Frozen Vision Foundation Models Narrow the Sensor-Shift Gap in Non-Visual Spectrogram Recognition

Anonymous authors
Paper under double-blind review

Abstract

WiFi-based micro-Doppler activity recognition offers a privacy-preserving approach to human sensing, but models trained under one RF acquisition setting often perform poorly when tested with different sensors, environments, or datasets. A key challenge is that micro-Doppler spectrograms contain both activity-related motion patterns and unwanted variation from the sensing and processing pipeline. Activity information includes temporal rhythm, Doppler spread, and time-frequency motion structure, while unwanted variation may arise from room clutter, multipath propagation, sensor response, preprocessing, rendering choices, and dataset-specific artefacts. Existing approaches often address this problem by reducing the mismatch between source, synthetic, and target recordings. This paper investigates whether external WiFi micro-Doppler recognition is better addressed by learning stable features than by adapting input spectrograms toward the target appearance.

We investigate this idea using self-supervised vision encoders pretrained on natural images and kept frozen during RF training. All models are trained on WiFi micro-Doppler data collected in our lab and evaluated on two external settings: OPERAnet, a public multimodal RF benchmark collected outside our lab, and a through-wall test set collected by adding a wall condition to the same lab setup. Despite using an encoder that is not fine-tuned on RF data, DINOv2-S/14 with a lightweight MLP classifier substantially improves recognition on OPERAnet, increasing class-restricted macro-F1 from 0.20 ± 0.01 to 0.46 ± 0.01 compared with the strongest convolutional baseline trained on the same lab-collected data. On the through-wall test set, DINOv2 remains competitive but is not uniformly superior, suggesting that pretrained representations are most beneficial when the test data differ more strongly from the training data. These results suggest that external RF activity recognition should be treated not only as a problem of adapting input appearance, but also as a problem of learning representations that preserve activity-relevant motion structure while being less sensitive to sensor, environment, and dataset-specific variation.

1 Introduction

WiFi micro-Doppler sensing is attractive for human activity recognition because it can capture motion without cameras or wearable devices. In controlled settings, CNN-based classifiers can fit spectrogram activity classes well. The harder problem is evaluation under new acquisition conditions: a model trained in one room, with one sensor placement and preprocessing pipeline, can fail when the room, propagation path, sensor placement, rendering pipeline, or dataset source changes.

The challenge is not only ordinary overfitting. A micro-Doppler spectrogram entangles activity-relevant motion structure, such as temporal rhythm, Doppler spread, duration, and time-frequency energy shape, with acquisition-specific appearance from room clutter, multipath propagation, sensor response, preprocessing, colour mapping, and dataset-specific rendering. A model can therefore fit source data by using cues that are predictive inside one collection but do not transfer when the acquisition setting changes.

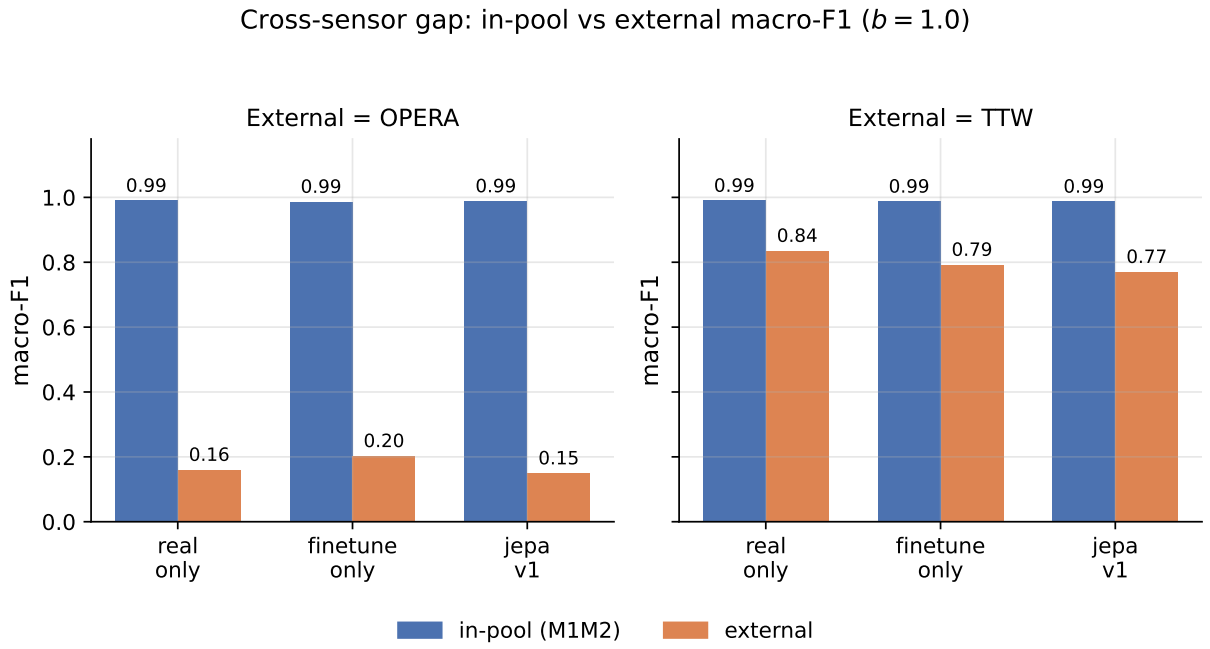

Figure 1: **Evaluation gap under unseen acquisition settings.** Three RF-specific baselines trained on lab-collected M1M2 spectrograms achieve high macro-F1 on a held-out M1M2 test split but degrade under held-out acquisition settings. "Real-only" is a supervised CNN trained only on real M1M2 data; "finetune-only" uses synthetic pretraining followed by supervised fine-tuning on M1M2; "JEPA-v1" uses JEPA-style self-supervised pretraining followed by supervised fine-tuning. The drop is largest on OPERA, a public benchmark collected outside our lab, and smaller on TTW, a through-wall setting collected using the same lab setup with a wall condition added.

We study this problem using two lab-collected source pools, M1 and M2, whose combined training set is denoted M1M2. Models are trained and selected only on this lab-collected source data, then evaluated on two held-out acquisition settings: OPERA, a subset of the public OPERAnet RF benchmark collected outside our lab, and TTW, a through-wall setting collected using the same lab setup with a wall condition added. OPERA therefore represents a stronger dataset and acquisition shift, while TTW represents a more controlled propagation shift.

Figure 1 illustrates the gap. A supervised CNN trained only on real M1M2 data ("real-only"), a CNN with synthetic pretraining followed by supervised fine-tuning on M1M2 ("finetune-only"), and a JEPA-style self-supervised variant followed by fine-tuning ("JEPA-v1") all achieve high in-pool macro-F1. Their external performance drops sharply on OPERA and less severely on TTW. This suggests that the bottleneck is not source-set fitting capacity alone, but whether the representation preserves activity-relevant time-frequency structure when acquisition appearance changes.

Much of the RF sensing literature addresses this mismatch by making source or synthetic spectrograms better aligned with the target setting. Simulation-based pretraining, style-transfer augmentation, adversarial adaptation, and RF-specific self-supervised learning all attempt to reduce the gap between training and test spectrograms Vishwakarma et al. (2022); Gatys et al. (2015); Zheng et al. (2019); Vishwakarma et al. (2021); Assran et al. (2023). We ask a complementary question: is external RF recognition better improved by adapting inputs toward the target appearance, or by using representations that are more transferable under acquisition shift?

We test this question with frozen vision foundation models. DINOv2 Oquab et al. (2024), CLIP Radford et al. (2021), and MAE He et al. (2022) are pretrained on natural images, not RF measurements; their pixels originally represent photographic appearance rather than signal energy over time and frequency. Strong

transfer to RF spectrograms is therefore not obvious. At the same time, these encoders may provide generic structural priors over shape, local contrast, continuity, texture, and global layout. If such priors transfer, frozen vision representations provide a direct test of whether external RF recognition benefits from stable feature spaces rather than only from input-level adaptation. Nevertheless, a frozen DINOv2-S/14 encoder with a lightweight MLP head improves class-restricted macro-F1 on OPERA from 0.20±0.01 for the strongest CNN baseline to 0.46±0.01, while remaining competitive but not uniformly superior on TTW. Matched CLIP and MAE probes also transfer non-trivially, but DINOv2 preserves more same-class neighbourhood structure in frozen feature space before any head training. Finally, a factorised generative augmentation baseline does not improve the frozen DINOv2 probe, suggesting that input-level augmentation alone does not solve this external-transfer regime.

The paper makes the following contributions:

1. We present a strict external-holdout study of WiFi micro-Doppler recognition on OPERA, a public RF benchmark collected outside our lab, and TTW, a held-out through-wall setting. The external pools are not used for training, validation, checkpoint selection, hyperparameter selection, or backbone selection.
2. We show that frozen natural-image encoders transfer non-trivially to RF spectrograms under this protocol. DINOv2-S/14 with a lightweight MLP head gives the strongest result on OPERA, while TTW shows that the gain is smaller when the target setting is closer to the source acquisition setup.
3. We diagnose why the gain occurs and where it does not. Class-restricted evaluation, feature-space distances, histogram statistics, and k-NN/retrieval probes show that DINOv2 preserves more activity-relevant neighbourhood structure across acquisition shift, while a strong generative environment-transfer baseline does not improve the frozen DINOv2 probe.

2 Related Work

Generalisation in WiFi and RF activity recognition. WiFi and RF micro-Doppler sensing have been widely studied for device-free human activity recognition, often by rendering motion as time-frequency spectrograms and training CNN-based classifiers Wang et al. (2015; 2017); Zheng et al. (2019). These models can perform well when training and test data share an acquisition setting, but performance often degrades when the room, sensor placement, hardware response, propagation path, preprocessing pipeline, or dataset construction changes. This failure is not only a classification problem: micro-Doppler spectrograms mix activity-related motion structure with acquisition-specific appearance. A model may therefore learn cues that fit the source collection but do not transfer reliably to new acquisition settings.

