READIT: RAPID AND EFFICIENT ATTENTION-BASED DISTILLATION FOR IRREGULARLY SAMPLED LIGHT CURVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-generation astronomical surveys like the Vera C. Rubin Observatory (LSST) will accumulate petabyte-scale data over the next decade, producing tens of terabytes each night, yet extracting representations via self-supervised learning currently demands prohibitive pre-training compute. To address this bottleneck, we introduce `READit`, an attention-based foundation model optimized via student–teacher self-distillation for asynchronous, multi-band light curves. Pre-trained on $\sim$ 2.1 million Zwicky Transient Facility (ZTF) light curves, `READit` reaches peak representational convergence in just 11 epochs ($\approx$ 15 GPU-hours)—a $\sim 95\%$ compute reduction ($18\times$ speedup) over prior multi-band contrastive frameworks. This rapid optimization is achieved by coupling multi-crop self-distillation with physics-informed temporal augmentations that stabilize the learning trajectory. On downstream benchmarks, `READit` consistently outperforms existing astronomical foundation models, general time-series architectures, and classical statistical baselines on 12-class variability classification. Moreover, `READit` exhibits remarkable label efficiency in extreme low-data regimes, clusters unseen out-of-distribution classes, and demonstrates robust zero-shot and few-shot cross-survey transfer to simulated LSST observations (PLAsTiCC). Consequently, `READit` establishes a computationally sustainable, "Green AI" foundation model for survey-scale time-domain discovery.

1 INTRODUCTION

The current era of time-domain astronomy is undergoing a paradigm shift driven by high-cadence, wide-field surveys. Next-generation facilities, most notably the Vera C. Rubin Observatory (conducting the Legacy Survey of Space and Time; LSST; Ivezić et al., 2019), will accumulate petabyte-scale volumes of data over the coming decade, generating tens of terabytes per night. Extracting meaningful scientific discoveries from this unprecedented data deluge is fundamentally bottlenecked by the scarcity of human-annotated labels and the prohibitive cost of spectroscopic follow-up. To fully leverage these vast observational archives, the astronomical community has increasingly adopted self-supervised learning (SSL) to extract robust, generalizable features directly from raw, unannotated light curves.

Recent foundational efforts have successfully adapted the Transformer architecture (Vaswani et al., 2017) for astronomical time-series. Frameworks like ASTROMER (Donoso-Oliva et al., 2023; 2026) demonstrated the efficacy of attention mechanisms on single-band light curves, while Astra-CLR (Majumder et al., 2026) pioneered the use of masked attention to generate multi-scale temporal "views" adapted from the multi-crop paradigm. Concurrently, advancements in computer vision—specifically self-distillation frameworks like DINO (Caron et al., 2021)—have demonstrated that pairing this multi-crop strategy with a teacher-student objective yields highly robust representations while bypassing the computational bottlenecks associated with negative pairs. Building upon the intersection of these milestones, we seek to address the computational overhead of large-scale pre-training. Scaling multi-band architectures to survey-level datasets currently demands exhaustive pre-training cycles, highlighting an urgent need for compute-efficient, domain-optimized solutions.

To address this, we introduce `READit` (**R**apid and **E**fficient **A**ttention-based **D**istillation for **I**rregularly Sampled Ligh**t** Curves). Pre-trained on a dataset of $\sim$ 2.1 million unlabeled multi-

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.

band light curves from the Zwicky Transient Facility (ZTF; Bellm et al., 2019) Zubercal Data Release 16, READit seamlessly integrates multi-filter temporal views with a self-distillation objective. Crucially, we curate a suite of physics-informed data augmentations specifically tailored to the irregular nature of photometric time-series. This domain-aware strategy stabilizes the optimization landscape, enabling READit to achieve peak representational convergence in just ~ 11 epochs—a substantial reduction in pre-training compute compared to prior contrastive frameworks. Furthermore, READit demonstrates strong label efficiency and robust out-of-distribution generalization on downstream classification tasks, establishing a compute-efficient foundation model for astronomical discovery.

2 Related Work

Recent advances in astronomical foundation models increasingly leverage multimodal data to capture rich astrophysical context across heterogeneous observations. Architectures such as AstroM^3 (Rizhko & Bloom, 2025), MAVEN (Wang et al., 2020), and AstroCLIP (Parker et al., 2023) employ cross-modal contrastive objectives to optimize a joint embedding space between photometric light curves and spectroscopic data. In the supervised domain, frameworks like ATAT (Cabrera-Vives et al., 2024) incorporate auxiliary tabular metadata (e.g., spectroscopic redshift and Milky Way extinction) directly alongside photometric sequences to guide representation learning. While these methods establish strong baselines, they inherently rely on observational regimes where rich multimodal or contextual metadata is readily accessible.

In regimes focusing strictly on raw photometric time-series, several powerful architectures have been developed. Operating on single-filter sequences, models such as FALCO (Zuo et al., 2026), ASTROMER (Donoso-Oliva et al., 2023; 2026), ASTRAFier (Gregory et al., 2026), and StarCLR (Ding et al., 2026) extract localized temporal representations directly from isolated passbands. To capture multi-band dynamics, ATCAT (Tung, 2025) implements multi-filter pre-training, while t^2 (Allam & McEwen, 2024) utilizes Gaussian Process interpolation and 1D convolutions to project multi-filter light curves onto a uniform input representation. Exploring multi-view paradigms, AstraCLR (Majumder et al., 2026) leverages multi-scale temporal crops within a contrastive framework to capture scale-invariant features. While Rui (2026) explore multi-view learning via a Joint-Embedding Predictive Architecture (JEPA), their approach relies on latent predictive modeling. In contrast, non-predictive self-distillation across asynchronous multi-band observations remains largely unexplored.

In parallel, the broader machine learning community has advanced domain-agnostic foundation models for sequential data. Architectures including Chronos-2 (Ansari et al., 2025) and MOIRAI (Woo et al., 2024) demonstrate strong zero-shot generalization across diverse time-series disciplines. Evaluating these paradigms within time-domain astronomy, the StarEmbed benchmark (Li et al., 2025) systematically evaluates generic sequence models against domain-specific architectures, extending them to multi-filter settings via post-hoc embedding concatenation.

Building upon these diverse advancements, READit is explicitly architected for purely self-supervised representation learning directly from raw, asynchronous multi-band photometry. By pairing an asymmetric student–teacher distillation objective with domain-aware temporal augmentations, READit bypasses the need for negative pairs, predictive latent heads, or regularized interpolation, establishing a highly sample-efficient and computationally sustainable foundation model tailored for survey-scale time-domain astronomy.

3 Method

Light Curve Parameters. A raw light curve $\mathcal{L}$ is defined as a sequence of L discrete photometric observations sampled at irregular intervals across all available passbands: $\mathcal{L} = \{(m_i, \sigma_i, t_i, \log \lambda_i)\}_{i=1}^{L}$. Here, m_i denotes the apparent magnitude, σ_i is the photometric uncertainty, t_i is the observational epoch in Modified Julian Date (MJD), and $\log \lambda_i$ is the log-effective wavelength of the ZTF filter ($\lambda_g = 4746.48$Å, $\lambda_r = 6366.38$Å, $\lambda_i = 7829.03$Å). Each sequence undergoes a domain-specific preprocessing and augmentation pipeline to produce multi-scale views: Global Views (GVs) and Local Views (LVs).

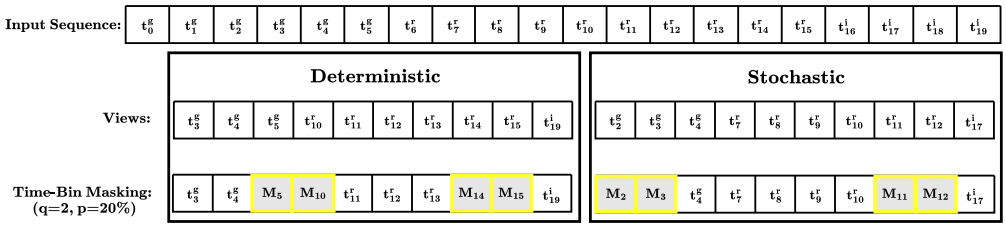

Figure 1: **Schematic of hierarchical multi-band windowing and time-bin masking.** To construct a fixed-length view ($l = 10$), continuous segments are extracted from each filter while preserving the survey's native cadence hierarchy ($D_r = 6, D_g = 3, D_i = 1$). The deterministic strategy extracts fixed anchors (e.g., the 'End' window), whereas the stochastic strategy samples random temporal windows (e.g., $[t_2^g \dots t_4^g]$, $[t_7^r \dots t_{12}^r]$, and $[t_{17}^i]$). Segments are concatenated in $[g, r, i]$ order. Additionally, dynamic time-bin masking ($q = 2, p = 20\%$) stochastically drops observations (e.g., masking positions M_5 and M_{10}) to simulate macroscopic instrumental or weather-induced observation gaps.

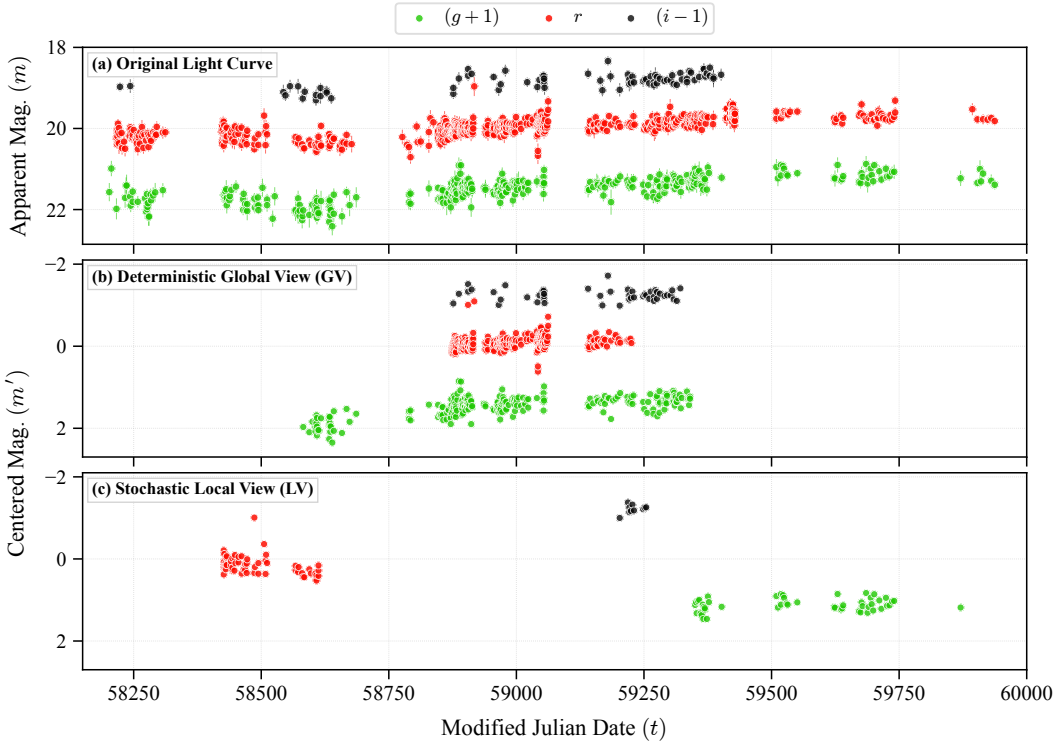

Figure 2: **Generation of multi-crop views via domain-specific data augmentations. (a)** The raw, unaugmented ZTF light curve of an Active Galactic Nucleus (AGN; Gaia DR3 ID 3445151224138710144, $\alpha = 87.652289°$, $\delta = +31.915072°$), serving as the baseline prior to transformation. **(b)** A Deterministic Global View (GV) of length $l = 600$ ($D_g = 250, D_r = 300, D_i = 50$). Extracted via deterministic windowing to capture long-term temporal trends, this sequence is augmented with time-bin masking ($q = 5, p = 5\%$) and photometric outlier injection. **(c)** A Stochastic Local View (LV) of length $l = 120$ ($D_g = 50, D_r = 60, D_i = 10$). Sampled via stochastic temporal windowing to isolate short-term variations, this sequence is augmented with time-bin masking ($q = 5, p = 1\%$) alongside synthetic photometric outliers.

