WHAT CLASSIFIER RECOVERY REVEALS ABOUT CROSS-VIEW GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A visual recognition system can fail after a change in camera viewpoint yet recover substantial accuracy by refitting its classifier on labelled target data while keeping the feature extractor frozen. Such recovery establishes accessible class information, but does not establish preserved representation geometry or a portable decision rule. We investigate this distinction through a diagnostic decomposition of local linear recoverability, class-relational organization, and cross-view decision compatibility across five task-trained convolutional architectures and frozen DINOv2 representations. Under cross-dataset transfer, convolutional classifiers retain 48.1–57.8% accuracy for a broadly comparable side-right view but fall to 6.9–12.1% for a frontal view; classifier-only frontal refitting recovers 47.1–59.4%, while DINOv2 rises from 11.8% to 94.1% under its reference protocol. Direct frontal–side-right geometry within 100-Driver shows a distinct structural regime: DINOv2 crop features retain substantially greater class-relational correspondence than the source-to-frontal cross-dataset comparison, with sensitivity to representation, input treatment, and covariance reference. Shared classifiers nevertheless remain view-asymmetric, and recovery decreases under driver-disjoint evaluation. The resulting diagnosis separates recoverable class information from class-relational change and from the functional compatibility of a fitted decision rule. These properties require separate measurements under explicit population contracts. This provides a diagnostic basis for evaluating single-source, closed-set cross-view methods beyond their target recovery score.

1 INTRODUCTION

Changes in camera viewpoint remain a persistent source of failure in visual activity recognition because the same semantic action can produce substantially different image evidence under different observation geometries. Cross-view action recognition has therefore been studied through cross-view modeling, view-invariant feature learning, contrastive representation learning, and domain adaptation (Wang et al., 2014; Kong et al., 2017; Zhang et al., 2023; Bhalla et al., 2025). The problem is particularly acute for fine-grained driver-behavior recognition, where camera placement changes which parts of the driver, hands, objects, and cabin are visible. Multi-camera benchmarks such as Drive&Act and 100-Driver explicitly expose cross-view evaluation settings (Martin et al., 2019; Wang et al., 2023), while recent driver-monitoring methods address viewpoint variation through pose-based invariance, domain adaptation, cross-modal adaptation, and cross-camera feature disentanglement (Martin et al., 2023; Li et al., 2024; Bhalla et al., 2025; Celona et al., 2026). Foundation-model studies further show that strong pretrained representations, including DINOv2, can still degrade substantially on previously unseen views (Ponbagavathi et al., 2024; 2025). These results establish camera viewpoint as a genuine generalization challenge, but they primarily ask how to improve recognition across views.

A different question arises when cross-view failure can be largely repaired without changing the feature extractor. Last-layer retraining has already shown that frozen representations may retain useful class information even when the original classifier relies on spurious correlations (Kirichenko et al., 2023). Work on linear probing and lightweight target adaptation likewise motivates separating classifier adaptation from changes to the feature extractor (Huang et al., 2024; Zhang et al., 2026). However, how classifier recovery should be interpreted under viewpoint change remains unresolved.

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.

Recovery alone does not locate the failure. It neither establishes preservation of class relationships nor tests whether an unchanged decision rule serves another view or unseen drivers. Existing cross-view methods and foundation-model probes largely report recognition or adaptation outcomes (Bhalla et al., 2025; Celona et al., 2026; Ponbagavathi et al., 2024). We instead ask what remains view-conditioned *after* local classifier recovery succeeds.

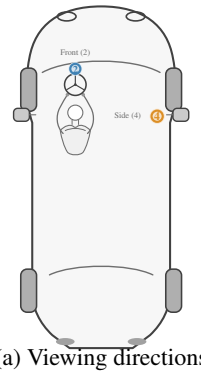
We investigate this question in fine-grained driver-behavior recognition through a nested cross-domain and cross-view design. State Farm Distracted Driver Detection (SF3D) (Montoya et al., 2016) provides the source collection and a right-windshield observation, while 100-Driver (Wang et al., 2023) provides the two target views used in this study: Cam4 (side-right) and Cam2 (frontal). Their arrangement creates three complementary comparisons (Figure 1): SF3D→Cam4 measures cross-domain transfer under broadly comparable viewing directions, SF3D→Cam2 combines cross-domain and cross-view change, and Cam4↔Cam2 compares views within the same parent collection, reducing collection-level confounding without matching participants or action instances. The distinction is important because SF3D and 100-Driver differ in people, interiors, and acquisition conditions as well as camera placement; the cross-dataset contrast alone therefore cannot identify viewpoint as the cause of a transfer loss.

The key finding is that classifier recoverability and cross-view generalization are distinct empirical properties. We operationalize this distinction through three complementary quantities—local linear recoverability, class-relational organization, and cross-view decision compatibility—measured under a fixed feature extractor and explicit population contracts (Figure 2). Direct within-100-Driver geometry qualifies the cross-dataset picture: frozen DINOv2 retains substantial frontal–side-right class-relational correspondence, while the functional tests still expose conditional classifier behavior. The contribution is the joint diagnostic design and its empirical application, not a new probing or RDM estimator. It distinguishes accessible information from preserved class relationships and compatible decisions, and specifies which additional evidence an invariance or generalization claim requires.

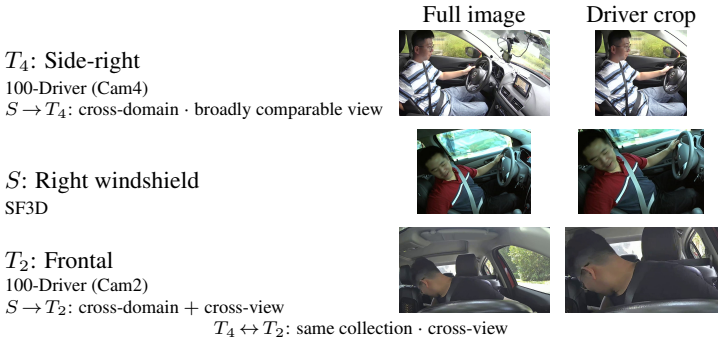
2 Related Work

Cross-view recognition and driver monitoring. Cross-view activity recognition has long treated viewpoint variation as a representation and generalization problem. Classical approaches model correspondences across views or learn view-invariant features (Wang et al., 2014; Kong et al., 2017), while more recent work uses contrastive objectives to encourage view-invariant action representations (Zhang et al., 2023). The challenge is particularly pronounced in fine-grained driver-behavior recognition, where multi-camera datasets such as Drive&Act (Martin et al., 2019) and 100-Driver (Wang et al., 2023) expose substantial changes in visible pose, object relations, and cabin context across camera viewpoints. Driver-specific approaches have addressed this variation through 3D pose-based invariance, domain adaptation, cross-modal adaptation, and cross-camera feature disentanglement (Martin et al., 2023; Li et al., 2024; Bhalla et al., 2025; Celona et al., 2026). Foundation-model probing further shows that broadly pretrained representations, including DINOv2, can remain sensitive to unseen camera views (Ponbagavathi et al., 2024; 2025). These works establish viewpoint variation as a persistent generalization problem, but their evidence is primarily expressed through cross-view recognition or adaptation outcomes; such outcomes do not by themselves determine what remains view-conditioned when performance can be recovered by changing only the classifier.

Classifier recovery and target-supported adaptation. A complementary literature asks how much predictive performance can be recovered from an existing representation when the feature extractor remains fixed or is only lightly adapted. Deep Feature Reweighting (DFR) shows that models trained under spurious correlations can retain useful core features even when the original final layer fails, allowing last-layer retraining on a small reweighting set to recover robustness (Kirichenko et al., 2023). LP++ demonstrates the strength of simple linear probing under few-shot adaptation, while FiSeR connects representation separability with few-shot classifier repair under domain shift (Huang et al., 2024; Zhang et al., 2026). Cross-domain few-shot learning further considers support-conditioned adapters, feature-space adaptation, foundation-model adaptation, and invariant-content reconstruction (Das et al., 2022; Li et al., 2022; Perera & Halgamuge, 2024; Zhou et al., 2025; Tian et al., 2026). This target-supported setting differs from single-domain generalization, which seeks robustness without supervised target-domain adjustment (Qiao et al., 2020; Xu et al., 2025; Hei-



(a) Viewing directions



(b) Separate Reach Behind examples ($c7/C20$)

Figure 1: Observation conditions and qualitative comparison paths. Reach Behind examples are separate instances; camera geometry is schematic, not calibrated.

dari & Guo, 2025). When classifier fitting succeeds, it establishes that the available representation supports discrimination under the corresponding fitting and evaluation conditions; it does not establish that class organization is preserved across views or that the recovered decision rule remains compatible with another viewpoint.

Representation structure and transfer diagnostics. Recognition accuracy alone does not reveal whether a change in viewpoint has altered the organization of classes in feature space or only the compatibility of a previously fitted decision rule. Representational Similarity Analysis (RSA) addresses the former by comparing relational structure across representations, commonly summarized through a representational dissimilarity matrix (RDM) of pairwise condition or class distances (Kriegeskorte et al., 2008). Other representation-comparison methods, including Singular Vector Canonical Correlation Analysis (SVCCA) and Centered Kernel Alignment (CKA), characterize similarity through complementary subspace- and kernel-based perspectives (Raghu et al., 2017; Kornblith et al., 2019). For class-relational analysis, subsequent RSA methodology provides crossvalidated dissimilarity estimates to reduce positive distance bias and covariance-aware RDM comparisons that account for dependence among pairwise distances (Walther et al., 2016; Diedrichsen et al., 2021; Schütt et al., 2023). A neighboring literature asks a different question by estimating transferability or out-of-distribution (OOD) performance from properties such as LEEP, LogME, or feature separability (Nguyen et al., 2020; You et al., 2021; Xie et al., 2023), while structure-consistent unsupervised domain adaptation (UDA) has explicitly exploited relational structure in driver-behavior recognition (Liu et al., 2023). These approaches therefore provide complementary evidence about representation structure and transfer, but neither representational similarity nor a transferability estimate establishes what classifier-only recovery means across camera views. We use class-relational geometry as one component of that diagnosis and measure local linear recoverability and cross-view decision compatibility separately rather than treating any one quantity as a proxy for the others.

3 Diagnostic Methodology

Classifier-only recovery is interpreted relative to a fixed representation and an explicit support–query population contract. For condition A, let $\mathcal{S}_{A,\pi}^{(k)}$ contain k labelled examples per class under protocol π, and let $h_{A,\pi}^{(k)}$ denote the fitted classifier-side pipeline: support-estimated feature standardization followed by a linear classifier on frozen features $z = f_\theta(x)$. When this pipeline is evaluated elsewhere, neither component is re-estimated. Figure 2 connects the classifier interventions to the three diagnostic readouts; Appendix C gives the complete fitting and population contracts.

Local linear recoverability. $L_A(k;\pi) = \mathrm{Acc}(h_{A,\pi}^{(k)} \circ f_\theta; Q_{A,\pi})$ asks whether class information remains linearly accessible after classifier-side refitting in the condition being evaluated. Image-disjoint evaluation removes support images but can retain the same drivers; driver-disjoint evaluation also excludes fitting drivers from the query. Their difference therefore measures population

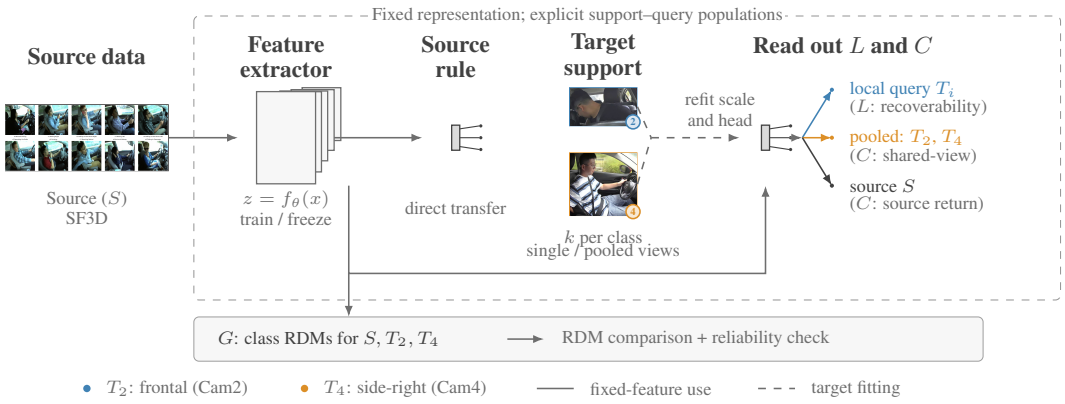


Figure 2: Diagnostic workflow on frozen CNN or DINOv2 features. Refit only the support-derived standardizer and linear head: single-view fits measure local recovery (L) and source return (C); pooled fits evaluate one rule on both views (C). RDMs measure class-relational organization (G) independently. The pooled budget is k total labels per class; support–query population contracts apply throughout.

sensitivity, not an isolated causal effect of identity leakage. Exact sampling, regularization selection, and query construction are specified in Appendix C, with population controls in Appendix H.

Class-relational organization. Recoverability does not determine whether class relationships are preserved. We compare class-level representational dissimilarity matrices using crossvalidated class distances and covariance-aware RDM correspondence (Kriegeskorte et al., 2008; Walther et al., 2016; Diedrichsen et al., 2021; Schütt et al., 2023). Independent within-condition RDMs provide reliability references. Cross-dataset $S \leftrightarrow T_2$ and $S \leftrightarrow T_4$ measurements remain compound contrasts; $T_2 \leftrightarrow T_4$ compares views within one collection. We retain both covariance-reference choices for the latter: $G(T_2, T_4; \Sigma_{T_2})$ and $G(T_4, T_2; \Sigma_{T_4})$. Their difference is estimator sensitivity, not directional classifier transfer. Appendix D gives the estimator details and Appendix G the full geometry analyses.

Cross-view decision compatibility. For a pipeline fitted in A and evaluated without refitting in B, $C_{A\to B}(k;\pi) = \mathrm{Acc}(h^{(k)}_{A,\pi} \circ f_\theta; Q_{B,\pi})$. The diagonal case is local recoverability; off-diagonal evaluation tests portability of the fitted rule. We additionally fit one classifier from pooled $T_2 \cup T_4$ support and return view-specific fitted pipelines to the source feature pool. Complete shared-view and source-return results are in Appendix F. These three quantities answer different questions and are not treated as interchangeable proxies or additive causal shares of cross-view error.

4 Experimental Design

Conditions and task. We study SF3D as source S, 100-Driver Cam4 as side-right target T_4, and Cam2 as frontal target T_2 (Montoya et al., 2016; Wang et al., 2023). Thus $S \to T_4$ is cross-dataset under broadly comparable viewing directions, $S \to T_2$ combines dataset and view change, and $T_4 \leftrightarrow T_2$ is the more focused within-collection view contrast. The datasets are normalized to a common ten-class task; one food/drink correspondence is imperfect and is retained with an evaluation-only sensitivity check. Primary CNN/DINOv2 experiments use detected driver crops. Appendix A gives the complete label mapping, view specification, sample counts, and supporting input treatments.

Representations. Five ImageNet-initialized CNNs are task-trained on S: ResNet-18/50, DenseNet-121, EfficientNet-B0, and SqueezeNet1.0 (He et al., 2016; Huang et al., 2017; Tan & Le, 2019; Iandola et al., 2017). We also use frozen DINOv2 ViT-S/14 features with and without registers (Oquab et al., 2024; Darcet et al., 2024). These seven common-coordinate feature spaces support local and transported classifier tests. DBMNet contributes raw and action-conditioned held-out-camera features as two additional within-condition comparators (Celona et al., 2026); its camera-specific checkpoints are not pooled across views. Model lineage, feature hooks, preprocessing, and training provenance are documented in Appendix B.

Fitting and evaluation. Finite-budget fits use $k \in \{2, 5, 10, 20, 50, 100\}$ labels per class and five support draws. The fitted standardizer and regularized linear classifier travel together when evaluated under another condition. We compare image-disjoint and driver-disjoint recovery, pooled two-view fitting and return to the source condition. Direct reciprocal target-view transport is a separate control, not inferred from RDM reference reversal. Historical CNN/DINOv2 target references retain their recorded protocols and are used as recoverability references rather than matched ceilings. Appendix C gives the exact optimizer, regularization, support construction, and query definitions; Appendix I records artifact and regeneration provenance.

5 RESULTS

5.1 CLASSIFIER REFITTING REVEALS RECOVERABLE TARGET DISCRIMINATION

Direct transfer depends strongly on observation condition. Across the five task-trained CNNs, source-to-frontal accuracy is only 6.9–12.1%, while the broadly comparable side-right view retains 48.1–57.8%. Refitting only the classifier-side pipeline raises frontal accuracy to 47.1–59.4% and side-right accuracy to 73.0–85.8%; DINOv2 rises from 11.8% to 94.1% frontally and from 40.3% to 98.7% side-right. Table 1 gives the nine-configuration reference view; complete per-model and per-treatment references are in Appendix E.

Recovery is not guaranteed by replacing the classifier: same-query source refitting reduces accuracy for every CNN in Table 1. More importantly, recovery depends on the evaluation population. Figure 3 shows the complete budget sweep for all nine feature configurations. Image-disjoint curves can remain high while driver-disjoint curves are much lower, especially on the frontal view; at $k = 100$, for example, ResNet50 changes from 94.3% to 30.8% and DINOv2 from 92.3% to 51.4%. The side-right condition shows the same population dependence. Because whole-driver support selection also changes support diversity and the remaining query population, this gap is interpreted as population sensitivity rather than a causal estimate of identity leakage. Appendix C specifies the sampling contract, and Appendix H reports query-population and sensitivity controls.

5.2 RECOVERABILITY DOES NOT IMPLY PRESERVED CLASS-RELATIONAL ORGANIZATION

Classifier recovery establishes accessible discrimination, not preservation of class relationships. For primary DINOv2 crop features, source-to-frontal class-relational correspondence is 0.021, compared with 0.306 for source-to-side-right; image-half reliability is 0.995 and 0.962, respectively. Figure 4 shows that high local recovery can coexist with a different recorded class-distance pattern. This remains a compound cross-dataset contrast, not a viewpoint-only effect.

Direct within-collection view comparison. Table 2 reports Cam2–Cam4 geometry for seven encoder configurations. The CNN extension is architecture-indexed; its checkpoint-level match to the recovery models is unverified (Appendix B.5). For DINOv2, correspondence is 0.727 with Cam2 as covariance reference and 0.647 with Cam4; with registers, the values are 0.782 and 0.698. These crop results retain substantially more correspondence than the corresponding source-to-frontal comparisons and rule out describing the latter as universal viewpoint-induced geometric collapse. The CNN crop results are more variable. The two reference choices are sensitivity measurements, not reciprocal classifier accuracies or independent replications.

Across all treatments, the source–side-right correspondence exceeds source–frontal in 19/21 cells, with SqueezeNet `seg`/`bboxseg` as the exceptions. The direct-versus-cross-dataset contrast is treatment-dependent, not uniformly positive. The full matrix, reference sensitivity, and population qualifications are in Appendix G; Appendix D specifies the estimator. Within-collection comparison reduces dataset confounding, but differing camera populations and unmatched instances prevent causal isolation of viewpoint.

5.3 RECOVERED DECISION RULES REMAIN CONDITION-DEPENDENT

A third question is whether a recovered decision rule remains useful beyond the condition on which it was fitted. With $k = 100$ pooled across both target views, the shared ResNet50 classifier reaches 18.8% on frontal queries but 63.1% side-right; the corresponding macro recalls are 20.5% and

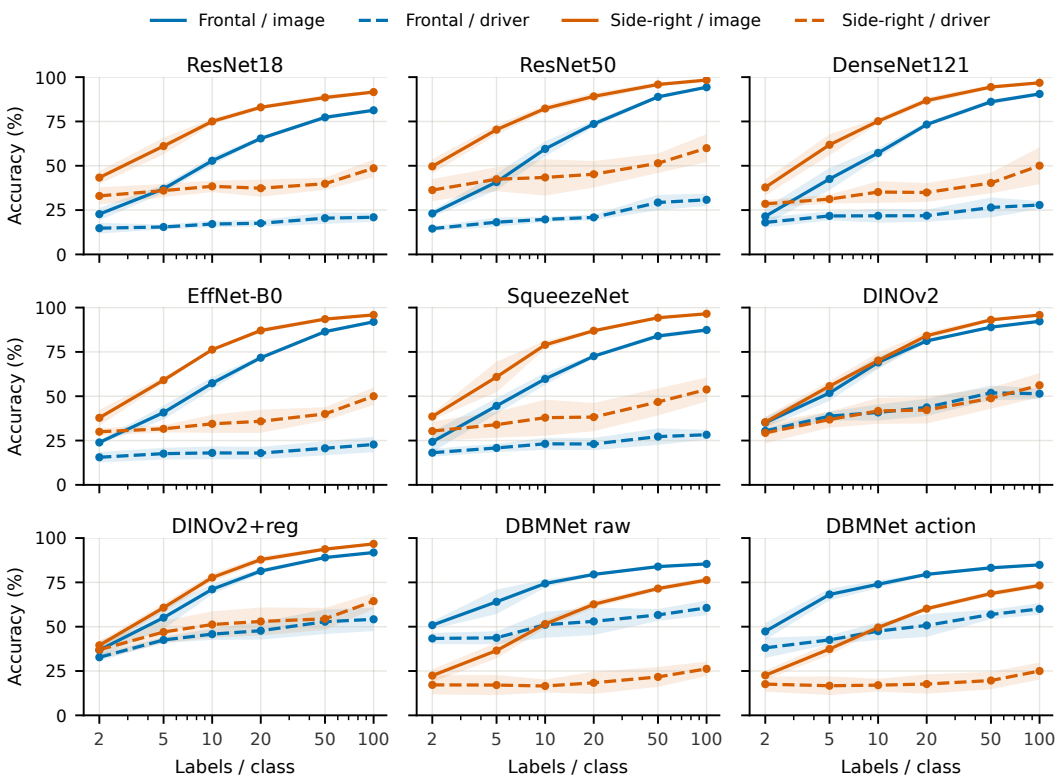


Figure 3: Local calibration can recover recognition, but the magnitude depends on the evaluation protocol. Both cameras and representations share axes; bands are standard deviations over five support draws. Driver separation is a supporting validity control, not a matched causal estimate of identity leakage.

