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 fail when evaluated under unseen acquisition conditions. A key challenge is that micro-Doppler spectrograms entangle activity-relevant motion structure with acquisition-specific appearance. Useful information includes temporal rhythm, Doppler spread, motion duration, and time-frequency energy structure, while nuisance 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 spectrograms. This paper asks whether external WiFi micro-Doppler recognition is better addressed by learning stable representations than by making spectrograms look more target-like.

We investigate this question using frozen vision foundation models pretrained on natural images. All models are trained on lab-collected WiFi micro-Doppler data and evaluated on two strictly held-out acquisition settings: OPERA, a subset of the public OPERAnet multimodal RF benchmark collected outside our lab, and TTW, a through-wall test set collected using the same lab setup with a wall condition added. Despite not being fine-tuned on RF data, a frozen DINOv2-S/14 encoder with a lightweight MLP classifier substantially improves recognition on OPERA, increasing class-restricted macro-F1 from 0.20±0.01 for the strongest convolutional baseline to 0.46±0.01. On TTW, DINOv2 remains competitive but is not uniformly superior, suggesting that frozen foundation features are most useful when the external acquisition shift is larger. These results suggest that external RF activity recognition should be treated not only as an input-matching problem, but also as a representation-transfer problem for preserving activity-relevant motion structure under held-out acquisition shift.

1 Introduction

Human activity recognition from WiFi micro-Doppler spectrograms is a promising route to privacy-preserving sensing because it can capture motion without cameras or wearable devices. In controlled settings, convolutional and self-supervised models can achieve high accuracy on held-out samples collected with the same sensor, room, placement, and preprocessing pipeline. However, this success often does not survive deployment-like evaluation. A model trained in one RF acquisition setting may fail when the room changes, the propagation path changes, the dataset pipeline changes, or the same activity is observed through a wall. This external acquisition gap remains one of the main obstacles to practical RF activity recognition.

The challenge is deeper than ordinary overfitting. A micro-Doppler spectrogram entangles two types of information in the same time-frequency image. The first is activity-relevant motion structure: rhythm, duration, Doppler spread, and the shape of energy over time and frequency. The second is acquisition-specific appearance: room clutter, multipath propagation, antenna placement, sensor response, wall attenuation, preprocessing, colour mapping, and dataset-specific rendering. These nuisance factors can be highly predictive

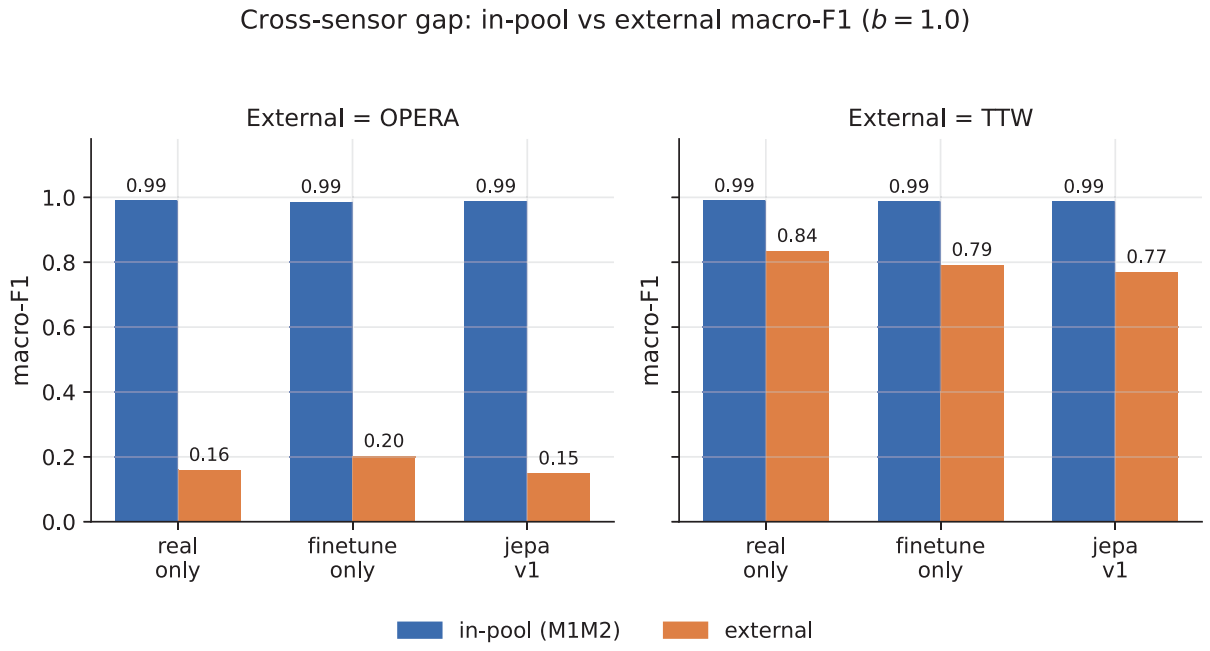

Figure 1: **Evaluation gap under unseen acquisition settings.** CNN-based RF models trained on lab-collected M1M2 spectrograms achieve high macro-F1 on a held-out M1M2 test split, but degrade when evaluated under unseen acquisition settings. The drop is largest on OPERA, a public benchmark collected outside our lab, and smaller on TTW, a through-wall test set collected using the same lab setup with a wall condition added. This gap motivates our focus on representations that preserve activity-relevant structure across acquisition and dataset shifts.

inside a training collection while being unrelated to the activity label. As a result, a model can learn a representation that fits the source data well but fails when the acquisition setting changes.