3.1 Data Preprocessing and Augmentation

Following Majumder et al. (2026), we preprocess ZTF light curves while preserving the native survey cadence hierarchy (r, g, i). We restrict the pre-training dataset to sources with ≥ 200 detections in both the g and r bands across 12 broad astronomical variability categories ($\geq 1,000$ sources per category), discarding all class labels during SSL pre-training. Detections are partitioned by filter $(g \to r \to i)$ and sorted chronologically within each band. To stabilize numerical throughput, MJD epochs are shifted by $-58,000$ and cast to 32-bit floating-point tensors. To ensure scale-invariant representation learning over relative magnitude variations, we apply variance-weighted standardization prior to view generation. For a sequence of length L, the weighted mean $\bar{m}_w$ and centered magnitudes m'_j are:

$$\bar{m}_w = \frac{\sum_{j=1}^{L} w_j m_j}{\sum_{j=1}^{L} w_j}, \quad m'_j = m_j - \bar{m}_w, \quad \text{where } w_j = \sigma_j^{-2}. \tag{1}$$

To learn scale-invariant features, we adapt the multi-crop paradigm (Caron et al., 2020; 2021; Majumder et al., 2026) to extract multi-scale Global Views (GVs) and Local Views (LVs). To simulate observational artifacts, we inject Gaussian magnitude noise ($\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, with $\sigma_n \in [0.0, 0.05]$) and synthetic photometric outliers based on the ZTF saturation thresholds. Fixed-length temporal views ($l = D_g + D_r + D_i$) are constructed by extracting continuous temporal windows per band and concatenating them in $[g, r, i]$ order, dynamically right-padding any band containing fewer than the required detections. As illustrated in Figure 1, segments are sampled either deterministically (anchored at the temporal start, middle, or end) or stochastically (randomized window origins), strictly maintaining the observational cadence hierarchy ($D_r > D_g > D_i$). Finally, dynamic time-bin masking (Russeil et al., 2024; Majumder et al., 2026) drops consecutive photometric bins to simulate macroscopic survey cadence gaps. The complete transformation pipeline is depicted in Figure 2.

3.2 Input Encoding and Self-Distillation Framework

Continuous Input Embedding. To process an augmented view, each multi-band light curve is mapped into a sequence of composite latent representations $\mathbf{X} \in \mathbb{R}^{l \times d_{\text{model}}}$. Centered magnitudes m' are linearly projected, while observational epochs t (MJD) are mapped via a continuous-time sinusoidal positional encoding to natively accommodate non-uniform survey cadences. Rather than assigning discrete categorical filter embeddings or simple linear mappings, we project the continuous log-effective filter wavelength ($\log \lambda$) through a non-linear two-layer MLP with ReLU activation, capturing complex chromatic and wavelength relationships. As verified empirically (Appendix D), these continuous temporal and non-linear wavelength inductive biases consistently outperform discrete tokenizations and linear projections. The composite input sequence $\mathbf{X}$ is formed by the element-wise sum of these three embeddings.

Encoder Backbone and Masked Pooling. The composite sequence $\mathbf{X}$ is processed by a Transformer encoder (Vaswani et al., 2017; Devlin et al., 2019; Majumder et al., 2026). To prevent right-padded tokens and bin-masked gaps from contaminating attention weights, we apply an additive attention mask $\mathbf{M}_{\text{attn}} \in \{0, -\infty\}^{l \times l}$ to the pre-softmax attention logits. The final contextualized token representations $\mathbf{Y} \in \mathbb{R}^{l \times d_{\text{model}}}$ are pooled into a single global vector $\mathbf{z} \in \mathbb{R}^{d_{\text{model}}}$ via a 1D Masked Global Average Pooling (GAP) layer:

$$\mathbf{z} = \left(\sum_{i=1}^{l} v_i \right)^{-1} \sum_{i=1}^{l} \mathbf{Y}_i v_i, \tag{2}$$

where $v_i \in \{0, 1\}$ denotes a binary validity indicator ($v_i = 0$ for padded or masked positions).

Self-Supervised Distillation Objective. To optimize the backbone without human annotations, `READit` adopts a student–teacher self-distillation objective (Caron et al., 2021). The training loss minimizes the cross-entropy between the predicted categorical distributions of multi-scale temporal views (GVs and LVs) originating from the same physical astronomical source. Full mathematical formulations of the asymmetric distillation loss, EMA parameter updates, and centering/sharpening dynamics are detailed in Appendix B.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Data and Benchmarks. We use multi-band (g, r, i) light curves from the Zwicky Transient Facility (ZTF; Bellm et al. 2019) Zubercal Data Release 16[1], cross-matched against the Gaia DR3 variability catalog (Gaia Collaboration et al., 2023; Rimoldini et al., 2023). Following Majumder et al. (2026) and our preprocessing protocol (Section 3.1), our primary catalog comprises 2,593,567 light curves across 12 broad variability classes (`2.6M-Full`). We partition `2.6M-Full` into an 80% training split ($2,073,458$ sources) for self-supervised pre-training and an isolated 20% validation pool from which three representative subsets are derived without data contamination: `6K-Proxy` and `26K-Proxy` for architectural ablations, and `120K-Eval` for downstream evaluations. Complete curation procedures, class distributions, and out-of-distribution dataset details are provided in Appendix A.

Architectural Search and Ablations. To identify the optimal configuration for large-scale pre-training, we conducted a systematic empirical search across data augmentations, inductive biases, multi-crop complexity, and model scale using the `6K-Proxy` dataset, as summarized in Table 1. Candidate configurations were subsequently validated on the `26K-Proxy` dataset to select the final architecture and ensure training stability across data scales. This validation revealed that `READit` achieves representational saturation rapidly, typically by the 8th epoch. Scaling to the full 2.6M light curves further refines the latent manifold while preserving these early convergence characteristics. Detailed epoch-by-epoch optimization dynamics, including the identification and rejection of high-scoring "shortcut" configurations (B1, B3), are provided in Appendix D.

Table 1: Comprehensive Ablation Study evaluated on the `6K-Proxy`-dataset. The proposed configuration (A1) provides the optimal balance of physical grounding, training stability, and parameter efficiency. For full epoch-by-epoch loss and Linear Probing Macro-F1 dynamics, see Appendix D.

Suite	ID	Configuration / Logic Tested	Peak Macro-F1
Anchor	**A1**	**Proposed (MJD, 9-view, Masking, GAP)**	$\mathbf{0.383 \pm 0.0065}$
A: Augmentations	A2	No Time-Bin Masking	0.370 ± 0.0062
	A3	No Photometric Noise/Outliers	0.370 ± 0.0051
B: Inductive Bias	B1	Integer Index PE	0.395 ± 0.0051 *(Shortcut)*
	B2	Label Encoding for Filters	0.378 ± 0.0063
	B3	Linear Layer for Filters	0.417 ± 0.0055 *(Shortcut)*
	B4	[CLS] Token (instead of GAP)	0.374 ± 0.0044
	B5	Stochastic Global Views	0.378 ± 0.0034
C: View Complexity	C1	6 Views (2 Global + 4 Local)	0.381 ± 0.0026
	C2	2 Views (2 Global + 0 Local)	0.346 ± 0.0066
D: Scale	D1	Medium Backbone ($L = 6, d = 384$)	0.378 ± 0.0061

Architecture and Multi-Crop Configuration. The `READit` encoder is a 4-layer Transformer with 4 attention heads, a feed-forward hidden dimension of 1024, and $d_{\text{model}} = 256$. This is coupled with a DINO-style projection head $h(\cdot)$ featuring a $65,536$-dimensional output (Caron et al., 2021). We utilize a 9-view multi-crop configuration: 3 deterministic GVs ($l = 600$) and 6 LVs ($l \sim \mathcal{U}(100, 200)$) sampled via a hybrid deterministic-stochastic strategy, maintaining the strict filter cadence hierarchy.

Optimization and Scheduling. Pre-training was conducted on the 80% training split of `2.6M-Full` (2.07M sources) for 250 epochs (with an early stopping patience of 20 epochs on validation loss) using 8 NVIDIA A100 GPUs with a global batch size of 800. We utilize the AdamW optimizer with a linear learning rate scaling rule (Goyal et al., 2017). To stabilize self-distillation, the learning rate linearly warms up over the first 10 epochs to a peak of 1.56×10^{-3} (scaled from a base rate of 5×10^{-4}) followed by cosine decay, while teacher temperature τ_t linearly increases from 0.04 to a constant 0.07 over the first 30 epochs. We fix student temperature $\tau_s = 0.1$, teacher EMA momentum to 0.996, and centering momentum to 0.9. This configuration drives rapid convergence, establishing `READit` as a sustainable alternative to compute-intensive SSL baselines.

[1] `http://atua.caltech.edu/ZTF/Zubercal.html`

Table 2: Downstream classification performance (Macro-F1) on the balanced `120K-Eval` benchmark. Models are evaluated using Linear Probing and Weighted k-NN on frozen representations. *For classical hand-crafted features, we report downstream performance using a Random Forest (RF) classifier (Macro-F1: 0.7024 ± 0.0015) instead of the Weighted k-NN metric.