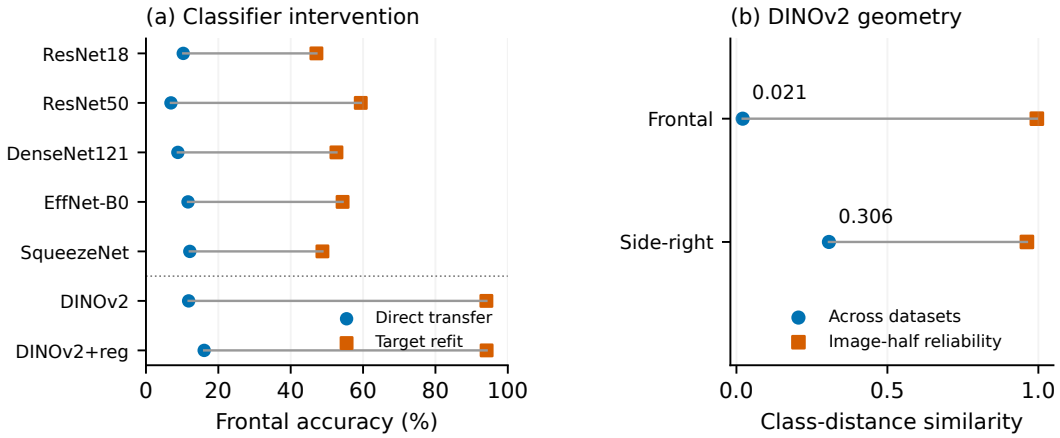


Figure 4: Recovered recognition need not preserve the source class-distance pattern. Panel (a) retains the distinct CNN and DINOv2 reference protocols; panel (b) shows cross-dataset DINOv2 correspondence with image-half reliability. Direct Cam2–Cam4 geometry is reported separately in Table 2; neither comparison isolates viewpoint causally.

	Source: SF3D			Target: frontal		Target: side-right	
Features	Native	Refit	Δ_S	Direct	Refit	Direct	Refit
ResNet18	82.3	78.3	-4.0	10.3	47.1	50.0	73.0
ResNet50	77.3	64.7	-12.5	6.9	59.4	57.8	78.3
DenseNet121	83.1	69.9	-13.2	8.8	52.6	55.6	81.5
EffNet-B0	77.4	65.4	-12.0	11.6	54.3	48.1	74.8
SqueezeNet	65.2	38.0	-27.2	12.1	48.8	50.2	85.8
DINOv2	–	96.2	–	11.8	94.1	40.3	98.7
DINOv2+reg	–	96.2	–	16.1	94.1	52.0	98.2
DBMNet raw	–	–	–	67.2	83.8	19.9	79.7
DBMNet action	–	–	–	67.2	84.3	19.9	76.1

Table 1: Reference accuracy (%) under direct transfer and classifier refitting. Reference fits retain their recorded protocols; DBMNet is a held-out-camera comparator rather than a matched representation ranking.

	Cross-dataset		Within 100-Driver	
Features	$G(S, T_2; \Sigma_S)$	$G(S, T_4; \Sigma_S)$	$G(T_2, T_4; \Sigma_{T_2})$	$G(T_4, T_2; \Sigma_{T_4})$
ResNet18†	0.356	0.574	0.438	0.123
ResNet50†	0.414	0.546	0.314	0.135
DenseNet121†	-0.046	0.544	0.176	0.210
EffNet-B0†	-0.144	0.571	0.113	-0.153
SqueezeNet†	-0.098	0.153	0.304	0.331
DINOv2	0.021	0.306	0.727	0.647
DINOv2+reg	0.295	0.591	0.782	0.698

Table 2: Crop-feature class-relational correspondence. The within-collection columns compare the same two camera conditions with different covariance references; they are not directions of classifier transfer. CNN rows (†) have unverified checkpoint-level matching to the recovery models. Complete treatment results appear in Table 21.

62.4%. DINOv2 reduces but does not eliminate this asymmetry (39.7/48.8%). The realized support averages 46.66% frontal and 53.34% side-right; this is not a strictly camera-balanced intervention. Macro-recall and no-$c6$ checks preserve the qualitative result (Appendix F, Tables 17–18).

Returning independently fitted target pipelines to the same source feature pool gives a complementary compatibility test. ResNet50 returns at 17.7% after frontal fitting versus 64.0% after side-right fitting, and the side-right-fitted pipeline returns better across all seven common-coordinate feature spaces. Because the backbone is fixed, this is condition dependence at the classifier interface rather than backbone forgetting. Figure 5 shows the two main compatibility views; complete seven-representation tables, support composition, confusion matrices, and source-return results are in Appendix F.

6 Discussion

Classifier recovery changes the interpretation of cross-view failure, but is not itself an invariance test. A frozen representation can retain substantial class information when a source decision rule fails, while its recovered behavior remains conditional on viewpoint, fitting support, and evaluation population. The diagnostic contribution is to expose these conditions through separate structural and functional measurements rather than assign additive causal shares to a "failed representation" and a "defective head."

The direct within-collection geometry is a consequential qualification, not a contradiction of this contribution. DINOv2 crop RDMs correspond much more closely across Cam2 and Cam4 than across SF3D and Cam2, whereas CNNs and masked treatments exhibit different patterns. Therefore the very low cross-dataset frontal score cannot be attributed to viewpoint alone. Conversely, substantial relational correspondence is not a guarantee of a common decision rule: class distances omit the coordinate placement and discriminative directions that matter to a fixed classifier. Appendix D

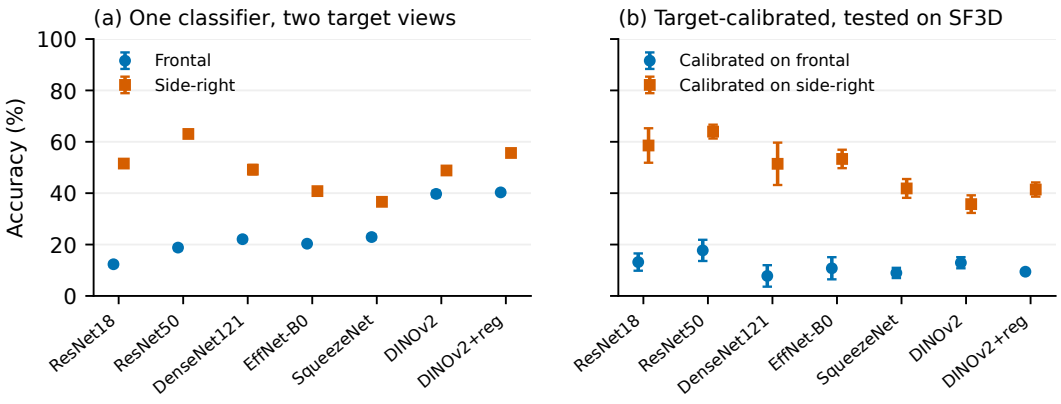


Figure 5: View dependence persists after classifier intervention. Left: one shared-view classifier evaluated on Cam2/frontal and Cam4/side-right. Right: independently view-calibrated pipelines returned to the same source query. Error bars show support-draw standard deviation. The panels test different rules and must not be read as a paired separate-versus-pooled comparison.

gives analytical boundary cases; the observed local, pooled, and source-return outcomes answer separate functional questions, not a matched causal test of geometry against accuracy.

This distinction guides evaluation of alignment, view-invariant learning, and classifier repair (Kong et al., 2017; Martin et al., 2023; Li et al., 2024; Bhalla et al., 2025; Kirichenko et al., 2023; Huang et al., 2024; Zhang et al., 2026). A method claiming recoverability needs a declared target-support protocol; a method claiming portable decisions additionally needs unchanged-rule evaluation on the intended views and populations. RDM correspondence is useful structural evidence, not a substitute for that functional test. Changing the covariance reference cannot establish a preferred transfer direction. The source-aligned side-right advantage of the CNN classifiers is compatible with source-task training bias, but support allocation, regularization, and visibility remain alternative contributors.

The scope is single-source, closed-set, static-image recognition, with a multi-camera-trained comparator evaluated under its distinct information contract. Direct Cam2–Cam4 measurements reduce collection-level confounding but do not match people or synchronized instances. Exact driver-disjoint fitting changes the remaining query population with k, so its curves are not pure sample-complexity estimates. Fixed held-out query drivers, strictly balanced shared support, reciprocal view-specific classifier transport, matched full-model adaptation, and additional multi-view benchmarks remain important controls. The evidence nevertheless supports a concrete conclusion: classifier recovery establishes accessible information under a declared protocol, not invariant geometry or decisions portable across people and viewpoints.

7 Conclusion

Classifier recovery reveals accessible class information in a frozen representation, but does not by itself establish preserved class relationships or a portable decision rule. Direct within-100-Driver geometry qualifies the cross-dataset evidence: DINOv2 crop features retain substantial correspondence across frontal and side-right views, while relational measurements vary with representation, input treatment, and covariance reference. Shared-view and source-return classifiers remain condition-dependent, and recovery depends on the support–query population contract. The resulting diagnostic decomposition separates recoverability, class-relational organization, and decision compatibility within a common evaluation design. Cross-view methods should be evaluated against the particular structural and functional properties they claim to preserve.

Reproducibility Statement

The paper and extended methods specify the diagnostic interventions, feature configurations, support budgets, evaluation populations, and measurement procedures. Numerical tables, prose values, and data-driven plots are traceable to recorded result exports and their regeneration scripts; aggregate statistics and shared-view confusion summaries are checked against the available draw-level records. The direct within-collection geometry results retain their estimator and covariance-reference settings. Historical feature projections are retained only as illustrative archived images where their original coordinates and sampling records are unavailable. Regeneration from recorded results is distinguished from an end-to-end rerun requiring the original datasets, model weights, and feature tensors.

Ethics Statement

This study reuses research datasets of in-cabin activity. Photographs illustrate predefined behavior categories, not claims about the people pictured. The observed accuracies do not validate safety-critical deployment or individual surveillance decisions. Population differences, imperfect label correspondence, and residual errors limit conclusions about use beyond the evaluated setting. Access to and redistribution of dataset material remain governed by the original dataset terms.

Statement on Use of Large Language Models

Large language models assisted literature discovery and comparison, hypothesis and experimental-design discussion, refinement of the diagnostic framework and analytical explanations, code inspection and development, analysis of recorded results, figure and table preparation, and manuscript writing and editing. Numerical findings originate from executed experiments and disclosed reanalysis of their records. Model-generated suggestions are not independent empirical evidence. The authors remain responsible for the final claims, citations, analyses, and verification.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Adv. Neural Inf. Process. Syst. NeurIPS*, volume 31. Curran Associates, Inc., 2018.

Aditi Bhalla, Christian Hellert, and Enkelejda Kasneci. Cross-view cross-modal unsupervised domain adaptation for driver monitoring system. In *2025 IEEEWIC Int. Conf. Web Intell. Intell. Agent Technol. WI-IAT*, pp. 73–80, 2025. doi: 10.1109/WI-IAT67162.2025.00018.

Luigi Celona, Simone Bianco, and Paolo Napoletano. Cross-Camera Distracted Driver Classification Through Feature Disentanglement and Contrastive Learning. *IEEE Trans. Intell. Transp. Syst.*, 27(5):5149–5162, May 2026. ISSN 1524-9050, 1558-0016. doi: 10.1109/TITS.2026.3675161.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Debasmit Das, Sungrack Yun, and Fatih Porikli. ConFeSS: A Framework For Single Source Cross-Domain Few-Shot Learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Jörn Diedrichsen, Eva Berlot, Marieke Mur, Heiko H. Schütt, Mahdiyar Shahbazi, and Nikolaus Kriegeskorte. Comparing representational geometries using whitened unbiased-distance-matrix similarity, August 2021.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. In *Int. Conf. Learn. Represent. ICLR*, 2022. doi: 10.48550/arXiv.2201.04234.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Marzi Heidari and Yuhong Guo. Target-Oriented Single Domain Generalization. *ArXiv Prepr. ArXiv250900351*, 2025.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Int. Conf. Learn. Represent. ICLR*, 2017.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP. In *2024 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 23773–23782, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02244.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. In *International Conference on Learning Representations (ICLR)*, 2017.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-training is Sufficient for Robustness to Spurious Correlations. In *Int. Conf. Learn. Represent. ICLR*, 2023.

Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply Learned View-Invariant Features for Cross-View Action Recognition. *IEEE Trans. Image Process.*, 26(6):3028–3037, June 2017. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2017.2696786.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proc. 36th Int. Conf. Mach. Learn. ICML*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, 2:4, 2008. ISSN 16625137. doi: 10.3389/neuro.06.004.2008.

Guofa Li, Guanglei Wang, Zizheng Guo, Qing Liu, Xiyuan Luo, Bangwei Yuan, Mingrui Li, and Lu Yang. Domain Adaptive Driver Distraction Detection Based on Partial Feature Alignment and Confusion-Minimized Classification. *IEEE Trans. Intell. Transp. Syst.*, 25(9):11227–11240, September 2024. ISSN 1524-9050, 1558-0016. doi: 10.1109/TITS.2024.3367665.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain Few-shot Learning with Task-specific Adapters. In *2022 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 7151–7160, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00702.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Adv. Neural Inf. Process. Syst. NeurIPS*, 2020.

Yuying Liu, Shaoyi Du, Qinbo Guo, Zhiyue Zhao, Zhiqiang Tian, and Nanning Zheng. Structure Consistent Unsupervised Domain Adaptation for Driver Behavior Recognition. In *2023 IEEE 26th Int. Conf. Intell. Transp. Syst. ITSC*, pp. 1038–1043, Bilbao, Spain, September 2023. IEEE. ISBN 979-8-3503-9946-2. doi: 10.1109/ITSC57777.2023.10422284.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiss, Michael Voit, and Rainer Stiefelhagen. Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In *2019 IEEECVF Int. Conf. Comput. Vis. ICCV*, pp. 2801–2810, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00289.

Manuel Martin, David Lerch, and Michael Voit. Viewpoint Invariant 3D Driver Body Pose-Based Activity Recognition. In *2023 IEEE Intell. Veh. Symp. IV*, pp. 1–6, Anchorage, AK, USA, June 2023. IEEE. ISBN 979-8-3503-4691-6. doi: 10.1109/IV55152.2023.10186682.

Anna Montoya, Dan Holman, SF_data_science, Taylor Smith, and Wendy Kan. State Farm Distracted Driver Detection, 2016.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A New Measure to Evaluate Transferability of Learned Representations. In *International Conference on Machine Learning (ICML)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Rashindrie Perera and Saman Halgamuge. Discriminative Sample-Guided and Parameter-Efficient Feature Space Adaptation for Cross-Domain Few-Shot Learning. In *2024 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 23794–23804, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02246.

Thinesh Thiyakesan Ponbagavathi, Kunyu Peng, and Alina Roitberg. Probing Fine-Grained Action Understanding and Cross-View Generalization of Foundation Models, July 2024.

Thinesh Thiyakesan Ponbagavathi, Kunyu Peng, and Alina Roitberg. T-MASK: Temporal Masking for Probing Foundation Models across Camera Views in Driver Monitoring, August 2025.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to Learn Single Domain Generalization. In *2020 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 12553–12562, 2020. doi: 10.1109/CVPR42600.2020.01257.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *Adv. Neural Inf. Process. Syst. NeurIPS*, 2017.

Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *eLife*, 12:e82566, August 2023. ISSN 2050-084X. doi: 10.7554/eLife.82566.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE Int. Conf. Comput. Vis. ICCV*, pp. 618–626, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.74.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019.

Hongduan Tian, Feng Liu, Ka Chun Cheung, Zhen Fang, Simon See, Tongliang Liu, and Bo Han. Cross-domain Few-shot Classification via Invariant-content Feature Reconstruction. *Int. J. Comput. Vis.*, 134(2):54, February 2026. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-025-02601-5.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, August 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2015.12.012.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *IEEECVF CVPR Workshop*, 2020.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-View Action Modeling, Learning, and Recognition. In *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2649–2656, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.339.

Jing Wang, Wenjing Li, Fang Li, Jun Zhang, Zhongcheng Wu, Zhun Zhong, and Nicu Sebe. 100-Driver: A Large-Scale, Diverse Dataset for Distracted Driver Classification. *IEEE Trans. Intell. Transp. Syst.*, 24(7):7061–7072, 2023. doi: 10.1109/TITS.2023.3255923.

Renchunzi Xie, Hongxin Wei, Lei Feng, Yuzhou Cao, and Bo An. On the importance of feature separability in predicting out-of-distribution error. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhipeng Xu, De Cheng, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Adversarial Domain Prompt Tuning And Generation For Single Domain Generalization. In *2025 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 18584–18595, 2025. doi: 10.1109/CVPR52734.2025.01732.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 12133–12143. PMLR, 2021.

Shan Zhang, Yongxin He, Mingming Zhang, Huiwen Tian, and Lei Ma. FiSeR: Fine-Grained Source Representations for Cross-Domain AI Image Detection, August 2026.

Yuexi Zhang, Dan Luo, Balaji Sundareshan, Octavia Camps, and Mario Sznaier. Cross-view Action Recognition via Contrastive View-invariant Representation, May 2023.

Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, Chen Ding, Guosheng Lin, and Yanning Zhang. Towards Effective Foundation Model Adaptation for Extreme Cross-Domain Few-Shot Learning. In *2025 IEEECVF Int. Conf. Comput. Vis. ICCV*, pp. 4582–4593, Honolulu, HI, USA, October 2025. IEEE. ISBN 979-8-3315-8775-8. doi: 10.1109/ICCV51701.2025.00436.

A Dataset, View, and Label Specification

A.1 Observation conditions and camera views

The experiments use three observation conditions drawn from two independently collected driver-behavior datasets. State Farm Distracted Driver Detection (SF3D) provides the source condition S, recorded from a right-windshield viewpoint (Montoya et al., 2016). The 100-Driver dataset provides two target conditions: Cam4, denoted T_4, observes the driver from the side-right direction, while Cam2, denoted T_2, provides a frontal observation (Wang et al., 2023). The resulting condition graph is

$$S \leftrightarrow T_4, \qquad S \leftrightarrow T_2, \qquad T_2 \leftrightarrow T_4.$$

The S–T_4 comparison is cross-dataset under broadly comparable viewing directions; S–T_2 changes both dataset and viewing direction; and T_2–T_4 provides the within-100-Driver view contrast used to reduce collection-level differences.

SF3D and 100-Driver are not calibrated observations of the same physical scenes. Cross-dataset comparisons additionally change people, vehicle interiors, acquisition pipelines, and other collection-specific factors. The two 100-Driver conditions share the parent collection and normalized task, but they are not synchronized observations of matched action instances and do not provide a measured camera-pose transformation between views. Differences in visible body regions, objects, and occlusion can therefore accompany the camera change. We consequently treat S, T_2, and T_4 as observation conditions rather than assigning them a numerical viewpoint-shift magnitude.

A.2 Closed-set label correspondence

All central experiments use the same closed $K = 10$ class set. SF3D and 100-Driver retain independent native category indices, so class correspondence is defined explicitly rather than inferred from index values. Lowercase c_i denotes the SF3D category and uppercase C_j the native 100-Driver category. Table 3 gives the complete normalization used by the experiments.

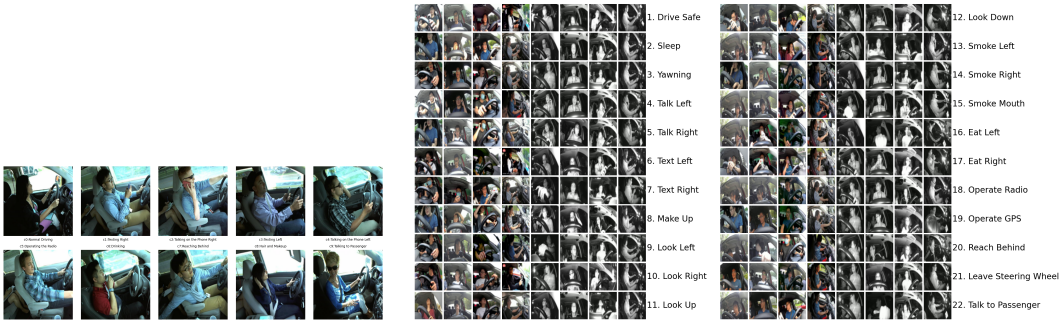
Twelve remaining 100-Driver categories are outside this common task and are excluded from the ten-class experiments. One retained correspondence requires particular care: SF3D class $c6$, *Drink*, is mapped to 100-Driver class $C16$, whose native label is `C16_Eat_Left`. This is a semantic harmonization assumption rather than an exact renaming. Analyses that test sensitivity to this correspondence remove true $c6$ examples only at evaluation time. The fitted classifier remains the original ten-class classifier, so these measurements are not nine-class retraining experiments.

A.3 Image treatments and example atlas

The primary CNN and DINOv2 experiments use detected driver crops, denoted `bbox`. Three additional image treatments are retained as supporting controls: the native full frame; `seg`, which retains detected-person foreground on the original canvas; and `bboxseg`, which applies foreground segmentation after cropping. Figure 7 illustrates these conditions across the three observation settings.

The supplied annotation implementation uses an automated YOLOv8s-seg pipeline, rather than manual behavior masks. Its defaults specify `yolov8s-seg.pt`, inference size 640, confidence threshold 0.25, IoU threshold 0.45, and the COCO person class (index 0). The crop branch selects the largest detected bounding box; the full-canvas segmentation branch renders detected foreground, and the crop-plus-segmentation branch performs a second prediction on the crop. These are source-code defaults, not a recovered invocation log or checkpoint hash for every historical image asset. The exact historical model build, command-line overrides, and failure-case provenance are not fully established by the reconstruction snapshot. Detector and mask imperfections therefore remain part of the treatment boundary.

A crop changes both field of view and image scale. Foreground masking changes the available pixels while preserving the canvas, and crop-plus-mask combines these operations. These transformations can remove behavior-relevant objects, hands, body regions, or pose context as well as background. Their effects are therefore interpreted as input-treatment sensitivity rather than isolated causal measurements of background dependence.



(a) SF3D native ten-class task (b) 100-Driver native 22-class, multi-camera task

Figure 6: Native dataset task spaces before imposing the common ten-class correspondence. The experiments use selected daytime camera conditions and the explicit class normalization in Table 3.

Canonical behavior	Dataset indices	Native 100-Driver label
Normal Driving	`c0 / C1`	`C1_Drive_Safe`
Texting Right	`c1 / C7`	`C7_Text_Right`
Phone Right	`c2 / C5`	`C5_Talk_Right`
Texting Left	`c3 / C6`	`C6_Text_Left`
Phone Left	`c4 / C4`	`C4_Talk_Left`
Operate Radio	`c5 / C18`	`C18_Operate_Radio`
Drink	`c6 / C16`	`C16_Eat_Left`
Reach Behind	`c7 / C20`	`C20_Reach_Behind`
Make Up	`c8 / C8`	`C8_Make_Up`
Talk Passenger	`c9 / C22`	`C22_Talk_to_Passenger`

Table 3: Canonical ten-class behavior correspondence. Lowercase indices denote SF3D categories and uppercase indices denote the independent 100-Driver categories. Native 100-Driver labels are retained to make the normalization explicit.

A.4 Dataset populations and usable pools

The original SF3D split contains 16,722 training images, 2,868 validation images, and 2,834 test images. The primary cropped-feature analysis contains 2,833 valid source test examples. For the corresponding target crop analysis, T_2 contains 5,239 images from 11 drivers and T_4 contains 5,489 images from 12 drivers. Table 4 records these population quantities separately from the support/query partitions constructed later by each classifier-fitting protocol.