Most prior work addresses this problem at the input level. Simulation-based pretraining, style-transfer augmentation, adversarial adaptation, and RF-specific self-supervised learning all aim to make source, synthetic, or augmented spectrograms more useful for a target acquisition condition Vishwakarma et al. (2022); Gatys et al. (2015); Zheng et al. (2019); Vishwakarma et al. (2021); Assran et al. (2023). These approaches are well motivated because collecting labelled RF data across many rooms, devices, and propagation conditions is expensive. However, they leave open a more basic question: under external acquisition shift, is the main bottleneck that the input spectrograms do not look target-like enough, or that the learned representation does not preserve activity-relevant structure when low-level appearance changes?

We study this question using frozen vision foundation models. Large pretrained encoders such as DINOv2 Oquab et al. (2024), CLIP Radford et al. (2021), and MAE He et al. (2022) are trained on natural images, not RF measurements. Their pixels originally represent photographic appearance, whereas micro-Doppler pixels represent signal energy over time and frequency. Strong transfer to RF spectrograms is therefore not guaranteed. At the same time, these encoders may provide generic structural priors over shape, continuity, local contrast, 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 target-like inputs.

Our evaluation uses two lab-collected WiFi micro-Doppler pools, M1 and M2, as the only source data for training, validation, hyperparameter selection, and checkpoint selection. Their combined training set, M1M2, is used as the primary source pool. We then evaluate under two strictly held-out acquisition settings. OPERA is a subset of the public OPERAnet multimodal RF benchmark collected outside our lab, representing a strong change in acquisition environment, dataset construction, and processing pipeline. TTW is a through-wall test

set collected using the same lab acquisition setup as M1 and M2 but with a wall condition added. OPERA and TTW are never used for training, validation, hyperparameter selection, checkpoint selection, backbone selection, or target-specific tuning.

Figure 1 illustrates the gap that motivates the paper. CNN-based RF baselines trained on M1M2 reach near-perfect in-pool performance, but degrade under held-out acquisition settings, with the largest drop on OPERA. This motivates our central hypothesis: external RF activity recognition is limited not only by input appearance mismatch, but also by whether the feature representation preserves activity-relevant motion structure across acquisition shift.

Our main finding is that frozen vision representations substantially narrow the harder external gap. A frozen DINOv2-S/14 encoder with a lightweight MLP classifier trained only on lab-collected RF data increases class-restricted macro-F1 on OPERA from 0.20 ± 0.01 for the strongest convolutional baseline to 0.46 ± 0.01. On TTW, DINOv2 remains competitive but is not uniformly superior to the best CNN baseline. This difference is important: it suggests that frozen foundation features are most useful when the external test data differ strongly from the source acquisition setting, rather than acting as a universal replacement for RF-specific models.

We then ask whether this effect is specific to DINOv2 or reflects a broader property of frozen vision encoders. Under the same frozen-backbone and MLP-head protocol, DINOv2, CLIP, and MAE all transfer non-trivially to WiFi micro-Doppler spectrograms, but their external performance differs substantially. DINOv2 gives the strongest results on the harder OPERA target, showing that the choice of pretrained feature space matters. Cross-domain nearest-neighbour probes further show that this ranking is already visible before the MLP head is trained: DINOv2 better preserves same-class neighbourhood structure across acquisition shift.

To separate representation effects from artefacts of the evaluation setup, we diagnose the external gap using frozen-feature distances, UMAP projections, luminance statistics, and cross-domain nearest-neighbour analysis. These diagnostics show that the source and held-out pools differ in both low-level appearance and feature geometry. We also use class-restricted macro-F1 when a held-out target does not contain the full source label set, preventing missing target classes from artificially depressing the reported score. This evaluation policy does not remove the external recognition problem, but it ensures that the reported failure reflects recognition on shared classes rather than label-set mismatch.

Finally, we test a stronger input-level alternative: a factorised environment-transfer generator designed to produce additional spectrograms with varied acquisition appearance using adversarial and MMD regularisation. Adding these generated spectrograms does not improve the frozen DINOv2 probe at any tested augmentation dose. We do not conclude that generative augmentation cannot help RF sensing in general. Rather, our result shows that, in this setting, more generated target-like spectrograms do not improve a strong frozen foundation-model baseline. This provides a useful methodological lesson: future augmentation methods for external RF recognition should be compared against strong frozen-representation baselines, not only against CNNs.