Paradigm	Model	Linear Probe	Weighted k-NN
Classical Expert	Hand-Crafted Features* (r-band)	0.5935 ± 0.0012	–
General Time-Series	Chronos-2 (r-band)	0.5804 ± 0.0022	0.4875 ± 0.0025
Astro Single-Band	ASTROMER-2 (r-band)	0.5789 ± 0.0016	0.5139 ± 0.0019
Astro Multi-Band	Astra-CLR (g, r, i-bands)	0.7027 ± 0.0011	0.6217 ± 0.0030
Proposed	**READit** (g, r, i-bands)	$\mathbf{0.7450 \pm 0.0014}$	$\mathbf{0.6921 \pm 0.0020}$

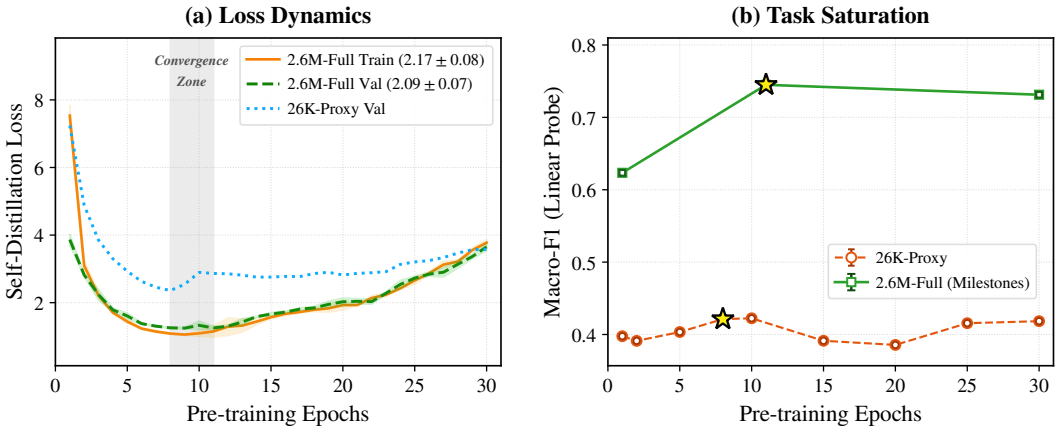

Figure 3: **Pre-training dynamics and computational efficiency of `READit`. (a) Loss dynamics:** Self-distillation loss trajectories across pre-training epochs on the `2.6M-Full` dataset (Train/Val, Mean $\pm$ Std across 3 runs) compared against the `26K-Proxy` validation loss. The shaded gray band marks the empirical convergence zone (Epochs 8–11). **(b) Task saturation:** Downstream Macro-F1 scores evaluated via linear probing on the `120K-Eval` benchmark. Performance saturates rapidly by Epoch 11, matching full pre-training milestones.

Evaluation Protocol. Representational quality is evaluated via *Linear Probing* on the frozen $\mathbf{z}$ vector and *Weighted k-NN* (Caron et al., 2021). While Linear Probing assesses the linear separability of features, weighted k-NN is utilized specifically during supervised fine-tuning to measure latent space compactness. Performance is reported as the Macro-averaged F1-score, with results computed as the mean and standard deviation across 100 bootstrap resamples.

4.2 Main Results and Baseline Comparisons

We benchmark `READit` against domain-specific (Astra-CLR, ASTROMER-2) and general-purpose (Chronos-2) foundation models, alongside classical hand-crafted statistical features (Malanchev et al., 2021). To ensure architecturally faithful comparisons, multi-band models (`READit`, Astra-CLR) process unified (g, r, i) sequences totaling $l = 600$ detections ($D_g = 250, D_r = 300, D_i = 50$). Conversely, univariate SSL baselines and statistical features are evaluated strictly on continuous r-band sequences ($D_r = 600$) to maximize signal-to-noise, intentionally avoiding uncalibrated post-hoc filter aggregation. Full extraction protocols and baseline implementation details are provided in Appendix C.

As summarized in Table 2, `READit` consistently outperforms all baselines on the `120K-Eval` benchmark. It achieves a Linear Probing Macro-F1 of $\mathbf{0.7450 \pm 0.0014}$ and a Weighted k-NN Macro-F1 of $\mathbf{0.6921 \pm 0.0020}$. The substantial performance margin in the non-parametric k-NN metric demonstrates that our self-distillation objective successfully organizes the 12 variabil-

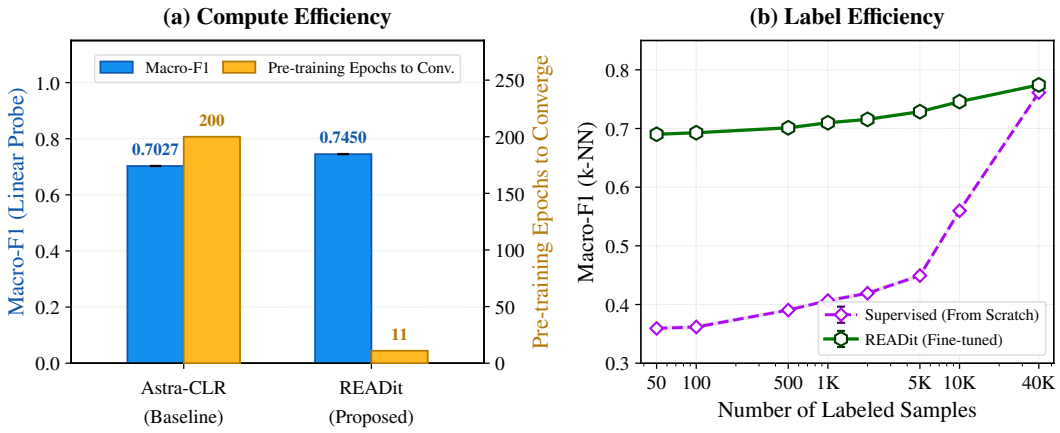

Figure 4: **Computational efficiency and low-resource label efficiency. (a) Compute efficiency:** Pre-training convergence and linear probe performance comparison between Astra-CLR and `READit`. `READit` reaches full convergence in only 11 epochs ($18\times$ speedup) while outperforming the baseline by $+4.23\%$ Macro-F1. **(b) Label efficiency:** Downstream Macro-F1 (k-NN) under limited labeled training samples on `120K-Eval`. `READit` demonstrates substantial sample efficiency, achieving ≈ 0.69 Macro-F1 with as few as 50 labels—a regime where supervised training from scratch fails (≈ 0.36).

ity classes into dense, well-separated clusters within the latent manifold, minimizing downstream classification complexity.

4.3 Rapid Convergence Analysis and Computational Efficiency

A critical bottleneck in large-scale self-supervised learning is identifying an optimal stopping criterion without expending excessive pre-training compute. Figure 3 illustrates that `READit` exhibits highly predictable and sample-efficient optimization dynamics. The model achieves task saturation—capturing the underlying astrophysical manifold of the light curves—by Epoch 11. Crucially, this convergence trajectory is closely mirrored on our smaller `26K-Proxy` validation split, providing an inexpensive proxy to track optimization. Pre-training beyond this point yields diminishing representational returns, as the student network begins to overfit to stochastic masking noise and instrumental artifacts rather than underlying astrophysical signals. Furthermore, we observe successful distillation with near-perfect student–teacher parity ($\Delta < 0.2\%$ Macro-F1) at Epoch 11.

Figure 4(a) compares this efficiency against the baseline multi-band contrastive framework, Astra-CLR. Pre-trained on the identical ZTF light-curve distribution, Astra-CLR requires 200 epochs to converge, whereas `READit`'s domain-aware inductive biases achieve superior representational quality in just 11 epochs. This $\sim 95\%$ reduction in pre-training compute establishes a computationally sustainable foundation model capable of meeting the processing demands of next-generation surveys such as LSST.

4.4 Label-Efficient Fine-Tuning

To evaluate `READit`'s robustness to label sparsity, we simulate low-data regimes across exponentially scaled subsets ($N \in [50, 40, 000]$) of labeled sources for bright objects ($m \leq 18$) drawn from the Gaia DR3 variability catalog (Gaia Collaboration et al., 2023; Rimoldini et al., 2023). To adapt task-specific decision boundaries without perturbing foundational representations, we unfreeze only the top two Transformer layers for fine-tuning on the N labeled instances; this partial fine-tuning adaptation grants the latent space sufficient flexibility to form compact clusters that fully frozen features cannot achieve. Input sequences ($l = 600$) strictly follow the multi-band Global View (GV) hierarchy.

Table 3: Zero-shot retrieval and few-shot classification on five held-out out-of-distribution (OOD) ZTF variability classes. The Random Baseline denotes an untrained backbone with identical architecture.

Protocol	Metric	Random Baseline	READit (Proposed)
Zero-Shot Retrieval	Macro-Recall@1	46.64%	**90.02%**
	Macro-Recall@5	70.64%	**93.26%**
Few-Shot (Centroid)	5-Way, 5-Shot (Macro-F1)	$26.72 \pm 4.65\%$	$\mathbf{73.17 \pm 4.67}\%$

Following adaptation, we freeze the encoder and evaluate the refined embeddings via Linear Probing and Weighted k-NN against a supervised Transformer trained from scratch. As shown in Figure 4(b), READit dramatically outperforms the supervised baseline in extreme low-data regimes ($N \leq 10^3$), where training from scratch fails to generalize. Weighted k-NN trajectories indicate that limited supervision primarily refines local latent topology rather than perturbing global representations. At $N = 40,000$ labels, READit reaches a peak Linear Probing Macro-F1 of 0.7873 ± 0.0011, consistently surpassing the fully supervised model (0.7656 ± 0.0013). We provide UMAP visualizations comparing the latent representations of the frozen pre-trained and fine-tuned READit encoders on the 2.6M-Full dataset in Appendix E.1.

5 Generalization and Transfer

5.1 Zero-Shot and Few-Shot Generalization

To evaluate the representational robustness of READit, we expose it to five completely unseen, out-of-distribution (OOD) astronomical classes. These classes are highly imbalanced, with some categories containing fewer than 15 light curves (see details in Appendix A.2).
As shown in Table 3, READit successfully generalizes to these unseen classes in both zero-shot retrieval and 5-way, 5-shot prototypical evaluation. While the randomly initialized baseline achieves a surprisingly high Macro-Recall@5 (70.64%)—a known artifact of raw input-statistic clustering in high-dimensional spaces—it degrades sharply at exact retrieval (Macro-Recall@1) and fails to form coherent decision boundaries in the few-shot regime. In contrast, READit maintains precision across all stringencies, proving its representations are geometrically dense and semantically meaningful. We provide UMAP visualizations of these OOD embeddings in Appendix E.2.

5.2 Cross-Survey Transferability

Table 4: Cross-survey transferability to simulated LSST observations (PLAsTiCC) across overlapping optical passbands (g, r, i). Frozen representations are evaluated via Zero-Shot Retrieval and 5-Shot Prototypical Classification across 100 randomly sampled episodes. Astronomical classes include Active Galactic Nuclei (AGN), Eclipsing Binaries (ECL), RR Lyrae (RR), and Type Ia/II Supernovae (SNIa, SNII).

