PATCHSHOT: RARE PRECIPITATION EVENTS ARE ENCODED YET HIDDEN FROM STANDARD READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rare precipitation events are difficult to recognize under severe label scarcity. We show that frozen 3DMAE radar representations already encode discriminative evidence for high-threshold events, yet standard readouts fail to recover it, revealing a *representation accessibility gap*. We introduce **PatchShot**, a case-disjoint inductive few-shot framework for parameter-efficient representation re-access. PatchShot freezes the weather encoder, learns a lightweight adapter through support-supervised contrastive learning, constructs support-derived prototypes, and calibrates the decision threshold using support data alone. Query patch scores are aggregated into case-level predictions without using query labels. We further propose **PatchShot-Env**, which conditions the re-access adapter on matched ERA5 low-mode context through lightweight expert routing. Across precipitation thresholds and support budgets, PatchShot substantially improves over the frozen QPE output and direct prototype readouts, while matched environmental routing provides further gains and shuffled-context controls weaken them. Probe and intervention analyses confirm that the improvement reflects recovered access to encoded rare-event evidence rather than encoder retraining.

1 INTRODUCTION

Deep learning has substantially advanced radar-based precipitation estimation and nowcasting. Early recurrent models learned nonlinear radar evolution from observation sequences (Xingjian, 2015; Shi et al., 2017), while generative and transformer-based approaches improved spatial organization, motion modeling, and forecast uncertainty (Ravuri et al., 2021; Gao et al., 2022). Recent systems further target intense precipitation through physical constraints, diffusion, and frequency-aware decomposition (Zhang et al., 2023; Gao et al., 2023; Yu et al., 2024; Lin et al., 2025). Despite this progress, reliable recognition of rare precipitation events remains unresolved.

Heavy and extreme rainfall occupy the tail of the data distribution, where positive examples are scarce and decisive evidence may be confined to only a few spatial regions. Objectives for continuous precipitation estimation are dominated by common weak and moderate cases; consequently, accurate precipitation fields do not necessarily yield reliable high-threshold event decisions. End-to-end adaptation is also fragile in this regime because the available rare-event labels are insufficient for stable retraining of a large weather encoder.

Pretrained weather representations make this failure more subtle. Masked autoencoders preserve transferable structure from partially observed inputs (He et al., 2022), and related pretraining has been extended to meteorological fields (Man et al., 2023). Large-scale weather models further demonstrate that learned representations can retain rich, reusable atmospheric information (Nguyen et al., 2023; Bi et al., 2023; Lam et al., 2023). However, such representations are typically assessed through a particular prediction head or after task-specific adaptation. Failure of the downstream prediction is therefore easily conflated with failure of the representation itself.

This conflation overlooks a critical distinction: information may be encoded in a representation without being accessible to the deployed readout. Few-shot methods use limited support labels to construct task-specific decisions (Vinyals et al., 2016; Snell et al., 2017), and simple prototype or nearest-neighbor rules can be effective when the embedding is already aligned with the target classes (Chen et al., 2019; Wang et al., 2019). A radar representation trained for reconstruction or quantitative precipitation estimation, however, may retain high-intensity structures while organizing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

them poorly for a binary rare-event boundary. Direct prototype construction then inherits rather than resolves this geometry.

We call this mismatch the *representation accessibility gap*. It separates two distinct failure modes: task-relevant information may be absent, requiring a better encoder, or it may be retained but hidden from the current decision rule. In the latter case, the appropriate objective is not to relearn the weather representation, but to recover access to it. This distinction is especially important under label scarcity, where encoder retraining is statistically fragile and computationally costly. It also reflects the broader view that probe performance depends not only on what a representation contains, but on how readily that information can be extracted (Voita & Titov, 2020).

We study this gap using frozen 3DMAE representations of multi-frame, three-dimensional weather radar. The original QPE readout is nearly ineffective at high precipitation thresholds, whereas lightweight supervised readouts applied to the same frozen patch tokens recover substantial discrimination. As illustrated in Figure 1, rare-event evidence can therefore remain encoded in the patch space while being hidden from both the original prediction head and a direct prototype boundary.

We introduce **PatchShot**, a case-disjoint inductive few-shot framework for representation re-access. PatchShot freezes the radar encoder and uses a small labeled support set to reorganize patch geometry through a lightweight adapter. Support-derived prototypes provide an event-oriented readout, and patch scores are aggregated into case-level predictions. Adaptation, prototype construction, and threshold calibration use support labels alone. We further introduce **PatchShot-Env**, which uses temporally matched ERA5 context (Hersbach et al., 2020) to condition how radar evidence is read under different atmospheric backgrounds.

Across heavy- and extreme-rain thresholds, PatchShot substantially improves over the frozen QPE output and direct prototype readouts, while matched environmental conditioning provides further gains. Probe analysis shows that PatchShot moves the prototype readout toward the strongest lightweight readout available on the same representation. Support-removal interventions further demonstrate that the recovered historical evidence materially changes query scores and decisions. Together, these results support a complete accessibility account: rare-event evidence is retained by the frozen representation, hidden from standard readouts, exposed by PatchShot, and used in the final decision.

Our contributions are threefold:

- We formulate rare precipitation recognition as a *representation accessibility* problem, distinguishing missing information from information encoded but hidden from standard readouts.
- We propose **PatchShot**, a case-disjoint few-shot framework that re-accesses frozen patch representations using limited support supervision, support-derived prototypes, and support-only decision construction.
- We introduce **PatchShot-Env**, an environment-guided extension, and provide predictive, probe-based, and intervention-based evidence that PatchShot recovers previously inaccessible rare-event information from frozen weather representations.

2 Related Work

2.1 Precipitation Modeling and Weather Representation Learning

Radar-based precipitation modeling has evolved from recurrent spatiotemporal prediction toward transformer-based and generative forecasting. ConvLSTM and TrajGRU established early deep-learning formulations for learning radar evolution (Xingjian, 2015; Shi et al., 2017). Subsequent work has improved the representation of spatial organization, nonlinear motion, forecast uncertainty, and high-intensity rainfall. Representative developments include deep generative radar forecasting (Ravuri et al., 2021), space–time transformers (Gao et al., 2022), physics-conditioned extreme-precipitation nowcasting (Zhang et al., 2023), and latent or residual diffusion models (Gao et al., 2023; Yu et al., 2024). Frequency-aware methods further decompose precipitation evolution into complementary amplitude and phase components (Lin et al., 2025).

These methods primarily aim to improve the predicted precipitation field. Even when extreme rainfall is explicitly considered, the main objective is usually to train a stronger end-to-end predictor.

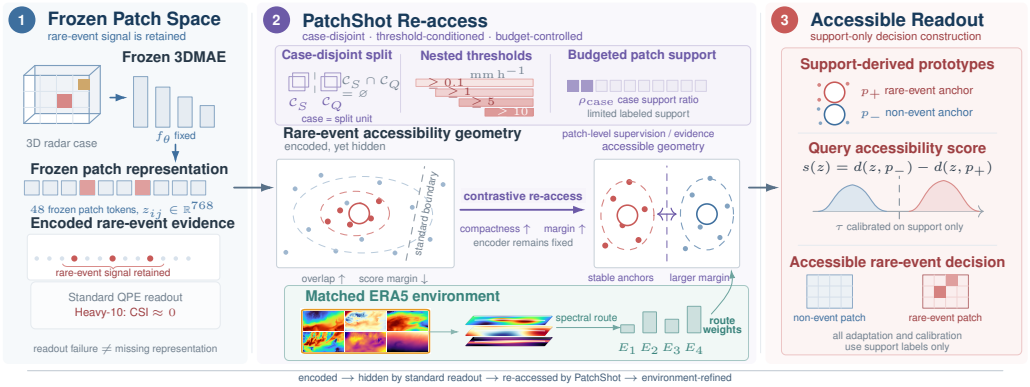


Figure 1: **PatchShot re-accesses rare precipitation events hidden from standard readouts.** A frozen 3DMAE representation is shared by both the failing standard QPE readout and the proposed support-guided path, isolating accessibility from encoder retraining. PatchShot is case-disjoint, threshold-conditioned, and budget-controlled: radar cases define the split unit, while patches provide localized supervision and evidence. Support-guided contrastive reorganization increases class compactness and decision margin, and a matched ERA5 environment provides a case-aligned refinement. All adaptation, prototype construction, and threshold calibration use support labels only.

Our work studies a different failure mode: a precipitation head may fail to recognize rare events even when its upstream representation still retains useful event evidence. This distinction shifts the question from how to improve the entire predictor to whether task-relevant information can be recovered from a frozen weather representation.

Masked pretraining and weather foundation models make this question increasingly relevant. Masked autoencoders learn transferable representations by reconstructing partially observed inputs (He et al., 2022), and the same principle has been extended to multivariable meteorological fields (Man et al., 2023). ClimaX (Nguyen et al., 2023), Pangu-Weather (Bi et al., 2023), and GraphCast (Lam et al., 2023) demonstrate that large-scale learning can capture rich meteorological structure from reanalysis data. More recent systems further combine learned representations with probabilistic or physically structured forecasting (Kochkov et al., 2024; Price et al., 2025; Bodnar et al., 2025).

Spectral operator learning provides another mechanism for representing large-scale environmental structure. Fourier neural operators learn global transformations efficiently in the frequency domain (Li et al., 2020), while FourCastNet and spherical Fourier operators extend spectral modeling to high-resolution and spherical weather dynamics (Pathak et al., 2022; Bonev et al., 2023). However, these representations are generally evaluated through dedicated forecasting heads or downstream training. Consequently, prediction quality alone does not reveal whether a failed readout reflects missing information or poor access to information already encoded in the representation.

2.2 Few-Shot Learning and Imbalanced Recognition

Few-shot learning transfers information from a small labeled support set to unseen queries. Matching Networks introduced support-conditioned prediction (Vinyals et al., 2016), while Prototypical Networks classify queries using distances to support-derived class representatives (Snell et al., 2017). Later studies showed that feature normalization, nearest-neighbor classification, and carefully designed baselines can remain competitive with more elaborate meta-learning methods (Chen et al., 2019; Wang et al., 2019; Hou & Sato, 2022). Diffusion-based approaches instead generate or refine prototypes and classifier parameters across sampled few-shot tasks (Du et al., 2023; Zhang et al., 2024).