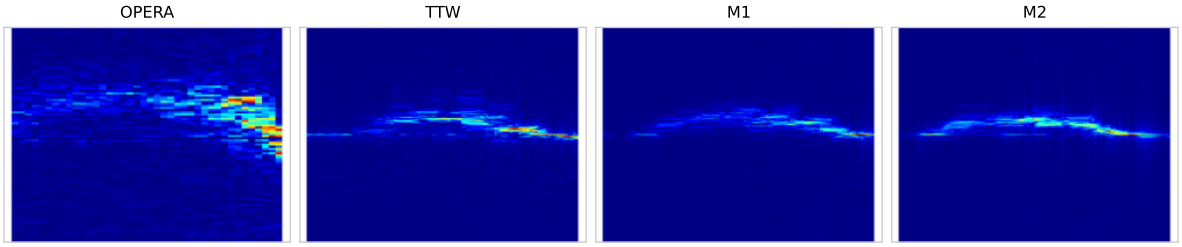

Figure 2: **Same activity, different acquisition appearance.** Representative walking-forward spectrograms from OPERA, TTW, M1, and M2 show that activity-related time-frequency structure is mixed with dataset- and acquisition-specific rendering differences. This motivates evaluating whether models preserve activity cues when acquisition appearance changes.

Input-level adaptation and generative augmentation. A common strategy is to reduce the mismatch between source, synthetic, and target spectrograms. Prior work has explored physics-grounded simulation and synthetic pretraining Vishwakarma et al. (2022), GAN-based noise or environment modelling Vishwakarma

et al. (2021), style-transfer augmentation Gatys et al. (2015), adversarial adaptation Zheng et al. (2019), semi-supervised adaptation with radar features Li et al. (2022), and JEPA-style self-supervision Assran et al. (2023). More general content-style and factorised generative models, including MUNIT Huang et al. (2018), FactorVAE Kim & Mnih (2018), StarGANv2 Choi et al. (2020), and adversarial or MMD-based regularisation Gretton et al. (2012); Ganin & Lempitsky (2015), provide related tools for separating task content from domain style. In RF spectrogram recognition, these methods correspond to adapting source or generated spectrograms toward a target acquisition condition. Our work tests whether stable pretrained representations provide an alternative to relying only on input-level matching.

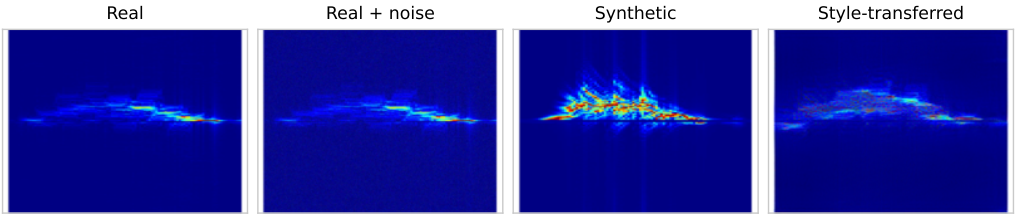

Figure 3: **Input-level adaptation examples.** Representative real, noise-perturbed, synthetic, and style-transferred spectrograms illustrate the family of input-space strategies used to increase appearance diversity or reduce source-target mismatch. The experiments in this paper ask whether such input-level changes are sufficient, or whether the feature representation itself is the main bottleneck under held-out acquisition shift.

Vision foundation models as frozen feature extractors. Large pretrained vision encoders such as CLIP Radford et al. (2021), DINO Caron et al. (2021), DINOv2 Oquab et al. (2024), MAE He et al. (2022), and I-JEPA Assran et al. (2023) provide strong frozen representations for many natural-image tasks. Prior probing studies Oquab et al. (2024); Darcet et al. (2024) show that such features can support classification, retrieval, segmentation, and depth estimation using linear or lightweight MLP heads. These models have also been used beyond standard RGB images, including medical imaging Zhang & Shen (2024) and satellite imagery Reed et al. (2023); Cong et al. (2022). WiFi micro-Doppler spectrograms provide a severe transfer test because their pixels represent signal energy over time and frequency rather than optical reflectance. They are therefore useful for testing whether natural-image representations encode structural priors that survive in non-photographic sensing data; strong transfer is not implied by natural-image performance alone.

Foundation models for wireless sensing. Recent work has begun to explore foundation-model ideas in wireless sensing. LVM4CSI Guo et al. (2025) studies the reuse of large vision models for CSI-based sensing tasks such as activity recognition and localisation, while MMSense Li et al. (2025) and WavesFM Aboulfotouh & Abou-Zeid (2025) explore broader foundation-model directions, including multimodal fusion and native wireless pretraining. Our paper asks a narrower question: whether frozen natural-image representations transfer to WiFi micro-Doppler spectrograms when the target acquisition setting is held out entirely. We therefore compare multiple frozen vision encoders with RF-specific and generative baselines under the same no-target-use protocol.

Comparability to prior RF and OPERA results. The comparison in this paper is not a claim of best reported accuracy on OPERAnet under any protocol. It is a protocol-specific external-transfer claim: models are trained and selected on lab-collected M1M2 data, OPERA is never used for training or model selection, and performance is reported with the class-restricted label set defined in Sec. 3.4. Many prior RF and wireless-sensing papers answer different questions, for example in-domain recognition, adaptation with access to target data, or foundation-model reuse on CSI rather than micro-Doppler spectrograms. Table 1 summarises the distinction. Under this reading, the relevant comparison is not against an OPERA leaderboard number, but against methods evaluated under the same source-only external-holdout constraint.

Evaluation under held-out acquisition settings. Testing external recognition requires separating ordinary held-out testing from recognition under acquisition shift. RF datasets can differ not only in acquisition conditions, but also in label sets and dataset composition. Prior work on domain generalisation

Table 1: **Protocol comparability of related RF and wireless-FM work.** The present paper evaluates M1M2-trained models on OPERA/TTW without using target data for training, validation, checkpoint selection, or backbone selection. Prior work is important context, but most entries are not direct baselines for this exact external-holdout protocol.

Work	Modality / task	Target dataset	Target data used during training or tuning?	Reported protocol	Directly comparable?
OPERAnet Bocus et al. (2022)	Multimodal RF / activity data	OPERAnet	Dataset-specific splits	Dataset introduction and benchmark setting	No, not M1M2→OPERA external transfer
Widar-style adaptation Zheng et al. (2019)	WiFi sensing / activity recognition	WiFi domains	Adaptation setting may use target-domain structure	Domain-adaptation framing	No, target access differs
SimHumalator Vishwakarma et al. (2022) and GANNoise Vishwakarma et al. (2021)	Radar / RF spectrogram augmentation	Synthetic plus real RF pools	Synthetic or generated data used for pretraining/augmentation	Input-level simulation or augmentation	Partially, reproduced here as RF-specific baselines where possible
Semi-supervised radar adaptation Li et al. (2022)	Radar HAR	Radar domains	Uses adaptation data	Semi-supervised adaptation	No, target access differs
LVM4CSI Guo et al. (2025)	CSI sensing with large vision models	CSI tasks	Task-specific protocol	Frozen/large-vision-model reuse for CSI	Adjacent, but not micro-Doppler OPERA/TTW transfer
MMSense Li et al. (2025); WavesFM Aboulfotouh & Abou-Zeid (2025)	Wireless foundation models	Wireless sensing tasks	Native or multimodal wireless training	Foundation-model construction or fusion	Adjacent, not the same frozen natural-image probe

has shown that benchmark design and model-selection choices can strongly affect conclusions Gulrajani & Lopez-Paz (2021). We therefore use held-out acquisition settings that are not used for training, validation, hyperparameter tuning, checkpoint selection, or backbone selection. We also use class-restricted macro-F1 when the held-out target does not contain the full training label set. The next section defines these data pools and evaluation rules.

3 Data and Evaluation Setup

3.1 Spectrogram image representation

All samples are WiFi micro-Doppler spectrograms represented as images. Each spectrogram is first rendered as a single-channel time-frequency map. It is then converted to an RGB PNG using a fixed colour map, cropped to a constant time-frequency window, and resized to 224×224 at model input time. This gives all methods the same image-shaped input and allows us to compare CNN baselines and frozen vision encoders within a common pipeline.

For DINOv2, CLIP, and MAE, we use the input normalisation required by each pretrained encoder. For CNN baselines, we follow the normalisation used in their original training recipe.

3.2 Source, synthetic, and held-out evaluation pools

We organise the data into pools according to how and where they were collected. The six source activity classes are Body Rotation, Picking Up, Sitting Down, Standing Up, Walking Backward, and Walking Forward.

- **Lab-collected source pools.** M1 and M2 are two WiFi micro-Doppler collections acquired in our lab using the source acquisition setup. Unless otherwise stated, models are trained on the combined M1+M2 pool, denoted *M1M2*. These pools provide the labelled source data used for training, validation, and in-pool testing.
- **Synthetic pretraining pools.** S is a physics-based micro-Doppler simulation pool generated following SimHumalator Vishwakarma et al. (2022). $SE1E2$ is a synthetic pool whose low-level appearance is matched to $M1 \cup M2$ using neural style transfer Gatys et al. (2015). These pools are used only in synthetic pretraining or augmentation experiments.

Table 2: **Per-pool sample counts.** M1M2 is the primary source pool. S and $SE1E2$ are synthetic pools used for pretraining or augmentation. OPERA and TTW are held-out evaluation pools governed by the rules in Sec. 3.5.

Pool	Total	Train (b=1.0)	Val	Test	Train (b=0.25)	Role
M1	500	349	50	101	89	lab source
M2	322	223	33	66	58	lab source
M1M2	799	559	80	160	142	**primary source**
S	1,105	773	109	223	—	synthetic pretraining
$SE1E2$	2,210	1,545	221	444	—	style-matched synthetic pretraining
OPERA	373	—	—	373	—	held-out benchmark
TTW	69	—	—	69	—	held-out through-wall

- **Held-out evaluation pools.** OPERA is a subset of the public OPERAnet multimodal RF/vision dataset Bocus et al. (2022), collected outside our lab and used as a held-out benchmark. TTW is a through-wall evaluation pool collected using the same lab acquisition setup as M1 and M2, but with a wall condition added.