The paper makes the following contributions:

1. We formulate external WiFi micro-Doppler recognition as a representation-transfer problem under held-out acquisition shift, rather than only as an input-level matching or augmentation problem.
2. We show that a frozen DINOv2-S/14 encoder pretrained on natural images, combined with a lightweight MLP head trained only on lab-collected RF data, substantially improves recognition on OPERA, a public RF benchmark collected outside our lab, compared with convolutional baselines trained on the same source data.
3. We compare DINOv2, CLIP, and MAE under a matched frozen-backbone protocol and show that transfer from natural-image encoders to RF spectrograms is non-trivial but strongly backbone-dependent.
4. We analyse the external evaluation gap using feature-space distances, UMAP projections, luminance statistics, cross-domain nearest-neighbour probes, and class-restricted macro-F1 for held-out targets with incomplete label overlap.

5. We report a controlled negative result for factorised generative environment transfer, showing that additional generated spectrograms do not improve the frozen DINOv2 probe at the tested doses, and that augmentation methods should be compared against strong frozen-representation baselines.

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 representing 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 testing data share the same acquisition conditions, 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. Our work studies this external generalisation problem directly, using held-out acquisition settings that are not used for training or model selection.

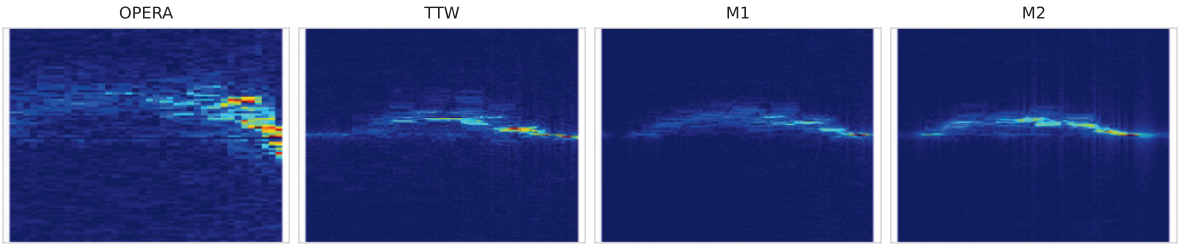

Figure 2: **Same activity, different acquisition appearance.** Representative walking-forward spectrograms from OPERA, TTW, M1, and M2 show that activity-related time-frequency structure is mixed with dataset- and acquisition-specific rendering differences. This motivates evaluating whether models preserve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

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 encoders have also been reused beyond ordinary RGB photography, including medical imaging Zhang & Shen (2024) and satellite imagery Reed et al. (2023); Cong et al. (2022). WiFi micro-Doppler spectrograms are a harder transfer test because their pixels represent signal energy over time and frequency rather than optical reflectance. This makes them useful for testing whether natural-image representations encode structural priors that survive in non-photographic sensing data.

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, while MMSense Li et al. (2025) and WavesFM Aboulfotouh & Abou-Zeid (2025) explore broader multimodal and native wireless pretraining directions. Our focus is narrower and complementary. We do not ask whether foundation models help wireless sensing in general; we test whether frozen natural-image representations improve WiFi micro-Doppler recognition under strict held-out acquisition shift, and whether this improvement exceeds input-level adaptation and generative augmentation baselines.

Input-level adaptation and generative augmentation. A common response to RF acquisition shift 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. These approaches mainly operate at the input or data-distribution level. Our experiments test whether a stronger frozen representation can produce larger external gains than making the spectrograms more target-like.

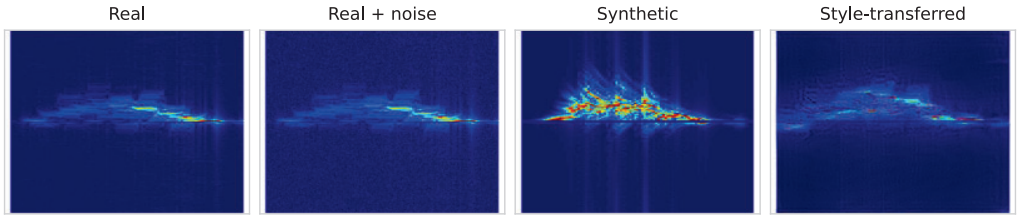

Figure 3: **Input-level adaptation examples.** Representative real, noise-perturbed, synthetic, and style-transferred spectrograms illustrate input-space strategies used to increase appearance diversity or reduce source-target mismatch. The experiments in this paper ask whether such input-level changes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

Evaluation under held-out acquisition settings. Testing external generalisation requires separating ordinary held-out testing from recognition under acquisition shift. RF datasets can differ in acquisition conditions, class coverage, and dataset composition, and prior work on domain generalisation has shown that benchmark design and model-selection choices can strongly affect conclusions Gulrajani & Lopez-Paz (2021). We therefore use OPERA and TTW only for final evaluation, never for training, validation, hyperparameter tuning, checkpoint selection, or backbone selection. We also use class-restricted macro-F1 when a held-out target does not contain the full source label set, so that models are evaluated on shared classes rather than penalised for absent target classes.

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.

Table 1: **Per-pool sample counts.** M1 and M2 are the lab-collected source pools, and M1M2 is their combined source pool. S and $SE1E2$ are used for synthetic pretraining or augmentation experiments. OPERA and TTW are held-out evaluation pools and are never used for training, validation, hyperparameter tuning, checkpoint selection, or backbone selection.

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

- **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. They are not used as held-out evaluation data.