The target counts above are properties of the primary cropped CNN/DINOv2 analysis and should not be transferred to DBMNet. DBMNet operates on full-frame inputs using its own camera-specific feature extraction and checkpoint regime. Likewise, the target pool sizes do not by themselves define an evaluation population: image-disjoint, driver-disjoint, and shared-view protocols construct different support/query relationships from these pools.

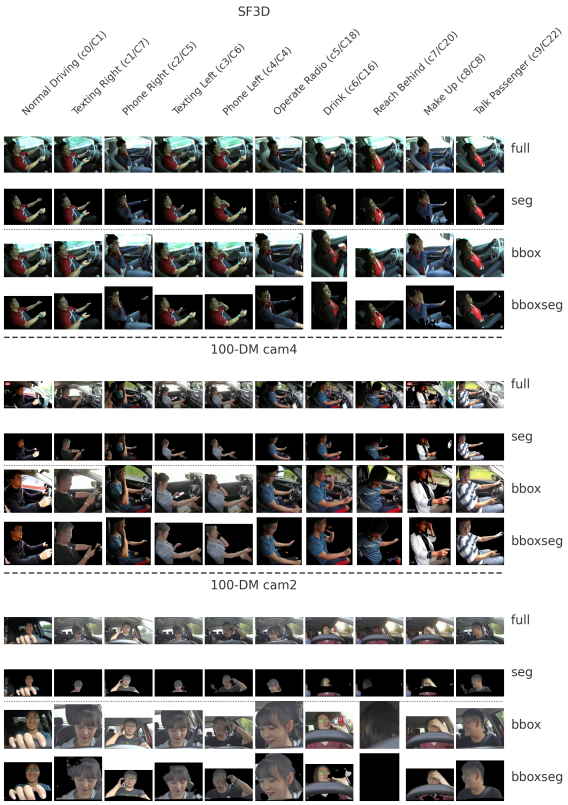


Figure 7: Behavior and preprocessing atlas across S, T_4, and T_2. Examples are separate instances of the normalized behavior classes rather than matched participants or synchronized actions.

Condition	Population	Images	Drivers
SF3D (S)	Original training split	16,722	–
SF3D (S)	Original validation split	2,868	–
SF3D (S)	Original test split	2,834	–
SF3D (S)	Valid cropped test features	2,833	–
100-Driver Cam2 (T_2)	Frontal cropped pool	5,239	11
100-Driver Cam4 (T_4)	Side-right cropped pool	5,489	12

Table 4: Dataset populations used by the primary CNN/DINOv2 analysis. Target counts refer to the cropped pools from which the finite-support protocols construct their support and query sets.

B Model, Training, and Feature Specification

B.1 Task-trained convolutional representations

The five convolutional feature extractors are ResNet18, ResNet50 (He et al., 2016), DenseNet121 (Huang et al., 2017), EfficientNet-B0 (Tan & Le, 2019), and SqueezeNet1.0 (Iandola et al., 2017). Each model is initialized from ImageNet weights and trained on the normalized ten-class SF3D source task. The recorded source-training configuration uses 224×224 inputs and multiclass cross-entropy with SGD at learning rate 0.01, momentum 0.9, and weight decay 5×10^{-4}. A two-epoch warmup precedes a multi-step schedule that multiplies the learning rate by 0.2 at epochs 40 and 80. The campaign configuration specifies batch size 32 and 100 training epochs.

Separate source checkpoints are retained for the full-frame, `bbox`, `seg`, and `bboxseg` treatments. Consequently, a CNN recovery comparison across input treatments changes the matched source-training treatment together with the target-input treatment; it is not a test in which one CNN checkpoint is exposed to four target-only perturbations. The registered checkpoint files terminate in

`final.pth`; the available provenance does not establish that these files are validation-best checkpoints.

Table 5 records the primary crop-trained checkpoint identities and the feature hooks used for the frozen-representation experiments. Spatial tensors extracted at those hooks are globally pooled before classifier fitting.

The source-trained CNN treatments use treatment-specific RGB statistics rather than one common ImageNet normalization. Table 6 records the normalization used by each source treatment.

B.2 Frozen DINOv2 representations

The self-supervised comparison uses the public DINOv2 ViT-S/14 backbone `dinov2_vits14` and its register-token variant `dinov2_vits14_reg`, pretrained on LVD-142M (Oquab et al., 2024; Darcet et al., 2024). Neither transformer backbone is fine-tuned during the classifier interventions. The global representation used by the recovery and compatibility experiments is the recorded 384-dimensional CLS feature.

The register variant introduces four learned register tokens into the transformer computation but retains the same 384-dimensional CLS interface used by the task classifier. Register tokens therefore define a distinct pretrained representation rather than a different behavior-label space. Both DINOv2 pipelines use the configured 224×224 transform with ImageNet RGB normalization.

The registered checkpoint SHA256 prefixes are `b938bf1bc15c` for the base ViT-S/14 model and `f433177089a6` for the register variant. A separate spatial diagnostic path excludes CLS and register tokens and aggregates patch features. That path is used only for spatial analyses and must not be conflated with the global CLS representation used for classifier fitting.

B.3 DBMNet multi-camera comparator

DBMNet provides a deliberately different supervision regime (Celona et al., 2026). Rather than receiving task supervision only from S, its leave-one-camera-out models are trained using three labelled 100-Driver cameras and evaluated on the fourth. For the daytime models used here, the `cam123` checkpoint holds out Cam4, while the `cam134` checkpoint holds out Cam2. This makes DBMNet an informative cross-camera comparator but not an information-matched competitor to the single-source CNN and frozen DINOv2 experiments.

The common ten-class analysis maps the relevant native 22-class DBMNet categories onto the task defined in Appendix A. Native 22-class performance and the remapped ten-class results are therefore different evaluation tasks and are not numerically interchangeable.

Two representations are exported from each DBMNet checkpoint. The `feat` channel is the raw GhostNet backbone representation, while `x_action` is the action-conditioned representation produced by DBMNet's view-conditioned computation. Both originate from the same held-out-camera network for a given target view, so duplicated native direct-classification values across those two feature rows are intentional.

Crucially, frontal and side-right DBMNet features are extracted from different leave-one-camera-out checkpoints. They therefore do not define one common coordinate system across T_2 and T_4. DBMNet can participate in within-condition recovery experiments, but its frontal and side-right representations are excluded from shared-view classifier fitting and other tests that require one common frozen feature space.

DBMNet uses full frames with ImageNet normalization in its native evaluation pipeline. Supporting crop or mask perturbations reuse the corresponding native DBMNet checkpoint rather than retraining a treatment-specific DBMNet. External repository commit and checkpoint-hash identifiers are not populated in the supplied provenance and remain an explicit reproducibility boundary.

B.4 Feature spaces and comparison units

Table 8 summarizes the representations entering the diagnostic experiments. For a given CNN architecture and matched source treatment, the same frozen source checkpoint is used to encode

Crop-trained CNN	Feature hook	Dimension	SHA256 prefix
ResNet18	`layer4`	512	0471917d229e
ResNet50	`layer4`	2048	ba89a90d90fc
DenseNet121	`features.norm5`	1024	89dfe1a792c7
EfficientNet-B0	`avgpool`	1280	a2605d2eabb9
SqueezeNet1.0	`features.12`	512	ec7a0a676b5f

Table 5: Primary crop-trained CNN feature provenance. SHA256 prefixes identify the registered source checkpoints. Spatial feature tensors are globally pooled before classifier fitting.

Source treatment	RGB mean	RGB std.
SF3D	[0.3097, 0.3799, 0.3736]	[0.2789, 0.3249, 0.3269]
SF3D bbox	[0.3184, 0.3886, 0.3879]	[0.2763, 0.316, 0.3181]
SF3D seg	[0.0925, 0.1154, 0.1163]	[0.1716, 0.2097, 0.2151]
SF3D bboxseg	[0.1297, 0.162, 0.1636]	[0.1921, 0.2341, 0.2407]

Table 6: Recorded RGB normalization for the CNN source treatments. These statistics are distinct from the ImageNet normalization used by the frozen DINOv2 and native DBMNet pipelines.

Train cameras	Held out	Author-reported	Local evaluation	Query n
2,3,4	C1	32.37	35.91	11,196
1,3,4	C2	62.51	58.34	10,890
1,2,4	C3	50.99	50.07	11,373
1,2,3	C4	11.29	10.93	11,689

Table 7: Native daytime DBMNet leave-one-camera-out provenance. Author-reported 22-class performance and locally recorded evaluation are shown separately rather than merged.

S, T_2, and T_4. Each DINOv2 variant likewise defines one frozen feature coordinate system across observation conditions. These seven feature spaces can therefore support transported and shared classifiers without introducing an additional representation-alignment operation.

DBMNet differs because the held-out Cam2 and Cam4 representations originate from different checkpoints. Its two feature channels increase the local-recovery comparison to nine feature configurations, but they are excluded from experiments whose interpretation requires a common coordinate system across target views.

The feature configurations are not independent experimental replications. The five CNNs share ImageNet initialization, the two DINOv2 variants share a pretrained model family, and the two DBMNet channels are outputs of the same camera-specific networks. Model-wise aggregates are therefore descriptive comparisons across representations rather than samples from an independent population of architectures.

B.5 Geometry-export feature identity

The extended geometry export names five CNN encoder registry keys and the two DINOv2 variants. The CNN keys resolve to full-frame SF3D model entries, whereas Table 5 identifies the crop-trained recovery models. However, the geometry loader locates cached tensors using only the architecture qualifier (the first component of the encoder key); it does not bind those tensors to the named registry checkpoint. The CNN feature writers likewise record an architecture-level encoder identity rather than a checkpoint hash. The registry key alone therefore establishes neither that full-frame weights produced the cached features nor that those features match the crop-trained recovery checkpoints.

We retain the CNN rows as architecture-indexed structural comparisons, with their exact numerical export, rather than present them as verified checkpoint-matched recovery–geometry pairs. The joint recovery–geometry interpretation relies on the fixed DINOv2 base/register representations, whose distinct public encoder identities are preserved across these experiments. Checkpoint-qualified

Representation	Feature interface	Dim.	Shared across T_2/T_4?
ResNet18	`layer4`	512	Yes, within matched source treatment
ResNet50	`layer4`	2048	Yes, within matched source treatment
DenseNet121	`features.norm5`	1024	Yes, within matched source treatment
EfficientNet-B0	`avgpool`	1280	Yes, within matched source treatment
SqueezeNet1.0	`features.12`	512	Yes, within matched source treatment
DINOv2 ViT-S/14	CLS token	384	Yes
DINOv2 + registers	CLS token	384	Yes
DBMNet raw	GhostNet `feat`	1280	No; camera-specific checkpoint
DBMNet action	`x_action`	1280	No; camera-specific checkpoint

Table 8: Feature-space identities used by the diagnostic experiments. "Shared across views" means that T_2 and T_4 are encoded by the same frozen checkpoint and can therefore be evaluated with one transported or jointly fitted linear classifier without an additional alignment map.

cache paths and feature manifests containing the generating weight hash are required to close the CNN provenance gap; the result archive does not supply that additional verification.

B.6 Provenance boundaries

The registered CNN checkpoints and DINOv2 weights provide checkpoint-level identities for the representations used by the primary experiments. DBMNet retains its documented leave-one-camera-out training configuration and local evaluation records, but the external repository commit and original checkpoint hashes are not available in the supplied evidence. We preserve this distinction rather than reconstructing missing provenance.

Model specification is also kept separate from classifier-intervention outcomes. Source native/refitted accuracy, target direct transfer, and target classifier-recovery results are empirical measurements and are reported with the recovery evidence rather than treated as properties of the feature extractor itself.

C Calibration and Evaluation Protocols

C.1 Information-access contracts

The experiments differ not only in observation condition but also in which labelled examples are available when a classifier is fitted. Table 9 makes these information-access contracts explicit. Direct transfer uses no labelled target examples. Finite-support classifier fitting uses exactly k labelled examples per class under the specified population protocol. Historical target-reference fits use different, larger fitting partitions and are therefore retained as protocol-specific reference measurements rather than endpoints of the finite-k recovery curves.

All classifier interventions keep the feature extractor fixed. A fitted classifier-side pipeline consists of the support-derived feature standardizer together with the multiclass linear classifier. Consequently, transporting a fitted classifier between conditions means transporting both components without recomputing either from the receiving condition.

C.2 Classifier fitting and support-dependent standardization

Let $z = f_\theta(x) \in \mathbb{R}^d$ denote a feature vector produced by a frozen extractor. For a support set, each feature coordinate is standardized using the support mean $\mu \in \mathbb{R}^d$ and nonzero coordinate scales collected in the diagonal matrix D:

$$T(z) = D^{-1}(z - \mu).$$

Protocol	Information used to fit	Evaluation population
Direct transfer	Previously fitted model; no target adaptation	Images from a target camera
CNN linear reference	Labelled target training partition	Fixed target query partition
DINOv2 linear reference	Class-stratified 70% of the target pool	Remaining 30%; drivers may recur
Image-disjoint calibration	Exactly k labelled images per class	Remaining images; drivers may recur
Driver-disjoint calibration	Exactly k per class from selected drivers	Only drivers excluded from fitting
Shared-view calibration	k per class in total across both views	Each target view evaluated separately
Return to source	Reuse saved classifier and standardizer; no new fit	Common SF3D feature pool

Table 9: Information available to the principal classifier-fitting and evaluation protocols. Historical target references use different partitions from the finite-support experiments and are not interpreted as population-matched or optimization-matched ceilings.

A multiclass linear classifier with weights $W \in \mathbb{R}^{K\times d}$ and bias $b \in \mathbb{R}^K$ is then fitted on $T(z)$. Its scores in the original frozen feature coordinates are

$$WT(z) + b = WD^{-1}z + \left(b - WD^{-1}\mu\right).$$

Thus the fitted pipeline remains affine in the original representation, but changing the support can change both its effective weights and its effective bias. This is why the fitted standardizer must accompany the classifier in shared-view and return-to-source tests.

The recorded fitting implementation minimizes multiclass cross-entropy with L_2 weight decay using SGD for 500 epochs, learning rate 0.01, batch size 512, and momentum 0.9. Inverse-regularization candidates are

$$\{0.001,\ 0.01,\ 0.1,\ 1,\ 10\},$$

implemented as reciprocal weight decay on both weights and bias. The selected value maximizes stratified support cross-validation accuracy. Five folds are used when feasible and two folds at $k = 2$.

Feature standardization is estimated from the complete external support before this internal cross-validation. Validation-fold feature statistics therefore enter the preprocessing stage, although no external query example enters standardization, classifier fitting, or regularization selection. The recorded finite-epoch procedure is treated as the implemented classifier intervention rather than as a certificate of the optimal attainable linear classifier.

For the finite-support experiments,

$$k \in \{2, 5, 10, 20, 50, 100\},$$

with exactly $Kk = 10k$ support images in a single-view fit. Five independent support draws are recorded for each model–view–budget combination. Reported variation across these draws is their standard deviation; it is not a confidence interval over independent drivers.

C.3 Image-disjoint fitting

Under the image-disjoint protocol, exactly k labelled examples are sampled for each of the $K = 10$ classes from one target condition. The external query consists of target images not selected into the support. Support and query are therefore disjoint at the image level, but the same driver may contribute images to both sets.

For target condition A, this protocol measures local recoverability using a fitted pipeline $h_A^{(k)}$ on unseen images from the same observation condition. Because driver identity is not excluded, this protocol answers whether the frozen representation supports classification of held-out images under the sampled target population; it does not by itself establish recovery on drivers absent from fitting.

C.4 DRIVER-DISJOINT EXACT-k FITTING

The driver-disjoint protocol preserves the same exact k-per-class support budget while separating support and query by driver identity. Driver groups are selected until their combined examples can supply every class at the requested budget. Exactly k images per class are then sampled from those selected drivers, and all remaining images belonging to those drivers are removed from evaluation. The query therefore contains only drivers absent from the support.

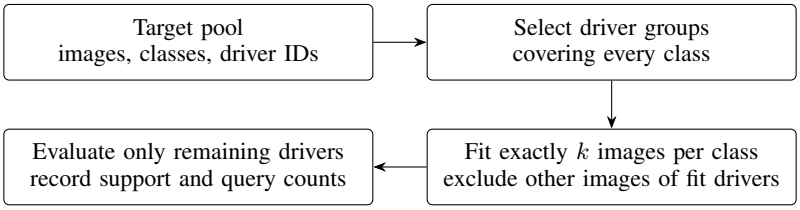


Figure 8: Whole-group exclusion enforces driver separation without relaxing the per-class budget. The selected groups can change the query population.

This is a coverage-constrained group-selection problem rather than an image-level random split. Increasing k can require additional driver groups to satisfy one or more classes, which in turn changes which drivers and how many images remain available for evaluation. The resulting driver-disjoint curves therefore combine changes in labelled-image budget, support-driver composition, and the induced query population. They are not interpreted as pure sample-complexity curves or as isolated causal estimates of identity leakage.

The protocol nevertheless preserves the nominal class budget exactly: every finite-support cell contains $10k$ support images. Five recorded draws use independent group/support selections. Query-image counts and the number of fitting drivers are retained with the per-draw records so that population changes can be inspected rather than hidden inside the accuracy curves.

C.5 SHARED-VIEW FITTING

The shared-view protocol asks whether one classifier-side pipeline can be fitted from labelled examples drawn from both target camera conditions and then serve each condition separately. For a nominal budget k, the support contains

$$Kk = 10k$$

images in total across $T_2 \cup T_4$, not $10k$ images from each camera. The fitted standardizer and classifier are shared; evaluation accuracy is then reported separately on the frontal and side-right target queries.

The recorded support sampler does not impose exact per-class camera balance. Camera allocation is inherited from the sampled joint support, so the completed shared-view experiment tests joint-support fitting rather than a controlled 50/50 camera-allocation intervention. Exact camera-balanced support is therefore a distinct control rather than an assumption of the existing protocol.

C.6 RETURN-TO-SOURCE EVALUATION

Return-to-source evaluation does not fit another classifier. A target-fitted or shared-view pipeline is saved together with its support-derived standardizer and then applied directly to the common S feature pool:

$$h_A^{(k)} \longrightarrow S, \qquad h_{T_2 \cup T_4}^{(k)} \longrightarrow S.$$

No source examples are used to update the standardizer, regularization, classifier weights, or bias during this return.

These scores measure the absolute compatibility of a target-fitted decision rule with the original source representation. They are not backbone-forgetting measurements because f_θ is unchanged. They are also kept separate from source-refit change: the available historical source references and target-fitted pipelines are not necessarily protocol matched in a way that would justify subtracting one from the other as a common source-regret quantity.

C.7 SOURCE-SIDE REFITTING CONTROL

For the task-trained CNNs, source-side refitting provides a matched-query control on the classifier intervention itself. A new classifier is fitted from the source training partition with the feature extractor frozen and is evaluated on the same held-out source query as the native source classifier. The signed difference

$$\Delta_S = 100\,[\mathrm{Acc}(h_{S,\mathrm{refit}}; Q_S) - \mathrm{Acc}(h_{S,\mathrm{native}}; Q_S)]$$

is therefore a same-query classifier-refit change in percentage points. Negative Δ_S denotes lower source accuracy after refitting.

This quantity answers a different question from target-to-source return. Source refitting compares two source-fitted decision rules on the same source query; return-to-source evaluates a rule fitted elsewhere on the source representation. The two interventions are consequently reported separately.

C.8 HISTORICAL TARGET-REFERENCE PROTOCOLS

The target-reference values used in the direct-versus-refitted comparison were produced under protocol families that precede the common finite-support study. For the task-trained CNNs, the target classifier is fitted on a labelled target training partition and evaluated on a fixed target query partition. For DINOv2, the target pool is split class-stratified into 70% fitting and 30% evaluation images, and drivers may recur across those partitions.

These references demonstrate that the corresponding frozen feature spaces can support substantially stronger target discrimination under their recorded fitting conditions. They are not interpreted as optimal linear ceilings, as $k = 100$ endpoints, or as matched comparisons of feature quality across CNN and DINOv2 protocol families. Finite-k recovery curves are therefore interpreted within their own support/query contracts rather than normalized to these historical references as if all fitting conditions were interchangeable.

D DIAGNOSTIC MEASUREMENT PROCEDURES

D.1 LOCAL LINEAR RECOVERABILITY

Let A denote an observation condition and let

$$z = f_\theta(x) \in \mathbb{R}^d$$

be the frozen feature representation of image x. For support budget k and population protocol π, let

$$\mathcal{S}_{A,\pi}^{(k)}$$

denote the labelled support and

$$\mathcal{Q}_{A,\pi}$$

the corresponding external query. The fitted classifier-side pipeline,

$$h_{A,\pi}^{(k)},$$

includes both the support-derived feature standardizer and the multiclass linear classifier defined in Appendix C.

Local linear recoverability is

$$L_A(k;\pi) = \mathrm{Acc}\Big(h_{A,\pi}^{(k)} \circ f_\theta; \mathcal{Q}_{A,\pi}\Big)\,.$$

Accuracy is

$$\mathrm{Acc}(h;\mathcal{Q}) = \frac{1}{|\mathcal{Q}|}\sum_{(x,y)\in\mathcal{Q}} \mathbf{1}[h(x)=y]\,.$$

The quantity therefore measures whether the frozen representation supports a linear decision rule under the specified support and query contract. It is not an optimal representation ceiling and does not imply that the fitted rule is useful under another observation condition.

For the image-disjoint protocol, $\pi = \mathrm{I}$, support and query contain different images but may contain the same drivers. For the driver-disjoint protocol, $\pi = \mathrm{D}$, support and query drivers are disjoint. Because exact-k group selection can change the remaining query population, differences

$$L_A(k; \mathrm{I}) - L_A(k; \mathrm{D})$$

are population-sensitivity measurements rather than isolated causal estimates of identity leakage.

For each finite-support cell, the reported mean is

$$\overline{L}_A(k; \pi) = \frac{1}{R} \sum_{r=1}^{R} L_A^{(r)}(k; \pi), \qquad R = 5,$$

and the accompanying dispersion is the standard deviation over the five support draws. These draws are not independent-driver replications.

D.2 Class-relational organization

For each observation condition A, we summarize relationships among the $K = 10$ behavior classes with a representational dissimilarity matrix

$$R_A \in \mathbb{R}^{K \times K}.$$

Its off-diagonal entries contain the estimated pairwise class dissimilarities. The unique class-pair entries are collected as

$$\mathbf{r}_A = \mathrm{vec}_{\triangle}(R_A) \in \mathbb{R}^P, \qquad P = \frac{K(K-1)}{2} = 45.$$

The recorded implementation constructs each ten-class RDM from six within-class image partitions using the crossvalidated-distance path of `rsatoolbox.calc_rdm_unbalanced`. Crossvalidation reduces the positive bias that can arise in ordinary distance estimates (Walther et al., 2016; Diedrichsen et al., 2021). No explicit empirical feature-space noise-precision matrix is supplied to this distance routine. The resulting distances should therefore not be described as if an independently estimated feature-noise precision matrix had been provided.

