WHEN DOES SELF-SUPERVISED LEARNING TRANSFER TO TIME-SERIES TASKS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) assumes that solving pretext tasks on unlabeled data yields representations that transfer effectively across downstream applications via linear probing or fine-tuning. While this paradigm has driven major progress in vision and language, its benefits for time series remain under-investigated and often confounded by inconsistent experimental controls. To address this gap, we benchmark seven representative methods from five key SSL paradigms across anomaly detection, classification, and forecasting under parameter- and data-matched budgets. We find that transfer efficacy depends heavily on the downstream task. SSL yields substantial gains in anomaly detection and provides effective initializations for classification under fine-tuning, but offers limited to no advantage over non-pre-trained controls in forecasting. Furthermore, linear probing does not reliably predict fine-tuning performance, synthetic pre-training is often competitive with real-world corpora, and scaling encoder depth degrades forecasting accuracy. We synthesize these empirical results into practical evaluation and development guidelines for time-series SSL. Our code is available in the supplementary material.

1 INTRODUCTION

The success of self-supervised learning (SSL) in natural language processing (Devlin et al., 2019) and computer vision (Caron et al., 2021; Assran et al., 2023) has motivated its widespread adoption for time-series analysis (Nie et al., 2023; Wang et al., 2025; Das et al., 2024; Ansari et al., 2024). SSL relies on the foundational premise that models can learn underlying structures and semantic relationships from unlabeled data that transfer effectively to downstream tasks. While this premise is well established in language and vision, its validity in time series remains less clear due to substantial cross-domain distribution shifts and diverse temporal dynamics.

To date, empirical evidence supporting time-series SSL remains fragmented. Much of the recent progress in time-series foundation models has been demonstrated in forecasting and, in particular, under zero-shot evaluation (Das et al., 2024; Ansari et al., 2024). Beyond this narrow task scope, newly proposed time-series SSL methods often introduce specialized backbones, making it difficult to attribute observed performance gains to the SSL objective itself. Existing surveys on time-series SSL provide informative high-level taxonomies of contrastive, generative, and reconstruction methods (Zhang et al., 2024; Trirat et al., 2026; Liu et al., 2024), but generally do not provide controlled, cross-task experiments that decouple the contribution of the pre-training objective from confounding variations in model architecture, training data, and evaluation protocols. Furthermore, latent-space alignment frameworks, such as invariance-based methods (e.g., DINO (Caron et al., 2021)) and Joint-Embedding Predictive Architectures (e.g., JEPA (LeCun, 2022; Assran et al., 2023), LeJEPA (Balestriero & LeCun, 2025)), remain largely underexplored in the time-series literature. Finally, while benchmarking libraries such as TSLib (Wang et al., 2024) provide valuable standardization of downstream evaluation, they do not natively support controlled pre-training workflows.

To address this gap, we empirically investigate the question: *When Does Self-Supervised Learning Transfer to Time-Series Tasks?* Specifically, we formulate three guiding research questions:

- RQ1: When does SSL pre-training improve performance in time-series forecasting, classification, and anomaly detection over random frozen features (linear probing) or random initialization (fine-tuning)?

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.

- RQ2: Do certain SSL methods consistently outperform others across different downstream tasks?
- RQ3: How do pre-training data source (real vs. synthetic), data scale, and backbone depth affect representation transfer?

To answer these questions, we benchmark seven representative methods across five key SSL paradigms: reconstruction, generative modeling, contrastive learning, invariance-based distillation, and joint-embedding predictive architectures, alongside non-pre-trained controls. We evaluate models pretrained on the real-world Monash repository (Godahewa et al., 2021), synthetic datasets (Taga et al., 2025), and a mixture of the two, as well as in-domain pre-trained controls. To isolate the effect of each SSL method, we keep the backbone architecture (a standard Transformer encoder), data splits, training budgets, and evaluation protocols strictly identical across comparisons.

Our empirical results show that transfer efficacy depends on the downstream task and evaluation protocol. SSL yields substantial gains in anomaly detection and provides effective initializations for classification under fine-tuning, but offers limited to no advantage over non-pre-trained baselines in forecasting. Furthermore, linear probing does not reliably predict fine-tuning outcomes. For example, when pre-trained on Monash, LeJEPA achieves the highest mean accuracy on five out of nine classification datasets under linear probing, but on only two after fine-tuning. Downstream rankings in anomaly detection also vary substantially across evaluation metrics. In data and compute scaling, scale-matched synthetic pre-training is often competitive with real-world corpora but increasing data volume through synthetic generation does not improve performance. Finally, scaling Transformer depth consistently degrades forecasting accuracy.

We distill our empirical findings into practical guidelines for time-series SSL, offering recommendations for selecting pretext objectives, establishing rigorous baseline controls, and standardizing downstream evaluation protocols.

2 PRELIMINARIES

2.1 SELF-SUPERVISED LEARNING PARADIGMS

Self-supervised learning (SSL) aims to learn generalizable representations from unlabeled data by optimizing proxy objectives designed around intrinsic structural priors (Chen et al., 2020). In this work, we systematically evaluate five core SSL paradigms: reconstruction, generative modeling, invariance-based distillation, contrastive learning, and joint-embedding predictive architectures (JEPA), selecting representative methods for each (Table 1). Following Nie et al. (2023), we adopt a channel-independent design in which each channel (variate) of a multivariate series $\mathbf{x} \in \mathbb{R}^{T \times C}$ is separately segmented into L non-overlapping patches $\mathbf{x}_i \in \mathbb{R}^P$ ($T = L \cdot P$). A shared Transformer encoder f_θ maps the patched series to latent representations $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^{L \times D}$, which are processed by method-specific heads and serve as the backbone features for downstream adaptation. Full architectural specifications and hyperparameters appear in Appendix A and C.4.

Reconstruction-Based SSL. Reconstruction methods learn contextual representations by recovering masked portions of the input signal. We evaluate Masked Auto-Encoding (**MAE**), popularized for time series by PatchTST (Nie et al., 2023). Given a randomly sampled subset of masked patch indices $\mathcal{M} \subset \{1, \ldots, L\}$, the masked patches are set to zero and the resulting sequence $\tilde{\mathbf{x}}$ is encoded and passed to a lightweight decoder ϕ to reconstruct the masked raw signal by minimizing the Mean Squared Error (MSE):

$$\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \tag{1}$$

where $\hat{\mathbf{x}}_i = [\phi(f_\theta(\tilde{\mathbf{x}}))]_i$.

Generative SSL. Generative methods learn the data distribution by conditioning on an observed context to estimate unobserved signal values. Within this paradigm, we evaluate two representative formulations: **Next-Token (Patch) Prediction (NTP)** and **Diffusion**. NTP adopts an autoregressive framework common in time-series foundation models such as TimesFM (Das et al., 2024). Given L_c

causal context patches, a shallow residual MLP head ψ_h predicts a continuous future horizon of h patches across N sliding windows by minimizing the prediction error:

$$\frac{1}{N}\sum_{j=1}^{N}\|\psi_h(f_\theta(\mathbf{x}^{(j)}_{1:L_c}))-\mathbf{x}^{(j)}_{L_c+1:L_c+h}\|_2^2. \tag{2}$$

The Diffusion representative, adapted from TimeDART (Wang et al., 2025), learns to reverse a forward Gaussian corruption process. At timestep t_i with noise schedule $\bar{\alpha}_{t_i}$ and noise vector $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a corrupted patch $\widetilde{\mathbf{x}}_i = \sqrt{\bar{\alpha}_{t_i}}\mathbf{x}_i + \sqrt{1-\bar{\alpha}_{t_i}}\boldsymbol{\epsilon}_i$ is embedded as $\hat{\mathbf{z}}_i^{in}$, and a denoising decoder ϕ reconstructs clean patches conditioned on clean preceding context by optimizing:

$$\sum_{n=0}^{N-1}\mathbb{E}_{\boldsymbol{\epsilon},q(\mathbf{x}_n)}\big[\|\mathbf{x}_{n+1}-\phi(\hat{\mathbf{z}}^{in}_{n+1},\, f_\theta(\mathbf{z}^{in}_{1:n}))\|_2^2\big]. \tag{3}$$

Invariance-Based SSL. Rather than predicting in the raw signal space, Invariance-based methods encourage representations of different views of the same sequence to map to consistent representations in the latent space. We evaluate **DINO** (Caron et al., 2021), which trains a student network to match centered and temperature-sharpened categorical distributions of an EMA teacher across view pairs $\mathcal{P}$ via cross-entropy:

$$-\frac{1}{|\mathcal{P}|}\sum_{(u,v)\in\mathcal{P}}\sum_{k} q_k^{(u)}\log p_k^{(v)}, \tag{4}$$

where $p_k^{(v)} = [\mathrm{softmax}(a_s^{(v)}/\tau_s)]_k$ and $q_k^{(u)} = \mathrm{sg}[\mathrm{softmax}((a_t^{(u)}-c)/\tau_t)]_k$ and $a_s^{(v)}$ and $a_t^{(u)}$ are the student and teacher encoding of views v and u, respectively. The running center $\mathbf{c}$ and temperature τ_t are used to prevent mode collapse.