- **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. OPERA and TTW are used only for final evaluation and are never used for training, validation, hyperparameter tuning, checkpoint selection, or backbone selection.

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 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. Separate random seeds are used for each experiment.

All learning curves, early stopping decisions, hyperparameter choices, checkpoint selection, and backbone selection use only the source validation split. OPERA and TTW are reserved for final held-out evaluation. They are not used at any stage of model development.

We report results at two source-data budgets. The full-budget setting, $b=1.0$, uses the full M1M2 training split with 559 spectrograms. The low-budget setting, $b=0.25$, 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. Unless otherwise stated, results are reported over three seeds, $\{0, 42, 7\}$, as mean $\pm$ standard deviation. Exact per-pool counts are shown in Table 1.

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 source 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 Strict holdout rules

OPERA and TTW are strictly held out. They are not used for training, validation, hyperparameter tuning, checkpoint selection, backbone selection, early stopping, augmentation selection, or test-time supervision. We do not perform per-target fine-tuning, target-specific class filtering, or target-specific calibration to improve results. The only target-dependent operation is the metric definition above, which follows the class set actually present in the held-out pool.

3.6 Motivating evaluation gap

Figure 1 quantifies the external gap for three CNN-based baselines trained on M1M2 at $b = 1.0$. All three achieve high macro-F1 on the held-out M1M2 test split 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 interventions that follow: we test whether external recognition improves more by correcting input appearance and synthetic-data mismatch, or by replacing the learned representation with a frozen pretrained feature space.

4 Diagnosing the External Evaluation Gap

Before testing interventions, we diagnose the external evaluation gap. The purpose is not to prove a single causal mechanism, but to determine whether the source and external pools differ in ways that are visible before task-specific training. We use four diagnostics: frozen-feature distances, feature-space visualisation, luminance statistics, and an evaluation-policy audit. Together, they show that the gap contains both an acquisition/rendering component and an evaluation-policy component, but that neither fully explains the recognition failure.

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

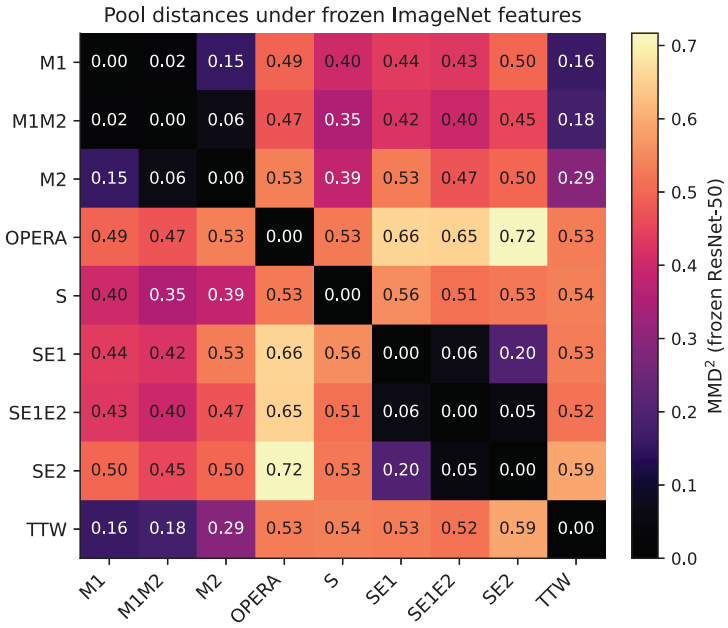

Figure 4: **Pool-to-pool RBF-MMD distances under frozen ImageNet ResNet-50 features.** OPERA is far from the lab-collected training pools (M1, M2) and from the synthetic pools (S, $SE1E2$). TTW is closer to M1M2, consistent with its smaller external evaluation gap in Fig. 1. $SE1E2$ is closer to M1M2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

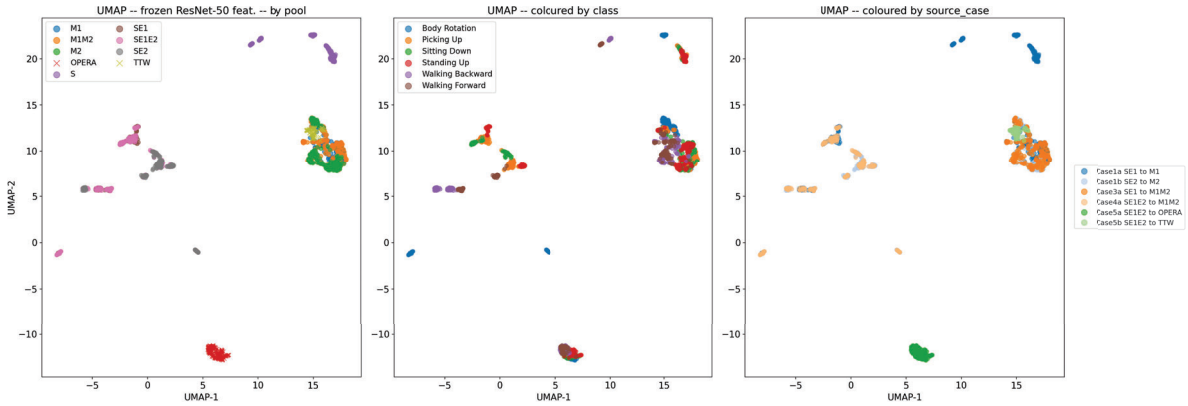

Figure 5: **Frozen ResNet-50 feature UMAP**, coloured by activity class (left) and source pool (right). Pool identity is more separable than activity identity at the frozen-feature level.