Two RDMs are compared using the covariance-aware `corr_cov` procedure (Schütt et al., 2023). We write the resulting class-relational correspondence as

$$G(A, B; \Sigma_A) = \mathrm{corr}_{\mathrm{cov}}(\mathbf{r}_A, \mathbf{r}_B; \Sigma_A)\,,$$

where A is the explicitly designated reference side. Here Σ_A denotes the covariance/noise estimate used by the RDM comparator and constructed from the reference-condition measurements with the corresponding class and partition descriptors. It is not the feature-space precision matrix omitted from the preceding crossnobis estimation.

The reference direction is retained because the covariance-aware comparison need not be treated as if its estimation procedure were direction-free. In the recorded source–target measurements, S is the reference condition. For within-condition split-half reliability, the first image half is the reference. Where both covariance-reference directions are evaluated, they are reported as sensitivity measurements rather than counted as independent replications.

For the direct target-view comparison, T_2 supplies the reference measurements for $G(T_2, T_4; \Sigma_{T_2})$, and T_4 supplies them for $G(T_4, T_2; \Sigma_{T_4})$. This reference reversal changes the covariance-aware weighting, not the direction in which a classifier is fitted and evaluated. The archive records six image partitions and the corrected/whitened paths for each measurement; it does not record a run-specific random seed. The supplied implementation defaults to seed zero, which is an implementation setting rather than independent provenance of every archived run.

All compared RDMs use the same ordered ten-class label set. Consequently, the evaluation-only removal of the imperfect food/drink class from selected classifier analyses does not retroactively alter the ten-class geometry measurements.

D.3 RELIABILITY REFERENCES FOR CLASS-RELATIONAL GEOMETRY

A cross-condition value $G(A, B; \Sigma_A)$ is interpreted relative to the reproducibility of the RDM construction within an observation condition. The original reliability analysis divides the images of a condition into two halves, rebuilds the RDM independently on each half, and compares them with the same covariance-aware procedure. High correspondence between these halves establishes image-split reliability of the class-relational estimate.

Image-half reliability is not a driver-independent estimate because the same driver may appear in both halves. The recorded stricter control therefore partitions observations by subject before rebuilding the RDMs. For source partition p and target partition q, the recorded within-minus-cross contrast is

$$\Delta_{pq} = \frac{1}{2}\left(\rho^{p}_{SS} + \rho^{q}_{TT}\right) - \rho^{pq}_{ST},$$

where ρ^{p}_{SS} and ρ^{q}_{TT} are within-condition reliability measurements and ρ^{pq}_{ST} is the corresponding cross-condition class-relational comparison. In the recorded implementation, the cross-condition term averages the available source-half/target-half comparisons for that partition pairing.

For the original image-random construction, the analogous contrast is

$$\Delta_{\text{image}} = \frac{1}{2}\left(\rho_{SS} + \rho_{TT}\right) - \rho_{ST}.$$

Neither Δ_{pq} nor Δ_{image} replaces the underlying cross-condition correspondence ρ_{ST}. The contrast only places that correspondence relative to a within-condition reliability reference.

The recorded subject-partition control is available only for the primary DINOv2 crop/frontal comparison. Its many partition pairings reuse a small number of valid source partitions and must therefore be interpreted as dependent sensitivity measurements rather than independent samples from a population. Complete values and distributions are reported with the geometry results rather than here in the measurement definition.

D.4 CROSS-VIEW DECISION COMPATIBILITY

Local recoverability asks whether a condition supports its own fitted rule. Cross-view decision compatibility instead asks whether a rule fitted under one condition remains useful under another while the frozen representation is held fixed.

For conditions A and B, define

$$C_{A\to B}(k;\pi) = \mathrm{Acc}\Big(h^{(k)}_{A,\pi} \circ f_\theta;\, \mathcal{Q}_B\Big).$$

The support-derived standardizer belonging to $h^{(k)}_{A,\pi}$ is transported together with the classifier. No statistics are recomputed from B. This is essential because changing the standardizer would constitute an additional fitting intervention at the receiving condition.

When the fitting and evaluation conditions coincide under the same population contract,

$$C_{A\to A}(k;\pi) = L_A(k;\pi).$$

The receiver-matched comparison

$$C_{B\to B}(k;\pi) - C_{A\to B}(k;\pi)$$

therefore measures how much accuracy is lost when the receiving query B is classified by a rule fitted under A rather than by its own locally fitted rule. Because both terms use the same receiving query, this comparison does not conflate classifier compatibility with differences in query difficulty.

A related test fits one pipeline using support from both target views. Let

$$h^{(k)}_{T_2\cup T_4}$$

denote the shared-view classifier-side pipeline fitted from k total support examples per class across $T_2 \cup T_4$. Its per-view compatibility measurements are

$$C_{T_2\cup T_4\to T_2}(k) \quad \text{and} \quad C_{T_2\cup T_4\to T_4}(k).$$

These quantities ask whether one jointly fitted affine rule can serve both target observation conditions. They are distinct from reciprocal transport of two independently fitted view-specific rules.

Return-to-source evaluation is another compatibility test:

$$C_{A\to S}(k) = \mathrm{Acc}\Big(h_A^{(k)} \circ f_\theta; \mathcal{Q}_S\Big).$$

It measures the absolute compatibility of a target-fitted rule with the source condition. Unless the source comparator is fitted under a matched protocol on the same query, this quantity is not subtracted from a historical source accuracy and called source regret.

D.5 Why the three diagnostics are non-equivalent

The three diagnostic properties answer different questions:

$$L \quad : \text{is class information locally linearly accessible?}$$
$$G \quad : \text{are class relationships organized similarly?}$$
$$C \quad : \text{does a fitted decision rule remain compatible elsewhere?}$$

Local recovery, relational correspondence, and off-diagonal decision compatibility are not interchangeable guarantees; the diagonal identity between compatibility and local recovery remains as defined above.

Preserved class distances do not preserve a fitted rule. Consider two one-dimensional classes centered at -1 and $+1$, separated by a decision threshold at zero. Translating both classes by $+3$ preserves their pairwise distance exactly but moves their centers to 2 and 4. A newly placed threshold at 3 separates the classes, while the original threshold does not. Preserved relational geometry therefore does not guarantee portability of a fixed classifier.

Camera-separable features need not impair classification. Let

$$z = (y, v),$$

where

$$y \in \{-1, +1\}$$

encodes class and

$$v \in \{-10, +10\}$$

encodes camera condition. The two cameras are strongly separated along the second coordinate, yet

$$h(z) = \mathbf{1}[z_1 > 0]$$

classifies both perfectly. A representation can therefore contain strongly camera-informative variation without that variation interfering with the discriminative direction.

Separate perfect fits need not admit one shared classifier. Suppose one condition maps the positive class to $+1$ and the negative class to -1, while another condition reverses those assignments. Each condition separately admits a perfect linear classifier. After pooling the two conditions, however, the feature value $+1$ is associated with contradictory labels. Strong local recoverability in both conditions therefore does not establish the existence of one compatible shared decision rule.

Classifier margins depend on displacement direction. For class c against competing class j, let

$$m_{cj}(z) = (w_c - w_j)^\top z + b_c - b_j.$$

A representation displacement Δz changes that margin by

$$m_{cj}(z + \Delta z) - m_{cj}(z) = (w_c - w_j)^\top \Delta z.$$

A large view-related displacement can therefore leave classification unchanged when it lies outside the relevant discriminative direction, while a smaller displacement aligned with that direction can cross a decision boundary. Feature displacement magnitude alone does not localize cross-view failure.

Accuracy differences are interventions, not additive causes. Differences among direct-transfer accuracy, locally refitted accuracy, shared-view accuracy, and a fully adapted model would describe different interventions only when their populations and fitting conditions are matched. They cannot in general be added or subtracted as causal shares of "classifier error" and "representation error." In particular, successful classifier-only recovery demonstrates accessible class information; it does not establish that classifier-only adaptation is sufficient for the best attainable target performance.

These constructions are analytical boundary cases rather than simulated evidence about the driver datasets. Their purpose is to show why local recoverability, class-relational organization, and decision compatibility require separate empirical measurements.

E Classifier Recovery Results

E.1 Direct-transfer and target-refit reference measurements

Table 10 records the reference measurements that motivate the finite-support recovery analysis. For the five task-trained CNNs, direct transfer to the frontal condition is substantially weaker than transfer to the side-right condition, while refitting only the classifier-side pipeline recovers substantial target discrimination in both conditions. The DINOv2 representations exhibit the same qualitative separation between direct transfer and target-supported refitting, with particularly high target-reference accuracy after refitting.

These reference measurements and the common-budget experiments below should not be joined into one learning curve. The CNN target references use labelled target training partitions with fixed queries, whereas the DINOv2 references use class-stratified target splits. The finite-support study instead enforces the common grid

$$k \in \{2, 5, 10, 20, 50, 100\}$$

with five support draws at every budget.

The source-side CNN results also show that fitting a new classifier is not automatically beneficial. Under the recorded crop protocol, all five CNNs lose source-query accuracy after source-side refitting, with Δ_S ranging from -4.0 to -27.2 percentage points. The recovery intervention is therefore interpreted as a measured fitting procedure rather than an oracle replacement of the native classifier.

E.2 Common-budget recovery under image- and driver-disjoint evaluation

The common-budget study evaluates local linear recoverability under both population protocols while holding the nominal labelled-image budget fixed. We use

$$\mathrm{F} \equiv T_2, \qquad \mathrm{S} \equiv T_4,$$

for frontal and side-right views, and

$$\mathrm{I} \equiv \text{image-disjoint}, \qquad \mathrm{D} \equiv \text{driver-disjoint}.$$

Every table entry is mean accuracy over five support draws with the corresponding support-draw standard deviation. The three groups below are a layout partition only; conclusions should be formed by comparing population protocols within the same representation and view before comparing architectures.

Task-trained residual and dense networks. Figure 9 and Table 11 contain the complete curves for ResNet18, ResNet50, and DenseNet121.

	Source: SF3D			Target: frontal		Target: side-right	
Features	Native	Refit	Δ_S	Direct	Refit	Direct	Refit
ResNet18	82.3	78.3	-4.0	10.3	47.1	50.0	73.0
ResNet50	77.3	64.7	-12.5	6.9	59.4	57.8	78.3
DenseNet121	83.1	69.9	-13.2	8.8	52.6	55.6	81.5
EffNet-B0	77.4	65.4	-12.0	11.6	54.3	48.1	74.8
SqueezeNet	65.2	38.0	-27.2	12.1	48.8	50.2	85.8
DINOv2	–	96.2	–	11.8	94.1	40.3	98.7
DINOv2+reg	–	96.2	–	16.1	94.1	52.0	98.2
DBMNet raw	–	–	–	67.2	83.8	19.9	79.7
DBMNet action	–	–	–	67.2	84.3	19.9	76.1

Table 10: Reference accuracy (%) for the source, frontal target, and side-right target conditions. The CNN and DINOv2 target-refit values originate from different historical fitting protocols and are therefore protocol-specific reference measurements rather than matched linear ceilings. DBMNet uses full frames and camera-specific held-out-view checkpoints. Δ_S is the same-query source-refit change in percentage points.

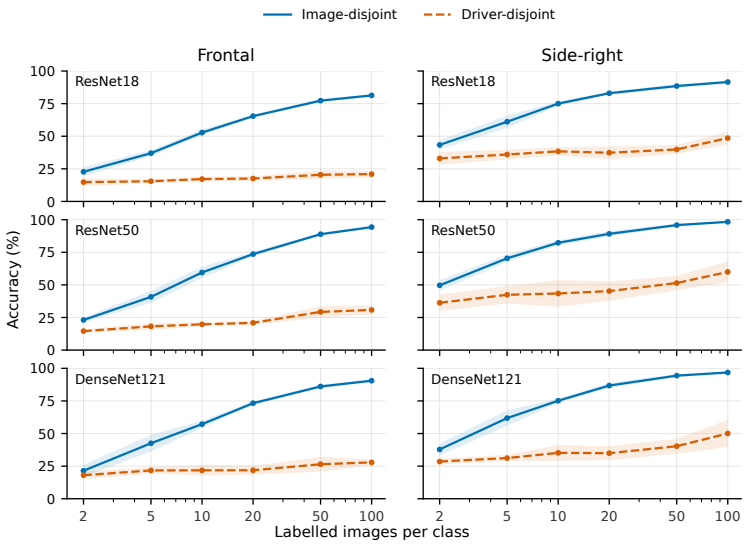


Figure 9: Both target viewpoints and population protocols under identical nominal class budgets. Bands show variation across five support draws.

Features	View/protocol	$k = 2$	$k = 5$	$k = 10$	$k = 20$	$k = 50$	$k = 100$
ResNet18	F/I	22.7±3.6	37.0±2.3	52.8±2.4	65.4±1.1	77.3±1.3	81.3±0.7
ResNet18	F/D	14.8±3.0	15.5±1.4	17.1±1.9	17.6±1.8	20.4±3.2	21.0±2.6
ResNet18	S/I	43.3±3.6	61.1±4.7	75.0±1.7	83.0±1.1	88.5±1.5	91.6±0.6
ResNet18	S/D	33.0±4.9	36.0±3.8	38.4±3.2	37.3±4.8	39.9±3.4	48.6±5.0
ResNet50	F/I	23.1±1.9	40.8±4.6	59.5±4.3	73.6±1.8	88.9±1.1	94.3±0.5
ResNet50	F/D	14.5±1.3	18.2±2.7	19.8±1.5	20.9±1.6	29.3±4.3	30.8±3.5
ResNet50	S/I	49.7±4.1	70.4±2.8	82.3±2.0	89.2±2.0	95.8±1.1	98.3±0.3
ResNet50	S/D	36.3±6.3	42.4±6.7	43.4±10.2	45.2±7.5	51.4±5.5	59.9±7.9
DenseNet121	F/I	21.5±4.6	42.5±6.3	57.2±2.9	73.3±1.4	86.1±1.3	90.5±1.1
DenseNet121	F/D	18.0±2.9	21.7±2.5	21.8±3.9	21.8±3.5	26.5±5.6	27.9±2.3
DenseNet121	S/I	37.8±4.6	61.8±5.9	75.2±1.6	86.8±1.9	94.4±1.2	96.8±0.4
DenseNet121	S/D	28.5±2.2	31.2±2.5	35.2±6.1	34.9±5.3	40.3±5.8	50.0±10.4

Table 11: Complete finite-support recovery curves for ResNet18, ResNet50, and DenseNet121. F/S denote frontal/side-right views and I/D denote image-/driver-disjoint evaluation. Entries are mean accuracy ± support-draw standard deviation (%).

Image-disjoint accuracy generally rises sharply with the labelled-image budget for both views. Driver-disjoint recovery is lower and can be nonmonotonic because increasing k can change the selected support drivers and the remaining query population. These curves should therefore not be interpreted as pure sample-complexity laws.

Compact CNNs and frozen DINOv2. Figure 10 and Table 12 report EfficientNet-B0, SqueezeNet1.0, and frozen DINOv2.

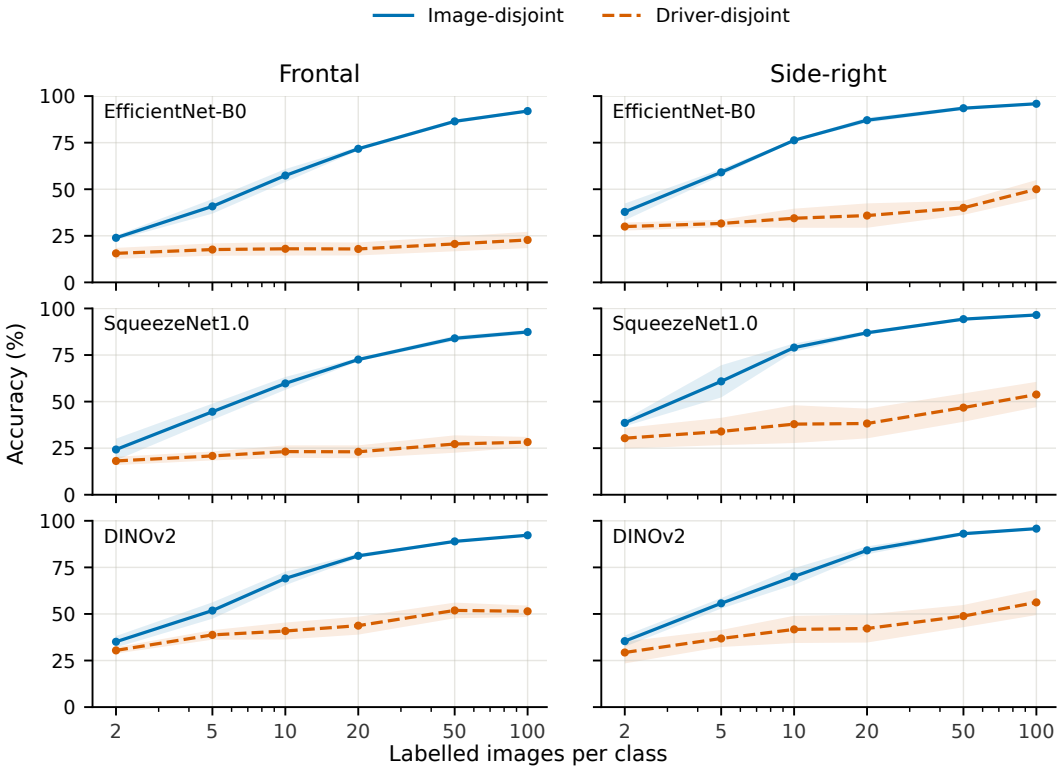


Figure 10: Both target viewpoints and population protocols under identical nominal class budgets. Bands show variation across five support draws.

Features	View/protocol	$k = 2$	$k = 5$	$k = 10$	$k = 20$	$k = 50$	$k = 100$
EffNet-B0	F/I	23.9±1.2	40.8±3.9	57.4±3.5	71.7±1.0	86.4±1.0	92.0±0.7
EffNet-B0	F/D	15.6±2.9	17.6±3.3	18.0±3.6	17.9±3.5	20.6±4.0	22.8±4.4
EffNet-B0	S/I	37.8±4.5	59.1±1.8	76.3±0.6	87.1±1.4	93.5±0.9	95.9±0.5
EffNet-B0	S/D	30.0±2.2	31.6±2.0	34.4±5.1	35.9±6.5	40.0±3.8	50.0±4.9
SqueezeNet	F/I	24.3±5.8	44.6±4.4	59.8±3.5	72.6±1.0	84.0±1.4	87.4±1.1
SqueezeNet	F/D	18.1±2.3	20.8±2.4	23.2±3.3	23.1±3.5	27.2±4.7	28.3±2.8
SqueezeNet	S/I	38.6±1.8	60.9±8.6	79.0±2.2	86.9±1.0	94.3±0.8	96.5±0.3
SqueezeNet	S/D	30.3±5.5	34.0±7.3	37.9±10.2	38.2±8.0	46.8±7.6	53.8±6.8
DINOv2	F/I	35.1±3.1	51.8±4.4	69.1±3.6	81.2±1.1	89.0±0.6	92.3±0.3
DINOv2	F/D	30.4±1.7	38.8±2.6	40.8±4.6	43.7±4.8	51.9±4.2	51.4±3.1
DINOv2	S/I	35.4±3.0	55.7±2.8	70.1±4.5	84.1±2.2	93.1±0.9	95.8±0.2
DINOv2	S/D	29.3±5.9	36.8±4.6	41.7±7.3	42.2±7.5	48.9±6.0	56.2±6.8

Table 12: Complete finite-support recovery curves for EfficientNet-B0, SqueezeNet1.0, and DINOv2. Notation follows Table 11.

DINOv2 provides a useful example of why image-level recovery and population-separated recovery must be reported separately. At $k = 100$, frontal image-disjoint accuracy reaches 92.3%, whereas driver-disjoint accuracy is 51.4%. The corresponding side-right values are 95.8% and 56.2%. High recognition of held-out images therefore does not specify the recovery obtained when the fitting drivers themselves are excluded from evaluation.

Register features and the multi-camera comparator. Figure 11 and Table 13 contain the DINOv2 register variant and the two DBMNet feature outputs.

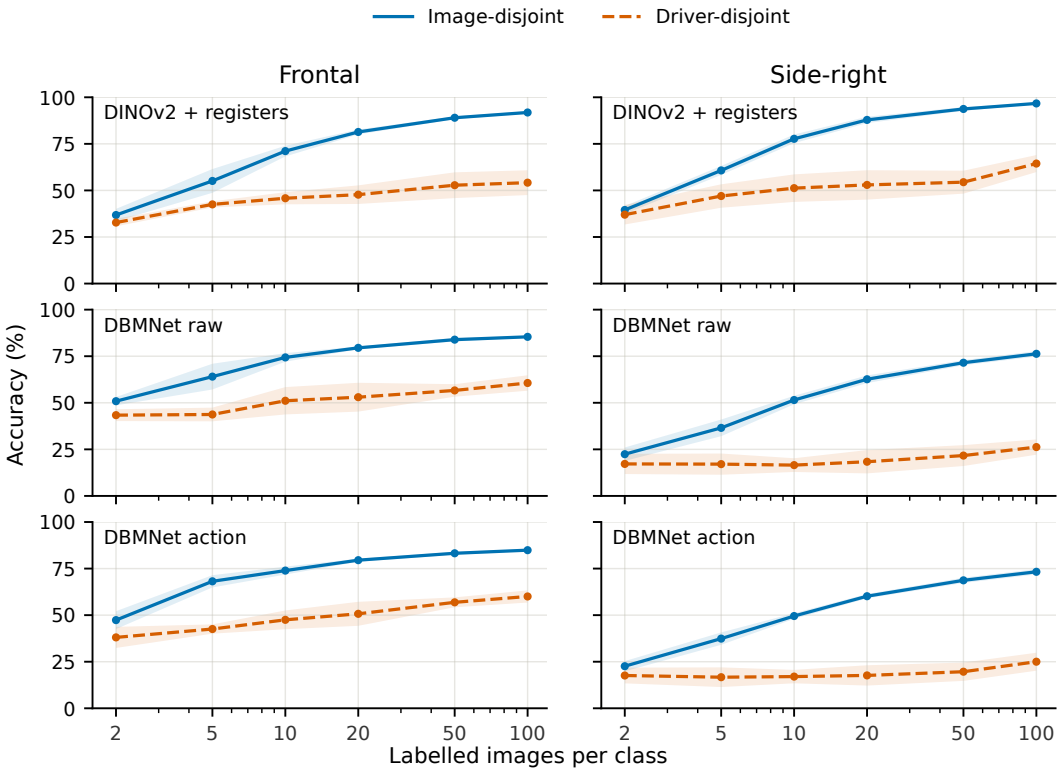


Figure 11: Both target viewpoints and population protocols under identical nominal class budgets. Bands show variation across five support draws.