Contrastive Learning. Contrastive frameworks pull representations of related views together while pushing unrelated samples apart. We evaluate **SoftCLT** (Lee et al., 2024), which generalizes InfoNCE by replacing rigid negative assignments with soft weights, optimizing:

$$\frac{1}{4NT}\sum_{i=1}^{2N}\sum_{t=1}^{2T}\Big(\lambda\,\ell_I^{(i,t)}+(1-\lambda)\,\ell_T^{(i,t)}\Big), \tag{5}$$

where ℓ_I is the instance-wise loss and ℓ_T is the soft temporal loss.

Joint-Embedding Predictive Architectures (JEPA). JEPA objectives predict target representations directly in latent space. We benchmark two formulations: **JEPA** and **LeJEPA**. Following LeCun (2022) and Balestriero & LeCun (2025), our JEPA implementation uses a student network f_θ and predictor head pred_ψ to infer masked patch targets $\mathrm{sg}[f_{\theta'}(\mathbf{x})]$ generated by an exponential-moving-average (EMA) teacher $f_{\theta'}$, minimizing the alignment error alongside VICReg (Bardes et al., 2022) variance and covariance penalties to prevent collapse:

$$\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\|\,\mathrm{pred}_\psi(f_\theta(\mathbf{x}_{\setminus\mathcal{M}}))_i-\mathrm{sg}[f_{\theta'}(\mathbf{x})_i]\|_2^2+\lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}}+\lambda_{\mathrm{cov}}\mathcal{L}_{\mathrm{cov}}. \tag{6}$$

LeJEPA (Balestriero & LeCun, 2025) eliminates the EMA teacher by processing distinct augmented views through a single shared encoder f_θ, preventing collapse via Sketched Isotropic Gaussian Regularization (SIGReg). SIGReg projects the high-dimensional latent representations along random 1D directions and matches their empirical characteristic functions to a standard normal distribution using an analytic statistical test $T(\cdot)$, optimizing:

$$\frac{1-\lambda}{L}\sum_{i=1}^{L}\|\mathbf{z}_i^{(1)}-\mathbf{z}_i^{(2)}\|_2^2+\frac{\lambda}{2}\big[T(\mathbf{z}^{(1)})+T(\mathbf{z}^{(2)})\big]. \tag{7}$$

While for DINO and SoftCLT using native augmentations yields stable results, we often observe collapse with LeJEPA. In order to support a fair comparison, we introduce an alternative augmentation protocol for LeJEPA based on the Discrete Wavelet Transform (DWT). We decompose the input signal into multi-scale approximation and detail coefficients and generate reference and perturbed views. A reference view applies soft-thresholding to detail coefficients while preserving low-frequency trends, whereas a perturbed view adds Gaussian noise. This encourages invariance to high-frequency sensor noise while preserving underlying temporal morphology. Implementation details and comparisons appear in Appendices A.5 and F.1.

Table 1: Taxonomy of SSL paradigms and representative methods evaluated in this work.

Paradigm	SSL Method	Target Space	Masking / Pretext Task
Reconstruction	MAE	Signal	Random patch masking (40%)
Generative	NTP	Signal	Causal masking, h future patches
	Diffusion	Signal	Denoising Gaussian corrupted signal from clean causal context
Invariance-based	DINO	Latent	Prediction of categorical probability distribution
Contrastive	SoftCLT	Latent	Distinguish similar pairs from dissimilar ones
Joint-Embedding Predictive Architectures	JEPA	Latent	Latent prediction
	Le-JEPA	Latent	Latent prediction

2.2 Evaluating SSL Transferability

We formalize the evaluation protocol used to assess the transferability of learned representations.

Downstream Adaptation and Baseline Controls. We evaluate representations under three protocols: linear probing, probing with a Multi-Layer Perceptron (MLP), and full fine-tuning. In linear and MLP probing, the pre-trained encoder backbone remains frozen while only a downstream task head is trained, isolating the linear or non-linear accessibility of task-relevant features in the learned latent space. Under full fine-tuning, the encoder and task head are updated jointly, assessing the quality of the pre-trained weights as an optimization prior. In all cases, pre-training pretext heads are discarded prior to downstream adaptation. To isolate the empirical transfer benefit attributable strictly to self-supervised learning, every pre-trained model is paired with a matched random control, sharing an identical backbone architecture, parameter count, and downstream optimization budget:

- *Random Frozen Encoder (Probing Control):* A randomly initialized, frozen backbone paired with the same trained downstream head, measuring performance attainable solely from architecture inductive priors.
- *Supervised from Scratch (Fine-Tuning Control):* An identical backbone randomly initialized and trained end-to-end with the downstream task head under supervised optimization.

Quantifying SSL Transfer Gain. To directly measure the net performance gain of pre-training relative to these baseline controls, we define the *SSL transfer gain* Δ for a given task $\mathcal{T}$ and adaptation protocol:

$$\Delta(\mathcal{L}_{\mathrm{SSL}}, \mathcal{T}) = s_{\mathcal{T}} \left[M(\theta_{\mathrm{SSL}}, \mathcal{T}) - M(\theta_{\mathrm{baseline}}, \mathcal{T}) \right], \tag{8}$$

where $M(\theta, \mathcal{T})$ denotes the downstream evaluation metric, and θ_{SSL} and $\theta_{\mathrm{baseline}}$ represent the pre-trained model and its corresponding control ($\theta_{\mathrm{baseline}}$ represents the frozen random encoder under probing, or supervised-from-scratch training under fine-tuning). The direction scalar $s_{\mathcal{T}} \in \{+1, -1\}$ aligns metric polarity: $s_{\mathcal{T}} = +1$ for higher-is-better metrics (e.g., accuracy) and $s_{\mathcal{T}} = -1$ for lower-is-better error metrics (e.g., mean squared error). Consequently, $\Delta(\mathcal{L}_{\mathrm{SSL}}, \mathcal{T}) > 0$ strictly indicates an empirical gain from pre-training over the baseline control.

Downstream Transfer Tasks and Metrics. We benchmark transfer across three core time-series task families:

- *Forecasting:* We report Mean Squared Error (MSE) and Mean Absolute Error (MAE), averaged across four prediction horizons $H \in \{96, 192, 336, 720\}$.
- *Classification:* We report top-1 classification accuracy.
- *Anomaly Detection:* Because metric choice substantially impacts model rankings, we report three complementary criteria: point-adjusted F1 ($\mathrm{F1}_{\mathrm{Pa}}$), which credits an entire contiguous anomaly segment upon detecting any timestamp within it; event-wise F1 ($\mathrm{F1}_{\mathrm{Ev}}$), which evaluates segment-level precision and recall; and affiliation F1 ($\mathrm{F1}_{\mathrm{Af}}$), which measures directed temporal distance between predicted detections and ground-truth events (App B.2.3).

We give full details of every metric in Appendix B.2.

3 Experimental Setup

Pre-Training and Evaluation Benchmarks. Our primary experiments pre-train models on the Monash dataset (Godahewa et al., 2021), comprising approximately 680K time-series samples. To isolate the effects of data origin and scale on downstream transfer quality, we also pre-train on scale-matched synthetic data, as well as on a scaled-up corpus (2.5M samples) consisting of either purely synthetic or mixed (Monash and synthetic) data. In addition, we evaluate an in-domain pre-training control for forecasting tasks. Synthetic data generation follows TimePFN (Taga et al., 2025) and is detailed in the Appendix C.1.

We evaluate on seven forecasting datasets (ETTh1/2, ETTm1/2 (Zhou et al., 2021), Weather, Electricity, and Traffic (Wu et al., 2021)), nine UEA multivariate classification datasets (Bagnall et al., 2018), and five anomaly-detection datasets (SMD (Su et al., 2019), PSM (Abdulaal et al., 2021), MSL and SMAP (Hundman et al., 2018), and SWaT (Mathur & Tippenhauer, 2016)), following the splits and evaluation protocols of TSLib (Wang et al., 2024).

Implementation Details. Unless otherwise specified, we use a shared 8-layer Transformer with embedding dimension 128, 16 attention heads, context length 336, patch size 16, and 20 pre-training epochs (full details in Appendix C.4). Main comparisons use method-specific learning-rate and batch-size tuning with comparable tuning budgets, retaining required heads, views, and optimization components. Appendix F.2 repeats the Monash comparison with identical learning rate, batch size, epochs, and context. Results report means and standard deviations over five seeds unless specified otherwise. Extended results and implementation details are provided in the appendix.

4 Results

We organize our findings in light of the three guiding research questions defined in the introduction (section headers give their shortened form).

4.1 RQ1: When does SSL pre-training improve performance on downstream tasks?

Table 2 reports how often pre-trained SSL models outperform non-pre-trained baselines across linear probing, MLP probing, and fine-tuning. Figure 1 details the dataset-level relative transfer gains over the frozen random baseline under linear probing. Across tasks, downstream transfer varies: anomaly detection benefits substantially, classification shows selective gains on specific datasets, and forecasting exhibits minimal to negative transfer. Absolute scores and standard deviations appear in Tables 10, 13, and 14 . We summarize the main trends per task below.