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, nor that pool identity alone explains external failure. Rather, it shows that activity-related motion structure is embedded inside a stronger pool-specific appearance geometry. This is exactly the setting in which a representation that preserves same-class neighbourhoods across acquisition shift should help.

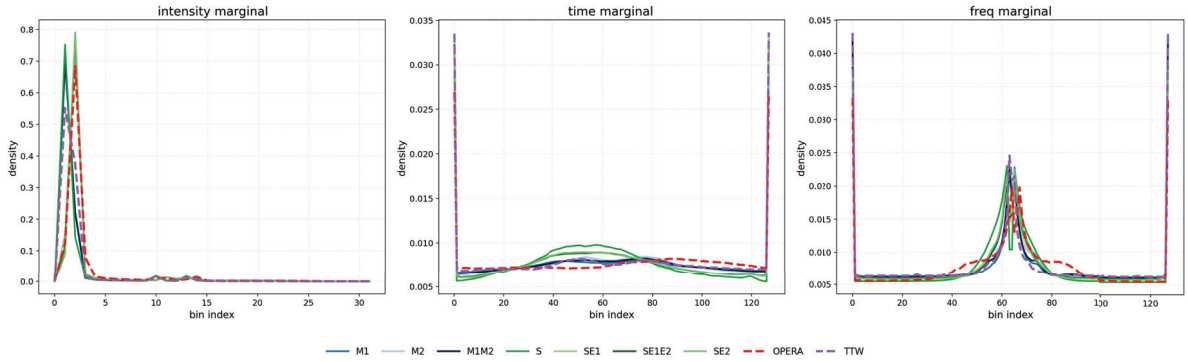

Figure 6: **Per-pool luminance marginals.** OPERA shows a rendering-induced shift relative to M1M2; TTW is closer but still offset. A single-pass CDF histogram match at test time reduces this offset; we quantify its effect in Sec. 5.1.

Table 2: **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

4.3 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.

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 2 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.

The diagnostics lead to three conclusions. First, OPERA is not merely another test split: it is farther from the source pools than TTW in frozen feature space, matching its larger recognition gap. Second, part of the gap is low-level and image-like, motivating a simple histogram-matching intervention. Third, evaluation policy matters because OPERA has incomplete label overlap. However, none of these findings alone solves external recognition. The next section therefore compares input-level corrections with a representation-level intervention.

Table 3: **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**

5 Interventions and Main Results

The diagnostic study in Sec. 4 showed that the external evaluation gap contains both low-level acquisition/rendering differences and a harder representation-transfer problem. We therefore compare three types of intervention. The interventions are deliberately ordered from weakest to strongest change: L1 changes only external image appearance at test time, L2 changes the synthetic pretraining partner, and A1 changes the feature representation itself.

L1 – 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.

L2 – Style-matched synthetic 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.

A1 – Frozen DINOv2-S/14 representation. 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 L1: 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 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.

5.2 L2: changing the synthetic 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 3 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 is useful, but its effect is much smaller than the representation change evaluated next.