Features	View/protocol	$k = 2$	$k = 5$	$k = 10$	$k = 20$	$k = 50$	$k = 100$
DINOv2+reg	F/I	36.8±3.3	55.1±6.4	71.1±2.9	81.4±1.3	89.0±0.7	91.8±0.7
DINOv2+reg	F/D	32.8±2.0	42.5±1.6	45.8±3.3	47.7±5.0	52.8±6.9	54.2±6.7
DINOv2+reg	S/I	39.5±2.1	60.7±2.4	77.7±2.3	87.9±1.7	93.7±1.0	96.7±0.4
DINOv2+reg	S/D	37.0±5.3	47.0±6.3	51.2±7.4	53.0±7.9	54.4±6.2	64.4±4.6
DBMNet raw	F/I	50.9±2.3	64.0±6.9	74.4±2.3	79.5±0.8	83.9±1.0	85.4±0.4
DBMNet raw	F/D	43.4±3.2	43.7±3.7	51.1±7.3	53.0±7.7	56.6±3.4	60.6±4.1
DBMNet raw	S/I	22.4±3.6	36.5±4.5	51.5±2.3	62.6±2.1	71.5±1.4	76.3±1.5
DBMNet raw	S/D	17.2±5.4	17.1±5.7	16.5±3.7	18.4±6.3	21.7±5.6	26.2±4.1
DBMNet action	F/I	47.3±4.8	68.2±3.3	73.9±2.1	79.5±0.8	83.2±0.8	84.9±0.5
DBMNet action	F/D	38.1±5.7	42.5±2.5	47.5±5.0	50.7±6.4	56.9±2.7	60.0±3.3
DBMNet action	S/I	22.6±2.7	37.4±3.3	49.5±1.5	60.2±1.2	68.7±1.4	73.3±1.4
DBMNet action	S/D	17.6±4.2	16.7±5.3	17.0±3.6	17.6±5.5	19.6±4.9	25.0±4.8

Table 13: Complete finite-support recovery curves for DINOv2 with registers and the two DBMNet feature outputs. DBMNet uses full-frame, camera-specific held-out-view checkpoints with additional multi-camera supervision and is therefore a contextual comparator rather than an information-matched single-source model.

The DINOv2 register representation retains the same common-coordinate interpretation as the base DINOv2 encoder. DBMNet does not: its frontal and side-right features originate from different held-out-camera checkpoints. Its rows therefore quantify within-condition recoverability only and should not be read as evidence that one DBMNet feature space supports classifier transport between T_2 and T_4.

E.3 Population sensitivity at the largest common budget

Table 14 summarizes the $k = 100$ endpoint for all nine feature configurations. The gap is defined descriptively as

$$\mathrm{Gap}_A = L_A(100; \mathrm{I}) - L_A(100; \mathrm{D}),$$

in percentage points.

Large positive gaps occur across the tested feature families, but their magnitude should not be interpreted as a causal estimate of identity leakage. The driver-disjoint construction simultaneously removes support-driver recurrence, changes the diversity of the support population, and can alter the remaining query population. The supported conclusion is therefore population conditionality of recovery rather than attribution of the full gap to one mechanism.

The contrast also prevents a representation ranking from being inferred from image-disjoint accuracy alone. A representation can perform strongly on unseen images of recurring drivers while producing a substantially different ordering when evaluation is restricted to drivers absent from fitting. Such ordering changes are descriptive properties of the tested population contracts, not independent architecture replications.

E.4 Complete driver-disjoint support-draw record

The aggregate driver-disjoint curves can hide substantial variation in which drivers remain available for evaluation. Table 15 therefore preserves the five recorded accuracy values for every model, view, and budget together with the range of external-query image counts and the number of fitting drivers realized across those draws.

Table 15: Every driver-disjoint draw: accuracy (%), query-image range, and fit-driver range.

		Draw accuracies (%)						
Model/view	k	1	2	3	4	5	Query n	Fit drivers
R18/F	2	18.8	15.6	16.5	13.0	10.0	3687–4334	2–3
R18/F	5	15.3	17.4	14.8	16.5	13.3	3687–4334	2–3
R18/F	10	19.6	17.8	17.9	16.4	14.0	3687–4334	2–3
R18/F	20	19.6	19.9	16.6	16.2	15.6	3687–4334	2–3
R18/F	50	18.7	21.0	18.0	18.0	26.5	2815–3778	3–5
R18/F	100	18.1	19.3	21.0	20.8	25.6	2185–3587	3–7
R18/S	2	35.1	34.2	35.2	23.5	37.0	3889–5026	1–3
R18/S	5	40.1	40.2	31.6	36.3	31.8	3889–5026	1–3
R18/S	10	38.6	42.6	34.7	41.3	34.8	3889–5026	1–3
R18/S	20	39.6	43.3	30.0	40.2	33.6	3889–5026	1–3
R18/S	50	34.1	44.2	38.8	42.1	40.1	3889–4440	2–4
R18/S	100	49.0	44.0	55.3	42.1	52.7	2710–3676	4–7
R50/F	2	13.2	15.3	12.8	15.4	16.1	3687–4334	2–3
R50/F	5	13.3	19.7	18.7	21.2	18.1	3687–4334	2–3
R50/F	10	19.4	22.7	19.4	18.6	18.6	3687–4334	2–3
R50/F	20	22.7	21.8	20.9	21.0	17.9	3687–4334	2–3
R50/F	50	33.3	22.2	30.5	26.8	33.7	2815–3778	3–5
R50/F	100	35.2	25.9	30.1	28.4	34.4	2185–3587	3–7
R50/S	2	30.6	45.4	29.9	42.0	33.4	3889–5026	1–3
R50/S	5	46.5	45.5	32.7	50.9	36.4	3889–5026	1–3
R50/S	10	29.3	46.2	37.1	60.0	44.5	3889–5026	1–3
R50/S	20	37.0	45.7	38.3	57.9	47.1	3889–5026	1–3
R50/S	50	43.4	48.5	50.0	57.8	57.4	3889–4440	2–4
R50/S	100	66.6	48.8	62.8	52.6	68.9	2710–3676	4–7
Dense121/F	2	16.5	23.2	15.0	16.2	19.2	3687–4334	2–3
Dense121/F	5	21.1	19.8	18.4	24.2	25.0	3687–4334	2–3

Continued on next page.

Table 15 continued.

Model/view	k	Draw accuracies (%) 1	2	3	4	5	Query n	Fit drivers
Dense121/F	10	24.5	16.9	17.2	25.6	24.6	3687–4334	2–3
Dense121/F	20	25.3	17.4	18.2	22.8	25.5	3687–4334	2–3
Dense121/F	50	33.3	16.1	26.7	28.5	27.6	2815–3778	3–5
Dense121/F	100	30.3	23.9	27.0	29.5	28.6	2185–3587	3–7
Dense121/S	2	30.5	30.4	28.9	28.4	24.5	3889–5026	1–3
Dense121/S	5	30.5	32.8	32.9	33.1	26.6	3889–5026	1–3
Dense121/S	10	33.9	39.9	28.9	44.3	28.9	3889–5026	1–3
Dense121/S	20	32.7	39.3	31.5	42.8	28.4	3889–5026	1–3
Dense121/S	50	32.6	40.1	50.5	40.2	38.1	3889–4440	2–4
Dense121/S	100	58.4	37.0	63.1	38.9	52.8	2710–3676	4–7
Eff-B0/F	2	15.8	10.7	19.9	14.9	16.6	3687–4334	2–3
Eff-B0/F	5	19.7	11.2	20.6	18.2	18.3	3687–4334	2–3
Eff-B0/F	10	19.8	12.4	22.8	19.5	15.6	3687–4334	2–3
Eff-B0/F	20	20.2	12.0	20.9	20.7	16.0	3687–4334	2–3
Eff-B0/F	50	20.7	13.3	21.3	23.4	24.7	2815–3778	3–5
Eff-B0/F	100	20.0	15.4	26.4	25.5	26.5	2185–3587	3–7
Eff-B0/S	2	30.2	27.2	29.1	33.9	29.5	3889–5026	1–3
Eff-B0/S	5	32.2	32.1	30.4	34.6	28.7	3889–5026	1–3
Eff-B0/S	10	32.3	36.0	29.3	43.7	30.8	3889–5026	1–3
Eff-B0/S	20	33.6	37.3	29.6	47.9	31.0	3889–5026	1–3
Eff-B0/S	50	33.6	38.9	41.0	45.1	41.4	3889–4440	2–4
Eff-B0/S	100	51.2	45.3	52.2	43.8	57.5	2710–3676	4–7
SqueezeNet/F	2	15.4	19.7	17.9	21.6	16.2	3687–4334	2–3
SqueezeNet/F	5	20.5	17.4	24.1	22.8	19.4	3687–4334	2–3
SqueezeNet/F	10	23.7	18.4	28.5	24.1	21.1	3687–4334	2–3
SqueezeNet/F	20	22.6	19.5	29.5	23.0	20.7	3687–4334	2–3
SqueezeNet/F	50	31.7	19.0	28.0	25.8	31.6	2815–3778	3–5
SqueezeNet/F	100	23.1	30.0	29.9	27.7	30.6	2185–3587	3–7
SqueezeNet/S	2	21.5	36.9	27.7	30.5	35.2	3889–5026	1–3
SqueezeNet/S	5	22.4	44.1	29.9	37.5	35.9	3889–5026	1–3
SqueezeNet/S	10	22.9	49.1	29.9	47.8	39.9	3889–5026	1–3
SqueezeNet/S	20	29.1	47.1	29.5	47.1	38.4	3889–5026	1–3
SqueezeNet/S	50	32.2	47.4	53.5	48.8	52.0	3889–4440	2–4
SqueezeNet/S	100	60.0	46.5	59.5	44.6	58.5	2710–3676	4–7
DINO/F	2	27.7	30.7	33.0	30.9	29.9	3687–4334	2–3
DINO/F	5	37.6	36.2	40.4	43.0	36.6	3687–4334	2–3
DINO/F	10	45.1	41.5	36.5	46.3	34.8	3687–4334	2–3
DINO/F	20	48.8	45.6	37.4	48.2	38.5	3687–4334	2–3
DINO/F	50	57.3	44.4	52.2	52.9	52.6	2815–3778	3–5
DINO/F	100	54.3	47.4	48.1	54.4	53.0	2185–3587	3–7
DINO/S	2	18.3	35.3	32.2	31.9	28.9	3889–5026	1–3
DINO/S	5	28.8	42.7	36.0	37.9	38.8	3889–5026	1–3
DINO/S	10	29.5	52.5	42.3	42.5	41.8	3889–5026	1–3
DINO/S	20	30.5	54.2	42.4	43.6	40.3	3889–5026	1–3
DINO/S	50	38.2	53.6	48.4	48.6	55.4	3889–4440	2–4
DINO/S	100	62.0	49.9	55.3	48.0	65.8	2710–3676	4–7
DINO+reg/F	2	30.1	33.9	35.9	32.6	31.3	3687–4334	2–3
DINO+reg/F	5	44.5	42.7	42.8	43.1	39.5	3687–4334	2–3
DINO+reg/F	10	51.8	45.4	43.0	45.9	42.8	3687–4334	2–3
DINO+reg/F	20	56.5	47.1	42.8	49.0	43.1	3687–4334	2–3
DINO+reg/F	50	57.9	41.4	52.8	50.3	61.6	2815–3778	3–5
DINO+reg/F	100	62.1	44.1	52.3	51.4	61.0	2185–3587	3–7
DINO+reg/S	2	32.1	41.4	44.5	30.7	36.2	3889–5026	1–3

Continued on next page.

Table 15 continued.

		Draw accuracies (%)						
Model/view	k	1	2	3	4	5	Query n	Fit drivers
DINO+reg/S	5	40.0	56.8	51.7	44.4	42.1	3889–5026	1–3
DINO+reg/S	10	39.3	59.9	58.2	50.2	48.5	3889–5026	1–3
DINO+reg/S	20	38.3	60.4	59.2	52.9	54.0	3889–5026	1–3
DINO+reg/S	50	43.8	63.0	55.1	56.1	54.1	3889–4440	2–4
DINO+reg/S	100	60.7	64.0	63.6	60.7	73.1	2710–3676	4–7
DBM-raw/F	2	46.7	45.5	38.3	45.3	41.1	3690–4339	2–3
DBM-raw/F	5	46.3	48.0	39.3	45.7	39.3	3690–4339	2–3
DBM-raw/F	10	53.6	57.3	48.0	58.3	38.3	3690–4339	2–3
DBM-raw/F	20	57.5	60.2	47.9	59.2	40.2	3690–4339	2–3
DBM-raw/F	50	57.3	53.1	52.4	59.5	60.9	2818–3793	3–5
DBM-raw/F	100	60.0	53.0	63.1	61.9	65.1	2195–3599	3–7
DBM-raw/S	2	11.1	22.3	10.0	21.8	20.6	3891–5028	1–3
DBM-raw/S	5	12.6	20.3	8.2	22.8	21.5	3891–5028	1–3
DBM-raw/S	10	15.4	19.3	9.8	20.0	18.2	3891–5028	1–3
DBM-raw/S	20	12.1	24.1	10.5	26.3	18.8	3891–5028	1–3
DBM-raw/S	50	10.7	25.3	22.8	26.0	23.5	3891–4440	2–4
DBM-raw/S	100	32.4	23.0	28.7	20.8	26.2	2710–3676	4–7
DBM-action/F	2	38.9	39.5	27.1	41.8	43.0	3690–4339	2–3
DBM-action/F	5	41.6	46.5	40.4	44.3	40.0	3690–4339	2–3
DBM-action/F	10	51.0	51.1	41.3	52.6	41.5	3690–4339	2–3
DBM-action/F	20	53.1	58.1	44.3	56.2	41.9	3690–4339	2–3
DBM-action/F	50	56.4	54.2	54.6	57.6	61.6	2818–3793	3–5
DBM-action/F	100	61.3	54.3	60.9	59.4	64.3	2195–3599	3–7
DBM-action/S	2	15.0	23.7	11.2	18.7	19.5	3891–5028	1–3
DBM-action/S	5	12.4	18.3	9.3	24.3	19.2	3891–5028	1–3
DBM-action/S	10	16.6	19.3	10.4	21.0	17.8	3891–5028	1–3
DBM-action/S	20	11.6	23.3	10.9	23.4	19.1	3891–5028	1–3
DBM-action/S	50	9.9	23.2	20.8	21.6	22.6	3891–4440	2–4
DBM-action/S	100	30.6	19.8	28.8	18.7	27.1	2710–3676	4–7

The query-count and fitting-driver columns are part of the experimental record rather than additional uncertainty measures. They show that identical nominal k-per-class budgets can be realized with different numbers of people and different remaining query populations. This is particularly relevant at larger budgets, where additional drivers may be required to satisfy class coverage.

The table contains the complete recorded driver-disjoint draw outcomes. The corresponding image-disjoint evidence is retained at the aggregate level in Tables 11–13. The manuscript does not infer unrecorded raw image-disjoint draw values beyond those available in the supplied experimental artifacts.

E.5 Interpretation boundary

The recovery experiments establish the following empirical statement: under a fixed feature extractor, labelled support from a target observation condition can make substantially more class-discriminative information accessible to a newly fitted linear decision rule than is exploited by the direct-transfer classifier. The amount of that recovery depends on the support budget, observation condition, representation, and support–query population contract.

They do not establish that the source class-relational organization is preserved, that one recovered classifier will serve another camera view, or that classifier-only refitting reaches the performance attainable by unrestricted feature adaptation. Those questions are measured separately by the class-relational and decision-compatibility analyses.

	Frontal			Side-right		
Features	I	D	Gap	I	D	Gap
ResNet18	81.3±0.7	21.0±2.6	60.3	91.6±0.6	48.6±5.0	43.0
ResNet50	94.3±0.5	30.8±3.5	63.5	98.3±0.3	59.9±7.9	38.4
DenseNet121	90.5±1.1	27.9±2.3	62.6	96.8±0.4	50.0±10.4	46.8
EffNet-B0	92.0±0.7	22.8±4.4	69.2	95.9±0.5	50.0±4.9	45.9
SqueezeNet	87.4±1.1	28.3±2.8	59.1	96.5±0.3	53.8±6.8	42.7
DINOv2	92.3±0.3	51.4±3.1	40.8	95.8±0.2	56.2±6.8	39.6
DINOv2+reg	91.8±0.7	54.2±6.7	37.7	96.7±0.4	64.4±4.6	32.3
DBMNet raw	85.4±0.4	60.6±4.1	24.8	76.3±1.5	26.2±4.1	50.1
DBMNet action	84.9±0.5	60.0±3.3	24.9	73.3±1.4	25.0±4.8	48.3

Table 14: Population sensitivity at $k = 100$ labelled images per class. I/D denote image-/driver-disjoint evaluation; Gap is their accuracy difference in percentage points. Every I/D entry is mean $\pm$ standard deviation across five support draws.

F Cross-View Classifier Analyses

F.1 Shared-view classifier behavior

The shared-view experiment asks whether labelled support from both target camera conditions is sufficient to fit one classifier-side pipeline that serves both views. The recorded experiment uses

$$k = 100$$

support images per class in total across $T_2 \cup T_4$, for a total support size of 1000 images. The same fitted standardizer and classifier are then evaluated separately on frontal and side-right target queries.

Table 16 reports the complete seven-representation compatibility summary. The analyses below provide the corresponding support composition, sensitivity checks, confusion structure, and source-return detail.

All five task-trained CNN representations favor the side-right condition under the shared classifier. ResNet50, for example, attains 18.8% on the frontal query and 63.1% on the side-right query. DINOv2 reduces but does not eliminate the disparity, reaching 39.7% and 48.8%, while the register variant reaches 40.3% and 55.6%.

The shared classifier is therefore not equivalent to fitting two locally optimal classifiers and averaging their behavior. It is one affine decision rule, including one support-derived standardizer, fitted to the union of labelled observations from both views. Strong local recoverability in each view does not guarantee that their pooled support admits one comparably effective linear rule.

F.2 Realized shared-support composition

The joint-support sampler does not impose exact camera balance within each class. Table 17 records the realized aggregate camera composition for the five $k = 100$ support draws.

Across the five draws, the frontal contribution ranges from 44.7% to 48.7%, with a mean of 46.66%; the remaining 53.34% is side-right support. The observed asymmetry therefore cannot be described as arising from a severely one-sided support set, but the current experiment also does not eliminate composition as an explanation. Exact camera balance would require controlling the per-class allocation rather than inspecting the realized aggregate fraction after sampling.

F.3 Class balance and label-mapping sensitivity

Raw accuracy can differ across views because of query class frequencies, and the normalized task contains one semantically imperfect correspondence between SF3D *Drink* and 100-Driver `C16_Eat_Left`. Table 18 therefore reports three summaries from the same saved shared-classifier predictions: ordinary accuracy, class-macro recall, and accuracy after removing true $c6$ examples from evaluation.

The frontal/side-right disparity remains visible under macro recall for all five task-trained CNNs. For ResNet50, accuracy changes from 18.8/63.1 to macro recall 20.5/62.4. Removing true $c6$ examples yields 20.3/63.7, again leaving the qualitative disparity intact.

	Shared classifier: target queries		View-specific classifier: source query	
Features	Frontal query	Side query	Frontal fit	Side fit
ResNet18	12.3±0.9	51.5±0.7	13.2±3.4	58.6±6.7
ResNet50	18.8±1.0	63.1±1.0	17.7±4.1	64.0±2.7
DenseNet121	22.1±0.9	49.2±1.8	7.8±4.2	51.4±8.2
EffNet-B0	20.3±0.5	40.8±1.3	10.8±4.3	53.3±3.6
SqueezeNet	22.9±1.1	36.6±1.0	9.0±2.0	41.8±3.7
DINOv2	39.7±1.2	48.8±1.2	12.9±2.2	35.7±3.4
DINOv2+reg	40.3±0.8	55.6±0.3	9.4±0.5	41.4±2.8

Table 16: Classifier compatibility at $k = 100$: pooled-view classifiers on target queries and view-specific classifiers returned to the common source query (mean ± support-draw standard deviation, %).

Draw	Frontal n	Side-right n	Frontal (%)
1	473	527	47.3
2	447	553	44.7
3	458	542	45.8
4	487	513	48.7
5	468	532	46.8
Mean	466.6	533.4	46.66

Table 17: Recorded camera composition of the five shared-view supports at $k = 100$. Each draw contains 1000 images in total, corresponding to 100 examples per class across both target views. Camera balance is inherited from sampling rather than constrained to 50/50.

Features	F acc.	F macro	F no-$c6$	S acc.	S macro	S no-$c6$
ResNet18	12.3	11.9	13.0	51.5	51.7	50.1
ResNet50	18.8	20.5	20.3	63.1	62.4	63.7
DenseNet121	22.1	21.8	21.3	49.2	49.0	49.7
EfficientNet-B0	20.3	19.8	20.0	40.8	41.5	40.8
SqueezeNet1.0	22.9	20.4	21.9	36.6	36.1	37.4
DINOv2	39.7	41.1	35.6	48.8	50.8	46.0
DINOv2 + registers	40.3	42.8	36.9	55.6	56.5	54.5

Table 18: Shared-view outcomes (%) at $k = 100$ total support examples per class. F/S denote frontal/side-right. "Macro" is class-macro recall. "No-$c6$" removes true examples of the imperfect food/drink correspondence from evaluation only; the original ten-class classifier is unchanged.

The no-$c6$ measurement does not add fitting data, remove an output class, or refit a nine-class classifier. It only tests whether one known semantic mismatch dominates the reported shared-view asymmetry. The resulting values indicate that the mismatch can affect individual scores but does not explain the principal CNN pattern.

F.4 Compatibility with the source condition

A second functional test asks how a classifier fitted under one target condition behaves when returned to the original source representation. The fitted target standardizer and classifier are reused together; neither is recomputed from source examples.

Table 19 reports absolute source-query accuracy for pipelines fitted on frontal support, side-right support, or shared-view support.

Across all seven common-coordinate feature spaces, the side-right-fitted pipeline returns higher source accuracy than the corresponding frontal-fitted pipeline. For ResNet50, the frontal-fitted classifier attains 17.7% on the source query whereas the side-right-fitted classifier attains 64.0%. The shared-view pipeline attains 65.8%.

This pattern is not backbone forgetting: f_θ is unchanged throughout. It instead shows that the effective affine decision rule induced by the fitted standardizer and classifier depends strongly on the condition from which support was obtained.

Features	Frontal-calibrated	Side-calibrated	Shared
ResNet18	13.2	58.6	45.7
ResNet50	17.7	64.0	65.8
DenseNet121	7.8	51.4	46.2
EfficientNet-B0	10.8	53.3	50.3
SqueezeNet1.0	9.0	41.8	22.6
DINOv2	12.9	35.7	30.4
DINOv2 + registers	9.4	41.4	38.4

Table 19: Return-to-source accuracy (%) for classifier-side pipelines fitted on frontal, side-right, or shared-view target support. The frozen feature extractor is unchanged. Values are absolute source-query accuracies rather than differences from historical source references.