Anomaly detection. Anomaly detection shows the clearest transfer gain, with SSL outperforming the random baseline in 54–94% of configurations under linear probing and up to 97% under MLP probing (Table 2). Relative gains are highest on MSL and SMAP (Figure 1a), where models such as JEPA and LeJEPA improve $\mathrm{F1}_{\mathrm{Pa}}$ by more than 250% over random features, driven by strong lifts in both $\mathrm{F1}_{\mathrm{Af}}$ and $\mathrm{F1}_{\mathrm{Ev}}$. However, these gains do not generalize to every benchmark: on SWaT, every SSL method underperforms random features on $\mathrm{F1}_{\mathrm{Pa}}$ and $\mathrm{F1}_{\mathrm{Af}}$, with only a marginal gain on event-wise detection (e.g., NTP at 2.86% $\mathrm{F1}_{\mathrm{Ev}}$).

Classification. SSL provides positive transfer on seven of nine classification benchmarks, though overall win rates remain moderate under frozen linear (46%) and MLP (40%) probes (Table 2). In Figure 1b, gains are heavily concentrated in handwriting, digit, and gesture recognition. On the Handwriting dataset, several objectives exceed a 50% relative improvement, with LeJEPA reaching +222%. Conversely, datasets such as EthanolConcentration and JapaneseVowels show negligible or negative transfer across nearly all methods. Full fine-tuning raises the SSL win rate to 67%, showing that in the majority of cases pre-trained weights offer a better starting point than random initialization once encoders can update their weights.

Table 2: Frequency of SSL outperforming non-pre-trained baselines across tasks and evaluation regimes, when pre-training on the Monash dataset. Each entry reports the count and percentage of objective–dataset pairs in which the pre-trained model beats the baseline. The baseline is a frozen random encoder for linear and MLP probing, and supervised training from scratch for fine-tuning.

	Linear probing		MLP probing		Fine-tuning	
Task	Count	%	Count	%	Count	%
Anomaly detection						
Point-adjusted F1 ($F1_{PA}$)	19/35	54	26/35	74	21/35	60
Event-wise F1 ($F1_{Ev}$)	33/35	94	34/35	97	24/35	69
Affiliation F1 ($F1_{Af}$)	19/35	54	29/35	83	24/35	69
Classification						
Accuracy	29/63	46	25/63	40	42/63	67
Forecasting						
MSE	9/49	18	11/35	31	3/35	9
MAE	8/49	16	11/35	31	2/35	6

Forecasting. Pre-training offers minimal advantage for forecasting. Under linear probing, pre-trained models beat the random frozen baseline in only 18% of MSE evaluations (Table 2), with most models yielding negative relative gains across the ETT, Weather, Electricity, and Traffic benchmarks (Figure 1c). Absolute MSE improvements are marginal, while LeJEPA severely degrades across all forecasting datasets (up to -101% relative to the baseline). Switching to an MLP probe raises the win rate to 31%. Under full fine-tuning, however, training directly from scratch outperforms pre-trained backbones in over 90% of runs, including all four ETT datasets.

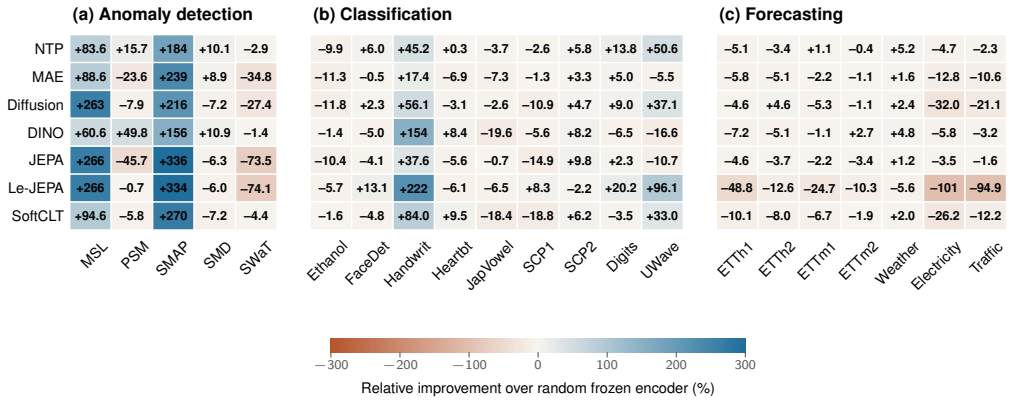


Figure 1: Relative SSL transfer gain (%) over a frozen random baseline under linear probing across downstream tasks (Monash pre-training). The relative gain quantifies the improvement compared to the respective random control, derived from Eq. 8 as $100 \times \frac{\Delta(\mathcal{L}_{\text{SSL}}, \mathcal{T})}{M(\theta_{\text{baseline}}, \mathcal{T})}$. Positive values indicate performance gains: MSE reduction for forecasting, a higher accuracy for classification, and a higher $F1_{Pa}$ scores for anomaly detection. Rounded values are computed from reported five-seed means.

4.2 RQ2: Do certain SSL methods consistently outperform others?

No single SSL method consistently outperforms the others. Instead, two main findings characterize the empirical results:

(a) *Under linear probing, performance splits by task target*: when evaluating with a linear probe, relative rankings depend on whether the downstream objective requires identifying discrete semantic states (classification and segment-level detection) or tracking continuous numerical trajectories (forecasting and point-wise residual scoring). LeJEPA achieves the highest accuracy on five of nine classification datasets (Handwriting, UWave, FaceDetection, SCP1, SpokenArabicDigits) and leads on MSL anomaly detection on both $F1_{Pa}$ and $F1_{Af}$, where

identifying a coarse regime shift suffices. Yet on forecasting, it never beats a frozen random encoder. A likely reason is that its regularizer (SIGReg) separates discrete classes well, but discards the point-wise continuity needed to extrapolate trajectories (Appendix F.4.3). Conversely, reconstruction and generative objectives perform better on continuous numerical tasks: Diffusion leads on ETTh2, and NTP leads on ETTm1 and on SMD ($F1_{Af}$). We also note that for anomaly detection, SSL performance and relative ranking vary depending on the evaluation metric (Fig. 2a).

(b) *Linear probing does not predict MLP or fine-tuning performance*: The ranking of SSL models under linear probing changes substantially with non-linear probing (MLP) or when unfreezing the encoder (Fig. 2b). In forecasting, Diffusion loses to a random encoder when using a linear head, but achieves the lowest MSE on roughly 90% of datasets when evaluated with a shallow MLP (with the remainder 10% within seed variance). The same trend is observed for fine-tuning. After fine-tuning, LeJEPA retains its classification advantage on only two datasets, as competing models close the performance gap in anomaly detection, Diffusion trails the random baseline on PSM under linear probing (43.78% vs. 47.55%), but beats every other SSL model after fine-tuning (48.30% vs. 28.29%). This suggests that while strongly regularized embeddings yield competitive frozen features, other representations serve as better starting points for fine-tuning.

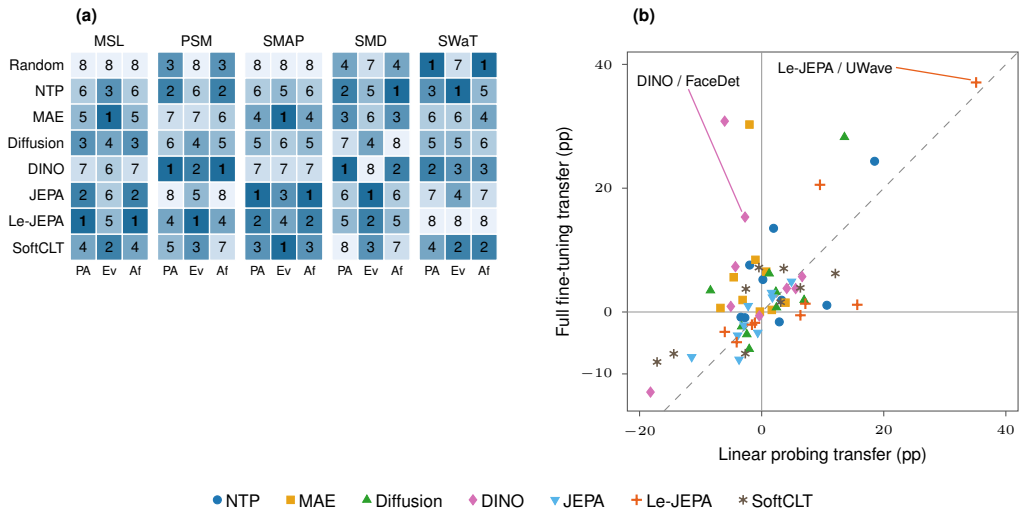


Figure 2: Downstream performance across evaluation protocols and metrics (pretraining on the Monash dataset). (a) SSL method performance rankings on Anomaly detection under linear probing, across point-adjusted (PA), event-wise (Ev), and affiliation (Af) F1. Ranks 1 to 8 include the seven SSL objectives and the random control; ties receive equal rank. (b) Classification gains in percentage points for all 63 objective–dataset pairs. Controls are a frozen random encoder for linear probing and supervised training from scratch for full fine-tuning. Solid axes mark baseline performance; the dashed diagonal indicates equal gain under both regimes. Complete scores and standard deviations are reported in Appendix Tables 13, 19, and 14.