Most prototype-based methods assume that the pretrained embedding is already organized around the target classes. This assumption is problematic for rare precipitation recognition. The frozen encoder is typically trained for reconstruction or continuous precipitation estimation rather than high-threshold event discrimination. Positive evidence is also sparse and heavily outnumbered by

ordinary patches. Directly constructing prototypes in the original space may therefore inherit a geometry that is informative but poorly aligned with the rare event boundary.

Long-tailed learning addresses related frequency imbalance using class-aware weighting, margin modification, or prior correction (Cui et al., 2019; Cao et al., 2019; Menon et al., 2020). Contrastive learning instead modifies representation geometry by encouraging within-class compactness and between-class separation (Chen et al., 2020; Khosla et al., 2020). Parametric contrastive learning and balanced contrastive learning further incorporate class centers or class-frequency control (Cui et al., 2021; Zhu et al., 2022). These approaches motivate the support-supervised adaptation used in PatchShot, but our objective is not general long-tailed representation learning. We use limited, case-disjoint event supervision to make a fixed weather representation accessible at a specified precipitation threshold.

2.3 Efficient Adaptation, Context Routing, and Representation Accessibility

Parameter-efficient adaptation changes downstream behavior while preserving most or all of a pretrained backbone. Adapter modules (Houlsby et al., 2019) and low-rank adaptation (Hu et al., 2022) show that small trainable components can efficiently specialize large pretrained models. Mixture-of-experts architectures provide a complementary mechanism in which inputs or tokens are routed through specialized transformations (Riquelme et al., 2021; Fedus et al., 2022; Zhou et al., 2022). PatchShot follows the parameter-efficient principle but uses a lightweight adapter specifically to reorganize frozen patch evidence. PatchShot-Env further conditions this re-access process on temporally matched ERA5 background information (Hersbach et al., 2020), rather than treating ERA5 as an independent prediction stream.

The difference between encoded information and usable information is closely related to representation probing. Lightweight probes can reveal whether task-relevant signals remain recoverable from a frozen feature space, although probe performance alone does not establish that a deployed decision rule actually uses those signals (Voita & Titov, 2020). Training-data attribution provides a complementary perspective by estimating which historical examples affect a particular prediction. Influence functions and gradient-tracing methods are representative approaches (Koh & Liang, 2017; Pruthi et al., 2020).

Our work connects these perspectives in rare-event weather recognition. We use lightweight probe envelopes to estimate how much discrimination is recoverable from each representation, and support-removal interventions to test whether identified historical evidence materially changes the query score and decision. This combination operationalizes the *representation accessibility gap*: rare-event information may be retained by a frozen weather encoder while remaining hidden from its standard downstream readout.

3 Rare-Event Accessibility

3.1 Multi-Frame Frozen Representation

Let $\mathbf{X}_i = (\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(T)})$ denote a multi-frame volumetric radar case, where each observation contains L vertical levels over an $H \times W$ horizontal grid. A pretrained 3D masked autoencoder encodes each frame independently. The resulting features are averaged over the historical window and spatially aggregated into P event patches:

$$\mathbf{Z}_i = F_{\Theta_0}(\mathbf{X}_i) = \left\{ \mathbf{z}_{i,j} = \Pi_j \left(\frac{1}{T} \sum_{t=1}^{T} f_{\theta_0}(\mathbf{x}_i^{(t)}) \right) \right\}_{j=1}^{P}, \qquad \Theta_0 \text{ fixed.} \tag{1}$$

Here Π_j denotes spatial aggregation for patch j, and $\mathbf{z}_{i,j} \in \mathbb{R}^d$. Each token is spatially indexed but summarizes vertically structured radar evidence over multiple historical frames. Temporal averaging retains multi-frame context while discarding explicit temporal ordering.

For precipitation threshold γ, patch- and case-level event labels are defined by

$$y_{i,j}^{(\gamma)} = \mathbb{I}\left[\max_{\mathbf{u} \in \Omega_j} r_i(\mathbf{u}) \geq \gamma \right], \qquad y_i^{(\gamma)} = \max_{1 \leq j \leq P} y_{i,j}^{(\gamma)}, \tag{2}$$

where r_i is the aligned precipitation target and Ω_j is the region associated with patch j. Radar cases are the statistical units, whereas patches are the localized units of rare-event evidence.

3.2 Case-Disjoint Readout Setting

Let $\mathcal{C}_S$ and $\mathcal{C}_Q$ denote the support and query case sets. Both are processed by the same frozen mapping, with disjointness enforced at the case level:

$$\mathcal{C}_S \cap \mathcal{C}_Q = \varnothing, \qquad \mathbf{Z}_i^S = F_{\Theta_0}(\mathbf{X}_i),\ i \in \mathcal{C}_S, \qquad \mathbf{Z}_i^Q = F_{\Theta_0}(\mathbf{X}_i),\ i \in \mathcal{C}_Q. \tag{3}$$

Support and query tokens therefore contain the same form of temporally pooled radar context. They differ only in label availability and their role in adaptation and evaluation.

For support ratio ρ, the available supervision is $\mathcal{S}_\rho^{(\gamma)} = \{(\mathbf{Z}_i^S, \mathbf{y}_i^{(\gamma)}) : i \in \mathcal{C}_S^\rho\}$. A support-constructed readout is represented by $\mathsf{R}_\mathcal{S} = (\ell_\mathcal{S}, \mathcal{A}_\mathcal{S}, \tau_\mathcal{S})$, where $\ell_\mathcal{S}$ scores patch evidence, $\mathcal{A}_\mathcal{S}$ aggregates it at the case level, and $\tau_\mathcal{S}$ is a support-calibrated threshold:

$$s_i^{\mathsf{R}} = \mathcal{A}_\mathcal{S}\Big(\{\ell_\mathcal{S}(\mathbf{z}_{i,j}^Q)\}_{j=1}^P\Big), \qquad \widehat{y}_i^{(\gamma)} = \mathbb{I}[s_i^{\mathsf{R}} \geq \tau_\mathcal{S}]. \tag{4}$$

All trainable components and decision rules are determined from the support set alone. Query labels are excluded from adaptation, readout construction, model selection, and threshold calibration.

3.3 The Accessibility Question

A failed rare-event prediction admits two distinct explanations: task-relevant discrimination may be absent from the frozen representation, or it may be retained but hidden from the deployed readout. Distinguishing these two regimes is the central scientific problem of this work.

Let $\mathfrak{R}_B$ denote a controlled family of support-trained readouts with bounded adaptation complexity B, and let $\mathcal{J}_Q(\mathsf{R}; \mathbf{Z})$ denote the query discrimination achieved by R. The corresponding readout envelope is

$$\mathcal{E}_B(\mathbf{Z}) = \sup_{\mathsf{R} \in \mathfrak{R}_B} \mathcal{J}_Q(\mathsf{R}; \mathbf{Z}), \tag{5}$$

which measures the discrimination empirically recoverable from $\mathbf{Z}$ under limited supervision and controlled readout capacity. It is a diagnostic frontier rather than an information-theoretic measure of everything encoded by the representation.

For a target readout R, its exposed and hidden discrimination satisfy

$$\underbrace{\mathcal{E}_B(\mathbf{Z})}_{\text{recoverable}} = \underbrace{\mathcal{J}_Q(\mathsf{R}; \mathbf{Z})}_{\text{exposed}} + \underbrace{[\mathcal{E}_B(\mathbf{Z}) - \mathcal{J}_Q(\mathsf{R}; \mathbf{Z})]}_{\mathcal{H}_B(\mathsf{R}|\mathbf{Z})\ :\ \text{hidden}}. \tag{6}$$

A substantial envelope together with a large $\mathcal{H}_B$ defines the *encoded-yet-hidden* regime: rare-event discrimination is recoverable from the frozen representation, but the current readout fails to expose it. The resulting scientific question is whether limited support supervision can construct a low-complexity readout that reduces this hidden component without changing the weather encoder:

$$\boxed{\mathsf{R}_\mathcal{S}^\star \in \arg\min_{\mathsf{R} \in \mathfrak{R}_B} \mathcal{H}_B(\mathsf{R} \mid \mathbf{Z}) \quad \text{s.t.} \quad \mathsf{R} = \mathcal{L}(\mathcal{S}_\rho^{(\gamma)}),\ \Theta = \Theta_0,\ \mathcal{C}_S \cap \mathcal{C}_Q = \varnothing.} \tag{7}$$

This is a readout problem rather than an encoder-retraining problem. PatchShot addresses it by reorganizing the frozen patch geometry so that a support-derived rare-event readout exposes more of the discrimination already retained in $\mathbf{Z}$.

4 PatchShot Re-Access

PatchShot addresses the accessibility problem in Eq. (7) by learning a compact support-conditioned readout over the frozen multi-frame radar tokens. As illustrated in Figure 2, the weather encoder is never updated. Support labels are used only to reorganize the patch geometry, construct event prototypes, and calibrate the final decision; query cases reuse the resulting readout without label access.

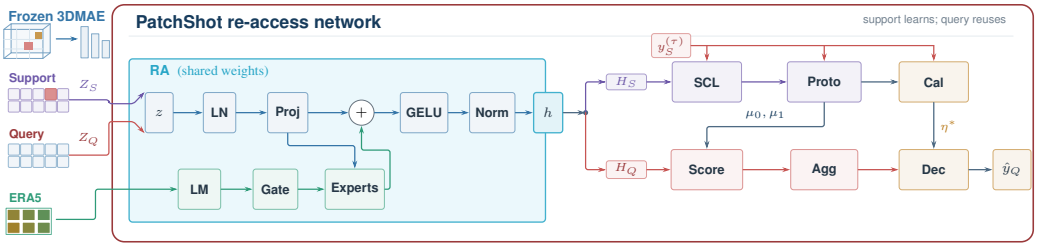


Figure 2: **Technical workflow of PatchShot.** RA: re-access adapter; SCL: supervised contrastive learning; Proto: prototype construction; Cal: support-only calibration; LM: low-mode encoding; Agg: case aggregation; Dec: decision.

