masking, pixel MSE), and data2vec-lite (masked-patch regression onto an EMA teacher's latents). Architecture, training data, and budget are fixed.

Full augmentation combines random resized crops, log-uniform brightness in $[0.4, 1.6]$, and mild colour jitter. The *no-brightness* condition removes the brightness term, while *none* removes all augmentation. Evaluation uses the fixed partition pipeline on the even-class pool (300 images, 20 draws), with each fitted rule applied to a fixed held-out pool. An initial configuration with a $\times0.35$ illumination gap was discarded because every run followed illumination at ARI $\approx$ 1, preventing discrimination between objectives. Only the stated testbed constants changed, and the predictions were retained.

Run	F1 (illum.)	F2 (shape)	F3 (patch)	self-cons.
supervised, SimCLR, BYOL, SwAV-lite, DINO-lite (full aug.)	0.00–0.01	≤ 0.06	≤ 0.01	0.20–0.43
SimCLR / BYOL, no-brightness aug.	**0.96 / 0.96**	0.00	0.01	0.96 / 0.99
SimCLR, no aug.	0.01	0.00	0.01	0.05 (collapsed)
MAE-lite, no aug. / full aug.	0.26 / **0.52**	0.00	**0.54** / 0.30	0.58 / 0.55
data2vec-lite	**0.93**	0.00	0.04	0.90

Table 31: Controlled testbed, mean over five seeds. Pairwise ARI between runs: cross-view runs (full aug.) vs. MAE-lite / data2vec-lite ≤ 0.01; SimCLR / BYOL without brightness augmentation vs. data2vec-lite 0.89, vs. MAE-lite (full aug.) 0.50; cross-view runs with full augmentation among themselves 0.00–0.32 (the supervised run agrees with none of the others).

Outcomes of the prespecified predictions. (P2) Masked objectives were predicted to follow a low-level factor and have low agreement with cross-view runs under full augmentation. This prediction was satisfied. (P3) Removing brightness augmentation was predicted to shift SimCLR and BYOL towards illumination. This occurred, and their agreement with data2vec-lite reached 0.89. Partition alignment therefore changed under the same cross-view loss. (P4) The registered predic-

patch (px)	var F1	var F3	sup.	SimCLR (cv)	BYOL (cv)	MAE-lite (msk-pix)	data2vec-lite (msk-lat)
4	6.2	1.8	F2	F2	F1	F1 .71	**F1** .92
6	7.0	4.1	F2	F2	F1	F3 .61	**F1** .93
8	8.2	7.2	F2	F2	F3	F3 .83	**F1** .77
10	9.5	**11.3**	F2	F3	**F3** 1.00	F3	**F1→F3** .64
12	11.0	16.3	F2	F3 .85	F3 1.00	F3	**F3** .81
16	14.4	28.9	F2	F3 1.00	F3 1.00	F3	**F3** .85
20	18.1	45.2	F2	F3 .63	F3 1.00	F3	**F3** .91

Table 32: Variance ladder over five seeds (cv: cross-view; msk-pix: masked-pixel target; msk-lat: masked-latent target). Each cell names the factor most aligned with the root partition; a reported ARI marks stronger alignment, whereas an unnumbered factor may be weak. Growing the F3 patch reverses the empirical F1–F3 variance ordering between the 8- and 10-pixel rungs. The data2vec-lite model gives the clearest crossover. BYOL becomes strongly F3-aligned near the reversal, while MAE-lite moves earlier. The transition is therefore objective-dependent.

tion was that augmenting MAE's input would not create invariance, since its reconstruction target is the augmented input itself. The observed partitions retained factor alignment, consistent with this prediction.

(P1) Cross-view models were predicted to agree at ≥ 0.5 on shape. This prediction failed. Their root partitions showed near-zero illumination alignment but did not converge on shape, and self-consistency was 0.20–0.43. Low stability also occurs for pretrained encoders on ISIC and CelebA (Appendix A.17). The testbed therefore supports suppression of an augmented factor in the root partition, not convergence on an arbitrary remaining factor.