The two held-out evaluation pools represent different degrees of acquisition shift. OPERA differs from the lab-collected source data in acquisition environment, dataset construction, and processing pipeline. TTW is closer to the source data because it uses the same lab acquisition setup, but it introduces a through-wall propagation condition. The pools also differ in class coverage: Picking Up is absent from OPERA, while TTW contains all six source classes.

This benchmark is intentionally modest rather than foundation-scale. The primary source pool M1M2 contains 799 spectrograms in total, with 559 training samples at the full training-data budget. The held-out evaluation pools contain 373 OPERA samples and 69 TTW samples. This scale is large enough to expose a reproducible external evaluation gap, but not large enough to support deployment-level claims. We therefore treat the study as an external-representation analysis rather than a complete RF activity-recognition system.

3.3 Splits, data budgets, and model selection

For the lab-collected source pools, we use a stratified 70/10/20 train/validation/test split per class. The training split is used to fit the model, the validation split is used for model selection, and the source test split is used for in-pool evaluation. Main trained-model results are reported over seeds $\{0, 42, 7\}$ as mean $\pm$ standard deviation. Frozen-feature diagnostics use fixed balanced samples and state the diagnostic sampling rule where applicable.

We report results at two source training-data budgets. The full-budget setting, b=1.0, uses the full M1M2 training split with 559 spectrograms. The low training-data budget, b=0.25, uses a per-class stratified subset with 142 spectrograms. The low-budget setting reflects the practical difficulty of collecting calibrated RF micro-Doppler data at scale. Exact per-pool counts are shown in Table 2.

3.4 Evaluation metrics and label-set mismatch

The primary metric is class-restricted macro-F1 on the held-out evaluation pools. This is the unweighted mean of per-class F1 over the activity classes present in the target pool. OPERA contains five of the six source classes because Picking Up is absent, while TTW contains all six classes. Class-restricted macro-F1 therefore avoids assigning an artificial zero to a class that is not present in the target data.

We also report full six-class macro-F1 when auditing the effect of evaluation policy. This distinction is important because a missing target class can depress macro-F1 even when the model is being evaluated on a pool where that class never occurs. Class-restricted evaluation is not used to cherry-pick favourable classes; it is determined by the class set present in the held-out target pool.

3.5 Holdout rules

All learning curves, early stopping decisions, hyperparameter choices, checkpoint selection, and backbone selection use only the source training and validation splits. OPERA and TTW are reserved for final held-out evaluation. We do not use them for training, validation, hyperparameter tuning, checkpoint selection, backbone selection, early stopping, per-target fine-tuning, test-time supervision, or target-specific class filtering to improve results. Class-restricted macro-F1 is computed only from the class set present in each held-out pool and is used to avoid penalising methods for classes that are absent from the target data.

3.6 Motivating evaluation gap

Figure 1 quantifies the evaluation gap for three CNN-based baselines trained on M1M2 at $b=1.0$. Here the gap means the difference between performance on the held-out M1M2 source test split and performance on the held-out OPERA or TTW evaluation pool. All three baselines achieve high source-test macro-F1 but degrade substantially on OPERA. The drop is smaller on TTW, consistent with TTW being collected using the same lab acquisition setup with an added wall condition, while OPERA is collected outside our lab with a different acquisition environment and dataset pipeline. This gap motivates the experiments that follow: we ask whether better performance under unseen acquisition settings comes mainly from correcting input appearance, or from using representations that preserve activity-relevant structure across acquisition shift.

4 Diagnosing the External Evaluation Gap

Before testing representation-transfer and augmentation methods, we first analyse the external evaluation gap. The goal is not to prove a single causal mechanism, but to identify which parts of the failure are visible before task-specific training and which parts come from the evaluation protocol itself. We use four diagnostics: frozen-feature distances, feature-space visualisation, luminance statistics, and an evaluation-policy audit. Together, these analyses show that the external gap has both an acquisition/rendering component and an evaluation-policy component.

4.1 Pool distances in a frozen feature space

We first ask whether the training and external pools are separated in a generic visual feature space. For each pool, we extract 2048-dimensional features from the training split using a frozen ImageNet-pretrained ResNet-50 and compute pairwise RBF-MMD Gretton et al. (2012) between pools (Fig. 4).

Two observations are important. First, OPERA is the most distant external pool from the data used for training, which is consistent with it being the hardest external target. Second, $SE1E2$ is closer to M1M2 than the purely synthetic pool S. Therefore, if synthetic pretraining is helpful, $SE1E2$ should be the more favourable synthetic partner. We test this in Sec. 5.

4.2 Pool identity versus activity identity

We next examine whether frozen features separate samples more strongly by activity class or by collection pool. We project the same frozen ResNet-50 features using UMAP McInnes et al. (2018), colouring samples by activity class and by pool identity (Fig. 5). Class structure is visible within pools, but pool identity is more clearly separated. This suggests that, even before any task-specific training, generic visual features capture acquisition-related differences more strongly than activity-related structure.

This does not mean that the activity signal is absent. Rather, it shows that activity-related motion structure is entangled with stronger pool-specific appearance cues. This supports the central motivation of the paper: external recognition requires features that keep activity-relevant motion information while reducing sensitivity to acquisition-specific variation.

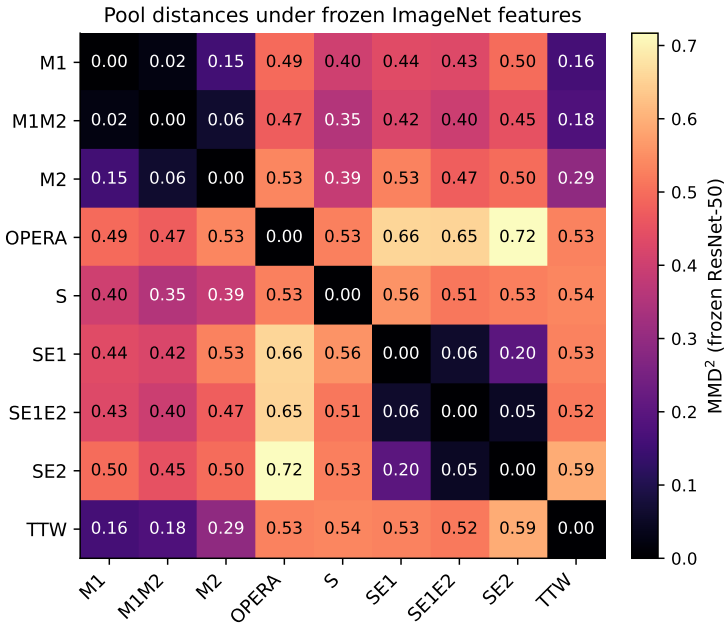

Figure 4: **Pool-to-pool RBF-MMD distances under frozen ImageNet ResNet-50 features.** OPERA is far from the lab-collected training pools (M1, M2) and from the synthetic pools (S, $SE1E2$). TTW is closer to M1M2, consistent with its smaller external evaluation gap in Fig. 1. $SE1E2$ is closer to M1M2 than S, motivating its use as the style-matched synthetic pretraining partner in later experiments.

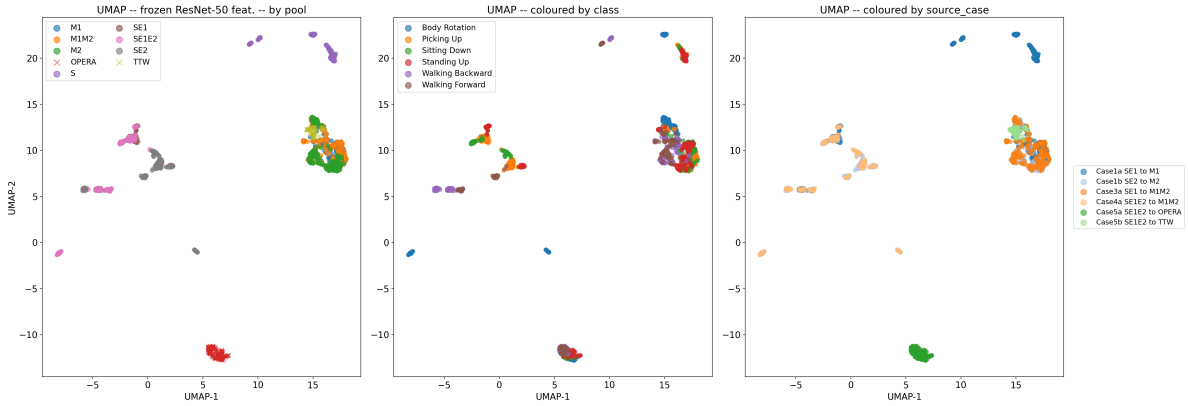

Figure 5: **Frozen ResNet-50 feature UMAP**, coloured by activity class (left) and source pool (right). Pool identity is more separable than activity identity at the frozen-feature level.

4.3 Low-level rendering differences

We then ask whether part of the external gap comes from simple low-level image statistics. For each image, we compute a 256-bin luminance histogram and compare pool-wise median histograms (Fig. 6). OPERA shows a clear shift relative to M1M2, consistent with differences in rendering, colour mapping, and global brightness. TTW is closer to M1M2 but still offset.

This diagnostic is important because it shows that some of the external shift is not semantic. It is visible in low-level image appearance. We therefore include a simple test-time histogram-matching intervention in Sec. 5 to measure how much of the recognition gap can be reduced by correcting this appearance shift.