4.3 RQ3: How Do Data Source, Scale, and Depth Affect Transfer?

We find that scaling pre-training data volume, training duration, and model depth does not yield consistent gains. Instead, transfer efficacy depends heavily on the alignment between the pre-training source and the downstream task structure.

Data source and domain relevance. Pairwise comparisons across all objective–dataset pairs show that the best pre-training source depends on the task (Table 3). For forecasting, in-domain data performs best, beating Monash and both synthetic sets. Compared to the scale-matched synthetic dataset, Monash gives a slight advantage, but adding synthetic data to Monash hurts performance. Mixing in synthetic data appears to wash out the real temporal patterns needed to forecast future points.

Table 3: Pairwise comparison of pre-training sources under linear probing. For each objective–dataset pair (49 for forecasting, 63 for classification, 35 per anomaly metric), entries report wins/losses/ties for Source A (row) against Source B (column); an entry is bolded when Source A achieves more wins than Source B. For example, **31/16/2** for Monash (680K) versus Synthetic (2.5M) on forecasting means that models pre-trained on Monash outperform those pre-trained on Synthetic (2.5M) in 31 of 49 objective–dataset comparisons, underperform in 16, and tie in 2. In-domain pre-training is evaluated for forecasting only.

		Forecasting	Classification	Anomaly detection		
Source A	Source B	MSE	Accuracy	$\mathrm{F1_{PA}}$	$\mathrm{F1_{Ev}}$	$\mathrm{F1_{Af}}$
Monash (680K)	Synthetic (680K)	**25/23/1**	26/37/0	17/17/1	**25/9/1**	16/18/1
	Synthetic (2.5M)	**30/17/2**	31/32/0	17/17/1	**24/10/1**	**19/15/1**
	Hybrid (2.5M)	**29/16/4**	29/34/0	16/16/3	**22/12/1**	17/18/0
Synthetic (680K)	Synthetic (2.5M)	**31/14/4**	**34/29/0**	14/19/2	13/22/0	**19/16/0**
	Hybrid (2.5M)	**33/15/1**	**34/29/0**	**17/16/2**	11/24/0	**18/17/0**
Synthetic (2.5M)	Hybrid (2.5M)	20/24/5	27/35/1	**18/16/1**	13/22/0	15/20/0
In-domain	Monash (680K)	**34/11/4**	–	–	–	–
	Synthetic (680K)	**35/14/0**	–	–	–	–
	Synthetic (2.5M)	**42/7/0**	–	–	–	–
	Hybrid (2.5M)	**41/6/2**	–	–	–	–

Table 4: Training duration and encoder depth (Monash pre-training, linear probing). For each objective–dataset pair, we count whether the first setting beats the second, and report wins, losses, and ties for the first setting; the larger of wins and losses is in bold. The depth comparisons are single-seed.

Comparison	Task	Metric	Pairs	Wins	Losses	Ties
40 vs. 20 epochs 40 vs. 20 epochs	Forecasting	MSE	49	13	**28**	8
	Classification	Accuracy	63	30	**33**	0
	Anomaly detection	$\mathrm{F1_{PA}}$	35	**20**	14	1
		$\mathrm{F1_{Ev}}$	35	**18**	17	0
		$\mathrm{F1_{Af}}$	35	**21**	14	0
2 vs. 8 layers	Forecasting	MSE	21	**14**	7	0
4 vs. 8 layers			21	**16**	5	0
12 vs. 8 layers			21	6	**15**	0
24 vs. 8 layers			21	3	**18**	0

For classification, this pattern flips. Scale-matched synthetic data beats Monash and large synthetic data. A likely reason is that clean synthetic curves expose basic class shapes without adding irrelevant sensor noise, making classes easier to separate. For anomaly detection, the preferred source depends on the metric. Pretraining on Monash yields better $\mathrm{F1_{Ev}}$, compared to the matched-scale synthetic dataset, and vice versa for $\mathrm{F1_{Af}}$.

Training duration and encoder depth. Extended pre-training yields negligible net benefits (Table 4, top panel). Doubling pre-training from 20 to 40 epochs does not change classification and degrades raw forecasting performance. While 40 epochs appears to favor anomaly detection, this advantage vanishes once filtering out seed-level variance (≤ 0.005 MSE or $\leq 1\%$ accuracy/F1), leaving mostly ties between the two training budgets.

Increasing the encoder depth degrades forecasting accuracy (Table 4, bottom panel). In a single-seed sweep over $N \in \{2, 4, 8, 12, 24\}$ layers, a 2-layer backbone outperforms an 8-layer encoder on forecasting, while the 24-layer variant outperforms the 8-layer model in only 3 of 21 comparisons. In classification, the 8-layer backbone similarly outperforms the 24-layer model. Under fixed training budgets, shallower networks consistently match or exceed deep architectures, indicating that parameter scaling practices from vision and NLP do not transfer directly to current time-series SSL objectives, at least under the dataset scales investigated in our work.

5 PRACTICAL TAKEAWAYS

We summarize the key findings, recommended practices and scope limitations in Table 5.

Table 5: Key empirical findings and recommended evaluation practices.

Finding (RQ)	Evaluation Practice	Implications and Scope
SSL gains are strongest in anomaly detection and selective in classification (RQ1)	Prioritize SSL pre-training for tasks requiring regime-shift or class separation rather than point trajectory extrapolation. Evaluate fine-tuning to measure full transfer potential.	Pre-trained representations reliably benefit anomaly detection and provide superior initializations for classification (reaching a 67% win rate under fine-tuning), even though linear probes on frozen features show more moderate gains.
Forecasting shows limited gains over non-pre-trained controls (RQ1)	Evaluate representations under linear probing and full fine-tuning rather than zero-shot transfer alone. Include both a frozen random encoder and supervised training from scratch controls across multiple seeds.	Zero-shot protocols do not evaluate adaptation capacity, as baselines are denied access to target supervision. Under linear probing and fine-tuning, pre-trained backbones underperform their randomly initialized controls in the majority of evaluated configurations.
Linear probing rankings do not reliably predict fine-tuning outcomes (RQ2)	Evaluate non-linear probing (MLP) and end-to-end fine-tuning alongside standard linear probes.	Linear probing can misrepresent downstream utility: some models require an MLP head or fine-tuning to show gains (e.g., Diffusion in forecasting), while others lose their relative lead once all models are fine-tuned (e.g., LeJEPA in classification).
Anomaly detection rankings vary substantially across evaluation metrics (RQ2)	Report event-wise ($F1_{Ev}$) and affiliation ($F1_{Af}$) scores alongside point-adjusted ($F1_{PA}$) metrics.	Point-adjusted F1 can mask poor event localization, as predicting a single anomalous time step marks the entire segment as correctly identified.
Target-domain proximity can outweigh corpus scale in forecasting (RQ3)	Prioritize domain-aligned data over larger, uncurated collections. Assess whether synthetic data aids or degrades transfer before expanding pre-training volume.	In-domain pre-training consistently provides the strongest forecasting performance. This may indicate that forecasting relies heavily on dataset-specific temporal patterns that transfer poorly across domains, with less universal shared structure than in language or vision. Whether substantially larger multi-domain corpora reverse this trend remains an open empirical question.
Increasing encoder depth and training duration shows diminishing returns or performance degradation (RQ3)	Benchmark deeper configurations against shallow backbones (e.g., 2–4 layers) and test whether improvements exceed seed variance (e.g., ± 0.005 MSE / $\pm 1\%$ F1) before expanding compute.	Our analysis evaluates standard Transformer encoders. Alternative model families (such as state-space models or convolutional networks) may exhibit different parameter-scaling behavior under self-supervised objectives.

6 CONCLUSION

We evaluated seven time-series SSL methods across forecasting, classification, and anomaly detection. Downstream transfer varies by task and evaluation protocol, with clear benefits for anomaly detection, selective gains for classification, and limited added value over non-pre-trained baselines in forecasting. Linear probing does not reliably predict fine-tuning performance, and no single SSL method dominates across all tasks. Furthermore, synthetic pre-training matches Monash on several benchmarks, whereas scaling Transformer depth and pre-training duration does not improve performance.

Our findings are bounded by the evaluated Transformer backbones and pre-training dataset scales. Key priorities for future work include evaluating scaling to larger multi-domain corpora, evaluating on additional architectures and developing hybrid objectives that balance semantic invariance with point-wise precision.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback to experimental data and results, and refine text writing. We have not used generative AI tools for: the ideas, the direction, paper structure and hypothesis. Additionally we used AI tools for some of the code tasks, which were reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work raises no direct ethical concerns. It involves no human subjects and no personal, private, or scraped data: every dataset is publicly available. This work studies the effects of SSL on time series tasks, and the effectiveness of the methods; we are aware of no dual-use or misuse risk specific to it. We report the compute used (App. C.5) in the interest of transparency about resource cost. Our work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have aimed to make every result reproducible. The data is publicly available, all parameters used for synthetic data generation and model hyperparameters are available in our code and in Appendix C. Unless noted (without std), all reported numbers are mean$\pm$std over the five seeds $\{123, 456, 789, 1337, 2003\}$. The code, data-generation scripts and experiment scripts are available in the supplementary material. All backbone checkpoints will be made publicly available upon paper acceptance.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *KDD*, 2021.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

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

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

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.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=jn2iTJas6h.