4.1 Support-Supervised Re-Access

Direct prototype construction assumes that the frozen geometry is already aligned with the rare-event boundary. PatchShot instead learns a lightweight adapter $g_\phi : \mathbb{R}^d \to \mathbb{R}^m$ that exposes threshold-relevant evidence:

$$\mathbf{h}_{i,j} = g_\phi(\mathbf{z}_{i,j}) = \frac{\sigma(\mathbf{W}\ \mathrm{LN}(\mathbf{z}_{i,j}) + \mathbf{b})}{\|\sigma(\mathbf{W}\ \mathrm{LN}(\mathbf{z}_{i,j}) + \mathbf{b})\|_2}, \qquad m = 128, \tag{8}$$

where σ is GELU. The same adapter is applied to support and query tokens, while gradients are computed from support patches only. Let $\mathcal{I}_S$ index the labeled support patches and $\mathcal{P}(a) = \{p \in \mathcal{I}_S \setminus \{a\} : y_p = y_a\}$ denote the same-event partners of anchor a. The adapter is optimized with the support-supervised contrastive objective

$$\mathcal{L}_{\mathrm{re}} = -\frac{1}{|\mathcal{I}_S|} \sum_{a \in \mathcal{I}_S} \frac{1}{|\mathcal{P}(a)|} \sum_{p \in \mathcal{P}(a)} \log \frac{\exp(\mathbf{h}_a^\top \mathbf{h}_p / \tau)}{\sum_{b \in \mathcal{I}_S \setminus \{a\}} \exp(\mathbf{h}_a^\top \mathbf{h}_b / \tau)}. \tag{9}$$

This objective does not relearn general radar dynamics. It contracts patches sharing the threshold-conditioned event state and enlarges the margin between rare-event and non-event evidence.

4.2 Support-Derived Readout

After adaptation, PatchShot constructs one normalized prototype for each event state:

$$\bar{\mu}_c = \mathrm{Norm}\left(\frac{1}{|\mathcal{I}_c|} \sum_{a \in \mathcal{I}_c} h_a \right), \qquad \mathcal{I}_c = \{a \in \mathcal{I}_S : y_a = c\}, \quad c \in \{0, 1\}. \tag{10}$$

For patch j of case i, the signed prototype margin is

$$d_{i,j} = \bar{h}_{i,j}^\top (\bar{\mu}_1 - \bar{\mu}_0), \qquad s_i = \mathrm{Agg}_j \left(\{d_{i,j}\}_{j=1}^P \right), \tag{11}$$

where a larger $d_{i,j}$ indicates stronger rare-event evidence and Agg maps localized patch evidence to a case-level score. The decision threshold is selected exclusively from support case scores and labels:

$$\eta^\star = \arg\max_\eta \mathrm{CSI}\left(\{(s_i, y_i^{(\gamma)}) : i \in \mathcal{C}_S^\rho\} \right), \qquad \hat{y}_q^{(\gamma)} = \mathbb{I}[s_q \geq \eta^\star]. \tag{12}$$

Thus, the deployed readout is completely specified by $\{g_\phi, \bar{\mu}_0, \bar{\mu}_1, \eta^\star\}$, all learned or constructed from support data alone; query labels are used only for final evaluation.

4.3 Environment-Conditioned Re-Access

Radar morphology can carry different event meanings under different large-scale atmospheric backgrounds. PatchShot-Env therefore uses a temporally matched ERA5 descriptor $\mathbf{e}_i$ to route each case through a small expert bank:

$$\boldsymbol{\alpha}_i = \mathrm{softmax}\big(r_\psi(\mathbf{e}_i)\big), \qquad \tilde{\mathbf{z}}_{i,j} = \mathbf{z}_{i,j} + \sum_{k=1}^{K} \alpha_{i,k}\, E_k(\mathrm{LN}(\mathbf{z}_{i,j}))\,. \tag{13}$$

Table 1: **PatchShot comparison for heavy rain (≥ 5 mm h^{-1}).** All methods use the same frozen 3DMAE patch representation, identical case-disjoint support/query splits, support-only model selection and threshold calibration, and five random seeds. CSI, F1, AP, and AUROC are reported in percent; Sep. is the standardized positive–negative separation of the final query score. Best and second-best completed results are shown in **bold** and underlined, respectively.

Method	Adaptation	CSI ↑	F1 ↑	AP ↑	AUROC ↑	Sep. ↑
Direct prediction						
QPE	Direct prediction	0.00±0.00	0.00±0.00	6.15±0.20	37.09±0.16	0.4833±0.0009
Standard support readouts						
ProtoNet	Dual prototypes	16.47±0.39	28.28±0.57	21.98±0.76	76.79±0.47	1.0144±0.0228
SimpleShot	Centering + normalization	17.74±0.60	30.13±0.87	21.80±0.57	77.24±0.43	1.0553±0.0202
Balanced Linear	Class-balanced linear classifier	22.35±0.42	36.54±0.56	27.86±0.32	84.22±0.42	1.3799±0.0213
Balanced CE	Balanced CE projection	20.61±2.42	34.10±3.34	26.02±2.09	80.56±1.62	1.1819±0.0697
Shrinkage QDA	Shrinkage quadratic classifier	17.13±0.46	29.25±0.68	22.00±0.39	77.28±0.45	1.0155±0.0083
Label-free representation adaptation						
Identity	Frozen representation	16.02±0.51	27.61±0.76	21.62±0.69	76.62±0.47	1.0042±0.0245
PCA	Linear compression	16.32±0.57	28.06±0.85	18.53±0.74	75.02±0.46	0.9774±0.0168
Autoencoder	Reconstruction projection	17.39±0.94	29.62±1.36	20.78±1.04	77.06±0.55	1.0788±0.0253
Adversarial	Adversarial projection	17.82±0.41	30.25±0.58	21.23±0.34	77.48±0.25	1.0898±0.0139
Denoising	Denoising projection	14.89±0.84	25.91±1.28	18.33±1.14	72.59±1.71	0.8548±0.0719
Spectral	Patch-grid spectrum	18.01±0.57	30.52±0.81	21.56±0.85	77.21±0.58	1.0757±0.0310
Instance Contrastive	Self-supervised contrastive	14.48±0.81	25.29±1.25	17.78±1.28	69.74±1.45	0.7292±0.0575
Support-aware adaptation						
BCL†	Balanced contrastive	21.96±0.87	36.00±1.18	27.01±0.65	81.47±0.71	1.2960±0.0480
PaCo-Proto†	Parametric contrastive + prototype	23.15±0.88	37.59±1.16	29.50±1.51	84.70±0.69	1.5397±0.0466
PaCo-Center†	Parametric contrastive + class center	22.93±0.94	37.30±1.25	28.92±1.63	84.67±0.77	1.5390±0.0486
PatchShot (Ours)	Event-supervised contrastive re-access	**23.29±0.57**	**37.78±0.75**	**29.54±0.66**	**84.88±0.54**	**1.5433±0.0401**
Meta-trained adaptation						
ProtoDiff‡	Diffused prototypes	16.94±2.70	28.88±4.06	21.67±4.76	76.02±4.80	1.0272±0.1969
MetaDiff‡	Diffused classifier weights	21.03±0.33	34.74±0.45	25.39±0.63	82.26±0.28	1.3666±0.0186

† denotes controlled binary adaptations of BCL or PaCo; ‡ denotes controlled diffusion-based meta-adaptation. PatchShot uses event-supervised contrastive re-access, support-derived prototypes, and support-only threshold calibration. Label-free methods do not use event labels to fit the representation adapter; query labels are used only for final reporting.

The routed token $\widetilde{\mathbf{z}}_{i,j}$ replaces $\mathbf{z}_{i,j}$ in Eq. (8); the remaining contrastive, prototype, aggregation, and calibration stages are unchanged. The unified PatchShot-Env configuration uses signed low-order ERA5 modes for every event threshold. ERA5 therefore acts as a case-aligned readout condition rather than an independent precipitation predictor.

PatchShot changes only how the frozen representation is accessed: F_{Θ_0} remains fixed, while the support-conditioned geometry and decision rule expose a larger fraction of its recoverable rare-event discrimination.

5 Evaluation

5.1 Experimental Setting

We evaluate whether limited support supervision can expose rare-event discrimination retained by a frozen radar representation. The radar observations cover $108.50°$E–$117.50°$E and $19.04°$N–$26.04°$N, with a 1-km horizontal grid spacing and a spatial grid of 700×900. Each case aggregates six historical radar volumes sampled at 12-min intervals, spanning a 60-min history. We use 3,000 development cases and 1,922 evaluation cases (1,918 with ERA5). Each case is represented by 48 temporally pooled 3DMAE patch tokens on a 6×8 grid, with token dimension $d = 768$. Heavy- and extreme-rain events are defined using thresholds of 5 and $10\,\mathrm{mm\,h^{-1}}$, denoted H5 and H10.

For each of five paired seeds, half of the evaluation cases form a fixed query set. Nested support subsets contain 19, 96, 192, 384, and 959 labeled cases, corresponding to nominal support budgets of 1%, 5%, 10%, 20%, and 50%, respectively. All methods within a seed share identical support and query cases. Cases, rather than patches, are the split unit.