Metric	Random Baseline	READit (Proposed)
Tier 1: Domain Adaptation (3-Way: AGN, ECL, RR)		
Zero-Shot Macro-Recall@1	54.88%	**75.31%**
Zero-Shot Macro-Recall@5	70.60%	**97.61%**
3-Way, 5-Shot (Macro-F1)	$34.89 \pm 4.91\%$	$\mathbf{60.16 \pm 4.87}\%$
Tier 2: Comprehensive Transfer (5-Way: Tier 1 + SNIa, SNII)		
Zero-Shot Macro-Recall@1	60.84%	**71.72%**
Zero-Shot Macro-Recall@5	75.92%	**93.44%**
5-Way, 5-Shot (Macro-F1)	$40.14 \pm 3.51\%$	$\mathbf{48.73 \pm 4.52}\%$

As detailed in Table 4, READit demonstrates robust transferability across both evaluation protocols. In Tier 1 (Domain Adaptation), the model effectively overcomes the ZTF-to-PLAsTiCC domain shift, outperforming the random baseline by over 25% in the 5-shot Macro-F1 metric. In Tier

2, we introduce completely unseen explosive transients (SNIa, SNII) to significantly increase task complexity. While the zero-shot topology remains highly structured (achieving 93.44% Recall@5), the 5-shot Macro-F1 highlights the inherent difficulty of photometric supernova classification. Differentiating between physically similar transient subtypes (e.g., SNIa vs. SNII) using only 5 support samples is a challenging decision boundary problem. Despite this severe data scarcity, READit outperforms the random baseline (48.73% vs. 40.14%), confirming that the self-distilled representations capture transferable morphological features even for novel, out-of-distribution observations.

6 Discussion and Limitations

Computational Sustainability and Optimization Stability. The transition to petabyte-scale time-domain surveys necessitates architectures that are both computationally sustainable and sample-efficient. READit achieves optimal representational convergence on $\sim$ 2.1M sources in just 11 epochs ($\approx$ 15 GPU-hours), reducing pre-training cycles by $\sim 95\%$ compared to prior multi-band contrastive baselines such as Astra-CLR and aligning with "Green AI" initiatives. Crucially, our proxy analyses reveal that while numerical shortcut configurations (e.g., integer positional encodings or linear filter projections; Appendix D) inflate short-term proxy scores by overfitting to majority-cadence artifacts, their validation loss diverges from downstream representational quality. By enforcing physically grounded priors (continuous MJD encoding and non-linear wavelength projections), READit ensures strict alignment between optimization convergence and peak downstream feature quality, rendering unguided self-supervised deployment viable without needing labeled early-stopping triggers.

Limitations and Future Work. A primary limitation of this work is that all downstream evaluation and fine-tuning labels are derived from automated machine-learning pipeline from Gaia DR3 variability catalog rather than human-verified or spectroscopically confirmed ground truth, potentially inheriting uncalibrated classifier biases. Additionally, while READit demonstrates robust zero-shot clustering, resolving fine-grained decision boundaries between physically similar transient subtypes (e.g., SNIa vs. SNII) under extreme few-shot constraints remains an open challenge. Finally, while transfer to simulated LSST observations (PLAsTiCC) is robust across overlapping optical bands, extending READit beyond optical filters (e.g., to X-ray or radio regimes) and deploying it on real-world heterogeneous archives (e.g., OGLE) will require survey-specific calibration to account for differing noise floors and sampling cadences, which we reserve for future investigation.

7 Conclusion

In this work, we introduced READit, a compute-efficient self-supervised foundation model tailored for irregular, multi-band photometric time-series. By coupling a compact Transformer backbone featuring continuous temporal and non-linear wavelength encodings with an asymmetric multi-crop self-distillation objective, READit successfully bridges deep sequence modeling with the physical constraints of astronomical cadence. Pre-trained on $\sim$ 2.1 million ZTF light curves, READit achieves peak representational convergence in just 11 epochs ($\approx$ 15 GPU-hours)—reducing pre-training compute by $\sim 95\%$ compared to prior multi-band baselines—while outperforming existing astronomical foundation models and classical statistical features on downstream classification benchmarks. Crucially, the learned representations demonstrate exceptional label efficiency under extreme data scarcity and transfer robustly to simulated LSST observations (PLAsTiCC) on novel, unseen transient phenomena. By demonstrating that domain-aware physical priors can dramatically accelerate optimization without sacrificing representational quality, READit establishes a sustainable foundation model for scientific discovery in the petabyte-scale era of the Vera C. Rubin Observatory.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools for LaTeX formatting, refining academic phrasing, and improving clarity and readability. We independently conceived the research methodology, executed all experiments, conducted the data analysis, and thoroughly reviewed all text and figures. We take full responsibility for the originality, scientific integrity, and final content presented in this work.

REFERENCES

Tarek Allam and Jason D. McEwen. Paying attention to astronomical transients: introducing the time-series transformer for photometric classification. *RAS Techniques and Instruments*, 3(1): 209–223, January 2024. doi: 10.1093/rasti/rzad046.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Eric C. Bellm, Shrinivas R. Kulkarni, Matthew J. Graham, Richard Dekany, Roger M. Smith, Reed Riddle, Frank J. Masci, George Helou, Thomas A. Prince, Scott M. Adams, C. Barbarino, Tom Barlow, James Bauer, Ron Beck, Justin Belicki, Rahul Biswas, Nadejda Blagorodnova, Dennis Bodewits, Bryce Bolin, Valery Brinnel, Tim Brooke, Brian Bue, Mattia Bulla, Rick Burruss, S. Bradley Cenko, Chan-Kao Chang, Andrew Connolly, Michael Coughlin, John Cromer, Virginia Cunningham, Kishalay De, Alex Delacroix, Vandana Desai, Dmitry A. Duev, Gwendolyn Eadie, Tony L. Farnham, Michael Feeney, Ulrich Feindt, David Flynn, Anna Franckowiak, S. Frederick, C. Fremling, Avishay Gal-Yam, Suvi Gezari, Matteo Giomi, Daniel A. Goldstein, V. Zach Golkhou, Ariel Goobar, Steven Groom, Eugean Hacopians, David Hale, John Henning, Anna Y. Q. Ho, David Hover, Justin Howell, Tiara Hung, Daniela Huppenkothen, David Imel, Wing-Huen Ip, Željko Ivezić, Edward Jackson, Lynne Jones, Mario Juric, Mansi M. Kasliwal, S. Kaspi, Stephen Kaye, Michael S. P. Kelley, Marek Kowalski, Emily Kramer, Thomas Kupfer, Walter Landry, Russ R. Laher, Chien-De Lee, Hsing Wen Lin, Zhong-Yi Lin, Ragnhild Lunnan, Matteo Giomi, Ashish Mahabal, Peter Mao, Adam A. Miller, Serge Monkewitz, Patrick Murphy, Chow-Choong Ngeow, Jakob Nordin, Peter Nugent, Eran Ofek, Maria T. Patterson, Bryan Penprase, Michael Porter, Ludwig Rauch, Umaa Rebbapragada, Dan Reiley, Mickael Rigault, Hector Rodriguez, Jan van Roestel, Ben Rusholme, Jakob van Santen, S. Schulze, David L. Shupe, Leo P. Singer, Maayane T. Soumagnac, Robert Stein, Jason Surace, Jesper Sollerman, Paula Szkody, F. Taddia, Scott Terek, Angela Van Sistine, Sjoert van Velzen, W. Thomas Vestrand, Richard Walters, Charlotte Ward, Quan-Zhi Ye, Po-Chieh Yu, Lin Yan, and Jeffry Zolkower. The Zwicky Transient Facility: System Overview, Performance, and First Results. *PASP*, 131(995):018002, January 2019. doi: 10.1088/1538-3873/aaecbe.

Guillermo Cabrera-Vives, D. Moreno-Cartagena, N. Astorga, I. Reyes-Jainaga, F. Förster, P. Huijse, J. Arredondo, A. M. Muñoz Arancibia, A. Bayo, M. Catelan, et al. ATAT: Astronomical transformer for time series and tabular data. *Astronomy & Astrophysics*, 2024. doi: 10.1051/0004-6361/202449475. URL https://www.aanda.org/articles/aa/full_html/2024/09/aa49475-24/aa49475-24.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, 2019.

Junyao Ding, Xiaodian Chen, Xinyi Gao, Xiaoyu Tang, Shu Wang, Yang Huang, Xinyu Qi, Guirong Xue, Ali Luo, and Jifeng Liu. StarCLR: Contrastive Learning Representation for Astronomical Light Curves. *ApJ*, 1003(2):141, June 2026. doi: 10.3847/1538-4357/ae64ef.

Cristobal Donoso-Oliva, Ignacio Becker, Pavlos Protopapas, Guillermo Cabrera-Vives, M. Vishnu, and Harsh Vardhan. ASTROMER: A transformer-based embedding for the representation of light curves. *Astronomy & Astrophysics*, 670:A54, 2023. doi: 10.1051/0004-6361/202243956.

Cristobal Donoso-Oliva, Ignacio Becker, Pavlos Protopapas, Guillermo Cabrera-Vives, Martina Cádiz-Leyton, and Daniel Moreno-Cartagena. Generalizing across astronomical surveys: Few-shot light curve classification with Astromer 2. *Astronomy & Astrophysics*, 707:A170, 2026.

L. Eyer, M. Audard, B. Holl, L. Rimoldini, M. I. Carnerero, G. Clementini, J. De Ridder, E. Distefano, D. W. Evans, P. Gavras, et al. Gaia data release 3 - summary of the variability processing and analysis. *Astronomy & Astrophysics*, 674:A13, 2023. doi: 10.1051/0004-6361/202244242. URL https://www.aanda.org/articles/aa/full_html/2023/06/aa44242-22/aa44242-22.html.