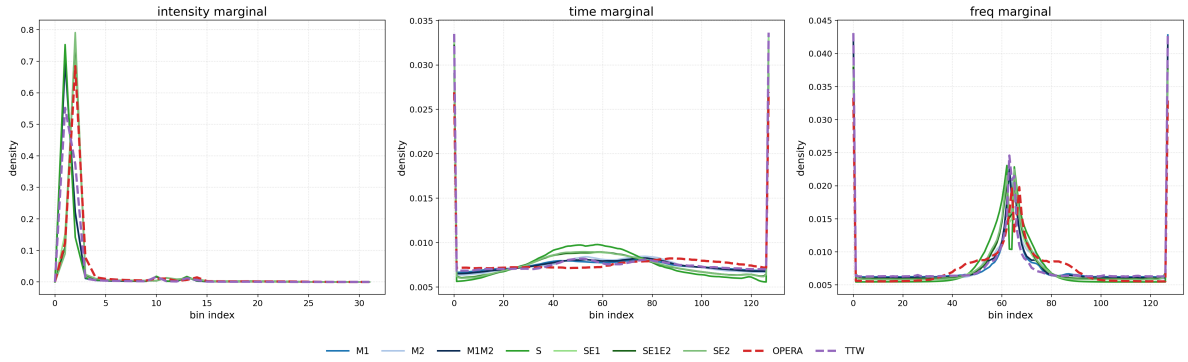

Figure 6: **Per-pool luminance marginals.** OPERA shows a rendering-induced shift relative to M1M2; TTW is closer but still offset. A single-pass CDF histogram match at test time reduces this offset; we quantify its effect in Sec. 5.1.

Table 3: **Effect of evaluation policy on OPERA.** OPERA external macro-F1 at $b=1.0$ under the naive full-six-class metric and the class-restricted five-class metric.

Method	Full (6)	Restricted (5)	Δ
real-only CNN	0.11	0.16	+0.05
finetune-only CNN	0.14	0.20	+0.06
JEPA-v1 CNN	0.15	0.29	+0.14

4.4 Effect of label-set mismatch

External evaluation can also be affected by label-set mismatch. OPERA does not contain the Picking Up class. If we report full six-class macro-F1 on OPERA, the missing class is assigned a per-class F1 of 0.00, which depresses the aggregate score even though the class is not present in the target pool. Table 3 shows the size of this effect. Switching from full six-class macro-F1 to class-restricted five-class macro-F1 increases reported OPERA performance, with the largest change for the JEPA baseline.

This correction does not make the recognition problem disappear, but it prevents the metric from penalising models for a class that is absent from the target data. For this reason, all main held-out results use class-restricted macro-F1, with full six-class macro-F1 reported only when auditing the evaluation policy.

Together, these diagnostics show that the evaluation gap is not caused by a single factor. OPERA is farther from the lab-collected source data than TTW; frozen features separate samples more strongly by pool than by activity; part of the shift is visible in low-level luminance statistics; and incomplete label overlap can distort macro-F1 if not handled carefully. None of these findings alone solves external recognition, but together they motivate the next section, where we test whether the gap is better reduced by correcting input appearance and synthetic-data mismatch, or by using frozen representations that preserve activity-relevant structure under acquisition shift.

5 Interventions and Main Results

The diagnostic study in Sec. 4 showed that the external evaluation gap, defined in Sec. 3.6 as the drop from held-out M1M2 source-test performance to held-out OPERA/TTW performance, contains both low-level acquisition/rendering differences and a harder representation-transfer problem. We therefore compare three kinds of change: input normalisation, synthetic-data alignment, and frozen representation transfer.

Input normalisation: test-time histogram matching. Each external test image is remapped using a 256-bin luminance CDF estimated from 300 lab-collected training images. This changes only the external test image appearance and introduces no learned parameters.

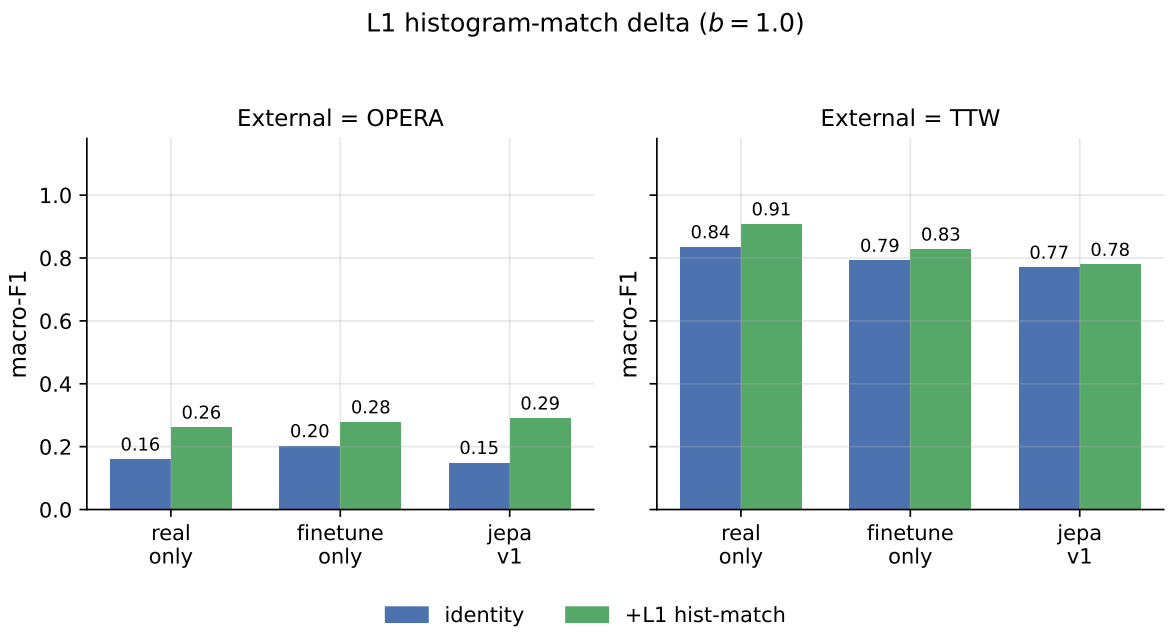

Figure 7: **Effect of test-time histogram matching.** External macro-F1 before and after applying the CDF histogram match to external test images for three CNN baselines at $b = 1.0$. Histogram matching improves OPERA more than TTW, consistent with the larger low-level appearance shift observed for OPERA.

Synthetic-data alignment: style-matched pretraining. CNN baselines are pretrained on $SE1E2$ instead of the purely synthetic pool S, motivated by the smaller frozen-feature distance between $SE1E2$ and M1M2 in Sec. 4.1.

Representation transfer: frozen DINOv2 probe. The CNN backbone is replaced with a frozen DINOv2-S/14 encoder pretrained on natural images. Only a lightweight MLP classifier is trained on the lab-collected RF data.

Class-restricted evaluation is used for the main external results, as defined in Sec. 3.4. It is not treated as a performance intervention, but as the required evaluation policy when an external target does not contain all source classes.

5.1 Input normalisation: correcting low-level appearance at test time

The luminance analysis in Sec. 4.3 showed that OPERA differs from the lab-collected training data in low-level appearance. We therefore test a simple parameter-free correction: CDF histogram matching from each external test image to a reference CDF computed from lab-collected training images. This is applied only at inference time and does not change the trained model.

Figure 7 shows that histogram matching improves CNN performance on OPERA by up to 14 percentage points. The gain on TTW is smaller, consistent with TTW being closer to the lab-collected source data in luminance statistics. This confirms that part of the M1M2-to-external gap is caused by low-level appearance differences. However, histogram matching does not close the gap, which suggests that correcting input appearance alone is insufficient.

Table 4: **Effect of synthetic pretraining partner.** External class-restricted macro-F1 at $b = 1.0$ when CNN baselines use the purely synthetic pool S or the style-matched synthetic pool $SE1E2$.

Method	Target	S	$SE1E2$
finetune-only CNN	OPERA	0.20	0.20
finetune-only CNN	TTW	0.79	0.79
JEPA-v1 CNN	OPERA	0.15	0.16
JEPA-v1 CNN	TTW	0.77	**0.80**

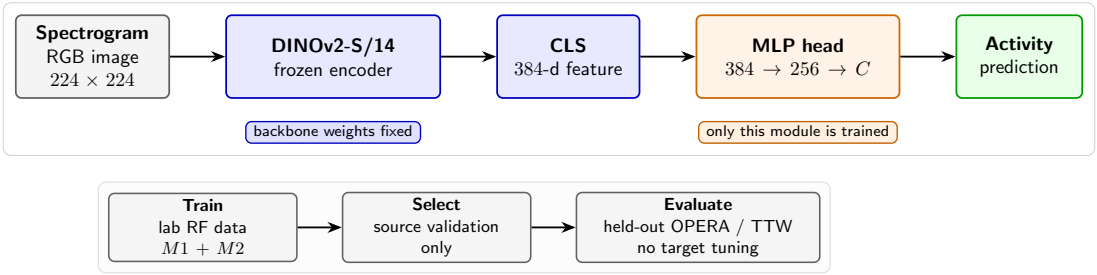

Figure 8: **Frozen DINOv2 probe.** Input spectrograms are resized to 224×224 and passed through a frozen DINOv2-S/14 encoder. Only a lightweight $384 \to 256 \to C$ MLP classifier is trained on the lab-collected RF data.

5.2 Synthetic-data alignment: changing the pretraining partner

The pool-distance diagnostic in Sec. 4.1 showed that the style-matched synthetic pool $SE1E2$ is closer to the lab-collected training pool than the purely synthetic pool S. We therefore test whether replacing S with $SE1E2$ improves CNN-based external recognition.

Table 4 shows that choosing the closer synthetic pool produces only small gains. The effect is most visible for the JEPA variant, where $SE1E2$ improves TTW by 0.03 F1 and OPERA by 0.01 F1. The finetune-only CNN is essentially unchanged. This result suggests that style-matched synthetic pretraining can help, but its effect is much smaller than the representation change evaluated next.