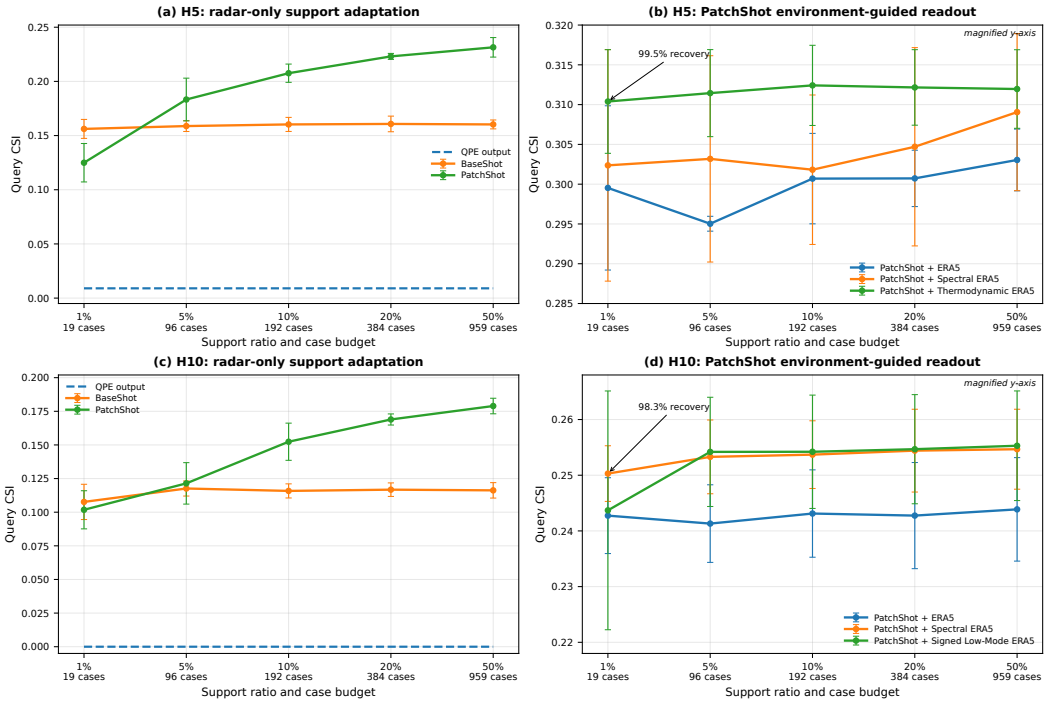


Figure 3: **Support efficiency of PatchShot.** Panels (a) and (c) compare radar-only support adaptation with the frozen QPE output and BaseShot at H5 and H10. Panels (b) and (d) compare environment-guided PatchShot readouts constructed from matched ERA5 variants. The horizontal axis reports the support ratio and corresponding case budget.

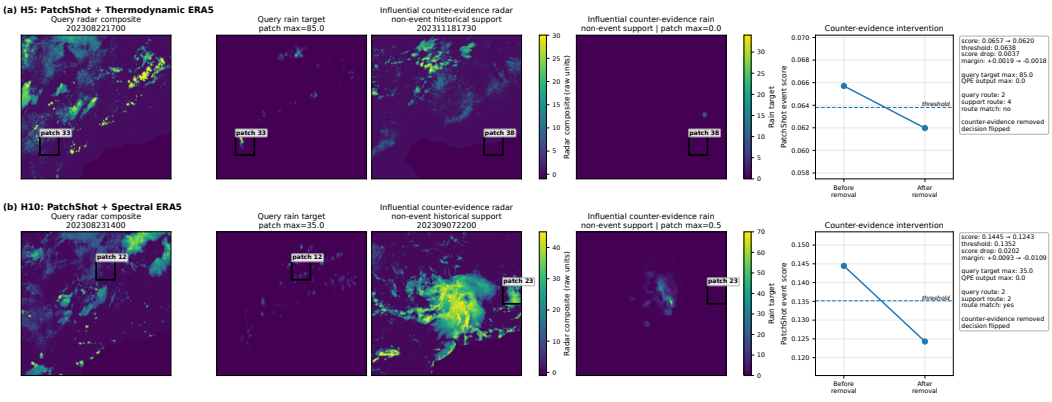


Figure 4: **Case-level traces of support-evidence interventions.** Each row shows a query radar composite and rain target, the traced influential historical support case, and the PatchShot score before and after removing that evidence. Black boxes identify the involved patches, and the dashed line denotes the support-calibrated decision threshold.

CSI is the primary metric, complemented by F1, AP, AUROC, and positive–negative score separation. Readout fitting, model selection, prototype construction, and threshold calibration use support data alone. Decision thresholds are selected from support case scores and labels; query labels are used only for final reporting. Complete baselines, implementation details, and additional protocol diagnostics are provided in the appendix.

5.2 Recognition and Environmental Re-Access

Tables 1 and 3 compare readouts constructed from the same frozen 3DMAE representation at 20% support. The original QPE output obtains zero CSI and F1 at both thresholds, although its nonzero AP indicates that some ranking information remains. A direct prototype readout already recovers meaningful discrimination, reaching 16.47 CSI at H5 and 12.22 CSI at H10. This establishes that readout failure does not imply that the frozen representation is empty of rare-event evidence.

PatchShot increases CSI to 23.29 ± 0.57 at H5 and 16.65 ± 1.58 at H10. At H5, it also achieves 29.54 ± 0.66 AP and the largest score separation among the completed baselines. At H10, PatchShot remains competitive with balanced and parametric contrastive readouts despite using no episodic meta-training. The result is therefore not universal dominance over every classifier; it is the recovery of a strong rare-event decision from a frozen encoder using only the local support set.

Figure 3 shows that radar-only PatchShot benefits consistently from additional supervision. Beyond the underdetermined 1% regime, its CSI increases with the support budget and separates progressively from the direct baseline at both thresholds. Environmental conditioning substantially improves the low-budget regime: the strongest matched ERA5 routes retain 99.5% of the H5 and 98.3% of the H10 performance obtained with 50% support while using only 1% support.

Table 4 isolates physical alignment from additional expert capacity. The unified PatchShot-Env model reaches 30.95 CSI at H5 and 25.59 CSI at H10, compared with 30.27 and 24.30 for the shared radar–ERA5 readout. Case-shuffled routing reduces CSI to 30.01 and 23.58, whereas uniform routing provides only small changes. The gains therefore depend primarily on case-aligned environmental context rather than on the presence of additional experts alone. The final model fixes signed low-mode routing for both thresholds; diagnostic threshold-specific routes are reported only in the appendix.

5.3 Accessibility and Evidence Use

Table 6 evaluates the accessibility quantities from Section 3. At H5, the frozen representation obtains a prototype CSI of 0.186 against a probe envelope of 0.236, leaving a hidden component of 0.050. After environment-conditioned re-access, the prototype score rises to 0.312 against an envelope of 0.316, reducing the hidden component to 0.004. The corresponding accessibility fraction increases from 0.788 to 0.987.

The same pattern appears at H10. The frozen representation exposes 0.150 CSI against a 0.183 envelope, whereas signed low-mode PatchShot exposes 0.255 against 0.258. Its hidden component decreases from 0.032 to 0.003, and accessibility rises from 0.823 to 0.988. PatchShot therefore improves not only absolute recognition performance, but also the extent to which a simple prototype rule approaches the best lightweight readout available on the same representation.

We finally test whether the recovered evidence materially affects the decision. Removing the ten most influential support cases and recomputing the prototypes causes event-score drops of approximately 0.032 at H5 and 0.038 at H10, with true-positive decision-flip rates near 40%. Random removal produces substantially smaller score changes and fewer than 8% flips. Figure 4 shows representative boundary flips, while Appendix Fig 6 reports the full intervention curves.

6 Conclusion

We studied rare precipitation recognition as a representation accessibility problem rather than an encoder-retraining problem. Frozen multi-frame 3DMAE representations retain substantial high-threshold event evidence, although standard precipitation outputs and direct prototype readouts expose only part of it. PatchShot uses limited, case-disjoint support supervision to reorganize the frozen patch geometry, construct event prototypes, and calibrate a support-only decision rule. PatchShot-Env further conditions this re-access process on matched large-scale atmospheric context. Across predictive, probe-based, and support-removal analyses, the results consistently support the *encoded-yet-hidden* interpretation. More broadly, our findings suggest that failures of pretrained weather models need not always be addressed by relearning the encoder: under label scarcity, the missing capability may instead lie in how the representation is read.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pp. 2806–2823. PMLR, 2023.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Yingjun Du, Zehao Xiao, Shengcai Liao, and Cees Snoek. Protodiff: Learning to learn prototypical networks by task-guided diffusion. *Advances in Neural Information Processing Systems*, 36: 46304–46322, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Mingcheng Hou and Issei Sato. A closer look at prototype classifier for few-shot image classification. *Advances in Neural Information Processing Systems*, 35:25767–25778, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. Alphapre: Amplitude-phase disentanglement model for precipitation nowcasting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17841–17850, 2025.

Xin Man, Chenghong Zhang, Jin Feng, Changyu Li, and Jie Shao. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*, 2023.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

SHI Xingjian. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28:1, 2015.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Baoquan Zhang, Chuyao Luo, Demin Yu, Xutao Li, Huiwei Lin, Yunming Ye, and Bowen Zhang. Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16687–16695, 2024.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619 (7970):526–532, 2023.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6908–6917, 2022.

A Additional Experiments and Diagnostics

Purpose and reading order. The main paper establishes the predictive claim that a frozen weather representation can retain rare-event evidence that is not adequately exposed by its standard readout. The appendix does not repeat that claim. Instead, it examines the mechanism through the following sequence of controlled questions:

1. Which parts of support-side re-access create the improvement?
2. Is the final decision robust to the calibration rule?
3. Does the deployed readout materially use identifiable support evidence?
4. Does the conclusion remain valid at the stricter H10 threshold?
5. What is contributed by matched ERA5 and AGRI context?
6. How much recoverable discrimination is exposed by the prototype readout?
7. Are conclusions stable across lightweight probes?
8. Are the evaluation episodes controlled against direct leakage?

This ordering forms an identification ladder rather than an additive decomposition of performance. Each experiment introduces a local counterfactual while preserving the remaining components as closely as possible.

Table 2: **Component attribution at the final Table-1 operating point.** All metrics are reported in percentage points as mean ± standard deviation over five paired splits. "Support" denotes a decision threshold selected exclusively from the current support set. A1 is the original classifier-based CE reference, whereas A2–A5 are controlled support-side re-access variants.