Gaia Collaboration, A. Vallenari, A. G. A. Brown, T. Prusti, J. H. J. de Bruijne, F. Arenou, C. Babusiaux, M. Biermann, O. L. Creevey, C. Ducourant, D. W. Evans, L. Eyer, R. Guerra, A. Hutton, C. Jordi, S. A. Klioner, U. L. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, C. Soubiran, P. Tanga, N. A. Walton, C. A. L. Bailer-Jones, U. Bastian, R. Drimmel, F. Jansen, D. Katz, M. G. Lattanzi, F. van Leeuwen, J. Bakker, C. Cacciari, J. Castañeda, F. De Angeli, C. Fabricius, M. Fouesneau, Y. Frémat, L. Galluccio, A. Guerrier, U. Heiter, E. Masana, R. Messineo, N. Mowlavi, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, W. Roux, G. M. Seabroke, R. Sordo, F. Thévenin, G. Gracia-Abril, J. Portell, D. Teyssier, M. Altmann, R. Andrae, M. Audard, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, P. W. Burgess, D. Busonero, G. Busso, H. Cánovas, B. Carry, A. Cellino, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, P. de Teodoro, M. Nuñez Campos, L. Delchambre, A. Dell'Oro, P. Esquej, J. Fernández-Hernández, E. Fraile, D. Garabato, P. García-Lario, E. Gosset, R. Haigron, J.-L. Halbwachs, N. C. Hambly, D. L. Harrison, J. Hernández, D. Hestroffer, S. T. Hodgkin, B. Holl, K. Janßen, G. Jevardat de Fombelle, S. Jordan, A. Krone-Martins, A. C. Lanzafame, W. Löffler, O. Marchal, P. M. Marrese, A. Moitinho, K. Muinonen, P. Osborne, E. Pancino, T. Pauwels, A. Recio-Blanco, C. Reylé, M. Riello, L. Rimoldini, T. Roegiers, J. Rybizki, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, E. Utrilla, M. van Leeuwen, U. Abbas, P. Ábrahám, A. Abreu Aramburu, C. Aerts, J. J. Aguado, M. Ajaj, F. Aldea-Montero, G. Altavilla, M. A. Álvarez, J. Alves, F. Anders, R. I. Anderson, E. Anglada Varela, T. Antoja, D. Baines, S. G. Baker, L. Balaguer-Núñez, E. Balbinot, Z. Balog, C. Barache, D. Barbato, M. Barros, M. A. Barstow, S. Bartolomé, J.-L. Bassilana, N. Bauchet, U. Becciani, M. Bellazzini, A. Berihuete, M. Bernet, S. Bertone, L. Bianchi, A. Binnenfeld, S. Blanco-Cuaresma, A. Blazere, T. Boch, A. Bombrun, D. Bossini, S. Bouquillon, A. Bragaglia, L. Bramante, E. Breedt, A. Bressan, N. Brouillet, E. Brugaletta, B. Bucciarelli, A. Burlacu, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, T. Cantat-Gaudin, R. Carballo, T. Carlucci, M. I. Carnerero, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, L. Chaoul, P. Charlot, L. Chemin, V. Chiaramida, A. Chiavassa, N. Chornay, G. Comoretto, G. Contursi, W. J. Cooper, T. Cornez, S. Cowell, F. Crifo, M. Cropper, M. Crosta, C. Crowley, C. Dafonte, A. Dapergolas, M. David, P. David, P. de Laverny, F. De Luise, and R. De March. Gaia Data Release 3. Summary of the content and survey properties. *A&A*, 674:A1, June 2023. doi: 10.1051/0004-6361/202243940.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Paul F. X. Gregory, Jeroen Audenaert, Mykyta Kliapets, Daniel Muthukrishna, Andrew Tkachenko, Marek Skarka, Marc Hon, and George R. Ricker. ASTRAFier: A Novel and Scalable Transformer-based Stellar Variability Classifier. *arXiv e-prints*, art. arXiv:2604.07437, April 2026. doi: 10.48550/arXiv.2604.07437.

Renée Hložek, Alex I. Malz, Richard Kessler, Tarek Allam, Anita Bahmanyar, Rahul Biswas, Alexandre Boucaud, Mi Dai, Lluís Galbany, Emille E. O. Ishida, et al. Results of the photometric LSST astronomical time-series classification challenge (PLAsTiCC). *The Astronomical Journal*, 165(6):268, 2023. doi: 10.3847/1538-3881/acd37e.

Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. LSST: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019. doi: 10.3847/1538-4357/ab042c.

Richard Kessler, Gautham Narayan, Andrés Avelino, Etienne Bachelet, Rahul Biswas, Peter J. Brown, David F. Chernoff, Andrew J. Connolly, Mi Dai, Scott Daniel, et al. Models and simulations for the photometric LSST astronomical time series classification challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, 2019. doi: 10.1088/1538-3873/ab26f1.

Weijian Li, Hong-Yu Chen, Nabeel Rehemtulla, Ved G. Shah, Dennis Wu, Dongho Kim, Qinjie Lin, Adam A. Miller, and Han Liu. StarEmbed: Benchmarking Time Series Foundation Models on Astronomical Observations of Variable Stars. *arXiv e-prints*, art. arXiv:2510.06200, October 2025. doi: 10.48550/arXiv.2510.06200.

Torsha Majumder, Konstantin Malanchev, and Emille E. O. Ishida. Multi-scale contrastive attention for light-curve representation learning, 2026. URL https://arxiv.org/abs/2606.31627.

K. L. Malanchev, M. V. Pruzhinskaya, V. S. Korolev, P. D. Aleo, M. V. Kornilov, E. E. O. Ishida, V. V. Krushinsky, F. Mondon, S. Sreejith, A. A. Volnova, A. A. Belinski, A. V. Dodin, A. M. Tatarnikov, S. G. Zheltoukhov, and (The SNAD Team). Anomaly detection in the Zwicky Transient Facility DR3. *MNRAS*, 502(4):5147–5175, April 2021. doi: 10.1093/mnras/stab316.

Liam Parker, Siavash Golkar, Miles Cranmer, Shirley Ho, François Lanusse, et al. Astroclip: A cross-modal foundation model for galaxies. *arXiv preprint arXiv:2310.03024*, 2023.

Lorenzo Rimoldini, Berry Holl, Panagiotis Gavras, Marc Audard, Joris De Ridder, Nami Mowlavi, Krzysztof Nienartowicz, Grégory Jevardat de Fombelle, Isabelle Lecoeur-Taïbi, Lea Karbevska, Dafydd W. Evans, Péter Ábrahám, Maria I. Carnerero, Gisella Clementini, Elisa Distefano, Alessia Garofalo, Pedro García-Lario, Roy Gomel, Sergei A. Klioner, Katarzyna Kruszyńska, Alessandro C. Lanzafame, Thomas Lebzelter, Gábor Marton, Tsevi Mazeh, Roberto Molinaro, Aviad Panahi, Claudia M. Raiteri, Vincenzo Ripepi, László Szabados, David Teyssier, Michele Trabucchi, Łukasz Wyrzykowski, Shay Zucker, and Laurent Eyer. Gaia Data Release 3. All-sky classification of 12.4 million variable sources into 25 classes. *A&A*, 674:A14, June 2023. doi: 10.1051/0004-6361/202245591.

M. Rizhko and J. S. Bloom. AstroM3: A self-supervised multimodal model for astronomy. *The Astronomical Journal*, 170(1):28, 2025. doi: 10.3847/1538-3881/adcbad. URL https://iopscience.iop.org/article/10.3847/1538-3881/adcbad.

Yicheng Rui. Domain-informed multi-view self-distillation for astronomical light-curve representation learning with jepa. *arXiv preprint arXiv:2606.28446*, 2026.

E. Russeil, K. L. Malanchev, P. D. Aleo, E. E. O. Ishida, M. V. Pruzhinskaya, E. Gangler, A. D. Lavrukhina, A. A. Volnova, A. Voloshina, T. Semenikhin, S. Sreejith, M. V. Kornilov, and V. S. Korolev. Rainbow: a colorful approach on multi-passband light curve estimation. *Astronomy & Astrophysics*, 683:A251, 2024. doi: 10.1051/0004-6361/202348158.

The PLAsTiCC team, Tarek Allam, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, et al. The photometric LSST astronomical time-series classification challenge (PLAsTiCC): Data set. *arXiv preprint arXiv:1810.00001*, 2018.

Zora Tung. ATCAT: Astronomical timeseries causal transformer. *arXiv preprint arXiv:2511.00614*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Peng, Yankai Lin, and Lei Hou. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1659–1671, 2020.

Gerald Woo, Chenghao Liu, Dripta Sahoo, Akshat Kumar, and Steven Hoi. Unified training of universal time series forecasting models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Xiaoxiong Zuo, Yihan Tao, Yang Huang, Zhixuan Kang, Huaxi Chen, Chenzhou Cui, Jiashu Pan, Xiao Kong, Yuan-Sen Ting, Xiaoyu Tang, Henggeng Han, Haiyang Mu, Yunfei Xu, Dongwei Fan, Guirong Xue, Ali Luo, and Jifeng Liu. FALCO: Foundation Model of Astronomical Light Curves for Time Domain Astronomy. Implementation and Applications on Kepler Data. *AJ*, 171 (1):10, January 2026. doi: 10.3847/1538-3881/ae1467.

A Dataset Details, Curation, and Class Distributions

All datasets in this study are derived from the cross-matched `Zubercal` × `Gaia` catalog curated and publicly released alongside the Astra-CLR framework (Majumder et al., 2026). Importantly, we note that the astronomical variability labels in this catalog are derived from automated machine-learning classification pipelines (specifically, the Gaia DR3 variability processing pipeline; Rimoldini et al. 2023; Eyer et al. 2023) rather than human annotation or exhaustive spectroscopic confirmation. Following established self-supervised learning benchmarks in time-domain astronomy (Majumder et al., 2026), these machine-generated annotations serve as a standardized, survey-scale proxy ground truth for downstream evaluation.

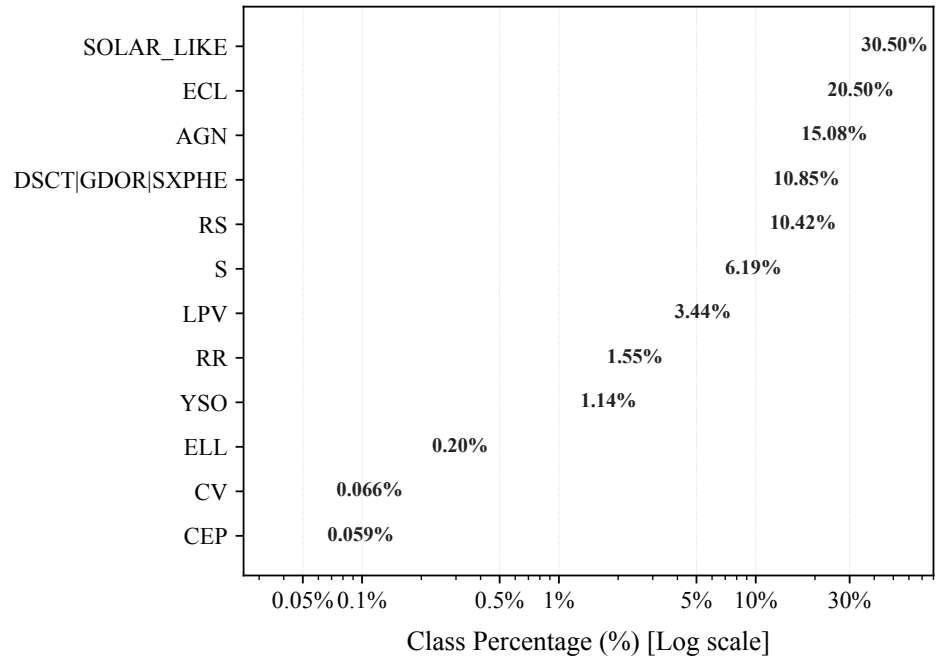

Figure 5: **Class distribution across in-distribution variability types.** Relative class frequencies across all 12 astrophysical variability categories shared identically between the `2.6M-Full` catalog and the `26K-Proxy` validation subset. The logarithmic vertical axis illustrates the severe class imbalance inherent to astronomical surveys, spanning over three orders of magnitude from majority Solar-like variables (`SOLAR_LIKE`, 30.50%) to rare classes such as Classical Cepheids (`CEP`, 0.059%) and Cataclysmic Variables (`CV`, 0.066%).

A.1 In-Distribution Datasets: Pre-training and Evaluation Benchmarks

Our primary in-distribution dataset comprises 2,593,567 multi-band light curves across 12 broad astronomical variability types, designated as `2.6M-Full`. We partition `2.6M-Full` into an $80/20$ split: the 80% training split ($2,073,458$ light curves) is used exclusively for self-supervised pre-training, while the isolated 20% validation pool ($\approx$ 520K sources) serves as the source from which all proxy and evaluation subsets are derived, guaranteeing zero data leakage.