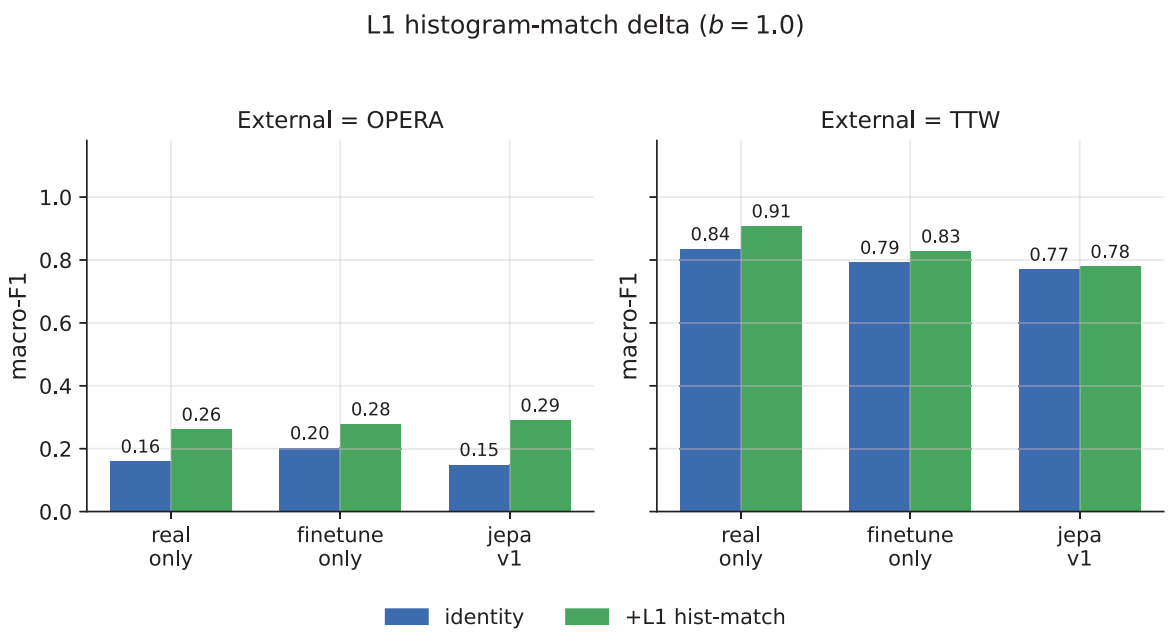

Figure 7: **Effect of test-time histogram matching.** External macro-F1 before and after applying the CDF histogram match to external test images for three CNN baselines at $b = 1.0$. Histogram matching improves OPERA more than TTW, consistent with the larger low-level appearance shift observed for OPERA.

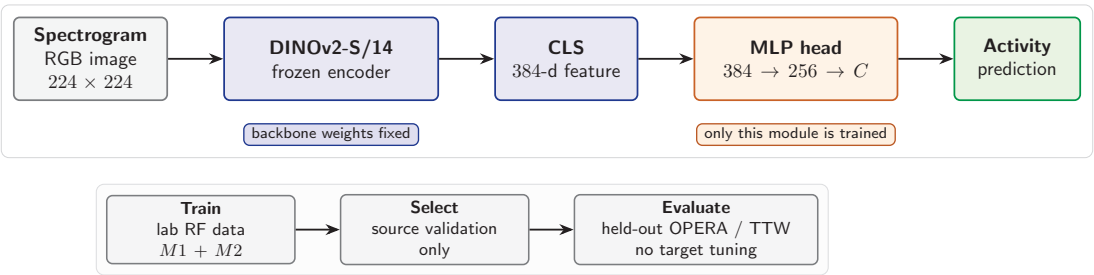

Figure 8: **Frozen DINOv2 probe.** Input spectrograms are resized to 224×224 and passed through a frozen DINOv2-S/14 encoder. Only a lightweight $384 \to 256 \to C$ MLP classifier is trained on the lab-collected RF data.

5.3 A1: replacing the CNN with a frozen DINOv2 representation

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 4 shows the central result. On OPERA, frozen DINOv2 more than doubles external macro-F1 relative to the strongest CNN baseline trained on the same source data. This improvement is larger than the gain from histogram matching and larger than the gain from changing the synthetic pretraining partner. On TTW, however, the best CNN remains stronger. This asymmetry is important: the frozen representation is most

Table 4: **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 + L1	0.29 ± 0.01	0.78 ± 0.03
DINOv2-S/14 frozen + L1	0.33 ± 0.03	0.57 ± 0.02
DINOv2-S/14 frozen	$\mathbf{0.46 \pm 0.01}$	0.70 ± 0.04

Table 5: **Input-level levers versus frozen representation transfer on OPERA ($b = 1.0$).** All entries are external class-restricted macro-F1 unless otherwise noted.

Configuration	OPERA F1
JEPA-v1 CNN baseline, S pretrain	0.15
+ L3 class-restricted policy	—
+ L1 test-time histogram matching	0.29
+ L2 $SE1E2$ pretraining	0.16
+ L1 + L2	0.30
A1 DINOv2-S/14 frozen representation	**0.46**

useful 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 shows that input normalisation is backbone-dependent. Histogram matching is a useful diagnostic and improves CNNs, but it is not a universally beneficial test-time adaptation step for frozen foundation models.

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 5 starts from a CNN baseline and adds class-restricted evaluation, 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 through L1. However, even the full input-level stack remains well below frozen DINOv2. This is the clearest comparison in the paper: correcting the input distribution helps, but replacing the representation helps more on the harder external target.

5.5 Low-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 6 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 useful when the external shift is large, whereas a well-tuned CNN can remain competitive when the test setting is closer to the training setup.

5.6 Comparison across frozen vision encoders

The previous results show that frozen DINOv2 provides the strongest external improvement on OPERA. We next ask whether this is a one-off model observation or a broader property of frozen vision encoders. We compare DINOv2, CLIP, and MAE under the same frozen-backbone and MLP-head protocol.

Table 6: **Low-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	$\mathbf{0.61 \pm 0.00}$
DINOv2-S/14 frozen	$\mathbf{0.29 \pm 0.05}$	0.53 ± 0.02