					Heavy rain at 5			Heavy rain at 10		
ID	Variant	Support labels	Objective / readout	Calibration	CSI	AP	AUROC	CSI	AP	AUROC
A0	BaseShot	True	None / prototype	Support	16.02±0.51	21.62±0.69	76.62±0.47	11.91±0.29	14.98±0.29	77.23±0.51
A1†	Balanced CE head	True	CE / classifier	Support	20.61±2.42	26.02±2.09	80.56±1.62	15.20±1.34	18.79±1.21	81.34±0.82
A2	Shuffled-label SupCon	Shuffled	SupCon / prototype	Support	16.64±1.07	17.50±1.92	77.68±1.45	11.10±0.96	11.74±1.19	77.49±1.67
A3	Matched CE re-access	True	CE / prototype	Support	22.55±1.37	29.49±2.46	84.45±0.65	**16.84±0.83**	**20.01±1.02**	**84.27±0.60**
A4	SupCon, zero margin	True	SupCon / prototype	Fixed zero	20.84±0.69	29.54±0.66	84.88±0.54	14.39±0.70	19.54±1.55	83.97±0.74
A5	**Full PatchShot**	True	SupCon / prototype	Support	**23.29±0.57**	**29.54±0.66**	**84.88±0.54**	16.65±1.58	19.54±1.55	83.97±0.74

†A1 is the original classifier-based baseline and serves as a reference rather than a decoder-matched ablation. A3 is the objective-matched CE control: it uses the same projection architecture, balanced sampling, optimization budget, prototype decoder, and support calibration as A5.

Protocol blocks. The supplementary experiments use three related but distinct case pools. The radar-only comparison and component-attribution experiment use the complete radar feature pool of 1,922 cases, yielding 384 support cases and 961 query cases at the 20% operating point. Experiments that require temporally aligned ERA5 descriptors use the 1,918-case aligned pool, yielding 384 support cases and 959 query cases. AGRI analyses use a smaller complete-observation subset and are interpreted only within that subset.

All numerical comparisons are paired within their own protocol block. Absolute values from different blocks should not be compared as though they were obtained from identical case populations. Unless otherwise stated, the 3DMAE encoder remains frozen, cases are the support/query split unit, patches are the supervision and evidence units, and query labels are used only for final reporting.

The unified PatchShot-Env model uses signed low-order ERA5 modes at both event thresholds. Threshold-specific thermodynamic, spectral, and dynamic routes are retained as diagnostic configurations rather than alternative definitions of the final model.

A.1 Component Attribution of Support-Side Re-access

The main comparison shows that support-guided adaptation improves over a frozen prototype readout, but that comparison alone cannot determine whether the gain is caused by additional parameters, event-label alignment, the adaptation objective, the prototype decoder, or the support-derived operating point. We therefore construct a controlled sequence that changes these factors separately.

A0 and A5 reproduce the radar-only BaseShot and PatchShot rows from the main H5 comparison in Table 1. A1 retains the original classifier-based Balanced CE method as an external supervised reference. A2–A5 use the same frozen tokens and paired support/query cases. A3 is the decoder-matched objective control for A5: the projection architecture, balanced sampling, optimization budget, prototype decoder, and support calibration remain fixed, while supervised contrastive learning is replaced by balanced cross-entropy.

The rows answer four distinct counterfactual questions. A0 versus A2 tests whether trainable capacity helps without valid labels. A0 versus A3 tests whether task-consistent CE supervision can re-access the representation. A3 versus A5 compares two matched re-access objectives. A4 versus A5 isolates support-side calibration while holding the score ranking fixed. The complete controlled comparison is reported in Table 2.

Capacity is not sufficient. The shuffled-label SupCon control preserves the projection network, training schedule, and prototype construction while breaking the correspondence between support labels and radar structure. It therefore tests whether a nonlinear adapter and additional optimization are sufficient by themselves. The reduction in AP, particularly at H10, shows that the gain cannot be attributed to parameter count or training steps alone. The support labels must remain task-consistent with the observed patch evidence.

Task-consistent supervision is the primary mechanism. Both matched CE and PatchShot recover substantially more discrimination than the frozen BaseShot representation. Thus, the central

mechanism is not exclusive to a particular contrastive loss. A small labeled support set can reorganize the accessible geometry of a frozen representation that was originally trained for reconstruction and precipitation estimation rather than threshold-conditioned event recognition.

The objectives nevertheless impose different geometric preferences. Balanced CE optimizes discrimination through a temporary classifier, whereas supervised contrastive learning directly contracts patches with the same event state and separates patches with different states. The latter geometry is naturally aligned with the prototype decoder used at inference. PatchShot should therefore be understood as a stable and prototype-compatible realization of support-side re-access, rather than as the only objective capable of exposing encoded evidence.

Calibration is an independent decision component. A4 and A5 share the same trained adapter, support prototypes, and query scores. Their AP and AUROC values are therefore identical. The difference in CSI isolates the effect of converting a continuous prototype ranking into a binary rare-event decision. Support calibration improves H5 and H10 CSI without altering the underlying ranking, demonstrating that representation organization and operating-point selection are related but distinct parts of the readout.

A.2 Calibration and Evidence-Use Diagnostics

PatchShot produces a continuous signed prototype margin before making a binary prediction. We therefore distinguish representation geometry, query-score ranking, and thresholded event decisions. AP and AUROC primarily characterize ranking, whereas CSI, precision, and recall also depend on the selected operating point.

A.2.1 Sensitivity to the operating-point rule

We compare the support-selected CSI/F1-optimal threshold with Youden's J statistic and a fixed zero-margin rule. CSI and F1 produce the same optimal threshold because, for a fixed confusion matrix,

$$\mathrm{F1} = \frac{2\,\mathrm{CSI}}{1 + \mathrm{CSI}},$$

which is monotonic in CSI. They are therefore treated as one operating rule rather than independent alternatives. Youden's J emphasizes the balance between sensitivity and specificity, whereas zero margin tests whether the prototype geometry is already centered well enough to require no support calibration.

A.2.2 Intervention on historical support evidence

Prototype construction makes the support set part of the deployed decision rule. We test whether this dependence is diffuse or concentrated through exact leave-support-out interventions. For each query, the trained adapter is held fixed, selected support cases are removed, the positive and negative prototypes are recomputed, and the query score is evaluated again:

$$\Delta s_q(\mathcal{K}) = s_q(S) - s_q(S \setminus \mathcal{K}),$$

where $\mathcal{K}$ denotes the removed support cases. This diagnostic measures dependence on the constructed readout; it does not retrain the frozen encoder or estimate influence on the original 3DMAE pretraining process.

Support cases are ranked using complete intervention arrays rather than visual appearance. Joint removal of the top $K \in \{1, 3, 5, 10\}$ cases is compared with an equally sized random removal control.

Figure 5 evaluates sensitivity to support-only threshold calibration, while Figure 6 reports the cumulative support-evidence intervention.

The threshold analysis shows that the PatchShot score remains informative across alternative operating rules, while the precision–recall point can change substantially. Support calibration therefore selects an appropriate rare-event operating point on an already informative score rather than manufacturing the ranking.

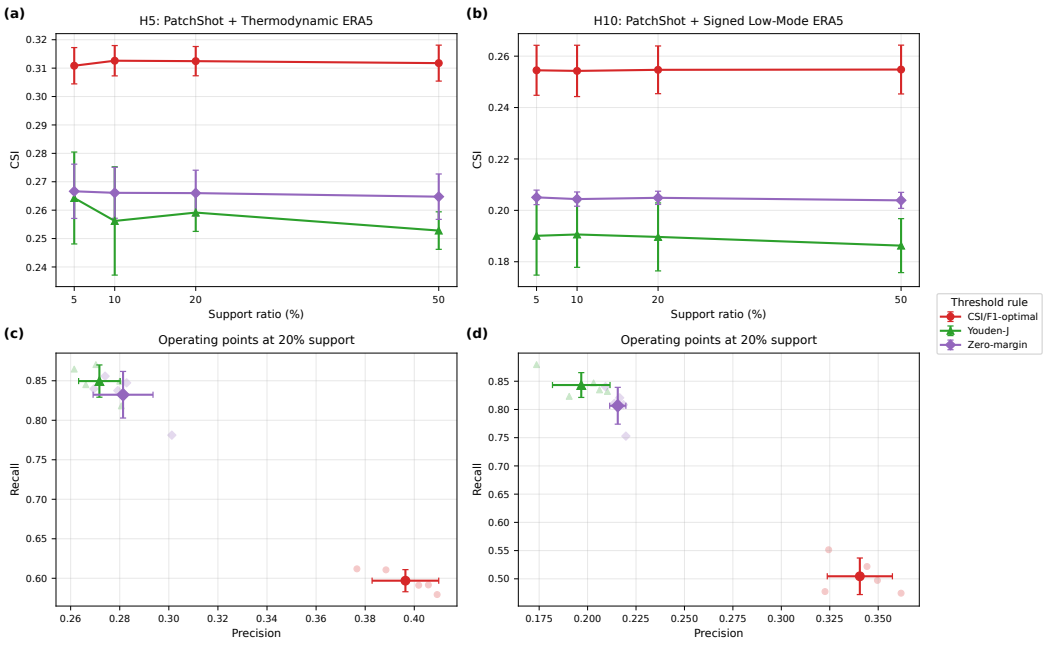


Figure 5: **Sensitivity to support-only threshold calibration.** Panels (a,b) compare query CSI across support ratios under three threshold rules; panels (c,d) show the corresponding precision–recall operating points at 20% support.

The cumulative intervention analysis shows that support examples have highly non-uniform effects. Removing influential cases produces larger score changes and more decision flips than random removal. Both positive evidence and counter-evidence are observed, indicating that PatchShot is not a nearest-positive-example rule: non-event support cases also shape the boundary through the negative prototype.

These interventions establish operational use rather than unique meteorological causation. An influential support case is one whose removal changes the constructed readout; it need not be the unique physical cause of the query event.