From this isolated validation pool, we extract three distinct experimental subsets:

1. **`6K-Proxy` (Ablation Benchmark):** A class-balanced subset containing 500 light curves per class across all 12 types ($6,000$ sources total). This dataset is partitioned into an $80/20$ train-validation split and is utilized for rapid convergence benchmarking and the extensive architectural ablation search detailed in Table 1.
2. **`26K-Proxy` (Scaling & Optimization Benchmark):** An imbalanced subset of 26,250 light curves replicating the natural class distribution of the full survey, used to validate training stability and ensure architectural findings hold across data scales.

Table 5: **Summary of Out-of-Distribution (`OOD-Eval`) and Cross-Survey (`PLAsTiCC`) evaluation datasets.** Sample counts reflect the final light curves retained after preprocessing. Classes with < 10 samples in PLAsTiCC are excluded from few-shot classification.

Dataset	Variability / Transient Category	Class Identifier	Sample Count
OOD-Eval	β Cephei Stars	BCEP	25
	Microlensing Events	MICROLENSING	13
	R Coronae Borealis Stars	RCB	13
	Supernovae	SN	136
	Variable White Dwarfs	WD	248
	Subtotal		**435**
PLAsTiCC	Active Galactic Nuclei	AGN	82
	Eclipsing Binary Stars	ECL	92
	RR Lyrae Stars	RR	57
	Type II Supernovae	SNII	93
	Type Ia Supernovae	SNIa	263
	Microlensing Events	MICROLENSING	5
	Tidal Disruption Events	TDE	6
	Subtotal		**598**

3. **`120K-Eval` (Downstream Evaluation Benchmark):** A standardized, balanced benchmark containing $10,000$ light curves per class ($120,000$ sources total), utilized for linear probing, weighted k-NN evaluations, baseline comparisons (Section 4.2), and label-efficiency experiments (Section 4.4).

Figure 5 illustrates the overall class distribution of the uncurated `2.6M-Full` catalog, highlighting the severe class imbalance that is preserved within the `26K-Proxy` subset.

A.2 Out-of-Distribution and Cross-Survey Datasets (OOD and PLAsTiCC)

ZTF Out-of-Distribution Benchmark (`OOD-Eval`). To assess zero-shot retrieval and few-shot adaptation on unseen physical phenomena, we construct an out-of-distribution evaluation set from the ZTF DR16 catalog comprising five classes excluded from all pre-training stages: β Cephei stars (`BCEP`), Microlensing events (`MICROLENSING`), R Coronae Borealis stars (`RCB`), Supernovae (`SN`), and Variable White Dwarfs (`WD`). The data curation and standardization pipeline follows the identical protocol established for the pre-training dataset (Section 3.1). Detailed class counts are summarized in Table 5.

Cross-Survey Transfer Benchmark (`PLAsTiCC`). To evaluate zero-shot and few-shot cross-instrument generalizability, we extract simulated light curves from the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC; The PLAsTiCC team et al., 2018; Kessler et al., 2019). To ensure high signal-to-noise and dense temporal sampling, we restrict our sample to objects located within the Deep Drilling Fields (`ddf = 1`). Integer class identifiers are mapped to physical astrophysical classifications in accordance with Kessler et al. (2019) and Hložek et al. (2023).

From this catalog, we construct two evaluation subsets: (i) a *Domain Adaptation* subset containing known physical classes present in ZTF pre-training (Active Galactic Nuclei [`AGN`], RR Lyrae stars [`RR`], and Eclipsing Binaries [`ECL`]); and (ii) a *Comprehensive Transfer* subset introducing explosive transients unseen during pre-training (Type Ia Supernovae [`SNIa`] and Type II Supernovae [`SNII`]).

To align with `READDit`'s multi-band architecture, we retain observations strictly in the g, r, and i passbands. Measurements are filtered to retain high-significance detections (SNR > 3). Raw fluxes and flux uncertainties are converted to apparent AB magnitudes using the PLAsTiCC zero-point $m_0 = 27.5$ mag via standard error propagation:

$$m = -2.5 \log_{10}(f) + 27.5, \quad \sigma_m = \frac{2.5}{\ln(10)} \frac{\sigma_f}{f}. \tag{3}$$

To ensure sufficient temporal support while accommodating finite transient lifespans, we enforce a minimum threshold of 10 detections in both the g and r bands, leaving the sparser i band un-

constrained. Light curves are subsequently processed via the standardization and random-window pipeline described in Section 3.1. Under these strict quality criteria, no kilonova candidates met the observational threshold, and they were consequently omitted. The final sample breakdown across all transfer classes is presented in Table 5.

B Self-Supervised Distillation Objective Formulation

Following the DINO self-distillation paradigm (Caron et al., 2021), the `READit` framework utilizes an asymmetric teacher–student architecture. Both the student network g_{θ_s} and teacher network g_{θ_t} share identical Transformer-encoder backbones and non-linear multi-layer perceptron (MLP) projection heads $h(\cdot)$, but maintain separate parameter sets θ_s and θ_t.

Parameter Updates via Exponential Moving Average (EMA). While the student parameters θ_s are optimized via stochastic gradient descent by minimizing the cross-entropy loss, the teacher parameters θ_t are updated via an Exponential Moving Average (EMA) of the student weights:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s, \tag{4}$$

where $\lambda \in [0, 1)$ is the momentum coefficient following a cosine schedule from 0.996 toward 1.0 during training.

Cross-View Distillation and Sharpening. Given an astronomical source, the input pipeline constructs a set of multi-scale views $\mathcal{V}$ consisting of Global Views (GVs, x_g) and Local Views (LVs, x_l). For any view x, the backbone extracts a global pooled vector $\mathbf{z}$, which is mapped by the projection head to a K-dimensional latent space, yielding the logit vector $g(x) = h(\mathbf{z}) \in \mathbb{R}^K$.

The networks output probability distributions P_s and P_t over K dimensions via temperature-scaled softmax functions:

$$P_s(x)^{(k)} = \frac{\exp\left(g_{\theta_s}(x)^{(k)}/\tau_s\right)}{\sum_{j=1}^{K}\exp\left(g_{\theta_s}(x)^{(j)}/\tau_s\right)}, \quad P_t(x)^{(k)} = \frac{\exp\left((g_{\theta_t}(x)^{(k)} - c^{(k)})/\tau_t\right)}{\sum_{j=1}^{K}\exp\left((g_{\theta_t}(x)^{(j)} - c^{(j)})/\tau_t\right)}, \tag{5}$$

where $\tau_s = 0.1$ denotes the fixed student temperature, and τ_t is the dynamic teacher temperature. Following standard distillation protocols, τ_t undergoes a linear warmup over the first $E_{\text{warmup}} = 30$ epochs from $\tau_{\text{start}} = 0.04$ to $\tau_{\text{base}} = 0.07$, after which it remains constant:

$$\tau_t(e) = \begin{cases} \tau_{\text{start}} + (\tau_{\text{base}} - \tau_{\text{start}}) \cdot \left(\frac{e}{E_{\text{warmup}}}\right), & \text{if } e < E_{\text{warmup}}, \\ \tau_{\text{base}}, & \text{if } e \geq E_{\text{warmup}}, \end{cases} \tag{6}$$

where e denotes the pre-training epoch. As established in Section 4.3, `READit` achieves peak representational saturation early at Epoch$\sim$ 11 (at which point $\tau_t \approx 0.051$), enabling early stopping before completing the planned 30-epoch warmup. Setting the teacher temperature lower than the student temperature ($\tau_t < \tau_s$) induces **sharpening**, which prevents the probability distribution from collapsing into a uniform distribution across all K dimensions.

Centering Dynamics (Preventing Collapse). To prevent representation collapse (where all samples map to an identical dimension regardless of input), teacher logits are centered by subtracting a running bias vector $\mathbf{c} \in \mathbb{R}^K$. The center vector $\mathbf{c}$ is updated across minibatch iterations of size B using momentum coefficient $m > 0$:

$$\mathbf{c} \leftarrow m\mathbf{c} + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i), \tag{7}$$

where $m = 0.9$. Centering ensures that the teacher's probability distribution does not dominate in any single dimension, while sharpening prevents it from collapsing to a uniform distribution.

Objective Function. The student is trained to predict the teacher's output distribution across asymmetric views. Specifically, the teacher receives only Global Views (x_g), while the student

processes both Global and Local Views ($x' \in \mathcal{V} \setminus \{x_g\}$). The self-distillation objective minimizes the cross-entropy loss $H(P_t, P_s)$:

$$\min_{\theta_s} \sum_{x_g \in \mathcal{V}_{\text{global}}} \sum_{x' \in \mathcal{V} \setminus \{x_g\}} H(P_t(x_g), P_s(x')), \quad \text{where } H(P_t, P_s) = -\sum_{k=1}^{K} P_t^{(k)} \log P_s^{(k)}. \tag{8}$$

Gradients are propagated exclusively through the student parameters θ_s, while the teacher weights are updated strictly via the EMA dynamics in Eq. equation 4.

C Baseline Implementation Details and Hand-Crafted Features

Classical Hand-Crafted Feature Baseline. To benchmark `READit` against domain-expert feature engineering, we extract statistical and morphological features using the open-source `light-curve` package (Malanchev et al., 2021). As outlined in Section 4.2, features are extracted exclusively from the ZTF r-band to maximize temporal cadence and signal-to-noise ratio. In total, 42 scalar features are computed per light curve across four core physical dimensions:

- **Statistical Moments & Distributional Metrics:** Mean, WeightedMean, Median, StandardDeviation, Skew, Kurtosis, MedianAbsoluteDeviation, InterPercentileRange, AndersonDarlingNormal, ReducedChi2, and MeanVariance.
- **Variability & Morphological Indices:** Amplitude, PercentAmplitude, ExcessVariance, Cusum, η, η_e, StetsonK, Roms, OtsuSplit, Beyond1Std, MagnitudePercentageRatio, MedianBufferRangePercentage, and PercentDifferenceMagnitudePercentile.
- **Temporal Sampling & Trend Properties:** ObservationCount, Duration, TimeMean, TimeStandardDeviation, MinimumTimeInterval, MaximumTimeInterval, MaximumSlope, LinearTrend, and LinearFit.
- **Periodicity & Spectral Power:** Lomb-Scargle Periodogram feature extraction estimating dominant period lengths and peak spectral powers.

The resulting 42-dimensional feature vectors are standardized (zero-mean, unit-variance) and classified using a Random Forest classifier with 250 estimators.