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 (NAACL-HLT)*, pp. 4171–4186, 2019.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=wEc1mgAjU-.

Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 635–645, 2022.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Söderström. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *KDD*, 2018.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022.

Yann LeCun. A path towards autonomous machine intelligence, version 0.9.2. OpenReview preprint, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/pdf/2312.16424.

Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised learning for time series: Contrastive or generative? *arXiv preprint arXiv:2403.09809*, 2024.

Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *CySWater*, 2016.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *KDD*, 2019.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Timepfn: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 20761–20769, 2025.

Patara Trirat, Yooju Shin, Junhyeok Kang, Youngeun Nam, Jihye Na, Minyoung Bae, Joeun Kim, Byunghyun Kim, and Jae-Gil Lee. Universal time-series representation learning: A survey. *ACM Computing Surveys*, 58(12), 2026. doi: 10.1145/3817600.

Daoyu Wang, Mingyue Cheng, Zhiding Liu, and Qi Liu. TimeDART: A diffusion autoregressive transformer for self-supervised time series representation. In *International Conference on Machine Learning (ICML)*, 2025. URL https://openreview.net/forum?id=v2G9HML7ep.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence*, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y. Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, and Shirui Pan. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6775–6794, 2024. doi: 10.1109/TPAMI.2024.3387317.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

APPENDIX

We provide extended results and discussion as part of this work, this includes detailed documentation of our experimental setup to support reproducibility:

- Section A gives the SSL objectives
- Section B gives the metrics and baselines
- Section C gives extended details of our experimental setup, encompassing:
  - Pretraining datasets details (Section C.1)
  - Synthetic data generation details (Section C.2).
  - Evaluation data details (Section C.3).
  - Additional implementation and training Details (Section C.4)
  - Compute and train time details (Section C.5)
- Section D gives the extended results to our three research questions:
  - Extended results for *Monash* pre-training under all three protocols (Section D.1)
  - Extended results for source and scale experiments (Section D.2)
- Section E provides results for extended training, enlarged depth, and convergence analysis.
  - Extended results for 40 epoch Evaluation (Section E.1)
  - Extended results for transformer depth scaling (Section E.2)
  - Convergence analysis (Section E.3)
- Section F Gives extended results for our controlled ablations:
  - Augmentations ablation study for LeJEPA (Section F.1)
  - Equal hyperparameter evaluation (Section F.2)
  - LeJEPA ablations (Section F.3)
  - Representation Analyses (Section F.4)

A SSL OBJECTIVE DEFINITIONS

We complement the objective definitions of Section 2.1 with implementation details and, where the main text states only the final loss, the full formulation. Subsections follow the paradigm order of Section 2.1, and each method is listed with its main-text equation. For clarity, the channel index is suppressed; the encoder operates independently on each channel.

A.1 RECONSTRUCTION-BASED SSL

Masked Autoencoding (MAE; Eq. 1). We mask 40% of patches uniformly at random, independently per channel and window. Masked patches are set to zero before encoding. The decoder is a single linear layer applied per patch (D=128 $\rightarrow$ P=16, preceded by dropout), and the squared error is averaged within each patch and then over masked patches only. MAE denotes masked autoencoding throughout; mean absolute error is a separate downstream forecasting metric.

A.2 GENERATIVE SSL

Next-Patch Prediction (NTP; Eq. 2). Each window of $L = L_c + h$ patches is split into $L_c = 21$ context and $h = 6$ target patches. The encoder sees only the context under causal attention, and the residual MLP head ψ_h maps the last h context representations to all h future patches in a single pass. The squared error is averaged over predicted patches, channels, and within-patch values; targets are continuous values rather than discrete tokens.

Diffusion (Eq. 3). Each patch receives an independently sampled step $t_i \in \{1, \dots, 1000\}$ under a cosine schedule. Clean patches are encoded with causal attention, and corrupted-patch embeddings query a single-layer cross-attention decoder that attends only to the preceding clean context and predicts the clean patch rather than the noise.

A.3 INVARIANCE-BASED SSL

DINO (Eq. 4). We define the remaining details for the implementation of DINO. This includes the Teacher-Student dynamics, augmentation strategies, and unique hyper-parameters.

View Defections. We define $\mathcal{G}$ to index the global views processed by the teacher, and let $\mathcal{V}$ index all views processed by the student. For teacher view $u \in \mathcal{G}$ and student view $v \in \mathcal{V}$. The set that defines them is:

$$\mathcal{P} = \{(u, v) : u \in \mathcal{G},\ v \in \mathcal{V},\ v \neq u\}. \tag{9}$$

meaning, all sets of non-equal views are proccessed.

Teacher–Student Dynamics. Teacher parameters and the output center are updated using exponential moving averages. Centering and temperature-based sharpening regulate the teacher targets during self-distillation. The teacher and projection heads are used during pre-training; downstream evaluation uses the learned encoder.

Augmentation Strategies. All reported DINO results use image DINO's multi-crop scheme, adapted to one dimension. The teacher receives $|\mathcal{G}| = 2$ global views spanning the full window: a Gaussian smoothing along the time axis with $\sigma \sim \mathcal{U}(0.5, 2.0)$, and a contrast/brightness jitter with contrast $\sim \mathcal{U}(0.6, 1.4)$, brightness offset $\sim \mathcal{U}(-0.2, 0.2)$, and per-sample jitter $\sim \mathcal{U}(0, 0.1)$, the 1-D analogue of colour jitter. The student additionally receives 6 local views, each a randomly positioned sub-window covering 40% of the sequence with additive Gaussian noise of standard deviation $\sim \mathcal{U}(0.05, 0.2)$. This gives $|\mathcal{V}| = 8$ student views and $|\mathcal{P}| = 14$ teacher–student pairs. Every view applies its transformation to all samples; only the magnitude parameter is resampled.

Hyper-parameters. The projection head h follows Caron et al. (2021): a three-layer MLP with GELU activations and hidden width 2048, a weight-normalized final layer with $K = 1024$ output dimensions, no batch normalization, and the last layer's weight norm frozen during training. Temperatures are $\tau_s = 0.1$ for the student and $\tau_t = 0.04$ for the teacher, held constant (no teacher-temperature

warm-up). The teacher-output center is updated by an exponential moving average with momentum 0.9, and the teacher weights by an EMA with base momentum 0.9995, cosine-annealed to 1.0 over training. The encoder uses a 1-epoch learning-rate warm-up. Local views crop $\lfloor 0.4 \times 336 \rfloor = 134$ steps at a random offset, of which the encoder's non-overlapping patching retains 8 whole patches (128 steps, 38% of the window), discarding the 6-step remainder.

A.4 Contrastive Learning.

SoftCLT (Eq. 5) We define the instance-wise and temporal loss terms of SoftCLT, and describe the augmentations and the hierarchical setup used for SoftCLT. We evaluate the temporal-soft variant of SoftCLT (Lee et al., 2024), combining hard instance-wise contrastive learning with soft temporal assignments.

Loss Definition. Given a batch of N series, two augmented views produce patch embeddings $\mathbf{z}_{i,t}$ and $\tilde{\mathbf{z}}_{i,t}$. Let $p_I((i,j),t)$ denote the softmax similarity between instance views at patch position t, normalized over candidate instance views other than the anchor. Let $p_T(i,(t,s))$ denote the corresponding temporal similarity within a series, normalized over candidate patch views other than the anchor. The indices $\tilde{i}$ and $\tilde{t}$ denote the matching instance and patch in the other view.

Instance-wise Loss. The instance branch uses the hard InfoNCE objective:

$$\ell_I^{(i,t)} = -\log p_I\big((i,\tilde{i}),t\big). \tag{10}$$

Our configuration uses $\tau_I = 0$ to select this hard branch. No instance-distance matrix or additional soft instance assignments are computed.

Temporal Loss. Additional temporal assignments decrease with the distance between patch positions:

$$w_T(t,s) = 2\,\sigma(-\tau_T|t-s|) = \frac{2}{1+\exp(\tau_T|t-s|)}. \tag{11}$$

Distances refer to the underlying temporal positions in the series. The temporal loss combines the matching-view positive with the additional weighted pairs:

$$\ell_T^{(i,t)} = -\log p_T\big(i,(t,\tilde{t})\big) - \sum_{s\neq t,\tilde{t}} w_T(t,s)\log p_T\big(i,(t,s)\big). \tag{12}$$

Combined Objective and Hierarchy. The complete objective combines the instance and temporal losses:

$$\mathcal{L}_{\text{SoftCLT}} = \lambda\mathcal{L}_I + (1-\lambda)\mathcal{L}_T, \tag{13}$$

where $\mathcal{L}_I$ and $\mathcal{L}_T$ aggregate the corresponding anchor losses across both views and the hierarchical representations.