A.3 Complete Comparison at the H10 Threshold

H10 represents a stricter rare-event regime than H5. Positive support patches are fewer, the positive prototype is estimated from less evidence, and a small score displacement can produce a larger change in the selected operating point. We therefore report the complete H10 comparison rather than showing only the strongest methods.

Table 3 reports the complete comparison at H10 and complements the main H5 results in Table 1.

Standard support readouts test whether a conventional classifier or prototype boundary can be fitted directly in the frozen feature space. Label-free adaptation tests whether generic compression, reconstruction, denoising, spectral organization, or instance-level contrast can improve the representation without event labels. Support-aware adaptation tests whether threshold-specific supervision is necessary to expose the event geometry.

Label-free adaptation does not consistently improve over the frozen identity representation. Generic feature reorganization is therefore insufficient for the strict H10 event boundary. Support-aware methods form a substantially stronger group, although their relative ordering is close.

PatchShot remains competitive with balanced linear and parametric contrastive alternatives but does not universally dominate every H10 metric. This is consistent with the scope of the paper: the aim is to recover a strong support-constructed decision from a frozen weather encoder, not to claim that one contrastive objective must outperform every classifier in every rare-event regime.

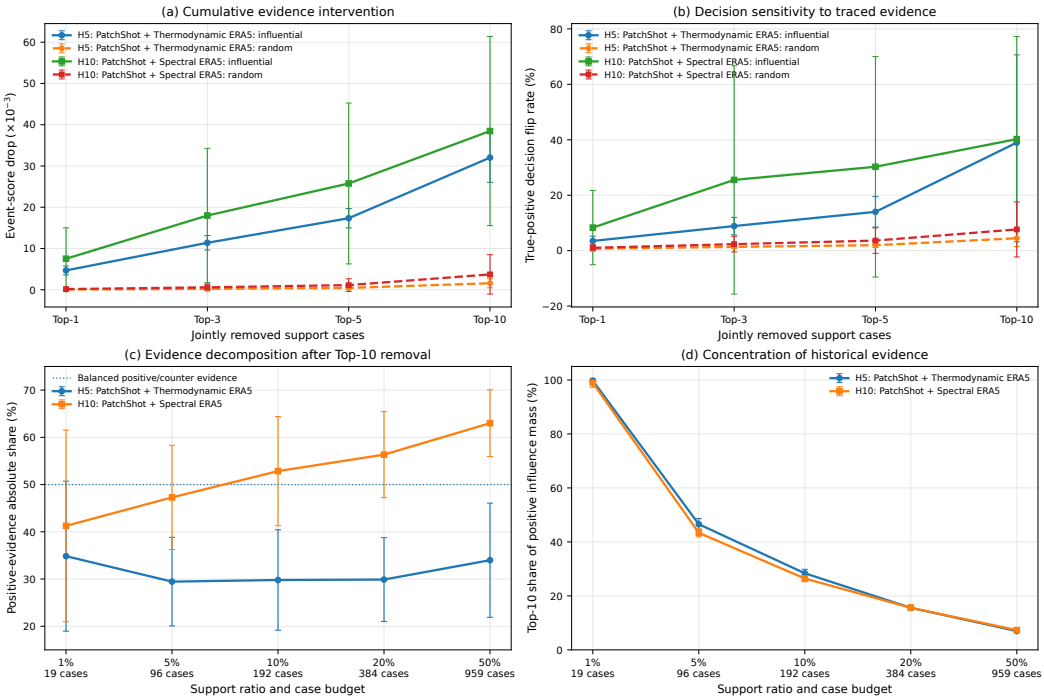


Figure 6: **Cumulative intervention on traced support evidence.** The panels report score drops, decision-flip rates, evidence decomposition, and concentration of historical influence under removal of influential or randomly selected support cases.

The larger H10 variation also reveals a practical limitation. As positive support evidence becomes rarer, prototype estimation and threshold calibration become more sensitive to the sampled support cases. Reporting paired splits, AP, and AUROC alongside CSI is therefore particularly important at H10.

A.4 Environment-Routing Controls

PatchShot-Env conditions how radar evidence is read under different large-scale atmospheric backgrounds. ERA5 is not treated as an independent precipitation predictor. The routing ablation separates three possible explanations for an environmental gain: expert capacity, case-level physical alignment, and environmental coordinate choice.

Table 4 separates expert capacity, case-level physical alignment, and environmental coordinate choice. The shared radar–ERA5 model introduces environmental information without case-dependent expert selection. Uniform Routing adds the same expert bank as the routed models but assigns equal weights to all experts, controlling for parameter count and additional nonlinear capacity. Shuffled Routing preserves the route architecture and the marginal distribution of routing weights while breaking the case-level correspondence between the ERA5 environment and the radar observation.

The difference between matched and shuffled routing therefore isolates the value of physical alignment more directly than the difference between routed and non-routed models. The results indicate that the gain depends primarily on case-aligned environmental context rather than on the presence of extra experts alone.

The remaining variants examine the coordinate system used to describe the environment. Raw routing retains pressure-level variables in their original coordinates, full-spectrum routing uses the broader spectral descriptor, and thermodynamic or dynamic variants restrict the route to selected variable groups. Low-mode routing tests whether broad environmental organization is sufficient without retaining the full high-frequency spectrum.

Table 3: **PatchShot comparison for extreme rain ($\geq$ 10 $\mathbf{mm\,h^{-1}}$).** All methods use the same frozen 3DMAE patch representation, identical case-disjoint support/query splits, support-only model selection and threshold calibration, and five random seeds. CSI, F1, AP, and AUROC are reported in percent; Sep. is the standardized positive–negative separation of the final query score. Best and second-best completed results are shown in **bold** and underlined, respectively.

Method	Adaptation	CSI ↑	F1 ↑	AP ↑	AUROC ↑	Sep. ↑
Direct prediction						
QPE	Direct prediction	0.00±0.00	0.00±0.00	3.97±0.14	36.02±0.18	0.4960±0.0036
Standard support readouts						
ProtoNet	Dual prototypes	12.22±0.31	21.78±0.50	15.08±0.22	77.41±0.51	1.0521±0.0254
SimpleShot	Centering + normalization	13.49±0.22	23.78±0.34	15.21±0.33	77.88±0.47	1.0920±0.0222
Balanced Linear	Class-balanced linear classifier	16.97±0.36	29.01±0.52	19.46±0.82	**84.52±0.44**	1.3923±0.0296
Balanced CE	Balanced CE projection	15.20±1.34	26.37±2.03	18.79±1.21	81.34±0.82	1.1645±0.0470
Shrinkage QDA	Shrinkage quadratic classifier	12.46±0.35	22.16±0.54	15.20±0.28	77.85±0.51	1.0400±0.0222
Label-free representation adaptation						
Identity	Frozen representation	11.91±0.29	21.28±0.46	14.98±0.29	77.23±0.51	1.0393±0.0262
PCA	Linear compression	11.23±0.60	20.18±0.97	11.35±0.81	74.42±0.68	0.9625±0.0195
Autoencoder	Reconstruction projection	12.78±0.27	22.66±0.43	14.14±0.67	77.65±0.60	1.1083±0.0310
Adversarial	Adversarial projection	13.84±0.41	24.31±0.64	14.87±0.30	78.13±0.34	1.1236±0.0152
Denoising	Denoising projection	10.60±0.55	19.16±0.90	12.12±0.49	72.89±1.37	0.8670±0.0581
Spectral	Patch-grid spectrum	13.60±0.39	23.94±0.60	14.76±0.46	77.85±0.59	1.1080±0.0329
Instance Contrastive	Self-supervised contrastive	10.51±0.90	19.01±1.48	12.17±0.83	70.09±1.46	0.7444±0.0591
Support-aware adaptation						
BCL†	Balanced contrastive	16.79±0.64	28.75±0.93	19.63±1.03	81.74±1.61	1.3185±0.1010
PaCo-Proto†	Parametric contrastive + prototype	**17.01±1.17**	**29.06±1.73**	**20.00±1.14**	84.26±0.66	1.4973±0.0526
PaCo-Center†	Parametric contrastive + class center	16.77±1.15	28.70±1.70	19.40±1.15	84.38±0.57	**1.5119±0.0467**
PatchShot (Ours)	Event-supervised contrastive re-access	16.65±1.58	28.51±2.37	19.54±1.55	83.97±0.74	1.4740±0.0546
Meta-trained adaptation						
ProtoDiff‡	Diffused prototypes	11.41±3.36	20.31±5.66	12.43±4.34	70.04±15.62	0.9536±0.2828
MetaDiff‡	Diffused classifier weights	15.73±0.40	27.18±0.60	18.15±0.68	82.95±0.57	1.4070±0.0299

† denotes controlled binary adaptations of BCL or PaCo; ‡ denotes controlled diffusion-based meta-adaptation. PatchShot uses event-supervised contrastive re-access, support-derived prototypes, and support-only threshold calibration. Label-free methods do not use event labels to fit the representation adapter; query labels are used only for final reporting.

Differences among matched coordinate systems are smaller than the degradation caused by shuffling the environment. We therefore avoid claiming that one physical coordinate is universally optimal. The stronger conclusion is that a temporally matched environment can refine the interpretation of radar morphology. For a unified final model, signed low-order ERA5 modes are fixed at both H5 and H10, while threshold-specific routes remain diagnostic ablations.

A.5 AGRI Cloud-State Extension

ERA5 represents a slowly varying dynamical and thermodynamic background. AGRI provides a complementary observation-side description of cloud morphology. The AGRI experiment asks whether a compact cloud-state descriptor contains case-aligned information that can further condition the interpretation of radar evidence.

Table 5 reports the matched and shuffled AGRI cloud-state controls on the common complete-observation subset. This analysis is performed on the complete-observation subset because AGRI coverage is not available for every radar case. All comparisons within this subsection use the same subset and paired splits. Their absolute values should not be compared directly with the full ERA5-aligned results. The multi-state descriptor provides a coarse representation of cloud organization, whereas the precursor variant isolates a more specific pre-event state.