A.20 Variance ladder: objective-dependent selection among surviving factors

The testbed above shows what a cross-view objective suppresses. To examine selection among the remaining factors, we run a controlled variance ladder in the same coloured-MNIST testbed, reusing the seven-objective machinery unchanged. We focus on F1 illumination, a global brightness gap covered by the brightness augmentation, and F3, a corner colour patch. F2 shape remains available as a residual factor. We raise F3's pixel variance by growing the patch from 4 to 20 pixels across seven rungs while holding the illumination parameter fixed. Because illumination is applied after the patch is rendered, its empirical variance also rises with patch size. We measure which factor is most aligned with each objective's root partition over five seeds. The empirical F1 and F3 variances reverse order between the 8- and 10-pixel rungs.

The objectives do not share one exact transition point (Table 32). The data2vec-lite model gives the clearest crossover, tracking F1 while $\mathrm{var}(\mathrm{F1}) > \mathrm{var}(\mathrm{F3})$ and moving towards F3 near the empirical reversal. BYOL is weakly top-aligned with F3 before that reversal but becomes stably aligned with F3 ($\mathrm{ARI} = 1.00$ at the 10-pixel rung). SimCLR remains weakly aligned with F2 until F3 becomes sufficiently strong. MAE-lite moves to F3 earlier, consistent with reconstruction-target variance rather than the raw input-variance ranking. The supervised arm has F2 as its top-aligned factor but shows no substantial alignment with any factor. Parity is independent of all three factors in the balanced evaluation pool, so predicting parity does not require separating these factors within a class. A repeat with an independent data seed, using a different MNIST subsample, factor assignment, and draws, reproduces the reported pattern and approximate transition region.

The variance sweep supports an objective-dependent response to the intervention, not a universal rule selecting the largest raw-variance factor. In particular, augmentation can prevent a stable illumination-aligned root partition even when illumination has the largest raw variance. Because patch size changes together with variance, the sweep does not separate those two properties. The pretrained observations are consistent with the proposed account but do not independently establish its mechanism. In the pretrained roster, each whole-image encoder's leading class-conditioned axis aligns with background ($|\cos|$ about 0.95), whereas MAE's leading axis does not (0.15, brightness). Thus, the observed leading axes align with different factors across objectives.

A.21 MONTAGES: THE SAME 300 IMAGES UNDER TWENTY PARTITIONS

Figures 13–17 show, for draw 180, the UrbanCars class-1 held-out pool split by each encoder's root partition (24 random images per side; side A is oriented to match the reference encoder's side A). Each block reports ARI against a reference encoder both as the 40-draw mean and for the draw shown, and the automatic caption names the labelled factor (background, co-occurring object) or mean thumbnail brightness that separates the two sides, when one does. Single draws are noisy: SwAV agrees with DINO at 0.69 over draws but 0.05 on this one, where DINO's split followed the co-occurring object instead of the background.



Figure 13: The core four encoders (reference CLIP-L/14). Left, UrbanCars: CLIP and VICReg split on background, DINOv2 partly follows, and MAE splits on brightness. Right, CelebA (Blond_Hair): MAE again splits on brightness and no root partition tracks gender, which appears one level deeper (Section 4.3).



Figure 14: Cross-view models, reference DINO-B/16.



Figure 15: Hidden-content models, reference MAE-L/16. data2vec is the one member whose split touches background as well as brightness.



Figure 16: Whole-image vs. hidden-content objectives, reference DINO-B/16.



Figure 17: Matched set, each run against its pretrained counterpart. BYOL has no pretrained counterpart in the roster, so SimSiam is used as a related self-distillation reference. These displayed draws do not show a consistent labelled factor across the matched-set runs. Mean factor-alignment results across draws appear in Appendix A.17.