Table 7: **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

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

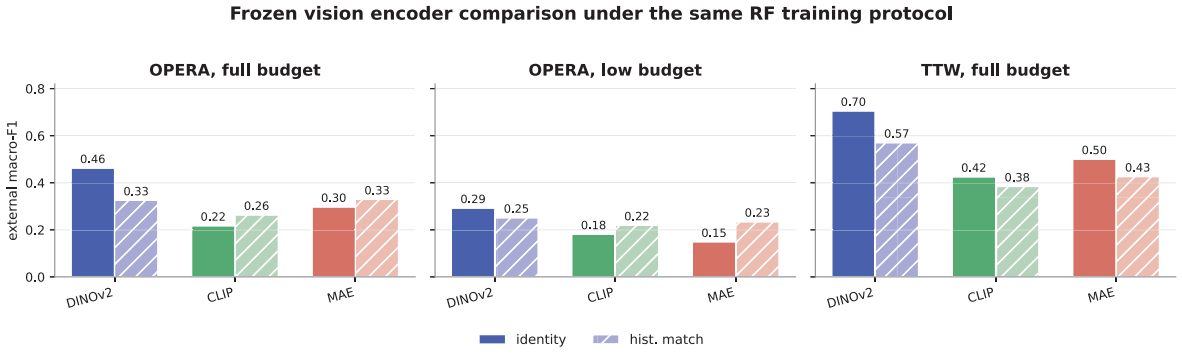

Figure 9: **Frozen encoder comparison.** External class-restricted macro-F1 for DINOv2, CLIP, and MAE under the same frozen-backbone and MLP-head protocol. The identity-input results show that all three natural-image encoders transfer non-trivially to RF spectrograms, but DINOv2 is strongest on OPERA 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 7 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.

This matters because it rules out the weakest interpretation of the result: that DINOv2 succeeds only because RF spectrograms are treated as ordinary images. If that were the whole explanation, the different pretrained encoders would be expected to behave more similarly. Instead, the large differences across DINOv2, CLIP, and MAE show that the learned feature space matters.

Histogram matching again interacts with the backbone. It helps CLIP and MAE modestly on OPERA but hurts DINOv2. This suggests that input-level normalisation is not universally beneficial for frozen foundation models. Whether it helps depends on how the pretrained encoder represents low-level appearance and structural information.

Table 8: **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 as a secondary summary of 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

5.7 Feature-space explanation of the DINOv2 advantage

Table 7 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 useful as a global geometry summary but does not explain the model ranking by itself. DINOv2 and CLIP have similar source-target distances, and MAE is not uniformly farther away. The stronger evidence is therefore local and class-conditional: in DINOv2 feature space, same-class external samples remain closer to source samples. This supports a representation-level explanation without claiming a single causal mechanism.

5.8 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 external 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.

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

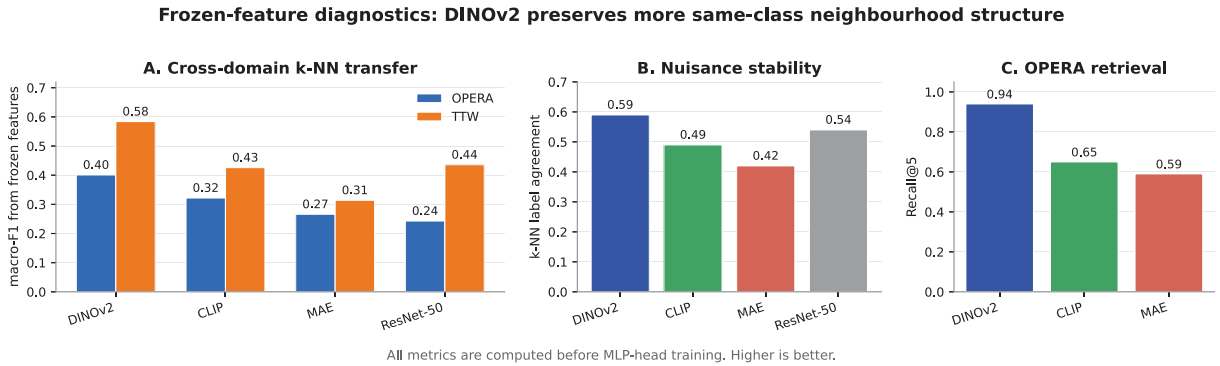

Figure 10: **Frozen-feature diagnostics for the DINOv2 advantage.** All metrics are computed before MLP-head training. Cross-domain k-NN transfer shows that the FM-family ranking is already present in frozen features. Nuisance perturbation agreement and OPERA retrieval Recall@5 provide additional evidence 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.

6 Does Generative Augmentation Improve the Frozen Probe?

The previous sections show that frozen DINOv2 is the strongest external baseline on OPERA. A natural question is whether this baseline can be improved by adding generated spectrograms with varied acquisition appearance. This is an important control because, if the main limitation were insufficient coverage of target-like inputs, then a strong generative augmentation pipeline should improve 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.

6.1 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.

6.2 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.