Month-shuffled controls preserve the seasonal frequency of cloud states while destroying their case-level correspondence with radar and precipitation. They therefore distinguish matched cloud morphology from a purely climatological month effect. Matched AGRI conditioning is more consistently favorable relative to its shuffled counterpart than relative to the unconditioned baseline. The experiment supports the presence of case-aligned cloud-state information, but it does not establish a

Table 4: **Environment-routing ablation of PatchShot.** PatchShot denotes the shared support-supervised contrastive re-access backbone. PatchShot-Env is the unified final configuration, using signed low-order ERA5 modes for every event threshold. Uniform Routing controls for expert capacity, whereas Shuffled Routing destroys case-level ERA5–precipitation alignment. Results are mean ± standard deviation over five paired seeds (%). Best and second-best values within each threshold block are shown in **bold** and underlined. * and † indicate one-sided exact sign-flip significance ($p \leq 0.05$) against PatchShot and Uniform Routing, respectively.

Method	Conditioning signal	CSI ↑	F1 ↑	AP ↑	AUROC ↑
		Heavy rain: ≥ 5 **mm h**$^{-1}$			
PatchShot	Radar–ERA5 early fusion	30.27±0.33	46.47±0.39	41.73±0.85	89.44±0.35
Uniform Routing	Equal expert weights	30.39±0.71	46.61±0.83	42.29±1.14	89.48±0.32
Shuffled Routing	Full spectrum, case-shuffled	30.01±0.54	46.16±0.64	39.86±0.73	88.86±0.36
Raw ERA5	Raw pressure-level variables	30.76±0.84	47.04±0.99	**42.36±1.41**	89.61±0.48
Full Spectrum	q, t, u, v, z spectrum	30.90±0.64	47.21±0.75	41.84±0.86	89.60±0.44
PatchShot-Env (Ours)	Signed background + broad vertical modes	30.95±0.63*†	47.27±0.74	42.27±0.82	89.58±0.36
Thermodynamic	q, t spectrum	**31.16±0.57***	**47.52±0.66**	42.10±0.50	**89.65±0.34**
Dynamic	u, v, z spectrum	30.89±0.35*	47.20±0.41	42.20±1.07	89.62±0.32
Normalized Low Modes	Low modes, position-normalized	30.94±0.77	47.25±0.90	41.96±0.71	89.56±0.30
		Extreme rain: ≥ 10 **mm h**$^{-1}$			
PatchShot	Radar–ERA5 early fusion	24.30±0.85	39.10±1.11	31.42±1.61	89.66±0.21
Uniform Routing	Equal expert weights	24.63±0.83*	39.52±1.07	31.82±1.38	89.58±0.44
Shuffled Routing	Full spectrum, case-shuffled	23.58±0.37	38.16±0.48	29.48±1.23	87.95±0.58
Raw ERA5	Raw pressure-level variables	25.16±0.63	40.20±0.81	32.77±0.46	89.59±0.29
Full Spectrum	q, t, u, v, z spectrum	25.31±0.83*†	40.39±1.06	32.47±1.28*†	89.50±0.41
PatchShot-Env (Ours)	Signed background + broad vertical modes	**25.59±0.87***†	**40.74±1.11**	33.06±1.47	89.56±0.75
Thermodynamic	q, t spectrum	25.13±0.72*†	40.16±0.92	32.70±0.96*†	89.73±0.37
Dynamic	u, v, z spectrum	25.34±0.69*†	40.43±0.89	32.82±1.32	**89.75±0.44**
Normalized Low Modes	Low modes, position-normalized	25.56±0.79*	40.71±1.01	**33.30±1.02**	89.68±0.39

Precision/recall, component-level spectral ablations, controlled gains, and routing diagnostics are reported in the appendix. PatchShot-Env is fixed to the signed low-mode route for all thresholds; no threshold-specific method selection is used.

uniform absolute gain for every metric and threshold. AGRI is consequently treated as a promising observation-side extension rather than as part of the unified final PatchShot-Env model.

A.6 Representation Accessibility

Predictive comparisons evaluate a particular deployed readout. Accessibility analysis instead asks how much query discrimination is recoverable from a representation by a bounded family of lightweight support-trained readouts, and how much of that recoverable discrimination is exposed by the prototype rule. For every paired split, Prototype, Linear, and k-NN probes receive the same representation and support/query cases. The empirical readout envelope is

$$\mathcal{E}_B(Z) = \max_{R \in \{R_{\text{proto}}, R_{\text{linear}}, R_{k\text{-NN}}\}} J_Q(R; Z).$$

The prototype accessibility fraction and hidden component are

$$\alpha_B = \frac{J_Q(R_{\text{proto}}; Z)}{\mathcal{E}_B(Z)}, \qquad \mathcal{H}_B = \mathcal{E}_B(Z) - J_Q(R_{\text{proto}}; Z).$$

The maximum is taken within each split before aggregation. It is not used to choose the final prediction method. Thus, $\mathcal{E}_B$ is a controlled diagnostic frontier over three low-complexity probes rather than an information-theoretic upper bound. The resulting accessibility quantities are reported in Table 6.

The frozen 3DMAE representation shows a substantial difference between its direct prototype CSI and its probe envelope. Rare-event discrimination is therefore recoverable from the frozen space but is not readily exposed by the original prototype geometry.

Table 5: **AGRI cloud-state extension of PatchShot.** The complete-observation subset contains 923 training and 603 validation cases. PatchShot is evaluated with and without matched AGRI cloud-state conditioning. Month-shuffled controls preserve seasonal cloud-state frequencies while destroying case-level AGRI–precipitation alignment. Results are mean ± standard deviation over five paired seeds (%). Best and second-best values within each threshold block are shown in **bold** and underlined. ‡ indicates a five-of-five-seed AUROC improvement over the corresponding shuffled control ($p = 0.03125$).

Method	AGRI conditioning	CSI ↑	F1 ↑	AP ↑	AUROC ↑
		Heavy rain: ≥ 5 **mm h**$^{-1}$			
PatchShot	None	28.43±0.96	44.27±1.16	**38.92±2.16**	**86.39±0.86**
PatchShot + Cloud States	Six-state one-hot, matched	**28.94±1.23**	**44.87±1.47**	38.11±1.13	85.69±1.30‡
PatchShot + Shuffled States	Six-state one-hot, month-shuffled	27.55±0.97	43.19±1.18	36.26±1.02	84.49±0.88
PatchShot + Precursor	State-5 indicator, matched	28.83±0.74	44.75±0.90	38.72±1.90	85.34±0.70
PatchShot + Shuffled Precursor	State-5 indicator, month-shuffled	28.76±1.28	44.66±1.53	37.47±1.56	86.16±1.17
		Extreme rain: ≥ 10 **mm h**$^{-1}$			
PatchShot	None	**27.52±1.22**	**43.15±1.50**	**34.72±1.14**	**86.03±0.65**
PatchShot + Cloud States	Six-state one-hot, matched	26.61±2.20	41.99±2.79	33.40±2.23	85.69±0.78‡
PatchShot + Shuffled States	Six-state one-hot, month-shuffled	26.02±1.60	41.27±2.05	32.60±1.54	84.19±0.94
PatchShot + Precursor	State-5 indicator, matched	27.13±1.65	42.65±2.06	33.79±1.43	85.52±1.78
PatchShot + Shuffled Precursor	State-5 indicator, month-shuffled	26.13±1.12	41.43±1.41	33.08±1.59	85.38±1.63

Absolute AGRI-subset values should not be compared directly with Table 4. These experiments extend the PatchShot backbone on the complete-observation subset; they do not yet evaluate AGRI on top of PatchShot-Env.

After support-supervised and environment-conditioned re-access, the prototype readout approaches the envelope much more closely. The improvement has two related forms: the representation supports a stronger lightweight readout, and a larger fraction of this recoverable discrimination becomes available to the deployed prototype rule.

A.7 Probe Stability

Envelope closure and probe stability answer different questions. Envelope closure asks whether the prototype readout approaches the strongest lightweight probe. Probe stability asks whether the conclusion about a representation changes substantially when the decoder is replaced.

Table 7 reports the individual probe values. We summarize stability using the three probe means and the normalized probe range,

$$\Delta_{\text{probe}} = \frac{\max_R J_Q(R;Z) - \min_R J_Q(R;Z)}{\max_R J_Q(R;Z)}.$$

A smaller value indicates less dependence on one decoder family. The profile column identifies which probe attains the largest mean CSI; "Directly accessible" denotes representations for which the prototype is close to the best probe despite small numerical differences.

The frozen representation is strongly decoder-dependent: its direct prototype readout trails the best lightweight probe, especially at H5. After re-access, Prototype and Linear readouts become much closer, while k-NN remains weaker for several environment-conditioned representations. Thus, PatchShot does not make every decoder identical, but it substantially reduces the specific failure of the direct prototype boundary that motivates the accessibility formulation.

The probe-stability result should not be read as a second leaderboard. Its role is diagnostic. The accessibility table measures how much of the empirical envelope is exposed by the deployed prototype rule, whereas Table 7 shows how that conclusion changes across global, prototype-based, and local neighborhood decoders.

Table 6: **Representation accessibility across lightweight probes.** Results are mean±standard deviation over five paired splits using 20% support. The empirical envelope $\mathcal{E}_B$ is the per-split maximum over Prototype, Linear, and k-NN. The accessibility fraction is $\alpha_B = J_Q(R_{\text{proto}}; Z)/\mathcal{E}_B(Z)$, and the hidden component is $\mathcal{H}_B = \mathcal{E}_B(Z) - J_Q(R_{\text{proto}}; Z)$.