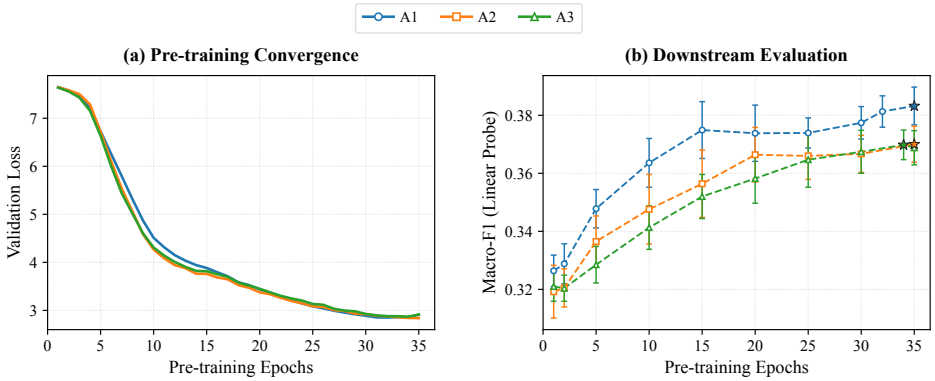

Figure 6: **Extended Training Dynamics: Suite A (Data Augmentations).** Evaluation on the `6K-Proxy` benchmark comparing the proposed `READit` configuration (A1) against removing time-bin masking (A2) and removing photometric noise (A3). **(Left)** Validation distillation loss. **(Right)** Downstream Linear Probing Macro-F1. Removing temporal masking or photometric perturbations fails to yield a superior peak Macro-F1. The combined augmentation suite (A1) acts as an essential regularizer, ensuring stable, monotonic convergence.

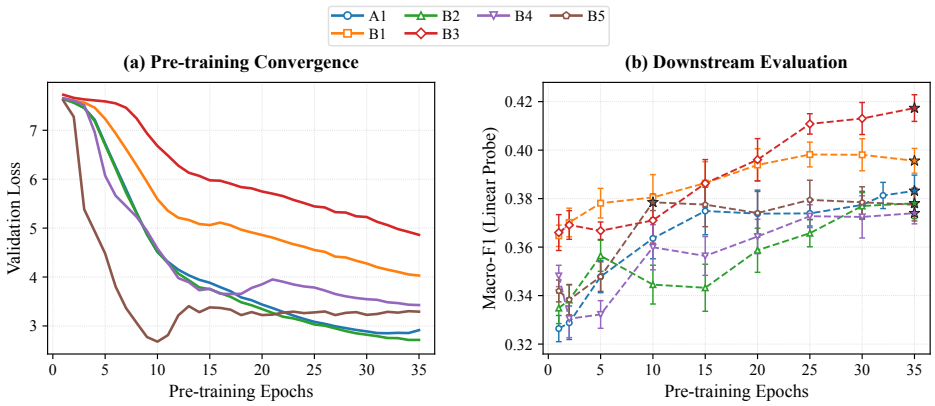

Figure 7: **Extended Training Dynamics: Suite B (Inductive Biases).** Benchmarking physical grounding against numerical shortcuts on the `6K-Proxy` dataset. **(Left)** Validation distillation loss. **(Right)** Downstream Linear Probing Macro-F1. Replacing continuous MJD with integer indices (B1) or substituting non-linear wavelength projections with linear layers (B3) allows the model to exploit survey-specific cadence regularities, yielding artificially inflated Macro-F1 scores. Furthermore, deterministic global views and 1D masked GAP provide superior training stability compared to stochastic global views (B5) and `[CLS]` tokens (B4).

D Systematic Ablation Studies and Analysis

D.1 Suite-Level Design Dynamics on the 6K-Proxy Benchmark

To systematically dissect the contribution of each design component, we benchmarked candidate configurations across four empirical suites on the `6K-Proxy` dataset (Table 1). In Figures 6–9, we present the epoch-by-epoch optimization dynamics across all suites, tracking validation distillation loss alongside downstream Linear Probing Macro-F1 across 30 epochs.

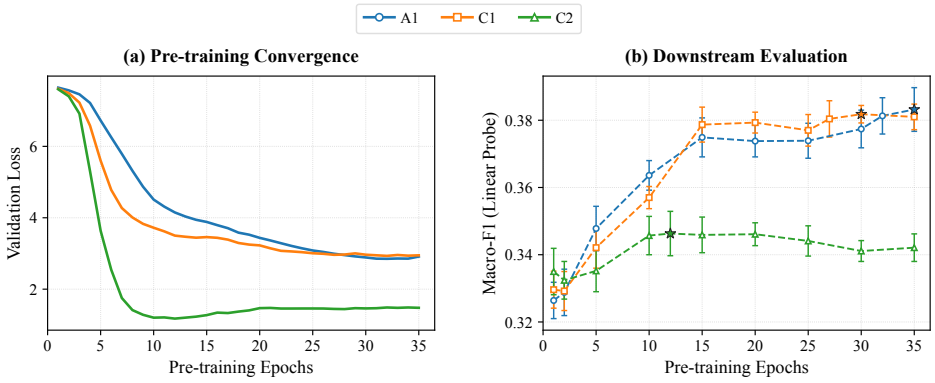

Figure 8: **Extended Training Dynamics: Suite C (View Complexity).** Impact of multi-crop density on representational capacity on the `6K-Proxy` dataset. **(Left)** Validation distillation loss. **(Right)** Downstream Linear Probing Macro-F1. The conventional 2-view baseline (C2) exhibits early task saturation, hitting an optimization ceiling near Epoch 10. Expanding to 6 views (C1) and 9 views (A1) systematically raises the representation quality ceiling, justifying the 9-view configuration to resolve complex multi-scale temporal dynamics.

D.2 Architectural Comparison Against Supervised Baselines

To evaluate whether self-supervised pre-training provides fundamental advantages over standard supervised models, we pre-trained `READit` on the `6K-Proxy` dataset and benchmarked it against supervised 1D-CNN, Bi-LSTM, and Transformer architectures trained from scratch on the identical

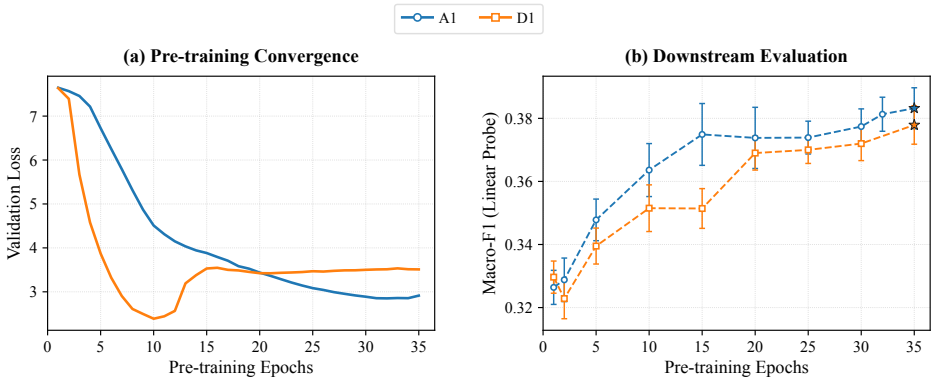

Figure 9: **Extended Training Dynamics: Suite D (Architectural Scale).** Benchmarking the proposed "Tiny" Transformer ($N = 4, d_{\text{model}} = 256$) against a "Medium" Transformer ($N = 6, d_{\text{model}} = 384$) on the `6K-Proxy` dataset. **(Left)** Validation distillation loss. **(Right)** Downstream Linear Probing Macro-F1. Scaling model capacity (D1) induces over-parameterization, causing early divergence in the validation loss without a corresponding peak in Macro-F1. This confirms that our compact 4-layer backbone is optimally calibrated for the astronomical manifold.

data. As illustrated in Figure 10, self-supervised representations outperform all supervised baselines trained from scratch, highlighting the value of label-free pre-training on irregular time-series.

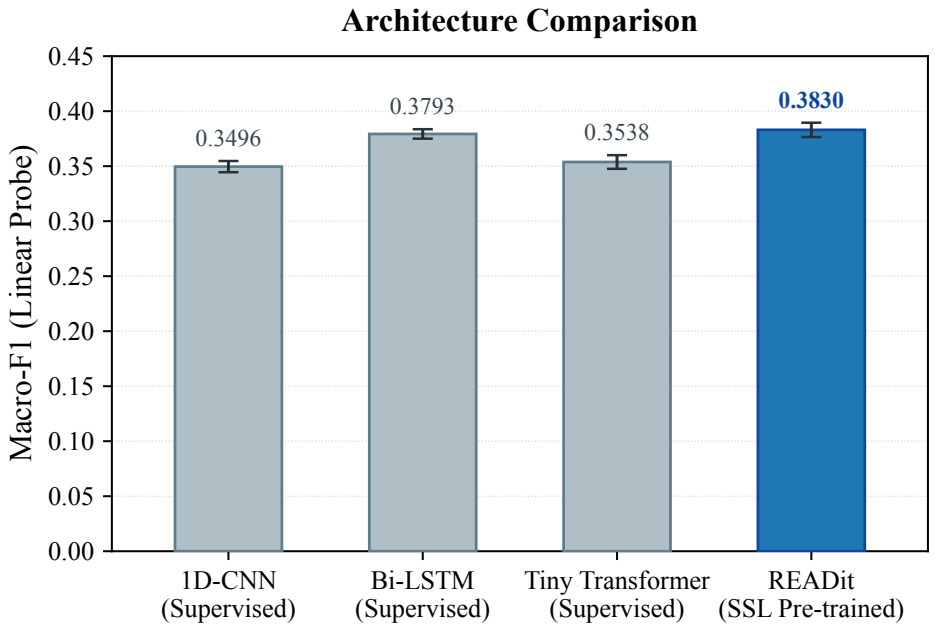

Figure 10: **Supervised Architectural Baselines vs. SSL Pre-training.** Downstream classification performance of supervised models trained from scratch compared against frozen representations from `READit`, evaluated on the `6K-Proxy` dataset. `READit` achieves superior representation quality (0.3830 ± 0.0055 Macro-F1), outperforming the strongest supervised baseline (Bi-LSTM, 0.3793 ± 0.0043). Error bars represent bootstrapped standard deviations over 50 resamples.

D.3 Scale Validation and the Shortcut Learning Hypothesis

To validate architectural choices prior to large-scale pre-training, we evaluated the proposed model (A1) against the top-scoring ablation runners-up from Suite B—Integer Index Positional Encodings (B1) and Linear Layer for Filters (B3)—using the imbalanced `26K-Proxy` dataset for pre-training.

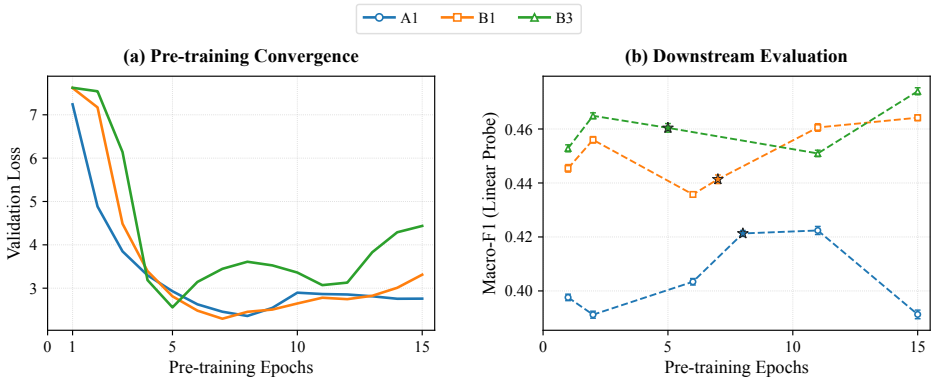

Figure 11: **Extended Training Dynamics: Proposed Architecture vs. Shortcut Runners-Up.** Epoch-by-epoch dynamics (Epochs 1–15) comparing the proposed `READit` configuration (A1) against shortcut runners-up: Integer Index PE (B1) and Linear Layer for Filters (B3), pre-trained on `26K-Proxy` and evaluated on `6K-Proxy`. **(Left)** Validation distillation loss. **(Right)** Downstream Linear Probing Macro-F1. For B1 and B3, downstream performance peaks after the validation loss basin, rendering automated early-stopping unviable. In contrast, the proposed A1 configuration exhibits strict alignment between the loss minimum and peak representation quality at Epoch 8.