5.3 Representation transfer: frozen DINOv2 probe

We next test whether a stronger frozen representation can improve external recognition. Each 224×224 RGB spectrogram is passed through a frozen `vit_small_patch14_dinov2` encoder Oquab et al. (2024). The encoder parameters are locked throughout RF training. Only a lightweight MLP classifier is trained on the lab-collected RF data, using the 384-dimensional CLS embedding as input.

The MLP head has dimensions $384 \to 256 \to C$, with GELU activation and dropout 0.2. We train it for 40 epochs using AdamW with learning rate 3×10^{-4}, cosine scheduling, batch size 64, and weight decay 1×10^{-4}. No RF fine-tuning is applied to the DINOv2 encoder.

Table 5 shows the central result. On OPERA, the external benchmark collected outside our lab, frozen DINOv2 more than doubles macro-F1 relative to the strongest CNN baseline trained on the same source data, increasing from 0.20 ± 0.01 to 0.46 ± 0.01. This gain is larger than the gain from histogram matching and larger than the gain from changing the synthetic pretraining partner. On TTW, the controlled through-wall setting collected from the same lab setup, DINOv2 is competitive but below the best CNN. This asymmetry is important: the frozen representation helps most under the stronger dataset-level acquisition shift, while a task-specific CNN can remain competitive when the target setting is closer to the source setup.

The interaction with histogram matching is also informative. Histogram matching helps CNNs on OPERA, but it hurts DINOv2. This indicates that low-level appearance correction is not universally beneficial. For

Table 5: **Backbone comparison.** External class-restricted macro-F1 at $b = 1.0$. DINOv2 substantially improves OPERA recognition relative to the best CNN baseline, while the CNN remains stronger on the closer through-wall setting.

Model	OPERA	TTW
Best CNN (finetune-only, $SE1E2$)	0.20 ± 0.01	$\mathbf{0.79 \pm 0.03}$
Best CNN + histogram match	0.29 ± 0.01	0.78 ± 0.03
DINOv2-S/14 frozen + histogram match	0.33 ± 0.03	0.57 ± 0.02
DINOv2-S/14 frozen, identity input	$\mathbf{0.46 \pm 0.01}$	0.70 ± 0.04

Table 6: **Input-level levers versus frozen representation transfer on OPERA ($b = 1.0$).** All entries are external class-restricted macro-F1.

Configuration	OPERA F1
JEPA-v1 CNN baseline, S pretrain	0.15
+ test-time histogram matching	0.29
+ $SE1E2$ pretraining	0.16
+ histogram matching + $SE1E2$ pretraining	0.30
DINOv2-S/14 frozen representation	**0.46**

DINOv2, forcing external spectrograms to match the source luminance CDF appears to disrupt a pretrained feature geometry that already preserves more cross-domain class structure than the CNN baseline.

5.4 Input-level levers versus representation transfer

We next compare the cumulative effect of input-level levers with the effect of replacing the backbone. Table 6 starts from a JEPA-style CNN baseline and adds histogram matching and the style-matched synthetic pretraining partner. We then compare the stacked CNN configuration with frozen DINOv2.

The stacked CNN levers improve OPERA performance, mainly because histogram matching corrects part of the low-level appearance shift. However, the full stack remains below frozen DINOv2. This is the clearest comparison in the paper: correcting the input distribution helps, but replacing the representation helps more on OPERA. The result supports the main claim that external RF recognition is not solved only by adapting input spectrograms toward the target appearance. A stronger stable representation can produce a larger gain than the input-level corrections tested here.

5.5 Low training-data budget behaviour

Finally, we test whether the same pattern holds when the amount of lab-collected training data is reduced. At $b = 0.25$, only 142 training spectrograms are used, sampled in a per-class stratified manner.

Table 7 shows the same qualitative pattern as the full-budget setting. DINOv2 retains a clear advantage on OPERA, while the CNN remains stronger on the closer TTW setting. This reinforces the interpretation that frozen pretrained representations are most beneficial when the external shift is large, whereas a well-tuned CNN can remain competitive when the test setting is closer to the training setup.

6 Understanding Frozen Foundation Representations

The previous section showed that frozen DINOv2 provides the strongest external improvement on OPERA. We now ask whether this result is specific to DINOv2 or reflects a broader property of frozen vision foundation models. We also analyse whether the performance ranking is already present in the frozen feature space before the MLP head is trained.

Table 7: **Low training-data budget external recognition.** External class-restricted macro-F1 at $b=0.25$ (142 training images).

Model	OPERA	TTW
Best CNN (finetune-only, $SE1E2$)	0.17 ± 0.01	$\mathbf{0.61} \pm \mathbf{0.00}$
DINOv2-S/14 frozen	$\mathbf{0.29} \pm \mathbf{0.05}$	0.53 ± 0.02

Table 8: **Frozen vision encoder comparison.** External class-restricted macro-F1 (mean±sd over seeds $\{0, 42, 7\}$). All backbones are frozen and use the same MLP head.

	OPERA b=1.0		OPERA b=0.25		TTW b=1.0	
Backbone	id	hist	id	hist	id	hist
DINOv2-S/14	0.462	0.325	0.292	0.251	0.705	0.570
CLIP ViT-B/16	0.217	0.263	0.181	0.219	0.425	0.384
MAE ViT-B/16	0.297	0.329	0.148	0.234	0.500	0.426

6.1 Comparison across frozen vision encoders

We compare three frozen vision encoders with different pretraining objectives: DINOv2-S/14, CLIP ViT-B/16, and MAE ViT-B/16. DINOv2 represents self-distillation, CLIP represents contrastive image-text pretraining, and MAE represents masked image reconstruction. In each case, the encoder is frozen and only the same MLP classifier is trained on the lab-collected RF data.

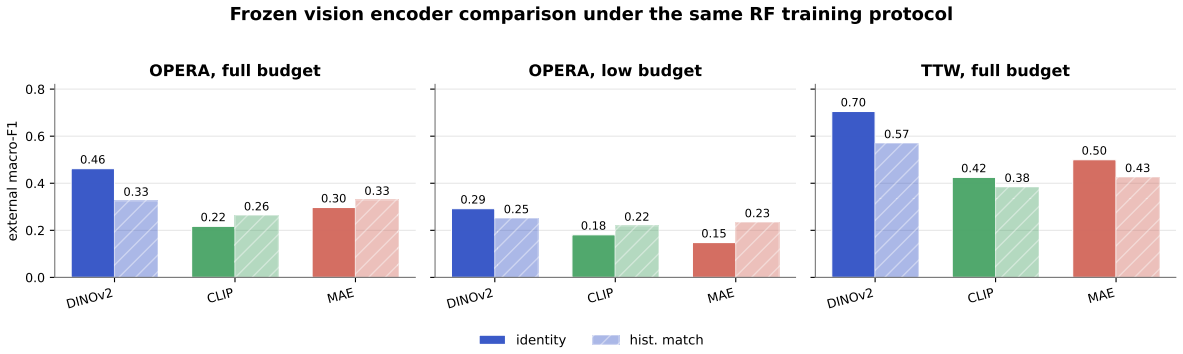

Figure 9: **Frozen encoder comparison.** External class-restricted macro-F1 for DINOv2, CLIP, and MAE under the same frozen-backbone and MLP-head protocol. "Identity input" means the external spectrogram is evaluated without histogram matching or any other test-time input transform. All three natural-image encoders transfer non-trivially to RF spectrograms, but DINOv2 is strongest on OPERA and TTW. Histogram matching helps CLIP and MAE on OPERA but hurts DINOv2, indicating that input-level normalisation interacts with the pretrained feature space.

Table 8 shows that transfer is not unique to DINOv2. All three frozen encoders transfer non-trivially to WiFi micro-Doppler spectrograms, despite being pretrained on natural images rather than RF data. However, the strength of transfer varies substantially. DINOv2 performs best across the external settings, with the clearest advantage on OPERA. This indicates that the choice of pretrained feature space matters: not all frozen vision representations provide the same degree of external robustness. It also rules out the weakest interpretation of the result, namely that RF spectrograms simply behave like ordinary images for any large pretrained image encoder.

The histogram-matching interaction makes this distinction sharper. MAE's masked-reconstruction objective encourages features that retain pixel-level information needed to reconstruct the input, so normalising luminance can help when the target rendering differs. CLIP's image-text objective organises features around

Table 9: **Feature-space diagnostics for frozen encoders.** Cosine k-NN macro-F1 is the primary diagnostic: a $k=5$ classifier is trained on M1M2 frozen features and evaluated on M1M2, OPERA, and TTW under the class-restricted policy, with no head training. MMD^2 is reported only as a secondary summary of global source-target geometry on L2-normalised CLS embeddings.

	Cosine k-NN macro-F1 (frozen features)			MMD^2 (L2-norm)	
Backbone	In-pool	OPERA	TTW	M1M2↔OPERA	M1M2↔TTW
DINOv2-S/14	0.809	0.401	0.584	0.517	0.168
CLIP ViT-B/16	0.724	0.322	0.426	0.518	0.170
MAE ViT-B/16	0.881	0.266	0.314	0.495	0.187

natural-image semantic alignment, which is only indirectly related to spectrogram activities. DINOv2's self-distillation objective, by contrast, is designed to make different augmented views map to similar representations. Our results are consistent with that view-invariant feature space being less dependent on the specific luminance statistics of OPERA. This is an interpretation of the observed pattern, not a causal proof of the pretraining mechanism.

6.2 Feature-space explanation of the DINOv2 advantage

Table 8 establishes that DINOv2 transfers better than the other frozen encoders we tested. We now ask whether this ranking is already visible in the frozen feature space before head training. Following the diagnostic logic of Sec. 4, we probe each backbone directly on a balanced sample of M1M2, OPERA, and TTW.

Our primary diagnostic is a cross-domain cosine k-NN probe. We train a $k=5$ classifier on M1M2 frozen features and evaluate class-restricted macro-F1 on OPERA and TTW, without training an MLP head and without any target-specific adaptation. As a secondary summary of source-target geometry, we compute RBF MMD^2 between M1M2 and each external pool on L2-normalised CLS embeddings.