The return values are not converted into a common source-regret quantity because the available historical source references do not provide a universally matched fitting protocol for these target-fitted pipelines. The matched source-side refit change Δ_S and target-to-source return therefore remain separate empirical quantities.

F.5 Class-conditional structure of shared-view errors

Aggregate accuracy cannot show whether the view disparity is diffuse across classes or concentrated in particular behaviors. Figures 12–18 report the class-conditional confusion structure of the $k = 100$ shared classifier for all seven common-coordinate representations.

Each confusion matrix is row-normalized within each support draw and then averaged across the five draws. Because the same fixed query is evaluated by each draw-specific classifier, repeated predictions across draws are repeated evaluations of the same query images rather than additional independent observations.

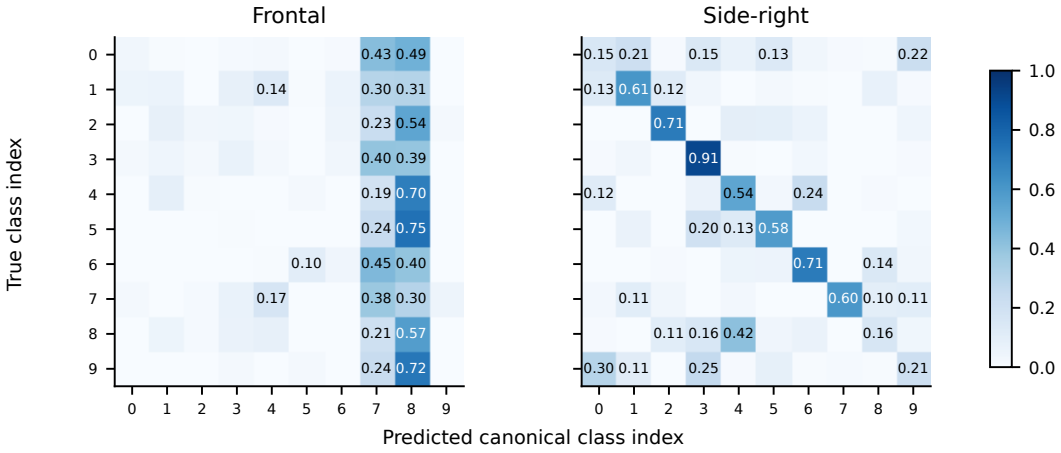


Figure 12: Class-conditional errors of the shared ResNet18 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

The matrices show that shared-view failure is not simply a uniform reduction of confidence across all classes. Different representations exhibit different class-specific confusion patterns, and the frontal and side-right conditions need not fail on the same behaviors. These matrices are therefore supporting localization evidence for the functional asymmetry, rather than independent tests of representation geometry.

F.6 Relationship among the completed compatibility tests

The completed classifier analyses probe different forms of decision compatibility:

$$C_{T_2 \cup T_4 \to T_2}, \qquad C_{T_2 \cup T_4 \to T_4}$$

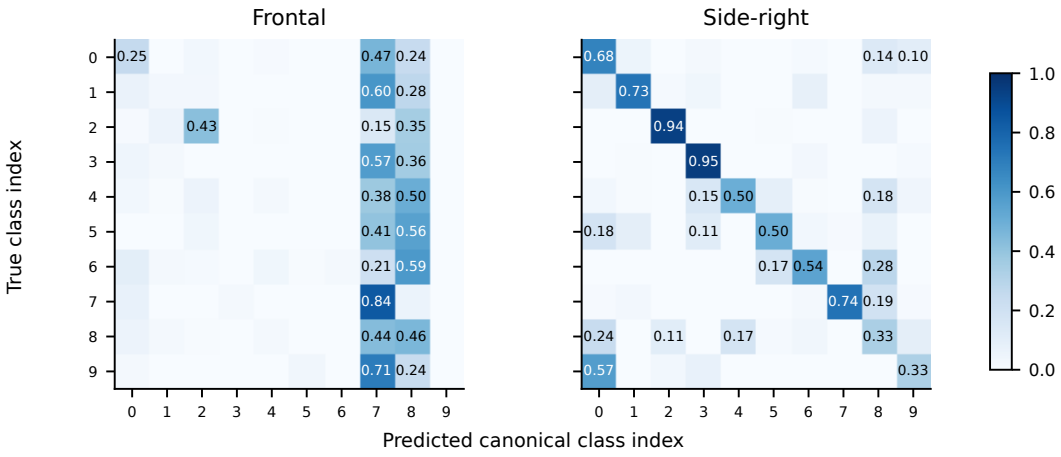


Figure 13: Class-conditional errors of the shared ResNet50 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

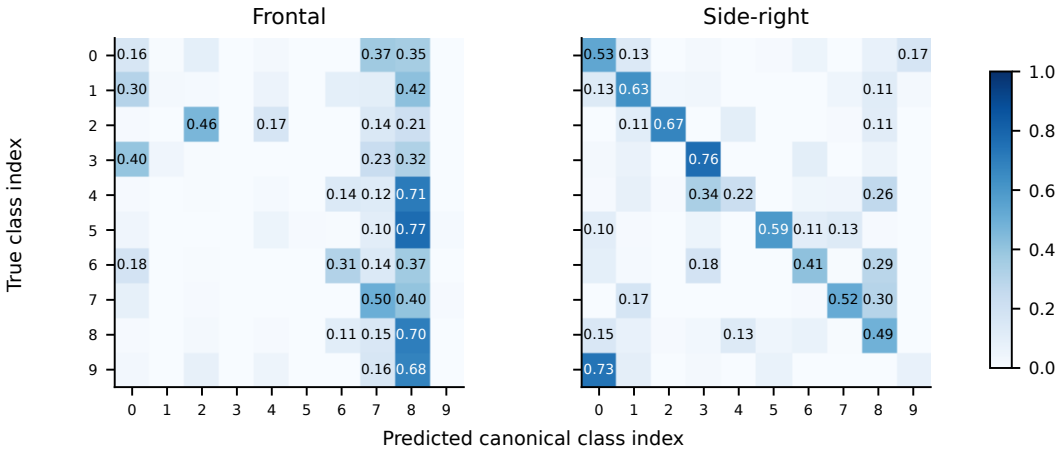


Figure 14: Class-conditional errors of the shared DenseNet121 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

measure whether one classifier fitted from joint target support serves each target view, while

$$C_{T_2 \to S}, \qquad C_{T_4 \to S}$$

measure how independently view-fitted rules behave when returned to the source condition.

These measurements provide converging evidence that classifier behavior depends on the condition under which the rule is fitted, but they do not substitute for direct reciprocal view-specific transport. The latter requires fitting on one target view and evaluating the unchanged pipeline on the other while holding the receiving query fixed.

F.7 Interpretation Boundary

The shared-view and return-to-source analyses establish that strong local classifier recovery does not by itself imply one condition-independent decision rule. One jointly fitted classifier can remain asymmetric across target views, and independently view-fitted classifiers can differ markedly in their compatibility with the same source representation.

These experiments do not by themselves identify why that incompatibility occurs. Source-view orientation, class-relational reorganization, support composition, regularization, finite optimization,

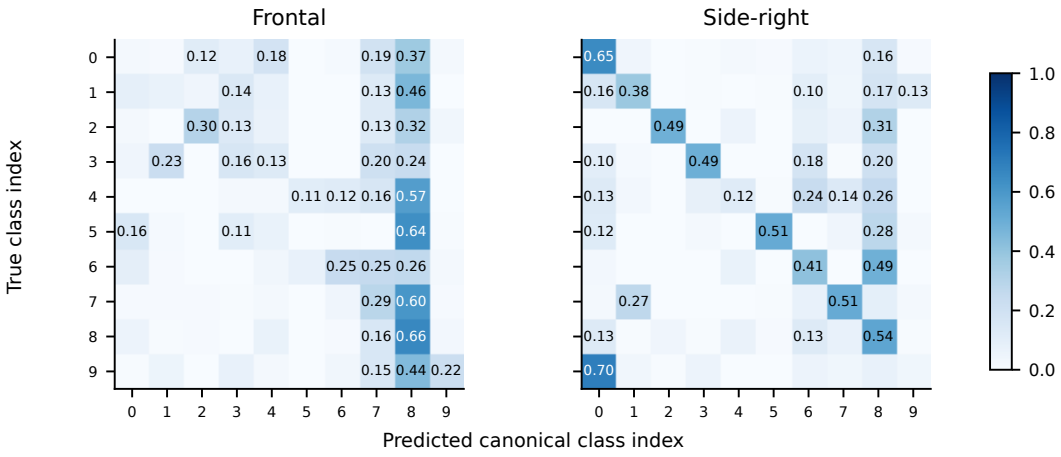


Figure 15: Class-conditional errors of the shared EfficientNet-B0 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

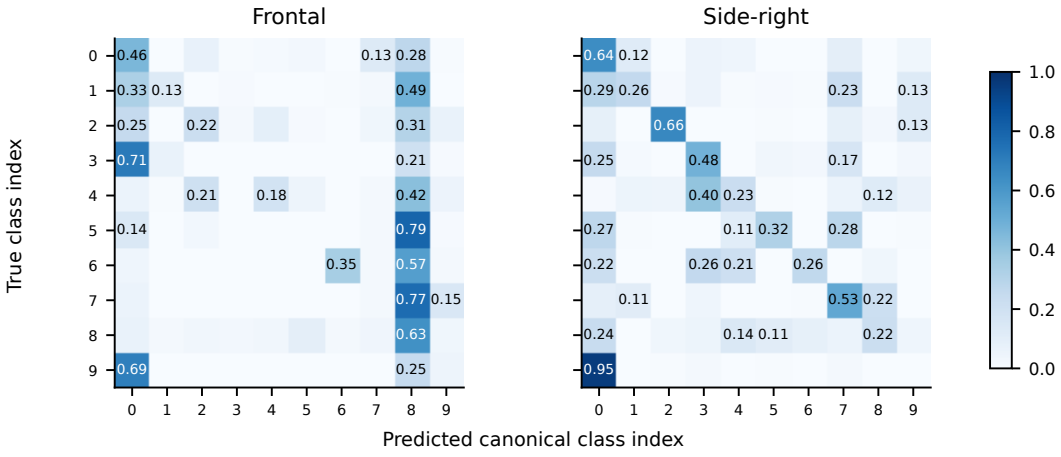


Figure 16: Class-conditional errors of the shared SqueezeNet1.0 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

and view-specific visibility remain possible contributors. Their role is therefore to establish the functional phenomenon—condition-dependent classifier compatibility—while the representation-geometry analyses examine a complementary structural property.

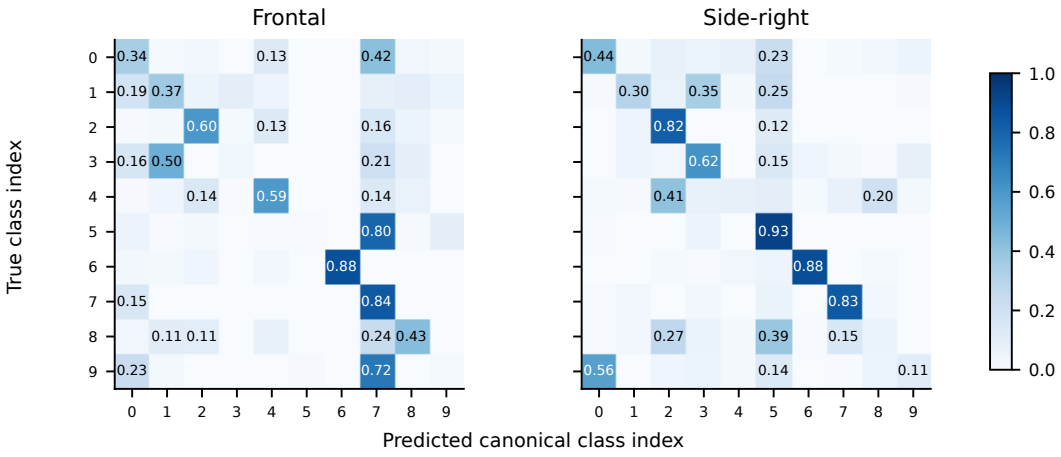


Figure 17: Class-conditional errors of the shared DINOv2 classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

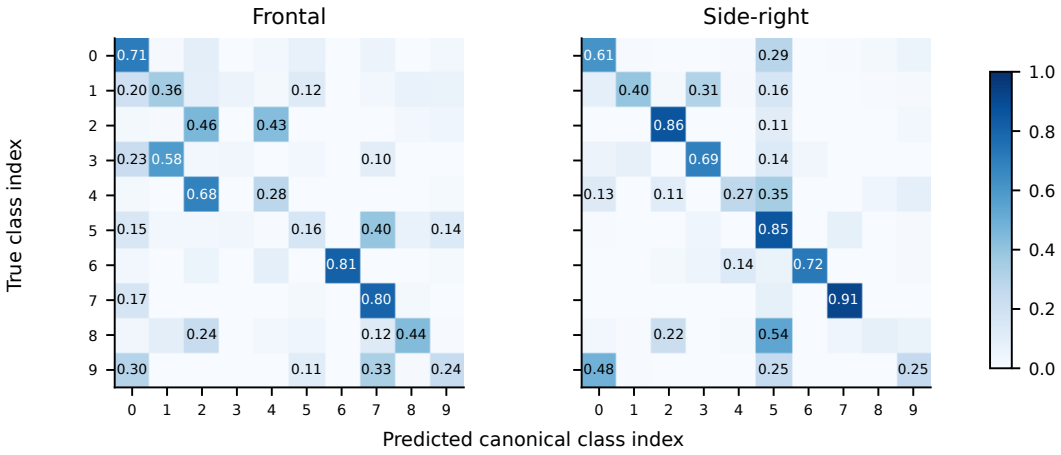


Figure 18: Class-conditional errors of the shared DINOv2 with registers classifier at 100 images per class across both views. Rows are normalized per draw and averaged over five support draws; query class counts are fixed.

G Representation-Geometry Analyses

G.1 Cross-dataset class-relational correspondence

The original class-relational analysis reports the two frozen DINOv2 representations, each of which provides one common feature coordinate system across S, T_2, and T_4. For each representation and input treatment, we construct ten-class RDMs using the measurement procedure defined in Appendix D and compare the source RDM with the corresponding target-condition RDM.

Figure 19 and Table 20 retain the original DINOv2 comparison and reliability references. The extended seven-representation matrix is reported below. "Across datasets" denotes the covariance-aware source–target RDM correspondence, with S as the reference side. "Within view" denotes the image-half reliability measurement for the corresponding target condition.

For the primary crop representation, base DINOv2 gives

$$G(S, T_2; \Sigma_S) = 0.021$$

for the frontal condition and

$$G(S, T_4; \Sigma_S) = 0.306$$

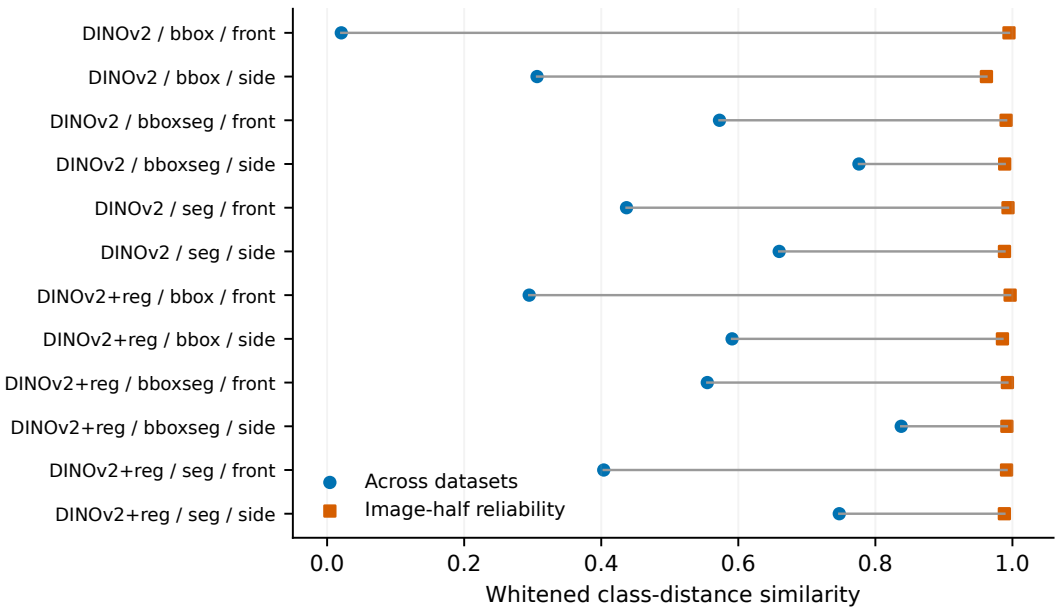


Figure 19: Image-split class-distance reliability remains high while cross-dataset correspondence varies with view, input treatment, and representation. The within-view reference can share drivers across its two halves.

Features	View	Input	Across datasets	Within view
DINOv2	Frontal	bbox	0.021	0.995
DINOv2	Side-right	bbox	0.306	0.962
DINOv2	Frontal	bboxseg	0.573	0.991
DINOv2	Side-right	bboxseg	0.776	0.989
DINOv2	Frontal	seg	0.437	0.994
DINOv2	Side-right	seg	0.660	0.988
DINOv2 + registers	Frontal	bbox	0.295	0.997
DINOv2 + registers	Side-right	bbox	0.591	0.985
DINOv2 + registers	Frontal	bboxseg	0.555	0.993
DINOv2 + registers	Side-right	bboxseg	0.838	0.992
DINOv2 + registers	Frontal	seg	0.404	0.991
DINOv2 + registers	Side-right	seg	0.747	0.988

Table 20: Class-relational correspondence for all recorded DINOv2 representations and input treatments. "Across datasets" compares S with the indicated target condition. "Within view" compares two image halves of that target condition using the same RDM procedure. The latter is an image-split reliability reference and may contain the same drivers in both halves.

for the side-right condition. The corresponding within-target image-half measurements are 0.995 and 0.962. The register representation yields higher cross-dataset correspondence under the same crop treatment, 0.295 for frontal and 0.591 for side-right, while retaining within-view values of 0.997 and 0.985.

These values should not be interpreted as percentages of semantic information retained or as classification accuracies. They measure correspondence between estimated pairwise class-distance structures under the stated RDM construction and covariance-aware comparison.

G.2 Direct Cam2–Cam4 geometry across all representations

Table 21 extends the source–target comparisons with direct within-100-Driver measurements for seven encoders and three input treatments. Each row compares cached features indexed by one encoder qualifier across observation conditions; the CNN checkpoint-binding limitation is specified in Appendix B.5. The two DINOv2 variants retain the fixed public-encoder identities of the original

geometry experiment. The archive contains 84 measurements: the two source–target comparisons and two covariance-reference choices for the direct camera pair, for each of 21 encoder–treatment cells. Every exported measurement uses ten classes, six within-class image partitions, the corrected distance path, and covariance-aware comparison without a recorded fallback.

What the within-collection comparison controls. Both target conditions belong to 100-Driver and use the same native behavior correspondence, including `C16_Eat_Left`. The cross-dataset Drink/Eat semantic mismatch is therefore absent between these two views, although all ten classes remain included. The comparison also avoids changing the parent dataset. It does not match the participants, samples, or synchronized action instances in the two extracted camera pools, and is not a calibrated camera-pose intervention. It consequently narrows the domain contrast without isolating a viewpoint-only causal effect.

Nested source–target ordering. Source–side-right correspondence exceeds source–frontal in 19/21 cells. The exceptions are SqueezeNet `seg` and `bboxseg`; the ordering holds for all six DINOv2 cells. This is a descriptive pattern across correlated model–treatment configurations, not a binomial test over independent architectures and not a numerical ranking of camera-angle distance.

Covariance-reference sensitivity. $G(T_2, T_4; \Sigma_{T_2})$ exceeds $G(T_4, T_2; \Sigma_{T_4})$ in 19/21 cells. The exceptions are DenseNet121 `bbox` and SqueezeNet `bbox`. Because the same class-relational comparison is weighted using a different reference covariance, this count does not show that a classifier transfers better from Cam2 to Cam4. It is also not an independent replication of the shared-classifier asymmetry. A fixed covariance metric or symmetric-comparator control would distinguish reference weighting from the underlying RDM relationship more directly.

Treatment-dependent comparison, without choosing the better reference. For each encoder a and treatment v, define the descriptive contrast

$$\delta_{a,v} = \tfrac{1}{2}\{G(T_2, T_4; \Sigma_{T_2}) + G(T_4, T_2; \Sigma_{T_4})\} - \tfrac{1}{2}\{G(S, T_2; \Sigma_S) + G(S, T_4; \Sigma_S)\}.$$

We average both direct reference choices rather than select their maximum. Table 22 reports the constituent means and the mean contrast across encoders. The means are +0.054 for `bbox`, +0.076 for `bboxseg`, and -0.052 for `seg`; positive contrasts occur for only 3/7, 4/7, and 2/7 encoders, respectively. These averages do not remove differences in covariance weighting and are not causal estimates of a dataset-shift contribution.

Both `seg` and `bboxseg` apply foreground masking; `bboxseg` must not be grouped with unmasked crops as if it retained background by design. Crop scale, detector or mask error, and removed objects can change the measured structure. The historical CNN recovery ablations use treatment-specific source checkpoints; the cached CNN geometry export does not independently establish that checkpoint binding, so its treatment differences are not attributed to a fixed-checkpoint background intervention. The strongest within-collection correspondence is therefore not a universal masking or background-removal effect. In particular, the high DINOv2 crop values qualify the severe source–frontal contrast without implying that every encoder retains the same geometry.

G.3 Image-half reliability and its interpretation

The within-view measurements in Table 20 serve as reliability references for the RDM estimator. They ask whether independently constructed image subsets from the same observation condition yield similar class-relational structure.

Across all twelve recorded DINOv2 cells, image-half correspondence lies between 0.962 and 0.997. This shows that the class-distance organization is highly reproducible under the particular image-half construction. It does not establish independence from driver identity, because drivers can recur across the two halves.

The distinction is important for interpreting the large difference between within-condition and cross-dataset correspondence. For the base-DINOv2 frontal crop cell, the image-half values produce a large within-minus-cross contrast, but part of that contrast can reflect the population structure used to estimate the reliability reference rather than only the cross-condition difference itself.