We use $\tau_T = 2$ and $\lambda = 0.5$. The hierarchy applies TS2Vec-style max-pooling to patch-level representations. The initial representation contains 21 non-overlapping patches of length 16 for a 336-step context. At pooling depth d, the temporal temperature is $\tau_T\,2^d$.

Augmentation Strategies. SoftCLT uses the random-cropping scheme of TS2Vec (Yue et al., 2022) and no other input transformation. Series longer than the context window are first reduced to a random 336-step window ($P = 21$ patches). For each batch we sample an overlap length $c \sim \mathcal{U}\{2,\dots,P\}$ in patch units, then two subsequences that share exactly those c patches: offsets $\ell \sim \mathcal{U}\{0,\dots,P-c\}$, $r = \ell + c$, $e_\ell \sim \mathcal{U}\{0,\dots,\ell\}$ and $e_r \sim \mathcal{U}\{r,\dots,P\}$ give the views $[e_\ell, r)$ and $[\ell, e_r)$, each additionally shifted by a per-sample offset $\delta \sim \mathcal{U}\{-e_\ell,\dots,P-e_r\}$. The contrastive loss is computed only on the c patch representations common to both views, so the two views differ in context while being aligned on the shared region. All crop boundaries are multiples of the patch length, so both views align with the encoder's patch grid. Because the backbone is the patched transformer rather than TS2Vec's dilated convolution, the binomial timestamp masking of the original implementation is not applied.

A.5 JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

JEPA (Eq. 6). We provide the details on the masking and predictor, collapse prevention (VicReg Bardes et al. (2022)), and unique hyperparameter details.

Masking and predictor. Targets $\mathcal{M}$ are sampled by multi-block masking (2 non-overlapping blocks, each covering 25% of the patches, so 10 of 21 patches are targets), and the context $\mathcal{C}$ is the complement of $\mathcal{M}$: the generator returns a disjoint index partition, so every patch is either context or target. The student encodes only the context patches, $\mathbf{Z}_{\mathcal{C}} = f_\theta(\mathbf{x}_{\mathcal{C}})$, and a 3-layer Transformer predictor (width 64, 4 heads) receives $\mathbf{Z}_{\mathcal{C}}$ together with positional queries for the target indices, $\widehat{\mathbf{Z}}_{\mathcal{M}} = \mathrm{pred}_\psi(\mathbf{Z}_{\mathcal{C}}, \mathcal{M})$. The teacher θ' is an exponential moving average of θ with momentum 0.996 and encodes the full, unmasked window.

Collapse prevention. The variance and covariance penalties of Eq. 6 follow VICReg (Bardes et al., 2022) but are computed on the context representations only. Writing $N = B \cdot C$ for the number of (sample, channel) pairs in a batch and $\mathbf{Z}_{\mathcal{C}} \in \mathbb{R}^{N \times |\mathcal{C}| \times D}$ for the context embeddings,

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{d=1}^{D} \max\left(0,\ \gamma - \sqrt{\mathrm{Var}_{p \in \mathcal{C}}(z_{n,p,d}) + \epsilon}\right), \tag{14}$$

$$\mathcal{L}_{\mathrm{cov}} = \frac{1}{D} \sum_{d \neq d'} \mathbf{C}_{d,d'}^2, \qquad \mathbf{C} = \frac{1}{M-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}}, \tag{15}$$

where $\tilde{\mathbf{Z}} \in \mathbb{R}^{M \times D}$ collects all $M = N|\mathcal{C}|$ context embeddings, centered columnwise. The variance hinge is taken *within* a sample across its context positions, penalizing positional collapse, while the covariance penalty acts on the pooled population and decorrelates feature dimensions. Neither term constrains the teacher's target embeddings, which are stop-gradiented, so the prediction loss is the only term evaluated at target positions. The terms use $\gamma = 1$ and $\epsilon = 10^{-4}$, and the objective is weighted as $\lambda_{\mathrm{pred}} = 1$, $\lambda_{\mathrm{var}} = 1$, $\lambda_{\mathrm{cov}} = 0.04$.

Hyper-parameters. Beyond the shared 8-layer backbone, JEPA introduces four groups of settings. *Predictor:* 3 Transformer layers of width 64 with 4 heads — a third the depth and half the width of the encoder, so that representation quality is not offloaded onto the predictor. It is optimized as a separate parameter group with learning rate 5×10^{-3} and weight decay 10^{-4}, under a one-cycle schedule with 5% warm-up and gradient clipping at 1.0. *Teacher:* an exponential moving average of the student with momentum 0.996, encoding the full unmasked window.

LeJEPA (Eq. 7) We provide the details of the views and objectives, anti-collapse mechanism and implementation (SIGreg Balestriero & LeCun (2025)), and unique hyper-parameters.

Views and objective. LeJEPA uses DWT-based views described in Section 2.1 (Ablations, studies and effects in App. F.1). In Eq. 7, $\mathbf{z}^{(1)} = f_\theta(\mathbf{x}^{(1)})$ is the encoding of the reference view $\mathbf{x}^{(1)} = \mathcal{A}_1(\mathbf{x})$, which soft-thresholds the wavelet detail coefficients at 30% of each band's largest coefficient, and $\mathbf{z}^{(2)} = f_\theta(\mathbf{x}^{(2)})$ that of the perturbed view $\mathbf{x}^{(2)} = \mathcal{A}_2(\mathbf{x})$, which adds Gaussian noise of scale $s \sim \mathcal{U}(0.1, 0.3)$ to every detail coefficient; both views leave the approximation band unchanged. The first term of Eq. 7 is the squared error between the two encodings averaged over the L patch positions, and the statistic $T(\cdot)$ is applied to each view independently and averaged. We use $\lambda = 0.008$. Both branches receive gradients through the shared encoder f_θ; no stop-gradient is used.

SIGReg Implementation. Sketched Isotropic Gaussian Regularization (SIGReg) uses an Epps–Pulley-based penalty to encourage Gaussianity along random projections. The regularizer treats every patch embedding as an individual sample: a batch of B windows with C channels and L patches yields $N = BCL$ vectors $\mathbf{v}_n \in \mathbb{R}^D$. At each training step we sample $M = 1024$ unit projection directions in $\mathbb{R}^D$ by normalizing the columns of a Gaussian matrix, seeded by the global step so that all workers draw the same directions.

For a direction $\mathbf{a}$ and the N embeddings, the empirical characteristic function is

$$\hat{\phi}_{\mathbf{a}}(t) = \frac{1}{N} \sum_{n=1}^{N} \exp(\mathrm{i}t\langle \mathbf{a}, \mathbf{v}_n \rangle).$$

We compare this function with the standard normal characteristic function $\exp(-t^2/2)$. The squared residual $|\hat{\phi}_{\mathbf{a}}(t) - e^{-t^2/2}|^2$ is evaluated at 17 grid points in $[-5, 5]$, weighted by $\exp(-t^2/2)$, integrated over t by the trapezoidal rule, scaled by N, and averaged across the M directions. For numerical stability we compute SIGReg in `float32`, since mixed-precision evaluation of complex exponentials caused numerical instability in our implementation.

Hyperparameters. LeJEPA shares the 8-layer backbone with the other objectives and, unlike JEPA, has no predictor and no EMA teacher: a single encoder processes both views and both branches are trained directly. Its distinctive settings are therefore confined to the views and the regularizer. *Views:* a `db4` wavelet decomposition over $J = 3$ levels, with the reference view soft-thresholding each detail band at $\sigma = 0.3$ of that band's largest coefficient and the perturbed view adding Gaussian noise of scale $s \sim \mathcal{U}(0.1, 0.3)$ to every detail coefficient. *Regularizer:* SIGReg with $M = 1024$ random projections, evaluated at 17 quadrature points on $[-5, 5]$ in `float32`. The default regularization weight is $\lambda = 0.008$; its downstream sensitivity is evaluated in Table 41. The encoder is optimized with AdamW at learning rate 5×10^{-4} and weight decay 10^{-2}, with a 5% warm-up and gradient clipping at 1.0.

B Evaluation Baselines And Metrics

This section details how SSL transferability is evaluated: the adaptation protocols, the baseline controls, and the metrics. For all tasks, we follow the dataset splits and evaluation protocols of TSLib (Wang et al., 2024).

B.1 Protocols and Baselines

Protocols. We assess the learned representations under three adaptation protocols, which differ in the capacity of the task head and in whether the encoder is updated:

- **Linear probing.** The pre-trained backbone is frozen and a single linear head is trained for each downstream task.
- **MLP probing.** The pre-trained backbone is frozen and a shallow MLP head (one hidden layer of width 512, GELU activation, dropout 0.2) is trained for each downstream task.
- **Full fine-tuning.** The pre-trained backbone and the task head are trained jointly on the downstream task.

Each protocol isolates a different aspect of transfer. Linear probing measures whether task-relevant information is linearly accessible in the frozen latent space, so that performance reflects the structure of the representation rather than the capacity of a non-linear head to compensate for it. MLP probing measures whether that information is present but only non-linearly accessible. Full fine-tuning measures the quality of the pre-trained weights as an initialization for task-specific optimization.