Representation	Prototype CSI ↑	Linear CSI	k-NN CSI	Envelope $\mathcal{E}_B$ ↑	Accessibility α_B ↑	Hidden $\mathcal{H}_B$ ↓
Heavy-rain events (≥ 5 mm h^{-1})						
Frozen 3DMAE	0.186 ± 0.006	0.233 ± 0.003	0.234 ± 0.008	0.236 ± 0.004	0.788 ± 0.022	0.050 ± 0.005
PatchShot + ERA5	0.302 ± 0.002	0.309 ± 0.005	0.275 ± 0.008	0.309 ± 0.005	0.977 ± 0.009	0.007 ± 0.003
PatchShot + Spectral ERA5	0.304 ± 0.012	0.316 ± 0.008	0.283 ± 0.008	0.316 ± 0.008	0.964 ± 0.018	0.011 ± 0.005
PatchShot + Thermodynamic ERA5	**0.312 ± 0.005**	**0.316 ± 0.006**	0.281 ± 0.006	**0.316 ± 0.005**	**0.987 ± 0.007**	**0.004 ± 0.002**
Extreme-rain events (≥ 10 mm h^{-1})						
Frozen 3DMAE	0.150 ± 0.007	0.182 ± 0.004	0.175 ± 0.010	0.183 ± 0.004	0.823 ± 0.034	0.032 ± 0.006
PatchShot + ERA5	0.243 ± 0.009	0.244 ± 0.012	0.219 ± 0.007	0.246 ± 0.010	0.988 ± 0.013	0.003 ± 0.003
PatchShot + Spectral ERA5	0.254 ± 0.008	**0.258 ± 0.004**	**0.222 ± 0.008**	**0.258 ± 0.004**	0.985 ± 0.015	0.004 ± 0.004
PatchShot + Signed Low-Mode ERA5	**0.255 ± 0.009**	0.255 ± 0.006	0.217 ± 0.008	0.258 ± 0.005	**0.988 ± 0.020**	**0.003 ± 0.005**

Table 7: **Probe stability of the representations used in the accessibility analysis.** CSI values are five-seed means at 20% support. The normalized range Δ_{probe} measures sensitivity to replacing the prototype decoder with a linear or local k-NN probe.

Threshold	Representation	Prototype	Linear	k-NN	Δ_{probe}	Probe profile
H5	Frozen 3DMAE	0.186	0.233	0.234	0.205	k-NN-dependent
H5	PatchShot + ERA5	0.302	0.309	0.275	0.110	Directly accessible
H5	PatchShot + Spectral ERA5	0.304	0.316	0.283	0.104	Linear-dependent
H5	PatchShot + Thermodynamic ERA5	0.312	0.316	0.281	0.112	Directly accessible
H10	Frozen 3DMAE	0.150	0.182	0.175	0.175	Linear-dependent
H10	PatchShot + ERA5	0.243	0.244	0.219	0.102	Directly accessible
H10	PatchShot + Spectral ERA5	0.254	0.258	0.222	0.137	Directly accessible
H10	PatchShot + Signed Low-Mode ERA5	0.255	0.255	0.217	0.149	Directly accessible

A.8 Unified Evaluation and Leakage Controls

Table 8 is an audit trail rather than an additional performance comparison. It specifies which information is available at each stage, which quantities are paired across methods, and which operations use support labels.

The case is always the statistical split unit. Patch labels become available only after an entire case enters the support set. Different patches from the same radar case therefore cannot appear in support and query. Within each seed, compared methods receive identical query cases and nested support subsets, removing split variation from paired comparisons.

Two calibration procedures must be distinguished. The deployed radar-only PatchShot readout selects its operating point from available support scores only. The probe-envelope diagnostic uses case-grouped out-of-fold support predictions to place Prototype, Linear, and k-NN probes on a common controlled footing. Probe OOF calibration is not used to select the final method and should not be conflated with the deployed PatchShot operating point.

Table 8: **Unified evaluation protocol and leakage controls.** All compared readouts use paired query sets and nested support subsets. No query label is used for representation fitting, prototype construction, deployed threshold calibration, model selection, or case attribution.

Component	Setting	Control
Evaluation pools	1,922 radar-only cases; 1,918 ERA5-aligned cases	Comparisons are paired only within the same observation-complete protocol block.
Patch definition	48 patches per case on a 6×8 grid	The same row-major patch layout is used by all methods and event thresholds.
Event labels	H5: ≥ 5 mm h^{-1}; H10: ≥ 10 mm h^{-1}	Labels are computed from the aligned patch-level precipitation target.
Query protocol	A fixed 50% query subset for each paired split	The same query cases are shared by all methods within a seed.
Support protocol	Nested 1%, 5%, 10%, 20%, and 50% subsets	Larger support subsets strictly contain the smaller subsets within each seed.
Randomness	Five paired seeds	Method comparisons use identical support and query cases for every seed.
Prototype readout	Positive and negative support prototypes	Only support labels are used to construct the local prototypes.
Main-readout calibration	Support-only score calibration	The deployed PatchShot threshold is determined without query-label access.
Probe calibration	Case-grouped out-of-fold support predictions	Used only to compare Prototype, Linear, and k-NN under a matched diagnostic protocol.
Probe envelope	Per-split maximum over the three probes	The maximum is computed before aggregation and is not used for final method selection.
Evidence intervention	Leave-support-out prototype recomputation	Selected support cases are removed and prototypes are recomputed without retraining the encoder or adapter.
Case selection	Boundary flips identified from complete intervention arrays	Representative cases are not selected from query performance or visual appearance alone.
Raw-data lookup	Exact timestamp search in annual radar H5 stores	The aligned feature index is not treated as a direct annual HDF5 row index.
Radar visualization	Vertical maximum over 11 stored radar levels	Negative values are treated as missing; displayed values are raw quantized units, not calibrated dBZ.

Evidence interventions are performed after adapter training. Selected support cases are removed and prototypes are recomputed exactly, while the encoder and adapter remain unchanged. Raw-data

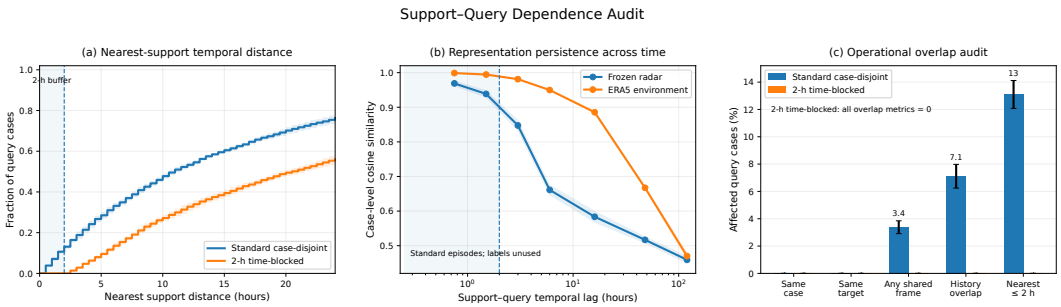


Figure 7: **Support–query dependence audit. (a)** Empirical cumulative distributions of the nearest temporal support distance under the standard case-disjoint protocol and the conservative 2-h time-blocked protocol. **(b)** Label-free cosine similarity of frozen radar and ERA5 representations as a function of temporal lag. Radar similarity decays more rapidly, whereas ERA5 retains longer-lived environmental continuity. **(c)** Operational overlap rates over case identity, target time, radar input frames, historical windows, and the 2-h neighborhood. The audit distinguishes prohibited sample leakage from admissible dependence between evolving atmospheric states.

visualization uses exact timestamp lookup in annual radar stores; aligned feature indices are not assumed to equal annual HDF5 row indices.

A.9 Support–Query Dependence Audit

Case disjointness prevents identical labeled cases from appearing in both support and query, but it does not imply marginal independence of the underlying atmosphere. Nearby observations may share an evolving precipitation process, and large-scale ERA5 environments may persist over substantially longer periods. We therefore separate prohibited information leakage from physically meaningful atmospheric dependence.

The standard protocol enforces zero shared case identity and zero shared target time. The additional 2-h time-blocked protocol removes support cases that are immediately adjacent to a query, including shared input frames and overlapping historical windows. The buffer is applied during episode construction and does not use precipitation labels. Figure 7 summarizes the temporal-distance, representation-persistence, and operational-overlap diagnostics.

In Figure 7, Panel (a) reports the degree of temporal proximity rather than reducing the audit to a binary overlap indicator. The standard episodes retain a limited short-range tail, whereas the buffered protocol removes this immediate proximity. Panel (b) shows that representation dependence persists beyond the buffer, particularly for ERA5. This is expected: local radar morphology evolves relatively quickly, while large-scale dynamical and thermodynamic environments remain coherent over longer time scales. Panel (c) converts these observations into an operational leakage audit. Both protocols remove exact case and target-time overlap. The 2-h protocol additionally removes shared radar frames, overlapping input histories, and immediately adjacent support cases. The time-blocked construction is deliberately conservative, but it is not treated as proof of storm independence. A fixed temporal buffer cannot determine whether two observations belong to a persistent, regenerated, split, or merged storm system. The audit therefore tests robustness to immediate process proximity, while the primary protocol retains physically meaningful atmospheric dependence.

A.10 Scope and Limitations

The experiments establish accessibility of threshold-conditioned rare-event evidence in temporally pooled radar representations. They do not establish accessibility of ordered storm dynamics. The current tokens average information over the historical window and therefore retain multi-frame context while discarding explicit temporal order.

Case-disjoint and time-buffered episodes prevent direct sample overlap and immediate historical overlap, but they do not define meteorological storm identity. Persistence, regeneration, splitting,

and merging cannot be resolved by a fixed time buffer. Extending support construction from independent cases to lineage-aware storm episodes is a natural next step.

PatchShot also uses threshold-specific binary event labels. H5 and H10 test accessibility at two operational boundaries, but they do not yet form a unified continuous precipitation estimator. A shared adapter with threshold-conditioned prototypes, an ordered multi-threshold readout, or a continuous severity coordinate remains an open extension.

Finally, the ERA5 and AGRI experiments use different complete-observation subsets. They demonstrate environmental and cloud-state conditioning separately but do not yet establish the gain of a joint radar–ERA5–AGRI readout under one common case pool.

A.11 AI USE STATEMENT

Generative AI tools were used for language editing, code development and debugging, and the organization of experimental materials. All scientific hypotheses, experimental designs, analyses, numerical results, and conclusions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.