The cross-domain k-NN result gives the clearest explanation. The same ranking observed after MLP training is already present in the frozen features. DINOv2 leads on OPERA (0.40 vs. 0.32 for CLIP and 0.27 for MAE) and on TTW (0.58 vs. 0.43 and 0.31). MAE has the highest in-pool k-NN F1 of the three but the largest cross-domain drop, indicating that its features are discriminative for M1M2 but do not place same-class external samples close to the M1M2 training samples. DINOv2, by contrast, preserves more class-neighbourhood structure across acquisition shift.

The MMD^2 measurement is less explanatory. DINOv2 and CLIP have similar source-target distances, and MAE is not uniformly farther away. Therefore, the DINOv2 advantage is not simply explained by smaller global source-target distance. The stronger evidence is the neighbourhood-level result: same-class external samples remain closer to M1M2 samples in DINOv2 feature space, and the retrieval diagnostics in Sec. 6.3 show the same pattern.

6.3 Supervised ImageNet control and mechanism checks

A natural concern is that the observed advantage may come from any large ImageNet-pretrained representation rather than from the self-supervised ViT backbones. To test this, we run the same k-NN diagnostic on a frozen supervised ImageNet ResNet-50 (`timm resnet50.tv_in1k`). The ResNet-50 has high in-pool k-NN F1, but performs worse than DINOv2 on the external targets, especially OPERA. This suggests that the cross-sensor advantage is not generic to ImageNet pretraining alone.

We also run four additional mechanism checks reported in the supplementary material: embedding stability under nuisance perturbations, a low-shot domain-versus-class probe, cross-domain retrieval Recall@5, and linear prediction of motion versus rendering descriptors. These analyses do not establish a single causal mechanism, but they support the same conclusion. DINOv2 preserves more activity-relevant neighbourhood structure under external shift than the other frozen encoders tested.

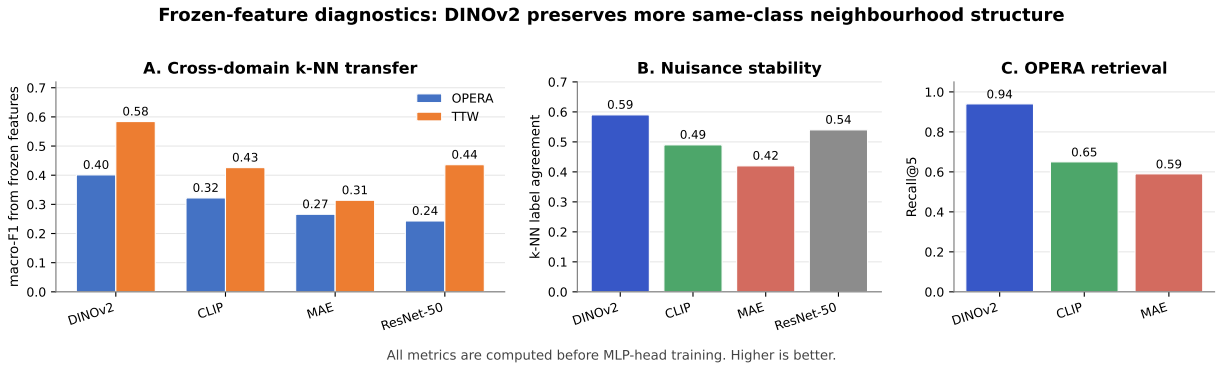

Figure 10: **Frozen-feature diagnostics for the DINOv2 advantage.** All metrics are computed before MLP-head training. Cross-domain k-NN transfer shows that the FM-family ranking is already present in frozen features. Nuisance perturbation agreement and OPERA retrieval Recall@5 provide additional evidence that DINOv2 preserves more same-class neighbourhood structure under acquisition shift. These diagnostics support a representation-level explanation, but do not establish a single causal mechanism.

Taken together, these results show that input-level corrections and style-matched synthetic pretraining provide limited gains. The largest improvement on the harder held-out target comes from replacing the CNN representation with a frozen DINOv2 encoder. This sets up the next section, where we test whether a strong generative augmentation pipeline can improve the frozen DINOv2 probe.

7 Does Generative Augmentation Improve the Frozen Probe?

The previous sections show that frozen DINOv2 provides the strongest external performance on the harder OPERA target. A natural follow-up question is whether this result can be improved further by adding generated spectrograms that vary acquisition or environment appearance. This tests the input-level view of the problem: if external recognition is mainly limited by insufficient coverage of target acquisition appearances, then a strong generative augmentation pipeline should help the frozen probe.

We therefore evaluate a factorised environment-transfer model designed to preserve motion content while changing acquisition or environment style. The goal is not to claim that all generative augmentation must fail, but to test whether a carefully engineered representative of this approach improves the strongest frozen foundation-model baseline.

7.1 Factorised environment transfer

The factorised environment-transfer (FET) model uses two latent codes: one intended to capture motion or activity content, and one intended to capture environment or rendering style. Given a content image x_c and an environment image x_e, the generator produces

$$\hat{x} = G\big(z_m(x_c),\, z_e(x_e)\big), \tag{1}$$

where z_m is the motion/content representation and z_e is the environment/style representation.

Training combines reconstruction, activity preservation, and disentanglement objectives. The reconstruction term encourages the generator to reproduce an input image from its own motion and environment codes. An activity critic encourages generated images to preserve the class of the content image. An environment critic with gradient reversal Ganin & Lempitsky (2015) discourages environment information from being encoded in z_m. We also use an MMD term Gretton et al. (2012) to align motion-code distributions across pools, a kinematic invariance loss on z_m, and an ℓ_2 disentanglement penalty between z_m and z_e. The critics use frozen ImageNet ResNet-50 features. Full hyperparameters are given in the supplementary material.

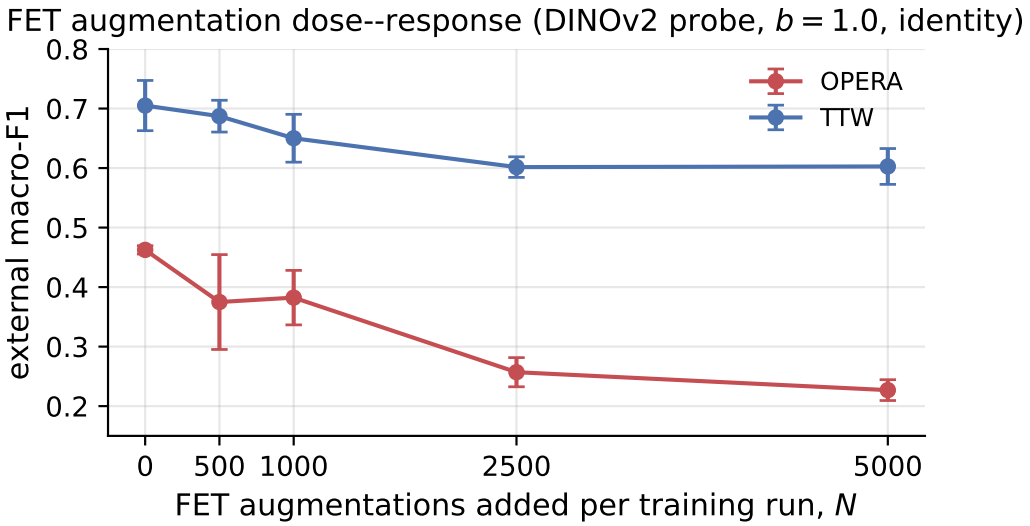

Figure 11: **Effect of FET augmentation dose.** External class-restricted macro-F1 at $b = 1.0$ as the number of FET-generated training samples increases. $N = 0$ is the unaugmented frozen DINOv2 baseline. Generated samples do not improve performance on either external target.

Table 10: **FET augmentation sweep.** External class-restricted macro-F1 at $b = 1.0$, mean $\pm$ standard deviation over three seeds. $N = 0$ is the unaugmented frozen DINOv2 baseline.

N	0	500	1000	2500	5000
OPERA	$\mathbf{0.46 \pm 0.01}$	0.37 ± 0.08	0.38 ± 0.05	0.26 ± 0.02	0.23 ± 0.02
TTW	$\mathbf{0.71 \pm 0.04}$	0.69 ± 0.03	0.65 ± 0.04	0.60 ± 0.02	0.60 ± 0.03

7.2 Integration with the frozen DINOv2 probe

We train FET using the lab-collected real pools M1 and M2, with $SE1E2$ used as the style-matched synthetic partner. We then generate additional training spectrograms by swapping motion and environment codes, assigning each generated sample the activity label of its content image. The generated samples are added to the training set of the frozen DINOv2 probe.

We test augmentation doses $N \in \{500, 1000, 2500, 5000\}$ generated samples per seed. The case $N = 0$ is the unaugmented frozen DINOv2 baseline from Sec. 5.3. The largest dose, $N = 5000$, more than doubles the amount of real training data.

7.3 Result: generated samples do not improve external performance

Figure 11 and Table 10 show that FET-generated samples do not improve the frozen DINOv2 probe at any tested dose. On OPERA, performance decreases from 0.46 ± 0.01 without generated data to 0.37 ± 0.08 at the smallest dose and 0.23 ± 0.02 at the largest dose. On TTW, the degradation is smaller but follows the same direction. Increasing the amount of generated data therefore does not compensate for the mismatch introduced by the generator.

The in-pool validation score increases slightly with more FET samples, while external performance decreases. This pattern suggests augmented-distribution mismatch: the generated samples help the model fit the lab-collected validation setting, but they do not move the frozen feature distribution closer to the external targets.

7.4 Implications

This negative result is informative because FET is designed around the same intuition that motivates many augmentation-based approaches: separate motion content from environment appearance, then generate additional examples with varied acquisition style. In our experiments, however, this input-level augmentation does not improve the strongest frozen representation. The result is consistent with the feature-space analysis in Sec. 6.2: DINOv2 already preserves more cross-domain activity neighbourhood structure than the other frozen encoders tested, and adding generated images does not further improve that structure.