Baselines. To quantify the benefit of self-supervised pre-training, every pre-trained model is compared against a non-pre-trained control with an identical Transformer backbone, parameter count, and downstream training budget. Under linear and MLP probing, the control is a randomly initialized, frozen encoder; under full fine-tuning, it is the same architecture trained from scratch on the downstream task. We regard pre-training as beneficial only when it outperforms the corresponding control, which is the condition $\Delta > 0$ in Eq. 8. These controls are strong rather than nominal: simple or untrained models have been shown to match more elaborate deep architectures on downstream time-series tasks (Zeng et al., 2023; Kim et al., 2022).

B.2 Metrics

We define the metrics reported for each task family. Forecasting is measured by error (lower is better), whereas classification and anomaly detection are measured by accuracy and F1 (higher is better); the direction scalar $s_{\mathcal{T}}$ in Eq. 8 accounts for this difference when computing the transfer gain Δ. All metrics are computed on the test split, and unless stated otherwise, reported values are means over five seeds.

B.2.1 FORECASTING

Metric Definitions And Implementation. For multivariate sequence-to-sequence prediction, where a fixed context window of length 336 is used to forecast the next H values of all C variates, with $H \in \{96, 192, 336, 720\}$. For a test set of N windows with targets $\mathbf{y}^{(n)} \in \mathbb{R}^{H\times C}$ and predictions $\hat{\mathbf{y}}^{(n)}$, we report the Mean Squared Error and the Mean Absolute Error,

$$\mathrm{MSE} = \frac{1}{NHC}\sum_{n=1}^{N}\left\|\hat{\mathbf{y}}^{(n)} - \mathbf{y}^{(n)}\right\|_2^2, \qquad \mathrm{MAE} = \frac{1}{NHC}\sum_{n=1}^{N}\left\|\hat{\mathbf{y}}^{(n)} - \mathbf{y}^{(n)}\right\|_1, \tag{16}$$

both computed on series standardized with training-set statistics. MSE is the primary metric and MAE the secondary one. Unless stated otherwise, reported values are averaged over the four horizons.

B.2.2 CLASSIFICATION

Metric Definitions And Implementation. For sequence-level classification, where each multivariate series is assigned one of K class labels. For a test set of N series with labels $y^{(n)}$ and predictions $\hat{y}^{(n)} = \arg\max_k \hat{p}_k^{(n)}$, we report top-1 accuracy,

$$\mathrm{Acc} = \frac{100}{N}\sum_{n=1}^{N}\mathbb{1}\left[\hat{y}^{(n)} = y^{(n)}\right], \tag{17}$$

expressed as a percentage. Since the number of classes differs across datasets, chance-level accuracy differs as well, and absolute accuracies are not comparable across datasets; relative gains over the matched control (Eq. 8) are used for that purpose.

B.2.3 ANOMALY DETECTION

Anomaly score. The anomaly score of timestamp t is the squared reconstruction error of a head g_ψ on the frozen encoder, $\hat{\mathbf{x}} = g_\psi(f_\theta(\mathbf{x}))$, averaged over the C variates:

$$s_t = \frac{1}{C}\sum_{c=1}^{C}\left(\hat{x}_{t,c} - x_{t,c}\right)^2, \tag{18}$$

The head is trained on the anomaly-free training split to minimize $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$.

Setup. We consider anomaly detection in multivariate time series, where timestamp t is flagged as anomalous, $\hat{y}_t = 1$, when its anomaly score s_t (Eq. 18) exceeds a threshold τ. Following the TSLib (Wang et al., 2024) protocol, τ is the $(100 - r)$-th percentile of the pooled train-and-test score distribution, where r is the benchmark's anomaly ratio (0.5% for SMD and 1.0% for all other datasets); no validation-based tuning is performed. Let $y_t \in \{0, 1\}$ denote the ground-truth label and $\mathcal{E} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$ the set of ground-truth anomaly events, i.e., the maximal contiguous runs of anomalous timestamps. Because point adjustment is known to inflate even weak detectors (Kim et al., 2022; Wu & Keogh, 2023), we report three complementary scores. All three are computed from the same binary sequence $\{\hat{y}_t\}$ and take the form $\mathrm{F1} = 2PR/(P + R)$; they differ only in how precision P and recall R are defined. Unless stated otherwise, reported values are percentages averaged over five seeds.

Metrics. Here we define the three anomaly detection metrics reported:

- **Point-adjusted F1** ($\mathrm{F1}_{\mathrm{Pa}}$): if any timestamp within an event E_k is flagged, all timestamps in E_k are counted as detected; precision, recall, and F1 are then computed point-wise over the adjusted predictions. This is the standard protocol of TSLib (Wang et al., 2024) and prior work, but is lenient toward sparse detections inside long events.
- **Event-wise F1** ($\mathrm{F1}_{\mathrm{Ev}}$): each ground-truth event counts as a single positive and each maximal run of predicted anomalies as a single detection. An event is recalled if at least one prediction falls within it; a predicted run is a true positive if it overlaps at least one event. Precision and recall are computed over events and detections rather than timestamps, so a long event contributes once regardless of its length.

- **Affiliation F1 (**$\mathrm{F1}_{\mathrm{Af}}$**)** (Huet et al., 2022): each timestamp is affiliated to its nearest ground-truth event, partitioning the time axis into affiliation zones. Within each zone, precision is the average distance from predicted anomalies to the event and recall is the average distance from event timestamps to the nearest prediction, both converted to probabilities against a uniformly random predictor over that zone and averaged across events. Unlike point adjusted and event-wise metrics, this metric is distance-based, parameter-free, and penalizes both early/late detections and false alarms far from any event.

C Experimental Setup

This section details the experimental setup: the pre-training corpora, the evaluation datasets, the implementation and hyperparameter details.

C.1 Pre-training Datasets

To isolate the effect of pre-training data source and scale on downstream tasks, we pre-train every objective on the five datasets. Dataset sizes are counted in training windows of length (336).

- *Monash (680K):* the real-world Monash corpus. It is the reference corpus for all main-text results and measures transfer from observed, multi-domain data.
- *Synthetic (680K):* a purely synthetic corpus matched to Monash in size. Comparing it with Monash isolates the effect of data source at fixed scale.
- *Synthetic (2.5M):* a larger draw from the same synthetic generator. Comparing it with Synthetic (680K) isolates the effect of scale at fixed source.
- *Hybrid (2.5M):* the full Monash corpus supplemented with $\sim$1.8M synthetic windows to match the size of Synthetic (2.5M).
- *In-domain:* the training data of the target forecasting benchmark itself. It is evaluated for forecasting only.

Monash. We use the Monash Time Series Forecasting Repository (Godahewa et al., 2021), a collection spanning diverse domains including finance, weather, and energy. The pre-training set consists of 683,982 training windows, with 77,365 windows held out for validation. It represents a realistic yet relatively small-scale pre-training scenario.

In-domain. For in-domain pre-training, each forecasting benchmark's own training and validation splits serve as the pre-training dataset, so that the encoder is pre-trained and evaluated on the same dataset. The test split is never used for pre-training.

Synthetic. To study the effect of scale in the absence of domain bias, we generate synthetic data following TimePFN (Taga et al., 2025). We create two datasets, of 680,000 and 2,512,000 training windows, respectively. These series provide diverse temporal "building blocks" such as trends and periodicities, but may lack the stochastic variability of real-world data. The generation procedure is detailed below.

Hybrid. The hybrid corpus combines the full Monash training set with $\sim$1.8M synthetic windows, for a total of 2.5M. It supports two comparisons. Against Synthetic (2.5M), which has the same size, it isolates the effect of replacing part of the synthetic data with real data. Against Monash (680K), which it contains, it isolates the effect of adding synthetic volume to a fixed set of real data.

C.2 Synthetic Data Generation

The synthetic corpus follows the structural-prior generation scheme introduced in TimePFN (Taga et al., 2025), which we instantiate using the original scripts with adjusted sequence length and dataset size. Each univariate time series is sampled as a single realization from a Gaussian Process (GP) prior on a uniform temporal grid of length T, where the covariance is constructed by randomly composing base kernels drawn (with replacement) from a fixed bank. The number of kernels per

series is sampled uniformly ($K \sim \mathcal{U}\{1, \ldots, 5\}$), and kernels are combined via random addition or multiplication operators (Bernoulli($\frac{1}{2}$)), applied in a left-folded manner.

The kernel bank includes periodic (ExpSineSquared) kernels spanning multiple temporal resolutions (e.g., daily, weekly, yearly), smooth kernels (RBF), multi-scale kernels (Rational Quadratic), linear kernels (Dot-Product), and noise components, enabling diverse temporal patterns such as trends, seasonality, and irregular fluctuations.

Multivariate series are generated using a Linear Coregionalization Model (LCM). A set of latent univariate GP factors is first sampled, and each channel is constructed as a simplex-weighted combination of these latent processes:

$$x_c(t) = \sum_{j=1}^{J} w_{c,j} f_j(t), \quad w_c \sim \text{Dirichlet}(\alpha \mathbf{1})$$

This induces structured cross-channel correlations with controllable coupling strength. The number of latent factors is sampled from a clipped Weibull distribution, while the Dirichlet concentration parameter controls the sparsity of channel assignments, interpolating between near-independent channels and dense entangled mixtures.