While B1 and B3 achieved marginally higher Macro-F1 scores evaluated on the balanced `6K-Proxy` benchmark (Table 1), our scaling analysis reveals this to be an artifact of *shortcut learning*. As shown in Figure 11, the downstream Macro-F1 of B1 and B3 continues to drift upward even as validation loss diverges, breaking the early-stopping alignment essential for self-supervised training.

Furthermore, evaluating these models on extreme minority classes ($< 2\%$ of the dataset) and on cross-survey transfer to simulated LSST observations (PLAsTiCC) reveals severe performance degradation for both B1 and B3 (Figure 12). This confirms that runners-up achieve inflated overall scores by overfitting to majority-class cadences at the expense of rare physical phenomena. In contrast, configuration A1 enforces continuous physical grounding via continuous MJD encod-

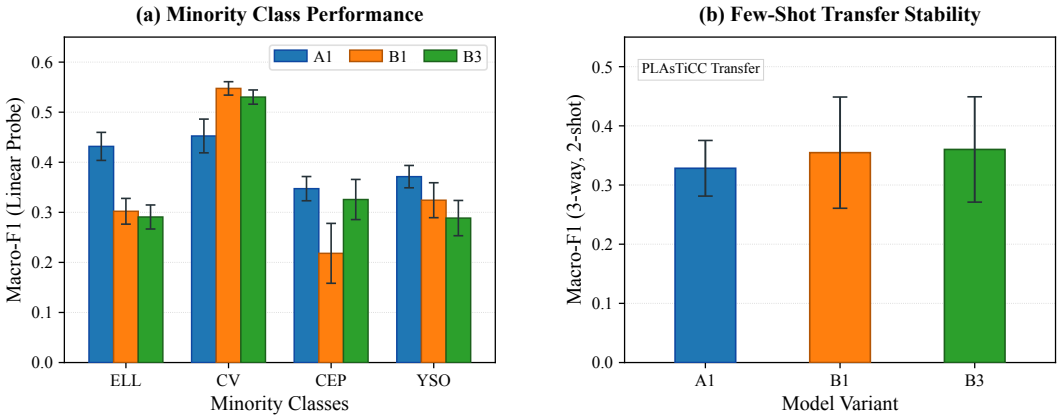

Figure 12: **Generalization and Stability: Proposed Architecture vs. Shortcut Runners-Up.** Evaluating configuration A1 against shortcut runners-up B1 (Integer Index PE) and B3 (Linear Layer for Filters). **(a) Minority Class Fidelity:** Pre-trained on the imbalanced `26K-Proxy` dataset and evaluated on `6K-Proxy`. While runners-up overfit to majority-class statistics, A1 maintains superior fidelity on extreme minority classes ($< 2\%$ of the sample), preventing representation collapse. **(b) Cross-Survey Transfer Stability:** 3-way, 2-shot prototypical evaluation on simulated LSST (PLAsTiCC) data over 100 random episodes. The proposed configuration (A1) exhibits significantly tighter variance and superior out-of-distribution stability.

ings and non-linear filter projections, preserving minority-class fidelity and ensuring robust out-of-distribution transferability.

E LATENT SPACE TOPOLOGY AND UMAP VISUALIZATIONS

E.1 GLOBAL MANIFOLD OF IN-DISTRIBUTION VARIABILITY (2.6M-FULL DATASET)

To qualitatively assess the geometry of the learned latent manifold, Figure 13 presents 2D UMAP projections of the feature space before and after partial fine-tuning.

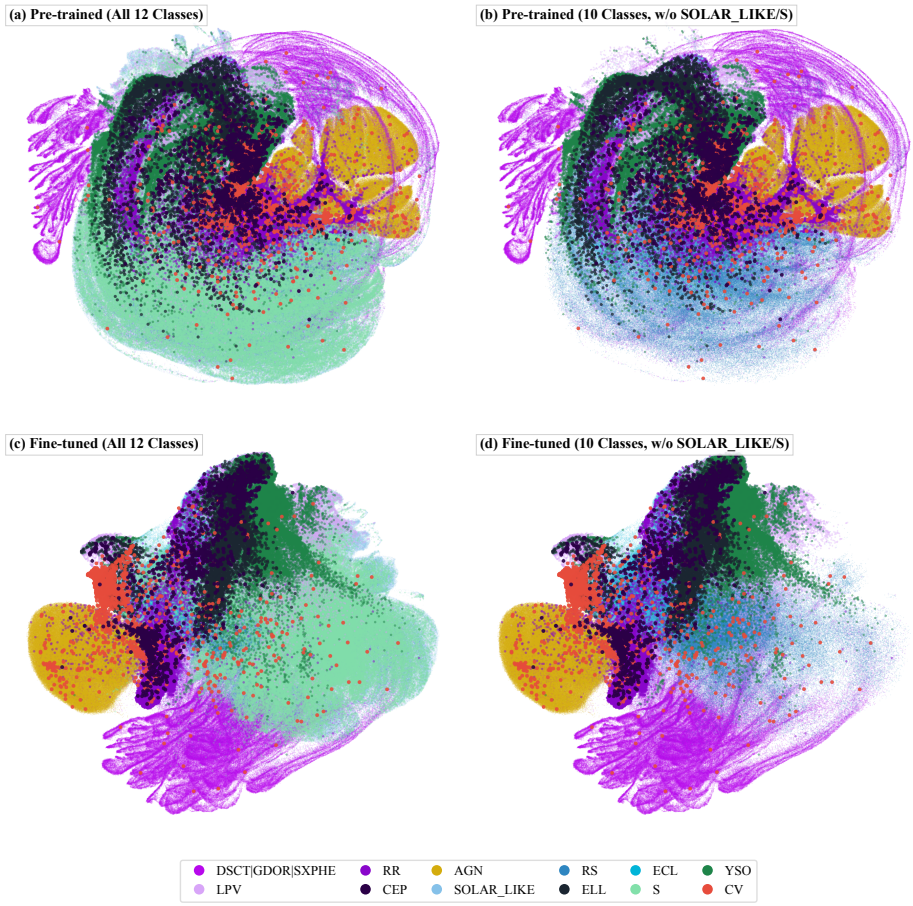

Figure 13: **UMAP projections of latent representation geometry before and after fine-tuning. (a) Pre-trained (All 12 Classes):** Frozen 2D latent space extracted from `READit` pre-trained on `2.6M-Full`. **(b) Pre-trained (10 Classes):** Feature distribution excluding the two dominant variability classes (`SOLAR_LIKE`, 30.50%, and `S`, 6.19%), unmasking fine-grained clustering and topological structure among minority classes. **(c) Fine-tuned (All 12 Classes):** Latent geometry following partial fine-tuning with $40,000$ labels, exhibiting tightened intra-class compactness and enlarged inter-class margin separation. **(d) Fine-tuned (10 Classes):** Partially fine-tuned manifold likewise excluding the two dominant classes (`SOLAR_LIKE` and `S`), demonstrating sharpened decision boundaries and tightly grouped clusters even among scarce variability types.

E.2 LATENT DISENTANGLEMENT OF UNSEEN OUT-OF-DISTRIBUTION SOURCES

Figure 14 illustrates the zero-shot representation geometry when projecting unseen out-of-distribution classes into the `READit` frozen feature space.

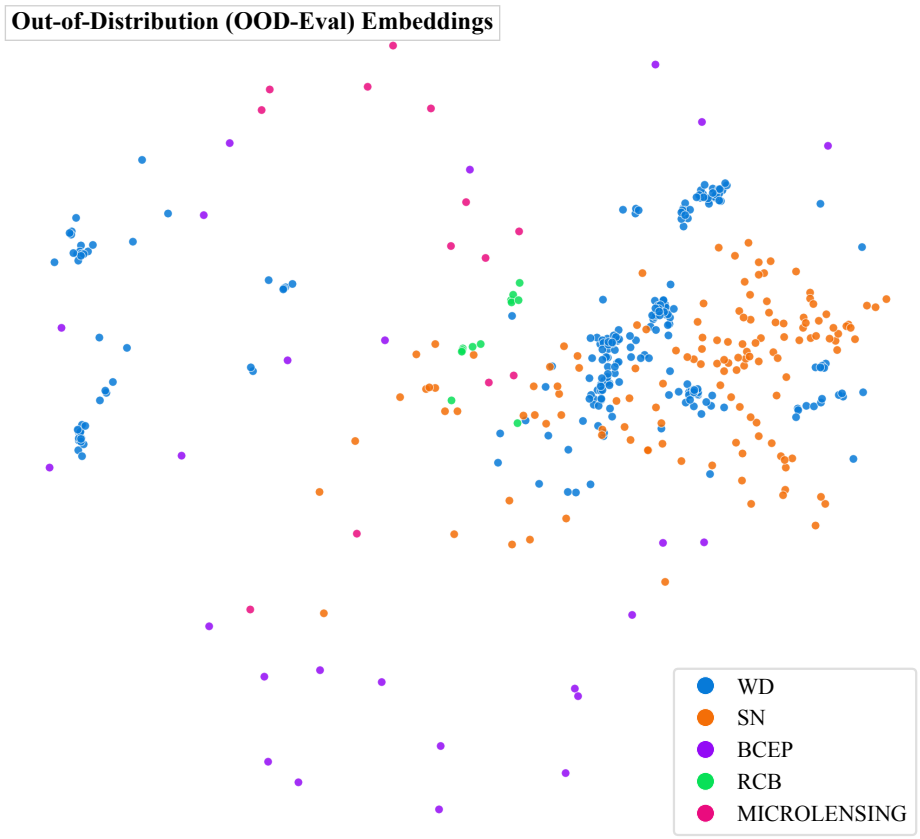

Figure 14: **Zero-shot feature projection of Out-of-Distribution (`OOD-Eval`) classes.** 2D UMAP projections of unseen astronomical variability classes and explosive transient phenomena (`BCEP`, `MICROLENSING`, `RCB`, `SN`, and `WD`) mapped into the frozen feature space of pre-trained `READit`. Despite receiving zero supervision on these specific categories during pre-training on `2.6M-Full`, the self-distillation objective naturally partitions these classes according to temporal morphology and multi-band photometric coherence, demonstrating strong representation transferability.