We do not claim that generative augmentation can never help RF activity recognition. The narrower conclusion is that this well-engineered factorised augmentation pipeline does not improve performance on top of the frozen DINOv2 probe across the augmentation doses tested. For future augmentation-based work in this setting, the relevant comparison should therefore include not only CNN baselines, but also a strong $N=0$ frozen-foundation-model baseline and a clear positive dose-response curve.

8 Discussion

8.1 External RF recognition is not only input matching

The main message of this study is that external WiFi micro-Doppler recognition should not be treated only as an input-matching problem. Adapting source or synthetic spectrograms toward the target appearance can help in some cases, as shown by the histogram-matching results for CNN baselines. However, the largest improvement on the harder external target comes from replacing the learned CNN representation with a frozen pretrained representation. This suggests that the key difficulty is not only the visual mismatch between spectrograms, but whether the feature space preserves activity-relevant motion structure while reducing sensitivity to acquisition and rendering differences.

This interpretation also explains why the benefit of DINOv2 is not uniform across targets. On OPERA, which is collected outside our lab and differs more strongly from the lab-collected training data, the frozen representation gives a substantial improvement. On the through-wall test set, which is collected by adding a wall condition to the same lab setup, the gain is smaller and the best CNN remains competitive. Thus, frozen foundation features should not be viewed as a universal replacement for task-specific models. They appear most helpful when the external test data differ substantially from the training acquisition setting.

8.2 What the negative augmentation result means

The generative augmentation experiment provides a boundary on the input-matching view. The factorised environment-transfer model was designed to preserve motion content while changing acquisition or environment style, but adding its generated spectrograms did not improve the frozen DINOv2 probe. This does not imply that generative augmentation can never help RF sensing. Rather, it shows that, in this setting, additional generated images did not improve the representation already provided by the frozen foundation model.

This has an important methodological implication. Future augmentation-based RF recognition work should compare not only against CNN baselines, but also against a strong frozen-foundation-model baseline without generated data. A successful augmentation method should show a positive dose-response curve on external targets, not only improved in-pool validation performance.

8.3 Limitations

External dataset coverage. We evaluate two external settings: OPERA and a through-wall test set. These are informative because they represent different degrees of acquisition shift, but they do not cover the full range of possible RF deployment conditions. Broader evaluation on additional datasets, rooms, devices, activities, and collection pipelines would strengthen the generality of the conclusions.

Dataset scale. The benchmark is modest in size. The primary lab-collected pool M1M2 contains 799 spectrograms, with 559 training samples at $b=1.0$, and the external evaluation pools contain 373 OPERA

samples and 69 through-wall samples. This scale is sufficient to expose a reproducible external gap, but larger datasets would be needed to make stronger deployment-level claims.

Pretraining asymmetry. The frozen foundation encoders arrive with much larger pretraining histories than the task-specific CNN and JEPA baselines. Our claim is therefore not that foundation models are a fair matched-scale replacement for RF-specific models. The narrower claim is that, when strong frozen natural-image representations are available, they provide a powerful baseline for external RF spectrogram recognition and should be included in future comparisons.

Absolute performance. Even the best OPERA result remains far below in-pool performance. The paper closes part of the external recognition gap, but does not solve external RF activity recognition. This should be read as a representation-learning result, not as a deployment-ready sensing system.

Foundation-model coverage. We evaluate DINOv2-S/14, CLIP ViT-B/16, and MAE ViT-B/16. These cover different pretraining objectives, but they are not exhaustive. Larger variants and newer encoders may change the ranking. The results should therefore not be read as showing that all frozen foundation models behave similarly.

Known target label set. Class-restricted macro-F1 assumes that the set of classes present in the external target is known. A more difficult setting would require detecting which classes are present at test time, or evaluating under open-set or partial-label conditions.

8.4 Broader implications

Although this paper focuses on WiFi micro-Doppler recognition, the broader issue applies to other non-photographic image-shaped modalities. Radar, sonar, ultrasound, audio spectrograms, seismic measurements, and some medical imaging formats all produce two-dimensional representations where task-relevant structure is mixed with sensor-, acquisition-, and processing-specific variation. Our results suggest that frozen visual representations may provide transferable structural priors for some of these settings, but this should be tested directly rather than assumed.

The broader lesson is therefore methodological. For spectrogram-like sensing tasks, it may not be enough to ask whether generated or adapted inputs better match the target acquisition appearance. We should also ask whether the learned representation preserves the task-relevant structure under external acquisition shift. In this sense, the paper is not only about WiFi sensing; it is evidence that representation quality can matter more than input-level appearance alignment when non-photographic signals are evaluated outside their training conditions.

9 Conclusion

This paper studied external WiFi micro-Doppler activity recognition under changes in acquisition setting. The central question was whether external recognition is better addressed by adapting source or synthetic spectrograms toward the target appearance, or by learning features that remain stable across sensors, environments, and processing choices. Using lab-collected WiFi micro-Doppler data for training and evaluating on two external settings, we found that a frozen DINOv2-S/14 encoder pretrained on natural images provides a strong representation for this non-photographic spectrogram task. With only a lightweight MLP classifier trained on RF data, DINOv2 substantially improves recognition on OPERAnet, the harder external benchmark collected outside our lab, while remaining competitive but not uniformly superior on the closer through-wall test set.

The results suggest that pretrained visual representations can preserve activity-relevant motion structure in RF spectrograms, even though the encoder is not fine-tuned on RF data. At the same time, the benefit is not universal: the choice of foundation model matters, and the advantage is strongest when the test data differ more substantially from the training data. Input-level corrections such as histogram matching and style-matched synthetic pretraining help in some cases, but they do not match the improvement from the

frozen DINOv2 representation on the hardest external target. A factorised generative augmentation pipeline also fails to improve the frozen DINOv2 probe, showing that adding more generated spectrograms is not sufficient when the learned representation is already strong.

These findings could reframe external RF activity recognition as a representation-learning problem, not only an input-matching problem. For micro-Doppler spectrograms, the important question is not simply whether the spectrograms match the target appearance, but whether the feature space preserves activity-relevant motion patterns while reducing sensitivity to acquisition- and dataset-specific variation. More broadly, the study shows that frozen natural-image encoders can provide strong baselines for non-photographic, image-shaped sensing modalities, and should be considered in future work on external generalisation for RF and related spectrogram-based tasks.

References

Ahmed Aboulfotouh and Hatem Abou-Zeid. Multimodal wireless foundation models. *arXiv preprint arXiv:2511.15162*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Mohammud J. Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, and Robert Piechocki. OPERAnet: A multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific Data*, 9:474, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 1180–1189, 2015.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.

Jiajia Guo et al. LVM4CSI: Enabling direct application of pre-trained large vision models for wireless channel tasks. *arXiv preprint arXiv:2507.05121*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. European Conf. on Computer Vision (ECCV)*, pp. 172–189, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 2649–2658, 2018.

Xinyu Li, Yuan He, Francesco Fioranelli, and Xiaojun Jing. Semisupervised human activity recognition with radar micro-Doppler signatures. *IEEE Trans. Geoscience and Remote Sensing*, 60:1–12, 2022.

Zhizhen Li et al. MMSense: Adapting vision-based foundation model for multi-task multi-modal wireless sensing. *arXiv preprint arXiv:2511.12305*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proc. Int. Conf. on Computer Vision (ICCV)*, 2023.

Shelly Vishwakarma, Chong Tang, Wenda Li, Karl Woodbridge, Raviraj S. Adve, and Kevin Chetty. GAN-based noise generation to aid activity recognition when augmenting measured WiFi radar data with simulations. In *IEEE Int. Conf. on Communications Workshops*, 2021.

Shelly Vishwakarma, Wenda Li, Chong Tang, Karl Woodbridge, Raviraj S. Adve, and Kevin Chetty. SimHumalator: An open-source end-to-end radar simulator for human activity recognition. *IEEE Aerospace and Electronic Systems Magazine*, 37(3):6–22, 2022.

Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of WiFi signal-based human activity recognition. In *Proc. Int. Conf. on Mobile Computing and Networking (MobiCom)*, pp. 65–76, 2015.

Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Device-free human activity recognition using commercial WiFi devices. *IEEE Journal on Selected Areas in Communications*, 35(5): 1118–1131, 2017.

Yichi Zhang and Zhenshuang Shen. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3.0: Zero-effort cross-domain gesture recognition with WiFi. In *Proc. Int. Conf. on Mobile Systems, Applications, and Services (MobiSys)*, 2019.

Supplementary Material

This supplementary material provides training details, per-seed results, additional frozen-feature controls, and a compact traceability summary for the quantitative claims in the main paper.

A Training Details

A.1 Frozen foundation-model probes

The same training recipe is used for all frozen foundation-model probes; only the backbone, embedding dimension, and input normalisation differ.

- **Backbones.** We evaluate three frozen ViT encoders loaded via `timm`: DINOv2-S/14, CLIP ViT-B/16, and MAE ViT-B/16. DINOv2-S/14 uses a 384-dimensional CLS embedding, while CLIP and MAE use 768-dimensional CLS embeddings. DINOv2 and MAE use ImageNet mean/std normalisation; CLIP uses the OpenAI CLIP normalisation statistics. All encoder weights are frozen and kept in `eval()` mode. No gradient flows into the backbone.
- **Classifier head.** A two-layer MLP $d \to 256 \to C$ is trained on top of the frozen CLS embedding, where d is the backbone feature dimension and $C = 6$ is the number of source training classes. The hidden layer uses GELU activation and dropout 0.2.
- **Optimisation.** The MLP head is trained with AdamW Loshchilov & Hutter (2019), learning rate 3×10^{-4}, weight decay 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.
- **Schedule.** We train for 40 epochs with a 2-epoch linear warm-up followed by cosine decay.
- **Batching and input.** Batch size is 64. RGB spectrogram PNGs are resized to 224×224 using bilinear interpolation and normalised using the corresponding backbone statistics. No training-time augmentation is used.
- **Model selection.** Checkpoints are selected by the best in-pool validation macro-F1. External pools are never used for checkpoint selection.