		Cross-dataset		Within 100-Driver	
Features	Treatment	$G(S, T_2; \Sigma_S)$	$G(S, T_4; \Sigma_S)$	$G(T_2, T_4; \Sigma_{T_2})$	$G(T_4, T_2; \Sigma_{T_4})$
ResNet18†	bbox	0.356	0.574	0.438	0.123
ResNet18†	seg	0.483	0.626	0.445	0.392
ResNet18†	bboxseg	0.137	0.271	0.706	0.449
ResNet50†	bbox	0.414	0.546	0.314	0.135
ResNet50†	seg	0.370	0.589	0.610	0.518
ResNet50†	bboxseg	0.194	0.443	0.560	0.439
DenseNet121†	bbox	-0.046	0.544	0.176	0.210
DenseNet121†	seg	0.252	0.359	0.489	0.422
DenseNet121†	bboxseg	-0.071	0.032	0.344	-0.055
EffNet-B0†	bbox	-0.144	0.571	0.113	-0.153
EffNet-B0†	seg	0.491	0.653	0.531	0.464
EffNet-B0†	bboxseg	0.095	0.413	0.427	-0.012
SqueezeNet†	bbox	-0.098	0.153	0.304	0.331
SqueezeNet†	seg	0.534	0.294	0.379	0.308
SqueezeNet†	bboxseg	0.378	0.086	0.569	0.325
DINOv2	bbox	0.021	0.306	0.727	0.647
DINOv2	seg	0.437	0.660	0.427	0.346
DINOv2	bboxseg	0.573	0.776	0.570	0.411
DINOv2+reg	bbox	0.295	0.591	0.782	0.698
DINOv2+reg	seg	0.404	0.747	0.469	0.365
DINOv2+reg	bboxseg	0.555	0.838	0.556	0.490

Table 21: Complete recorded source–target and direct target-view class-relational correspondence. S, T_2, and T_4 denote SF3D, frontal Cam2, and side-right Cam4. The covariance-reference argument is retained explicitly: reversing it changes the measurement procedure, not a fitted classifier. Values are correlations, not accuracies. CNN rows (†) are structural extensions rather than verified checkpoint-matched counterparts of the CNN recovery experiment.

Treatment	S, T_2	S, T_4	T_2, T_4	T_4, T_2	$\overline{\delta}$	$\delta > 0$
bbox	0.114	0.469	0.408	0.284	+0.054	3/7
seg	0.424	0.561	0.478	0.402	-0.052	2/7
bboxseg	0.266	0.408	0.533	0.293	+0.076	4/7

Table 22: Descriptive means across seven encoders, with covariance references as in Table 21. $\overline{\delta}$ averages both direct references before comparison; the final column counts positive encoder-level contrasts. No best-reference selection or independent-architecture significance claim is made.

G.4 Subject-partition reliability control

A stricter recorded sensitivity analysis reconstructs the base-DINOv2 crop/frontal geometry after partitioning source and target observations by subject rather than by random image halves. Table 23 and Figure 20 compare the resulting within-minus-cross contrast with the original image-random reference.

Under the image-random construction, the frontal within-minus-cross contrast is 0.963. Under subject partitioning, the recorded mean is 0.389 and the median is 0.442, with observed range

$$[-0.083,\ 0.601].$$

There are 1,383 matched partition pairings, but these do not constitute 1,383 independent replications. Only three valid source subject partitions are available, while the target side contributes 461 valid partitions, so the same source partitions are repeatedly paired with different target partitions.

This control materially reduces the apparent magnitude of the geometry contrast. It does not replace the underlying cross-dataset correspondence $G(S, T_2; \Sigma_S)$, nor does it make source–target and within-target geometry equivalent quantities. Its role is to show that the apparent contrast depends on the population unit used to construct the reliability reference.

Estimator	Contrast	Observed range	Pairings
Image-random halves	0.963	–	one split
Driver-partition mean	0.389	[-0.083, 0.601]	1383
Driver-partition median	0.442	same range	same pairs

Table 23: Sensitivity of the within-minus-cross geometry contrast for base DINOv2, crop input, S versus frontal T_2. The subject-partition range describes the observed dependent partition pairings and is not a confidence interval.

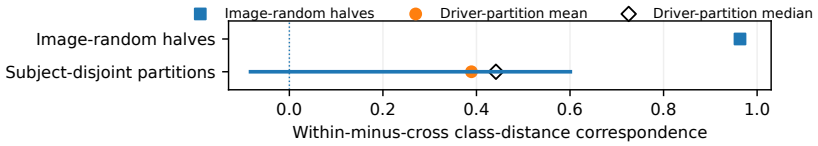


Figure 20: Changing the sampling unit alters the geometry reliability reference. The image-random contrast is compared with the mean and median across subject-partition pairings. The min–max range summarizes dependent pairings, not independent-replication uncertainty.

The stricter subject-partition analysis has been recorded only for the base-DINOv2 crop/frontal cell. The remaining rows in Table 20 retain image-half reliability and must not be described as driver-independent geometry estimates.

G.5 Input-treatment and representation sensitivity

The full geometry table shows that class-relational correspondence depends not only on target view but also on how the image is represented before feature extraction. For base DINOv2, frontal cross-dataset correspondence changes from 0.021 under `bbox` to 0.437 under `seg` and 0.573 under `bboxseg`. The corresponding side-right values are 0.306, 0.660, and 0.776.

The register representation follows a related but not identical pattern. Under frontal observations, correspondence is 0.295 for `bbox`, 0.404 for `seg`, and 0.555 for `bboxseg`; under side-right observations the corresponding values are 0.591, 0.747, and 0.838.

These comparisons show that the measured class-relational structure is sensitive to both representation and input treatment. They do not identify a universal mechanism by which cropping, segmentation, or register tokens improve cross-view recognition. In particular, higher RDM correspondence should not be equated with higher classifier compatibility unless the latter is measured separately.

The input treatments also change different things. Cropping changes field of view and image scale, segmentation removes pixels while preserving the full canvas, and crop-plus-segmentation combines both operations. The geometry results therefore characterize treatment-dependent relational structure rather than isolating one causal factor such as "background removal."

G.6 Covariance-reference and estimator sensitivity

The geometry implementation keeps the covariance-reference side explicit. In the reported source–target comparisons, S supplies the reference measurements used to estimate the covariance term for `corr_cov`. In within-view image-half comparisons, the first half supplies that reference.

This directionality is conceptually separate from the crossvalidated distance construction itself. The RDM estimator does not receive an explicit empirical feature-space precision matrix; covariance information enters subsequently in the RDM comparison. The terms "crossvalidated distance" and "covariance-aware RDM correlation" therefore refer to two different stages of the analysis.

G.7 Archived projected feature structure

The quantitative geometry analysis operates in the original frozen feature spaces. PCA and t-SNE are retained only as illustrative projections of selected ResNet50 feature configurations.

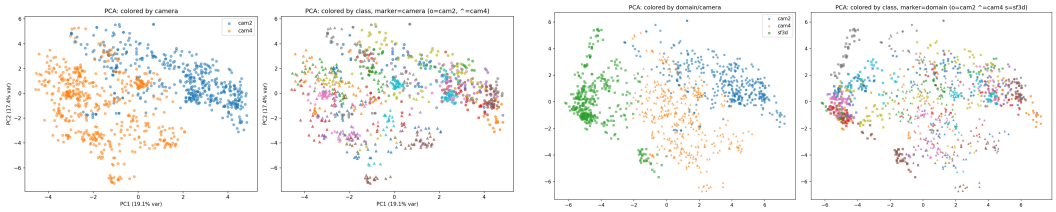
Figure 21 shows the archived PCA views. The two-target-view projection reports 19.1% and 17.4% variance explained by its first two principal components. Figure 22 shows the corresponding t-SNE visualizations.

The original feature coordinates and exact sampling indices underlying these archived projections are not present in the reconstruction record, so the displayed points have not been regenerated from current feature tensors. The quantitative RDM tables remain independently traceable and constitute the reproducible geometry evidence used for the manuscript's numerical claims.

G.8 Interpretation boundary

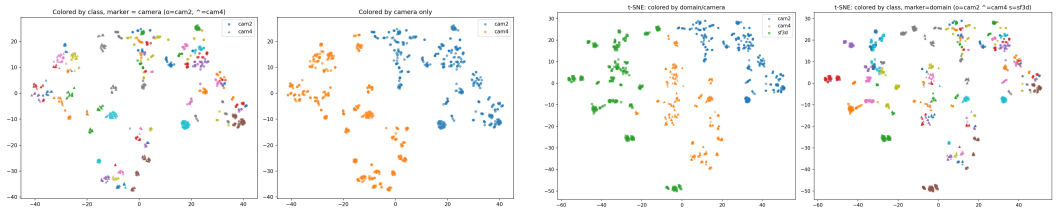
The geometry experiments support non-equivalence between local recoverability and class-relational correspondence, not universal geometric collapse from a change in camera viewpoint. Direct Cam2–Cam4 measurements provide the more focused within-collection comparison and reveal substantial DINOv2 crop correspondence alongside encoder- and treatment-dependent variation. Cross-dataset comparisons remain compound contrasts, while unmatched target populations and reference-dependent weighting limit the interpretation of the direct measurements.

The new direct results have no additional driver-independent RDM reliability analysis or raw-feature rerun in the publication reconstruction. The existing subject-partition control applies only to its specified source–frontal cell. Neither reference reversal nor relational correspondence establishes classifier portability; that functional property requires the separate interventions reported in Appendix F and the additional matched-query controls identified there.



(a) Target-view PCA (b) Three-condition PCA

Figure 21: Archived ResNet50 PCA projections for the two target views and the three observation conditions. Coloring illustrates camera/domain and behavior organization in the projected space; projected distances do not estimate physical camera displacement.



(a) Target-view t-SNE (b) Three-condition t-SNE

Figure 22: Archived ResNet50 t-SNE projections corresponding to the two-view and three-condition comparisons. Local cluster layouts are illustrative and are not measurements of cross-view decision compatibility, physical camera distance, or global linear separability.

H Population, Robustness, and Sensitivity Controls

H.1 Query-population variation under driver-disjoint fitting

The driver-disjoint protocol enforces exact k-per-class support while excluding every remaining image of the selected fitting drivers. Because additional driver groups can be required as the class budget increases, the external query population is not fixed over the recovery trajectory.

Figure 23 makes this population change explicit. For each model, view, and support budget, the plotted range spans the query-image counts realized over the five recorded support draws.

Similar query-count ranges do not imply identical query identities. They instead show that a nominal increase from, for example, $k = 20$ to $k = 100$ can simultaneously change the number of labelled images, the number and identity of fitting drivers, and the people remaining in the query. The nonmonotonic driver-disjoint accuracy curves in Appendix E should therefore not be interpreted as measurements of sample complexity on one fixed evaluation population.

The complete driver-disjoint draw record additionally reports the range of fitting-driver counts associated with every model–view–budget cell. Those counts are part of the population specification rather than uncertainty estimates.

H.2 Input-treatment sensitivity

The primary CNN analysis uses driver crops, but additional source-trained treatments test whether the direct-transfer and classifier-refit pattern is specific to that image construction. The recorded CNN treatments are crop (`bbox`), full-canvas foreground mask (`seg`), and crop-plus-mask (`bboxseg`). Each treatment uses its corresponding source-trained CNN checkpoint.

Table 24 gives the complete direct/refitted comparison for all five CNNs and both target views.

Figures 24 and 25 provide the corresponding visual comparison.

Classifier refitting improves frontal accuracy substantially under all recorded CNN treatments, but the magnitude of both direct transfer and recovery depends on the treatment and architecture. The side-right condition similarly varies across crop and masking choices.

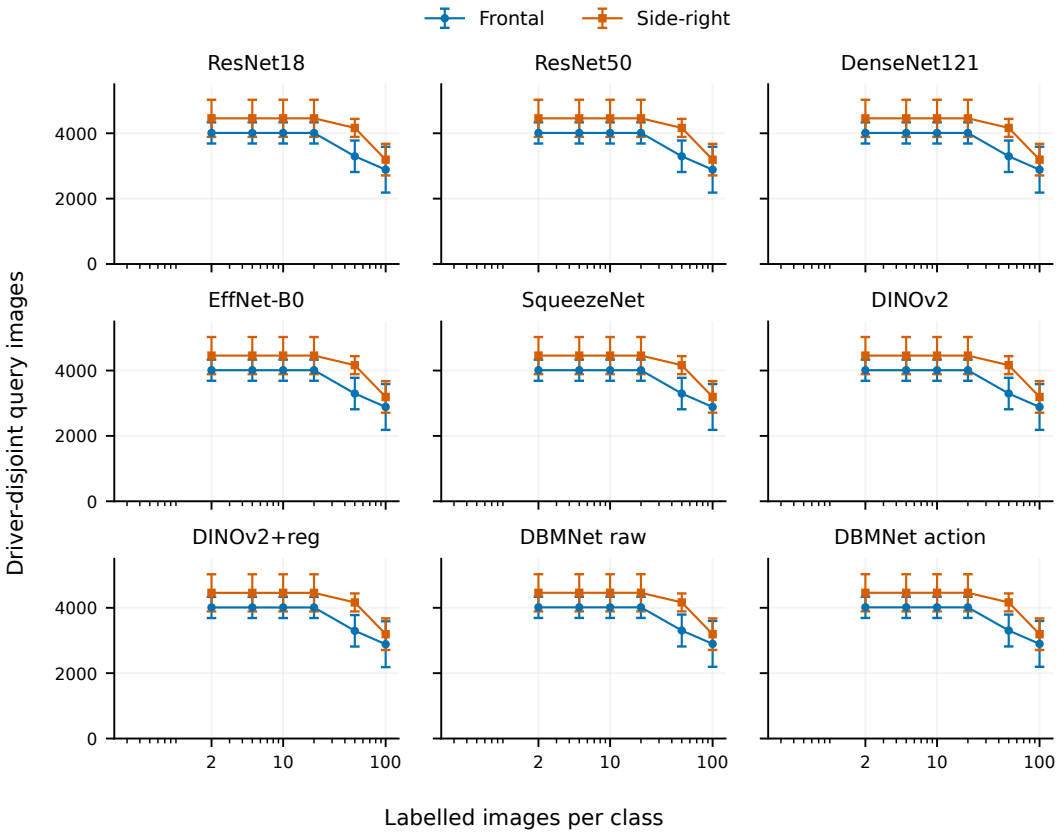


Figure 23: Whole-driver exclusion changes the evaluation population as the class budget grows. Bars span the query-image counts across support draws.

		Frontal		Side-right	
Features	Input	Direct	Refit	Direct	Refit
ResNet18	Crop	10.3	47.1	50.0	73.0
ResNet18	Mask	7.1	46.6	46.6	53.0
ResNet18	Crop+mask	6.4	47.9	52.6	66.8
ResNet50	Crop	6.9	59.4	57.8	78.3
ResNet50	Mask	6.6	52.7	57.2	74.6
ResNet50	Crop+mask	6.4	62.5	60.4	74.9
DenseNet121	Crop	8.8	52.6	55.6	81.5
DenseNet121	Mask	5.9	59.2	49.1	58.0
DenseNet121	Crop+mask	5.8	60.7	57.1	74.8
EffNet-B0	Crop	11.6	54.3	48.1	74.8
EffNet-B0	Mask	6.8	53.2	48.1	73.4
EffNet-B0	Crop+mask	5.4	54.6	52.4	64.4
SqueezeNet	Crop	12.1	48.8	50.2	85.8
SqueezeNet	Mask	8.8	56.4	59.5	66.6
SqueezeNet	Crop+mask	7.3	61.6	53.9	81.1

Table 24: Direct-transfer and target-refitted accuracy (%) for the crop, mask, and crop-plus-mask CNN treatments. Each row uses the source checkpoint matched to that input treatment; these are not three target-time perturbations of one fixed CNN checkpoint.

These differences do not isolate a causal effect of "background." Cropping changes image scale and field of view; segmentation can remove objects, hands, and body evidence together with background; and the CNN checkpoint itself is treatment-specific. The control therefore establishes robustness of the broad recovery phenomenon to several image constructions while retaining the treatment dependence of the exact numerical result.

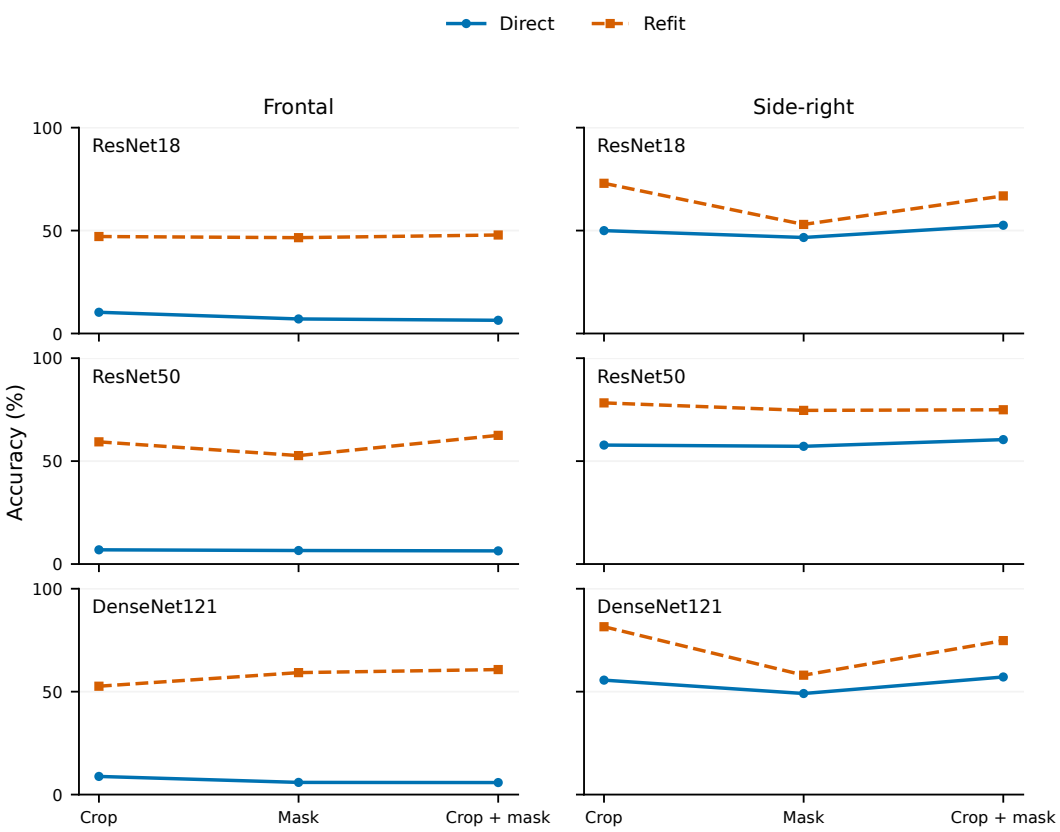


Figure 24: Classifier recovery under cropped, masked, and cropped-and-masked inputs. Each transformation changes the available visual evidence as well as its distribution.

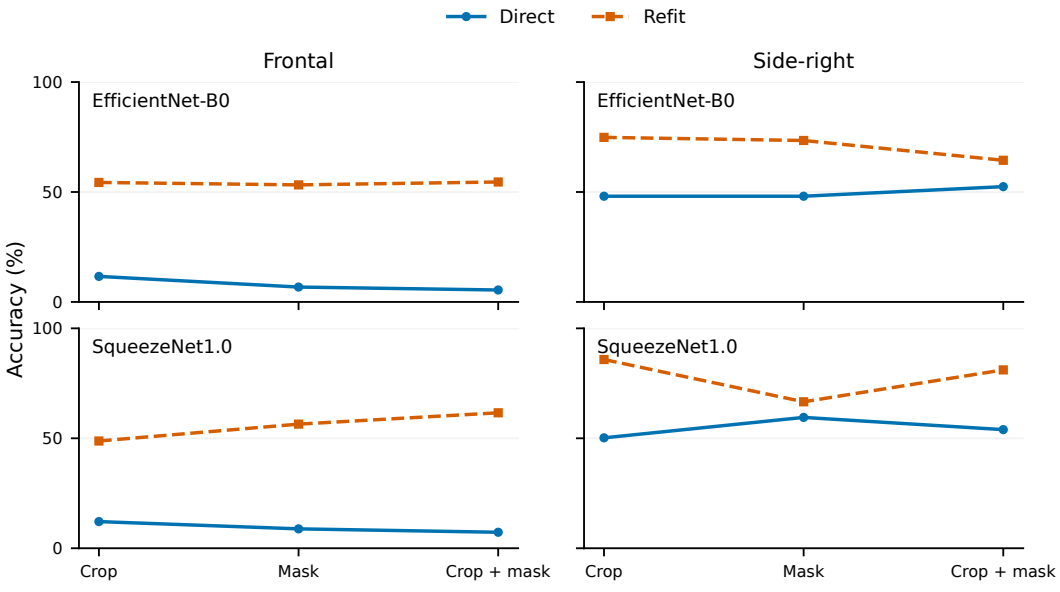


Figure 25: Classifier recovery under cropped, masked, and cropped-and-masked inputs. Each transformation changes the available visual evidence as well as its distribution.

H.3 NOMINAL PERMUTATION SENSITIVITY OF CLASS-RELATIONAL CORRESPONDENCE

The class-relational analysis in Appendix G treats the measured correspondence itself and its reliability reference as the primary quantities. The recorded experiment additionally includes a permutation comparison against unrelated class geometry.

Table 25 reports the twelve observed source–target correspondences together with the recorded nominal permutation probabilities from 10,000 permutations.

Nine of the twelve recorded conditions have nominal $p < 0.05$; the three conditions above that threshold are all crop (`bbox`) comparisons. A small value in this test means that the measured source–target correspondence is distinguishable from the recorded unrelated-geometry null. It does *not* mean that source and target geometry are invariant, nor does a large value establish complete geometric destruction.

The permutation result is therefore subordinate to the effect-size and reliability interpretation. In particular, the subject-partition analysis shows that the apparent within-minus-cross contrast changes substantially when the sampling unit changes. Nominal significance under the image-level construction does not remove that population dependence or isolate camera viewpoint from the cross-dataset contrast.

H.4 DBMNET AS A DIFFERENT SUPERVISION-REGIME CONTROL

DBMNet receives labelled supervision from three 100-Driver cameras and evaluates the fourth as held out (Celona et al., 2026). Figure 26 and Table 26 retain its native 22-class leave-one-camera-out measurements.

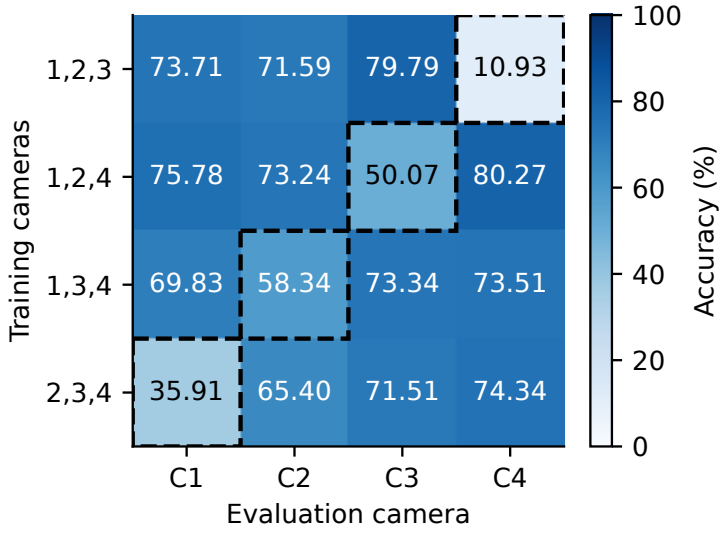


Figure 26: Native 22-class DBMNet accuracy. Rows identify training-camera sets; dashed outlines mark the held-out camera in each row. All cells show measured accuracies, not omitted conditions.