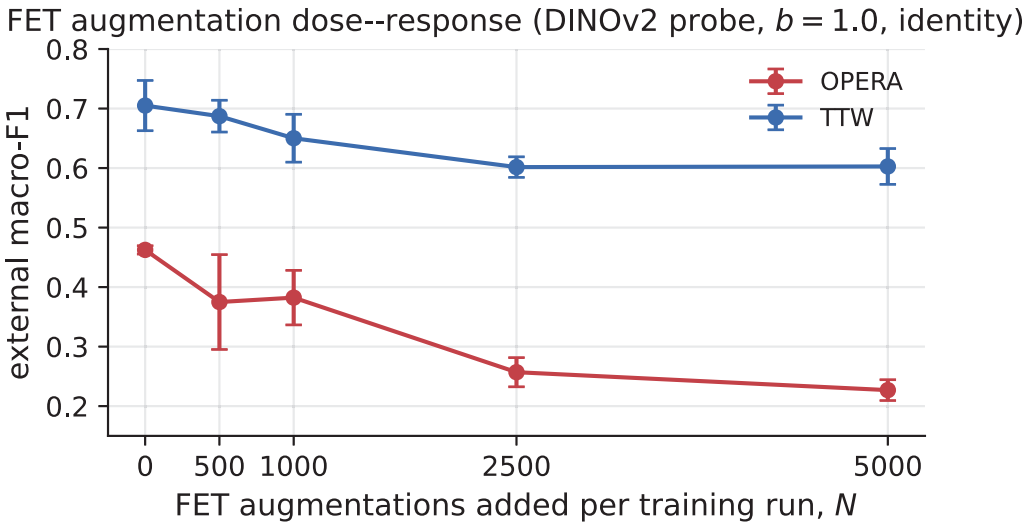

Figure 11: **Effect of FET augmentation dose.** External class-restricted macro-F1 at $b = 1.0$ as the number of FET-generated training samples increases. $N = 0$ is the unaugmented frozen DINOv2 baseline. Generated samples do not improve performance on either external target.

Table 9: **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

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.

6.3 Result: generated samples do not improve external performance

Figure 11 and Table 9 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.

6.4 Implications

This result is informative because FET implements the same intuition that motivates many augmentation-based RF approaches: preserve motion content while varying acquisition or environment appearance. In our experiments, however, adding such samples improves neither OPERA nor TTW on top of the frozen DINOv2 probe. The in-pool validation score increases slightly while external performance decreases, suggesting augmented-distribution mismatch rather than useful external diversity.

We do not claim that generative augmentation can never help RF activity recognition. The narrower conclusion is that this factorised augmentation pipeline does not improve a strong frozen foundation-model baseline across the tested doses. This changes the appropriate baseline for future augmentation work: it is not enough to outperform a CNN trained on real data; an augmentation method should also improve over an $N = 0$ frozen-foundation-model probe and ideally show a positive dose-response curve on external targets.

7 Discussion

7.1 External RF recognition is not only input matching

Our results suggest that external WiFi micro-Doppler recognition should not be treated only as a problem of making spectrograms look more target-like. Histogram matching and style-matched synthetic pretraining help in some cases, but their gains are smaller than the gain from replacing the learned CNN representation with a frozen DINOv2 feature space on OPERA. This does not make input-level methods irrelevant. Rather, it shows that input matching and representation quality should be evaluated separately.

7.2 When frozen foundation features help

The benefit of frozen DINOv2 is strongest on OPERA, where the acquisition and dataset shift is largest. On TTW, which is closer to the source acquisition setup, a CNN remains competitive. This target-dependent behaviour is important. Frozen vision features should not be presented as a universal replacement for RF-specific models; they are a strong baseline and diagnostic tool for external generalisation, especially when the target data differ substantially from the source collection.

7.3 What the negative augmentation result means

The FET result does not show that generative augmentation is impossible. It shows that generated target-like spectrograms did not improve a strong frozen representation in this setting. This suggests that future augmentation work should report not only visual realism or in-pool validation gains, but also external performance against a frozen-foundation-model baseline and a dose-response curve.

7.4 Limitations

The benchmark is modest in size, especially TTW. The conclusions should therefore be read as evidence about representation transfer under the acquisition settings studied here, not as a deployment-ready RF activity-recognition system. We evaluate three frozen vision encoders, but larger or newer models may change the ranking. We also assume that the target label set is known for class-restricted macro-F1; open-set and partial-label external evaluation remain important future directions.

7.5 Broader implications

Although the experiments focus on WiFi micro-Doppler sensing, the same issue appears in other non-photographic image-shaped modalities, including radar, sonar, ultrasound, audio spectrograms, seismic measurements, and some medical imaging formats. In these domains, useful task structure is often entangled with acquisition and rendering variation. Frozen visual representations may provide useful structural priors, but their value should be tested through strict external evaluation rather than assumed from in-domain accuracy.

8 Conclusion

This paper studied external WiFi micro-Doppler activity recognition under acquisition shift. We asked whether the gap is better addressed by making source or synthetic spectrograms look more target-like, or by using representations that remain stable when acquisition appearance changes. Using lab-collected WiFi micro-Doppler data for training and two held-out external settings for evaluation, 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 head trained on RF data, DINOv2 substantially improves recognition on OPERA, the harder external benchmark, while remaining competitive but not uniformly superior on the closer through-wall TTW setting.

The results show that frozen visual representations can preserve useful motion structure in RF spectrograms, but the benefit is not universal. The choice of foundation model matters, and the advantage is strongest when the target data differ substantially from the source acquisition setting. 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 OPERA. A factorised generative augmentation pipeline also fails to improve the frozen DINOv2 probe, indicating that adding more generated spectrograms is not sufficient when the representation baseline is already strong.

These findings could reframe external RF spectrogram recognition as a representation-transfer problem, not only an input-matching problem. For WiFi micro-Doppler and related non-photographic sensing modalities, future work should evaluate whether feature spaces preserve task-relevant structure under held-out acquisition shift, and should compare augmentation methods against strong frozen-foundation-model baselines.

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.