A.2 CNN baselines

- **Backbone.** The CNN baselines use an ImageNet-pretrained ResNet-50 He et al. (2016), fully fine-tuned on the downstream data.
- **Head.** The classifier is a linear layer $2048 \to C$.
- **Optimisation.** We use AdamW with learning rate 1×10^{-4} for the backbone, 3×10^{-4} for the head, and weight decay 1×10^{-4}.
- **Schedule.** Models are trained for 60 epochs with a 3-epoch warm-up and cosine decay.
- **Variants.**

  real-only Trained only on M1M2.

  finetune-only Fine-tuned on M1M2 together with either S or $SE1E2$.

  JEPA-v1 Self-supervised pretraining on S or $SE1E2$ using the target-block/context-block masking strategy of I-JEPA Assran et al. (2023), followed by supervised fine-tuning on M1M2.

A.3 Test-time histogram matching

At inference time, each external test image is remapped using a 256-bin luminance CDF. The reference CDF is computed once from 300 M1M2 training images. The implementation uses a single-pass interpolation with fixed parameters and no learned components.

A.4 Evaluation policy

Class-restricted macro-F1 is computed as the unweighted mean of per-class F1 over the class set present in the target pool. For OPERA, this excludes Picking Up; for TTW, all six source classes are retained.

B Per-Seed Results

Tables 11–15 report the per-seed external class-restricted macro-F1 values behind the means and standard deviations in the main paper. All entries are computed once on the held-out external pool after training.

Table 11: CNN baselines, external class-restricted macro-F1 per seed.

Method	Target	b	seed 0	seed 42	seed 7
Real-only	OPERA	0.25	0.159	0.146	0.167
Real-only	OPERA	1.0	0.137	0.203	0.141
Real-only	TTW	0.25	0.801	0.765	0.672
Real-only	TTW	1.0	0.882	0.777	0.847
Finetune-only	OPERA	0.25	0.153	0.183	0.163
Finetune-only	OPERA	1.0	0.190	0.214	0.204
Finetune-only	TTW	0.25	0.599	0.688	0.638
Finetune-only	TTW	1.0	0.832	0.781	0.764
JEPA-v1	OPERA	0.25	0.085	0.161	0.119
JEPA-v1	OPERA	1.0	0.073	0.195	0.180
JEPA-v1	TTW	0.25	0.735	0.547	0.613
JEPA-v1	TTW	1.0	0.764	0.804	0.743

Table 12: Frozen DINOv2-S/14 + MLP head, external class-restricted macro-F1 per seed, for both identity and histogram-matched inputs.

Target	b	transform	seed 0	seed 42	seed 7
OPERA	0.25	identity	0.254	0.258	0.365
OPERA	0.25	hist	0.214	0.270	0.268
OPERA	1.0	identity	0.453	0.465	0.470
OPERA	1.0	hist	0.282	0.354	0.340
TTW	0.25	identity	0.553	0.519	0.503
TTW	0.25	hist	0.397	0.444	0.444
TTW	1.0	identity	0.679	0.671	0.764
TTW	1.0	hist	0.541	0.568	0.600

Table 13: Frozen CLIP ViT-B/16 + MLP head, external class-restricted macro-F1 per seed.

Target	b	transform	seed 0	seed 42	seed 7
OPERA	0.25	identity	0.174	0.172	0.196
OPERA	0.25	hist	0.220	0.204	0.234
OPERA	1.0	identity	0.217	0.207	0.227
OPERA	1.0	hist	0.228	0.231	0.329
TTW	0.25	identity	0.261	0.337	0.362
TTW	0.25	hist	0.281	0.311	0.401
TTW	1.0	identity	0.426	0.393	0.455
TTW	1.0	hist	0.407	0.350	0.395

Table 14: Frozen MAE ViT-B/16 + MLP head, external class-restricted macro-F1 per seed.

Target	b	transform	seed 0	seed 42	seed 7
OPERA	0.25	identity	0.124	0.168	0.152
OPERA	0.25	hist	0.216	0.248	0.237
OPERA	1.0	identity	0.313	0.295	0.284
OPERA	1.0	hist	0.342	0.326	0.320
TTW	0.25	identity	0.238	0.278	0.325
TTW	0.25	hist	0.270	0.263	0.340
TTW	1.0	identity	0.513	0.513	0.476
TTW	1.0	hist	0.456	0.434	0.389

Table 15: FET augmentation dose-response on the DINOv2 probe, external class-restricted macro-F1 per seed.

Target	N	seed 0	seed 42	seed 7
OPERA	0	0.453	0.465	0.470
OPERA	500	0.361	0.479	0.285
OPERA	1000	0.356	0.447	0.345
OPERA	2500	0.287	0.227	0.257
OPERA	5000	0.252	0.215	0.214
TTW	0	0.679	0.671	0.764
TTW	500	0.671	0.665	0.725
TTW	1000	0.678	0.593	0.679
TTW	2500	0.625	0.596	0.584
TTW	5000	0.642	0.569	0.597

C Additional Frozen-Feature Controls

Supervised ImageNet CNN control. A natural concern is whether the cross-sensor advantage reflects any large ImageNet-pretrained representation rather than a property of the self-supervised ViT backbones. To test this, we run the same frozen-feature diagnostic pipeline on a supervised ImageNet ResNet-50 (`resnet50.tv_in1k`). The ResNet-50 has high in-pool k-NN F1, but its OPERA transfer is below all three frozen ViT backbones and its TTW transfer remains below DINOv2. This supports the narrower conclusion that the observed advantage is not generic to ImageNet pretraining alone.

Table 16: **Supervised ImageNet CNN control.** Same balanced-sample cosine k-NN diagnostic as Table 9, with a frozen supervised ImageNet ResNet-50 added as a control.

Backbone	In-pool (M1M2)	OPERA	TTW
DINOv2-S/14	0.809	0.401	0.584
CLIP ViT-B/16	0.724	0.322	0.426
MAE ViT-B/16	0.881	0.266	0.314
ResNet-50 (sup. IN-1k)	0.836	0.243	0.436

D Mechanism Ablations

The main paper uses cross-domain k-NN as the primary feature-space diagnostic. We add four follow-up checks here. We treat them as mechanism evidence, not as a causal proof: they identify properties of the frozen feature space that are consistent with the DINOv2 ranking.

Appearance-sensitivity stability. For a balanced sample from M1M2, OPERA, and TTW, we perturb each image with controlled pixel-space nuisances: luminance shift, contrast scaling, Gaussian blur, Gaussian

noise, JPEG compression, and grayscale collapse. We then ask whether the M1M2-indexed 5-NN majority label is preserved after perturbation. Table 17 shows that DINOv2 has the highest mean label agreement.

Table 17: **k-NN label agreement under nuisance perturbations.** Higher values indicate that the frozen neighbourhood label is more stable under controlled pixel-space perturbations.

Backbone	Mean k-NN agreement
DINOv2-S/14	**0.59**
ResNet-50 (sup. IN-1k)	0.54
CLIP ViT-B/16	0.49
MAE ViT-B/16	0.42

Low-shot domain-vs-class separability. We also train low-shot pool classifiers to test whether stronger backbones make acquisition identity less recoverable than activity identity. In practice, the pool probe saturates for all backbones, so this diagnostic does not rank the foundation-model families. We report it as a null result rather than evidence for the mechanism.

Cross-domain retrieval. For each OPERA query, we retrieve the top-5 cosine neighbours from the M1M2 feature index and measure whether at least one retrieved neighbour has the same activity label. Table 18 shows that DINOv2 has substantially higher OPERA Recall@5 than CLIP or MAE, which is a direct retrieval-based restatement of the neighbourhood structure claim in Sec. 6.2.

Table 18: **Cross-domain retrieval on OPERA.** Recall@5 measures whether an OPERA query retrieves at least one same-class M1M2 sample within its top-5 cosine neighbours.

Backbone	OPERA Recall@5
DINOv2-S/14	**0.94**
CLIP ViT-B/16	0.65
MAE ViT-B/16	0.59

Physical-descriptor predictability. Finally, we fit linear ridge probes from frozen features to simple spectrogram descriptors, including motion descriptors and rendering descriptors. The result is mixed: all backbones encode nuisance appearance to some degree, but DINOv2 and CLIP have lower linear rendering coupling than MAE and ResNet-50. We therefore use this diagnostic only as supporting context.

Reading. The perturbation and retrieval diagnostics agree with the main cross-domain k-NN result: DINOv2 preserves more activity-relevant neighbourhood structure under acquisition and appearance changes. The low-shot domain probe saturates and the physical-descriptor probe is mixed. We therefore do not claim a single causal mechanism; the supportable conclusion is that DINOv2's frozen feature space keeps more same-class external samples near the M1M2 source samples than the alternatives tested.

E Quantitative Traceability Summary

Table 19: **Traceability summary.** Main quantitative claims and where the supporting evidence appears in the paper.

Claim	Supporting evidence	Value
DINOv2 improves OPERA at full training-data budget	Main Table 5; per-seed values in Table 12	0.46 ± 0.01
Best CNN baseline on OPERA at full training-data budget	Main Table 5; per-seed values in Table 11	0.20 ± 0.01
DINOv2 is competitive but not best on TTW	Main Table 5; per-seed values in Table 12	0.70 ± 0.04
Histogram matching helps CNNs but hurts DINOv2 on OPERA	Fig. 7 and Table 5	$0.46 \rightarrow 0.33$
Frozen FM ranking is present before head training	Main Table 9	DINOv2 leads OPERA/TTW
Supervised ImageNet ResNet-50 is not enough	Table 16	OPERA 0.243
FET augmentation does not improve the frozen probe	Main Table 10; per-seed values in Table 15	No dose improves