This supervision regime differs fundamentally from the SF3D-only training of the task-trained CNNs. Its native 22-class results are also different from the remapped ten-class local-recovery experiments. The two sets of measurements are therefore retained as contextual evidence rather than merged into one information-matched leaderboard.

The distinction also constrains which downstream analyses are valid. Cam2 and Cam4 DBMNet representations come from different held-out-camera checkpoints, so they cannot be pooled or transported as if they occupied one common feature coordinate system. DBMNet consequently supplies an additional-supervision recovery comparator, not a direct realization of the shared-view compatibility experiment.

DINOv2	Input	View	r_{ST}	Image-half r_{TT}	Nominal p
Base	bbox	F	0.021	0.995	0.3360
Base	bbox	S	0.306	0.962	0.0846
Base	seg	F	0.437	0.994	0.0083
Base	seg	S	0.660	0.988	0.0001
Base	bboxseg	F	0.573	0.991	0.0008
Base	bboxseg	S	0.776	0.989	0.0001
+registers	bbox	F	0.295	0.997	0.0966
+registers	bbox	S	0.591	0.985	0.0033
+registers	seg	F	0.404	0.991	0.0145
+registers	seg	S	0.747	0.988	0.0001
+registers	bboxseg	F	0.555	0.993	0.0024
+registers	bboxseg	S	0.838	0.992	0.0001

Table 25: Recorded source–target class-relational correspondence and nominal permutation probabilities under a 10,000-permutation null of unrelated geometry. These probabilities are supporting sensitivity measurements; they do not replace the observed correspondence or its population-aware reliability analysis.

Training cameras	Test 1	Test 2	Test 3	Test 4
123	73.71	71.59	79.79	10.93
124	75.78	73.24	50.07	80.27
134	69.83	58.34	73.34	73.51
234	35.91	65.40	71.51	74.34

Table 26: Native 22-class DBMNet cross-camera accuracy. Rows identify the three cameras used for training and columns the evaluation camera; the corresponding held-out camera differs by row.

H.5 Spatial attribution as an auxiliary diagnostic

Spatial attribution asks where a fitted CNN assigns evidence in an image, not whether a frozen feature space admits a portable classifier. The recorded Grad-CAM/Score-CAM analysis summarizes normalized attribution weights on an 8×8 grid (Selvaraju et al., 2017; Wang et al., 2020).

For normalized attribution weights p_i, the entropy is

$$H = -\sum_i p_i \log p_i,$$

with maximum $\log 64$ for a uniform grid. Foreground attribution mass is compared with the foreground area available under the corresponding image condition.

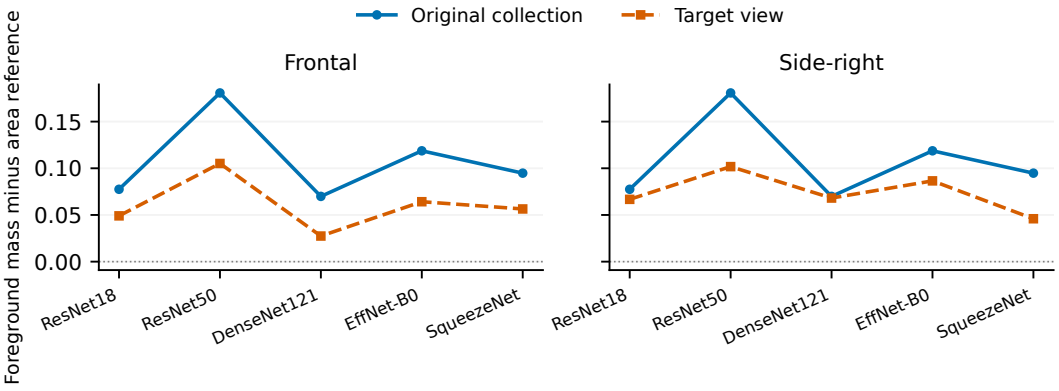


Figure 27: Foreground attribution relative to the available foreground area. Differences characterize spatial evidence allocation, not causal sufficiency of the highlighted region.

The target conditions generally show different foreground-mass allocation from the source condition, but such changes do not establish that the highlighted regions are causally sufficient for classification. Attribution maps can change without identifying which internal feature directions

CNN	View	H_S	H_T	m_S	m_T
ResNet18	F	4.019	4.055	0.527	0.196
ResNet50	F	3.739	3.863	0.630	0.252
DenseNet121	F	4.045	4.066	0.519	0.174
EffNet-B0	F	3.901	4.006	0.568	0.211
SqueezeNet	F	3.945	4.011	0.544	0.203
ResNet18	S	4.019	4.094	0.527	0.301
ResNet50	S	3.739	3.876	0.630	0.336
DenseNet121	S	4.045	4.082	0.519	0.302
EffNet-B0	S	3.901	4.011	0.568	0.320
SqueezeNet	S	3.945	4.045	0.544	0.280

Table 27: Recorded cropped-input attribution summaries. H_S and H_T are source- and target-condition attribution entropies in nats; m_S and m_T are foreground attribution masses.

determine the decision, and saliency methods require independent sanity checks (Adebayo et al., 2018). These measurements are therefore descriptive support for image-evidence change rather than a fourth localization test.

H.6 Prediction concentration does not localize failure

Another recorded auxiliary diagnostic compares recognition accuracy with the fraction of predictions assigned to the most frequent predicted class.

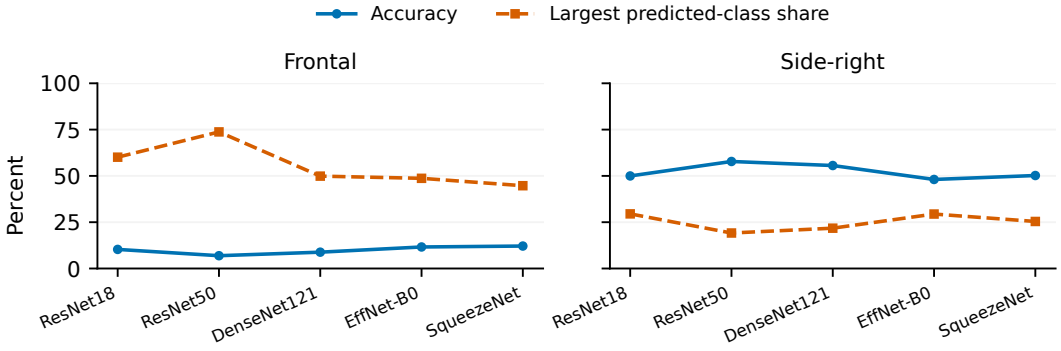


Figure 28: Prediction concentration and accuracy can disagree under the same camera shift. A concentrated output distribution does not uniquely locate the failure.

Highly concentrated predictions can accompany severe transfer failure, but concentration and accuracy need not move together across architectures or views. A collapsed output distribution is therefore a symptom that can help characterize failure, not evidence that the cause lies specifically in the classifier or feature extractor.

This distinction parallels the separation between distribution-shift detection, target-error estimation, and the diagnostic questions studied here (Hendrycks & Gimpel, 2017; Liu et al., 2020; Garg et al., 2022). Prediction concentration is retained as supporting evidence rather than substituted for L, G, or C.

H.7 Calibration-cost association and baseline control

An earlier exploratory analysis asked whether an image-treatment diagnostic could predict the amount of labelled target support required for recovery. The evaluated diagnostic is based on relative cross-condition accuracy loss under the two foreground-masking treatments:

$$S' = 1 - \frac{1}{2}\left(\frac{A^{\text{seg}}_{\text{cross}}}{A^{\text{seg}}_{\text{within}}} + \frac{A^{\text{bboxseg}}_{\text{cross}}}{A^{\text{bboxseg}}_{\text{within}}}\right).$$

Here A denotes classification accuracy, and the superscripts identify full-canvas person masking (`seg`) and crop-plus-mask (`bboxseg`). For CNNs, A^{v}_{within} is the source-trained classifier's source-query accuracy under treatment v, and A^{v}_{cross} is its direct target-query accuracy under the corresponding recorded treatment. For DINOv2, the two values are the source-fitted linear pipeline's recorded source and target accuracies, respectively. Neither ratio uses a target-refitted classifier. The ratios require nonzero denominators and use the same accuracy scale.

The inherited DBMNet proxy is less directly comparable: its numerator is remapped ten-class held-out-camera accuracy under the masking treatment, while its denominator is the median native 22-class full-frame accuracy among leave-one-camera-out checkpoints trained on the evaluation camera. That denominator is shared across masking treatments and differs in checkpoint and label space from the numerator. The DBMNet proxy values are therefore only part of the recorded exploratory association, not a matched within/cross transfer ratio. The raw/action rows share this task-head proxy. These quantities use class labels and observed classification performance; they are not unsupervised estimators of camera displacement or representation damage.

For a recovery curve $A(k)$, direct-transfer accuracy A_0, and historical reference accuracy A_{ref}, the recorded normalized recovery is

$$R(k) = \frac{A(k) - A_0}{A_{\text{ref}} - A_0}.$$

For the ordered budgets $k_j \in \{2, 5, 10, 20, 50, 100\}$, the corresponding response is the trapezoidal normalized area

$$U = \frac{\sum_{j=1}^{m-1} (\log_2 k_{j+1} - \log_2 k_j)\,[R(k_{j+1}) + R(k_j)]/2}{\log_2 k_m - \log_2 k_1}, \qquad m = 6.$$

The recorded normalization is not clipped to $[0, 1]$; $A_{\text{ref}} \neq A_0$ is required. Because A_0 appears directly in this normalization, a diagnostic associated with direct-transfer performance can also become associated with normalized recovery even when it adds little independent information about the calibrated curve.

Figure 29 shows the unadjusted and baseline-adjusted rank relationships.

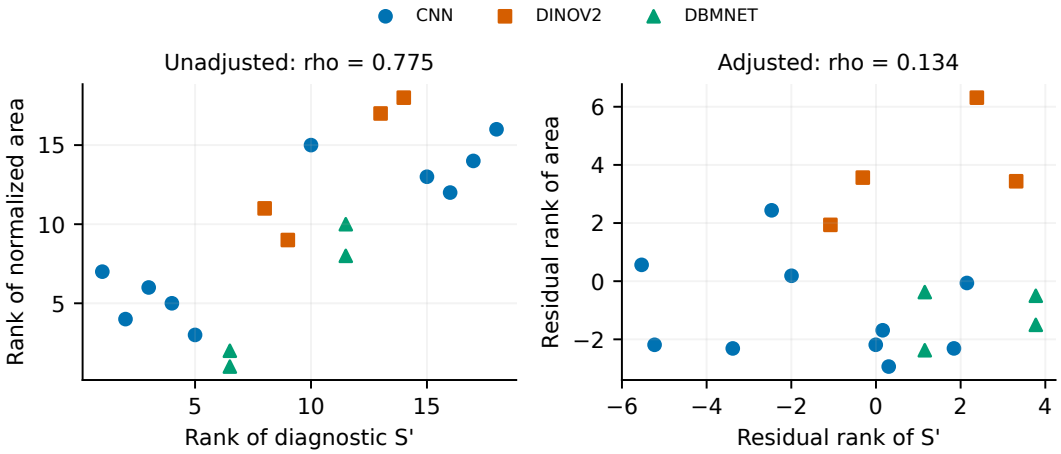


Figure 29: The diagnostic association is strongly attenuated after accounting for initial accuracy. The right panel plots residual ranks, not a new prospective prediction test.

For the driver-disjoint normalized-area response, the unadjusted Spearman coefficient is 0.775, but the partial coefficient after conditioning on direct-transfer accuracy is 0.134. Alternative response definitions likewise produce materially different associations.

The attenuation prevents a strong predictive interpretation. Historical A_{ref} values also come from protocol-specific populations, and all eighteen driver-disjoint curves remain below the historical $R = 0.5$ crossing at $k = 100$; the corresponding crossing budgets are therefore right-censored rather than observed sample-complexity estimates. The present evidence does not establish a population-matched predictor of classifier-recovery cost.

Response	Cells	Spearman	Partial Spearman
Image-disjoint: historical k_{50}	18	-0.473	0.280
Image-disjoint: raw accuracy area	18	-0.799	-0.499
Driver-disjoint: raw accuracy area	18	-0.485	0.236
Driver-disjoint: normalized area	18	0.775	0.134

Table 28: Recorded associations between the exploratory image-treatment diagnostic and alternative recovery responses. Partial Spearman coefficients condition on direct-transfer accuracy. The 18 model–view cells share architectures and training lineages and are not independent dataset replications.

Transferability and target-error estimation remain related research questions (Nguyen et al., 2020; You et al., 2021; Xie et al., 2023). This control is retained because it illustrates the same methodological principle as the main paper: a diagnostic associated with one property of transfer should not be treated as evidence for another property without a direct, appropriately controlled test.

H.8 Scope of the robustness evidence

The controls in this appendix qualify different parts of the empirical argument rather than forming one additional diagnostic axis. Query-population analysis qualifies the interpretation of driver-disjoint recovery; input-treatment experiments test sensitivity to available visual evidence; DBMNet changes the supervision regime; permutation and subject-partition analyses qualify geometry; and attribution, prediction concentration, and calibration-cost analyses characterize auxiliary symptoms or proposed proxies.

None replaces the three central measurements. The manuscript's principal claims continue to rest on local linear recoverability, class-relational organization, and cross-condition decision compatibility measured under their explicitly stated protocols.

I Experimental Reproducibility

I.1 Randomness and repeated trials

Finite-support recovery experiments use five recorded support draws for each model, view, population protocol, and support budget. The driver-disjoint draws use seeds

$$42,\ 43,\ 44,\ 45,\ 46.$$

Each draw performs the complete support-selection and classifier-fitting procedure rather than re-evaluating one fitted classifier under five reporting seeds.

The reported uncertainty for these experiments is the standard deviation across the five support draws. It therefore quantifies sensitivity to the recorded support construction and fitting procedure. It is not a confidence interval over independent drivers, datasets, or model-training runs.

For driver-disjoint fitting, randomness acts jointly on group selection and within-group support sampling. Different draws can consequently use different numbers of fitting drivers and leave different query populations. The query-image and fitting-driver ranges retained with the draw records are part of the reproducibility record rather than auxiliary metadata.

The shared-view experiments likewise retain draw-specific fitted pipelines and predictions. Their reported variation therefore reflects repeated support construction and refitting, while the external query for a given view remains fixed across those fitted classifiers.

I.2 Model and checkpoint provenance

The task-trained CNN representations are tied to registered source checkpoints, feature hooks, feature dimensions, training settings, and input-treatment-specific normalization described in Appendix B. For the primary crop experiment, checkpoint SHA256 prefixes provide compact identities for all five CNNs.

The two DINOv2 representations are tied to the public ViT-S/14 base and register variants and to their recorded checkpoint SHA256 prefixes. The classifier experiments use the 384-dimensional CLS representation from these frozen backbones.

DBMNet retains a different provenance boundary. Its camera-specific leave-one-camera-out configuration and locally recorded evaluation are available, but the supplied evidence does not contain the external repository commit identifier or original checkpoint hashes. These missing fields are reported as unavailable rather than reconstructed retrospectively.

Checkpoint provenance and classifier provenance are kept separate. A frozen feature extractor can be reused by multiple support-specific standardizers and linear classifiers; the latter are identified by their experimental condition, support protocol, budget, and draw rather than by the backbone checkpoint alone.

I.3 Experiment configuration provenance

Each reported result is associated with an explicit combination of

$$(\text{representation}, \text{observation condition}, \text{input treatment}, \text{population protocol}, k, \text{draw}),$$

with additional fields where an experiment requires a source condition, receiving condition, or geometry-reference direction.

For the finite-support study,

$$k \in \{2, 5, 10, 20, 50, 100\},$$

the semantic task remains the same ordered $K = 10$ class set, and the population protocol is recorded separately as image-disjoint or driver-disjoint. Shared-view fits record that k is the total number of labelled examples per class across both target views rather than a per-camera budget.

Geometry records retain the encoder variant, image treatment, target condition, RDM construction, comparison method, and covariance-reference side. This distinction is necessary because the cross-validated distance estimator and the covariance-aware RDM comparison are separate stages of the measurement procedure.

The experimental record therefore treats protocol identity as part of the result itself. Measurements obtained under different support populations or reference partitions are not merged merely because they use the same architecture and target camera.

I.4 Numerical aggregation and independent checks

The publication results are generated from committed numerical exports rather than manually transcribed from figures. The recorded publication pipeline retains an artifact registry containing hashes and dependencies, a table-cell ledger for empirical tables, and numerical selectors for values inserted into manuscript prose.

Several high-level outputs are independently recomputed from lower-level records:

- finite-support means and standard deviations are recomputed from the five recorded support draws;
- the $k = 100$ image-/driver-disjoint population table is derived from the same recovery records rather than maintained as an independent hand-entered table;
- shared-view accuracy and class-macro statistics are reconstructed from saved predictions;
- shared-view confusion matrices are reconstructed from the corresponding draw-specific prediction records;
- evaluation-only no-$c6$ summaries reuse the same fitted classifiers and predictions rather than triggering a new fitting run;
- return-to-source scores reuse the saved target-fitted classifier-side pipelines without refitting on source data.

These checks reduce the risk that manuscript tables, figures, and prose drift apart numerically. They do not constitute independent replications of the underlying model inference; they verify consistency of the recorded experimental outputs.

I.5 Traceability of reported evidence

Table 29 summarizes the level of traceability available for the main evidence families.

The strongest traceability is available for numerical recovery, population, shared-view, return-to-source, confusion, and class-relational tables because their numerical result records are preserved. Model lineage is additionally supported by checkpoint identifiers for the CNN and DINOv2 representations.

The distinction between an experimental rerun and a publication-level reconstruction is maintained throughout. Recomputing a table from saved draw records verifies aggregation; it does not reproduce feature extraction from the original images. Reconstructing a confusion matrix from saved predictions verifies the reported class-conditional outcome; it does not rerun the classifier.

I.6 Archived and non-regenerable visual evidence

Historical ResNet50 PCA/t-SNE panels are retained as archived image artifacts. Their original projected coordinates, exact sample identities, and sampling indices were not present in the supplied reconstruction record. Those panels therefore cannot be reproduced point-for-point from the current experimental package.

They are retained only as qualitative projections. None of the manuscript's quantitative class-relational claims depends on extracting measurements from those archived images. The numerical RDM correspondence records are preserved independently.

Photographic dataset examples are similarly source assets rather than generated experimental outputs. Their role is to illustrate observation conditions and normalized behavior categories, not to supply quantitative evidence.

I.7 Boundary of end-to-end reproducibility

The available reconstruction package supports regeneration and verification of manuscript-level numerical results from recorded experimental outputs. It does not provide a complete end-to-end rerun beginning from raw datasets.

In particular, the supplied reconstruction does not bundle:

- the original raw datasets;
- all trained model checkpoint files;
- the complete feature tensors used by every historical experiment;
- the original coordinate and sampling records for the archived PCA/t-SNE panels.

Consequently, the following claims are intentionally distinct:

$$\text{publication-result reproducibility} \neq \text{raw-data end-to-end experimental reproducibility}.$$

The former is directly supported by the committed numerical records and their aggregation checks. The latter additionally requires the original datasets, model weights, feature extraction environment, and experiment-specific intermediate tensors.

The direct T_2–T_4 geometry has now been incorporated from its own versioned 84-row export, not inferred by combining source–target values. The dedicated regeneration script checks cell completeness, class counts, partition counts, estimator paths, and the absence of recorded fallbacks, then derives the main and appendix tables, numeric prose selectors, and descriptive summaries. The same export is present in both supplied result snapshots, with identical SHA256. The checkpoint-binding qualification for the CNN geometry cache is recorded separately in Appendix B.5; matching

Evidence family	Preserved evidence	Reproducibility level
Finite-support recovery	Per-draw accuracies, budgets, protocols, query counts	Numerically regenerable
Population summaries	Underlying recovery records	Numerically regenerable
Shared-view classifier	Saved predictions and support composition	Numerically regenerable
Confusion matrices	Draw-specific saved predictions	Numerically regenerable
Return to source	Saved fitted pipelines and recorded scores	Numerically traceable
Class-relational geometry	Recorded RDM-comparison results and estimator metadata	Numerically traceable
CNN / DINOv2 lineage	Training configuration and checkpoint identifiers	Checkpoint traceable
DBMNet lineage	Camera split and evaluation records; hashes unavailable	Partially traceable
PCA / t-SNE panels	Archived rendered images only	Not pointwise regenerable
Raw-image end-to-end study	Original datasets/checkpoints/features not bundled	Not supplied

Table 29: Traceability of the principal experimental evidence. "Regenerable" refers to regeneration from the supplied recorded results, not necessarily rerunning raw-image inference or model training.

architecture names are not treated as verified checkpoint identities. This is an export-level and implementation audit, not an independent rerun of the feature-space calculations. Reciprocal target-view classifier transfer, balanced shared support, and fixed-query population controls require separately executed result records.

I.8 Software-environment boundary

The recorded manuscript package includes a Python dependency specification and a recorded environment for publication-level numerical regeneration. The supplied manuscript evidence, however, does not establish a complete historical software and hardware manifest for every original training, feature-extraction, and diagnostic run.

We therefore report implementation settings that are directly recorded—such as the CNN training configuration, classifier-fitting hyperparameters, RDM construction, support seeds, and checkpoint identities. The recorded regeneration environment is stated minimalistically: Python 3.10 with PyTorch 2.7, torchvision 0.22, NumPy 2.2, SciPy 1.15, scikit-learn 1.7, and rsatoolbox 0.3.3; unrecorded system-level details are not inferred for historical runs.

A future end-to-end release should preserve these environment identifiers alongside the raw experiment configuration and checkpoint manifests so that numerical determinism can be assessed independently of publication-level artifact regeneration.

I.9 Reproducibility scope

The reproducibility record is aligned with the scientific claims of the paper. It preserves the distinctions among observation condition, population protocol, feature identity, classifier intervention, geometry estimator, and evaluation population that determine the meaning of each reported number.

This is particularly important for the central diagnostic decomposition: reproducing a high local recovery score does not reproduce a geometry measurement, and reproducing either does not establish classifier compatibility across conditions. Each quantity must be reconstructed from the experiment and evaluation contract that defines it.