This construction yields synthetic data with controlled temporal structure and inter-channel dependencies, enabling large-scale pre-training while preserving key statistical properties such as multi-scale periodicity and smoothness. While we do not explicitly match higher-order statistics, the generated series exhibit diverse spectral and autocorrelation patterns consistent with real-world signals. However, the simplex-based mixing induces predominantly non-negative correlations, which differs from real benchmarks and represents a known limitation of the generator. We refer to Taga et al. (2025) for the full generation algorithm and parameterization.

We follow the default TimePFN generation setup with minor adjustments to sequence length and dataset size. Specifically, each dataset contains $N = 4000$ series of length $T = 2500$. For univariate generation, the number of kernels per series is sampled as $K \sim \mathcal{U}\{1, \ldots, 5\}$. The kernel bank consists of 33 atoms, including periodic (ExpSineSquared) kernels at multiple temporal resolutions, RBF and Rational Quadratic kernels with varying length scales, Dot-Product kernels, noise components, and a Constant kernel.

For multivariate generation, we use $C = 160$ channels constructed via a Linear Coregionalization Model with J latent factors sampled from a clipped Weibull distribution. Each latent factor is defined by its own kernel composition, where the number of kernels is sampled independently as $K \sim \mathcal{U}\{1, \ldots, 5\}$. Channel mixing weights are drawn from a Dirichlet distribution with concentration parameter $\alpha \sim \mathcal{U}[0.1, 1.0]$, sampled per series. We use the original TimePFN implementation with these parameter settings, modifying only sequence length and dataset size.

C.3 Evaluation datasets

Forecasting. We evaluate on seven forecasting datasets ETTh1/2, ETTm1/2 (Zhou et al., 2021), Weather, Electricity, and Traffic (Wu et al., 2021). Details of size, split and channels are given in the Table 6

Classification. We evaluate classification on nine datasets from the UEA multivariate archive (Bagnall et al., 2018), spanning spectroscopy, neural recordings, speech, and motion sensing (Table 7). The datasets differ widely in size, from 120 to 6,599 training series, and in the number of classes, from 2 to 26.

Anomaly Detection. We evaluate anomaly detection on five multivariate benchmarks (Table 8): two server-monitoring datasets, SMD (Su et al., 2019) and PSM (Abdulaal et al., 2021); two spacecraft-telemetry datasets, MSL and SMAP (Hundman et al., 2018); and the SWaT water-treatment testbed (Mathur & Tippenhauer, 2016). The datasets differ substantially in size and in the fraction of anomalous test timestamps, from about 4% on SMD to about 28% on PSM.

Table 6: Forecasting benchmarks. Splits are chronological and follow TSLib: 12/4/4 months for the ETT datasets and 70%/10%/20% for the others. Split sizes are given in timesteps; the number of evaluation windows additionally depends on the context length and the horizon.

Dataset	Variates	Frequency	Timesteps	Train	Validation	Test
ETTh1	7	1 hour	17,420	8,640	2,880	2,880
ETTh2	7	1 hour	17,420	8,640	2,880	2,880
ETTm1	7	15 min	69,680	34,560	11,520	11,520
ETTm2	7	15 min	69,680	34,560	11,520	11,520
Weather	21	10 min	52,696	36,887	5,270	10,539
Electricity	321	1 hour	26,304	18,412	2,632	5,260
Traffic	862	1 hour	17,544	12,280	1,756	3,508

Table 7: Classification benchmarks from the UEA multivariate archive (Bagnall et al., 2018), with the archive's predefined train/test splits as used in TSLib. Series length is the maximum length for the two variable-length datasets (JapaneseVowels and SpokenArabicDigits).

Dataset	Domain	Variates	Length	Classes	Train	Test
EthanolConcentration	Spectroscopy	3	1,751	4	261	263
FaceDetection	MEG	144	62	2	5,890	3,524
Handwriting	Accelerometer	3	152	26	150	850
Heartbeat	Heart sound	61	405	2	204	205
JapaneseVowels	Speech	12	29	9	270	370
SelfRegulationSCP1	EEG	6	896	2	268	293
SelfRegulationSCP2	EEG	7	1,152	2	200	180
SpokenArabicDigits	Speech	13	93	10	6,599	2,199
UWaveGestureLibrary	Accelerometer	3	315	8	120	320

Table 8: Anomaly-detection benchmarks, with the predefined train/test splits as used in TSLib. Training data are anomaly-free and unlabeled; labels are available for the test split only. Sizes are given in timesteps, and the anomaly ratio is the fraction of anomalous timestamps in the test split.

Dataset	Domain	Variates	Train	Test	Anomaly ratio (%)
SMD (Su et al., 2019)	Server monitoring	38	708,405	708,420	4.2
PSM (Abdulaal et al., 2021)	Server monitoring	25	132,481	87,841	27.8
MSL (Hundman et al., 2018)	Spacecraft telemetry	55	58,317	73,729	10.5
SMAP (Hundman et al., 2018)	Spacecraft telemetry	25	135,183	427,617	12.8
SWaT (Mathur & Tippenhauer, 2016)	Water treatment	51	495,000	449,919	12.1

C.4 Implementation And Training Details

All methods share the Transformer encoder backbone of Section 2.1. The default configuration has 8 layers, embedding dimension $D = 128$, 16 attention heads, feed-forward dimension 512, and patch length $P = 16$. The input context is 336 timesteps ($L = 21$ patches) for pre-training, forecasting, and anomaly detection, and 1,152 timesteps ($L = 72$ patches) for classification. Method-specific components are added only where the SSL objective requires them: a linear reconstruction decoder for MAE, a residual MLP prediction head for NTP, a cross-attention denoising decoder for Diffusion, a projection head and an EMA teacher for DINO, and a latent predictor and an EMA teacher for JEPA; LeJEPA and SoftCLT use the encoder alone. All such components are discarded after pre-training, so downstream evaluation uses the encoder only.

Pre-Training Details All objectives are pre-trained for 20 epochs on windows of 336 timesteps with patch length $P = 16$. Each objective constructs its training signal as described in Appendix F.2: masking for MAE and JEPA, causal prediction for NTP, noise corruption for Diffusion, and augmented views for DINO, SoftCLT, and LeJEPA. Learning rate and batch size are tuned per objective under comparable tuning budgets, and all objective-specific hyperparameters are listed in Table 9; the matched-budget comparison of Appendix F.2 instead fixes them across objectives. We saw our experiments, JEPA collapsed to a constant representation under both Adam and AdamW, whereas SGD trained stably. For every run, we keep the checkpoint with the lowest validation loss, and all pre-training runs are repeated over five seeds.

Downstream Training Details For classification and anomaly detection, downstream hyperparameters are shared across all SSL objectives and the non-pre-trained controls, so that differences in the result tables reflect the encoder weights rather than the adaptation budget. Classification uses a learning rate of $1\mathrm{e}{-3}$ with a batch size of 64 for 20 epochs. Anomaly detection uses Adam with a learning rate of $1\mathrm{e}{-3}$ and cosine annealing for 10 epochs, with early stopping at a patience of 3. Forecasting uses dataset-specific batch sizes (16–256) and learning rates in $[1\mathrm{e}{-4}, 4\mathrm{e}{-4}]$, and is trained for 20 epochs.

C.5 Compute and Train-time Details

All experiments were conducted on NVIDIA RTX PRO 6000 Blackwell GPUs (96GB VRAM, 600W TDP), with each backbone trained on a single GPU. Peak memory usage during pre-training ranged between approximately 8–32GB depending on model depth and dataset. On the Monash collection, pre-training runtime scaled with backbone depth: 2–4 layer models required approximately 2–3 hours, 8-layer models 4–5 hours, 12-layer models 6–7 hours, and 24-layer models around 10 hours per run. Synthetic and mixed-data variants approximately doubled these runtimes, with the deepest 24-layer configurations requiring up to 28 hours.

Downstream evaluation is substantially lighter than pre-training. Classification completes within approximately 20 minutes per dataset, and anomaly detection within 15 minutes per dataset, depending on model depth. Forecasting on smaller benchmarks (ETTm1/2, ETTh1/2, Weather) typically completes within 15 minutes per dataset, while larger datasets scale more significantly: Electricity requires up to 4 hours and Traffic up to 12 hours for the deepest configurations. Memory usage during downstream tasks remains below 20GB for most datasets, increasing to approximately 40GB for Electricity, while certain Traffic configurations may exceed available VRAM (out-of-memory).

D Extended Results

This section reports the complete results behind the main-text findings, together with a brief analysis of each table. It is organized around our three research questions, and for each one we present the extended results that support the answers given in the main text.

Table 9: Objective-specific pre-training hyperparameters. The backbone is identical across objectives (8 layers, $d_{\text{model}} = 128$, 16 heads, $d_{ff} = 512$, patch length 16, 336-step context; 1152 for the classification protocol) and all objectives pre-train for 20 epochs, so only the settings that differ are listed. Learning rates and batch sizes are each objective's own tuned values; under the equal-budget protocol both are standardized.

Objective	Optimizer	LR	Batch	Wd	Specific settings
NTP	Adam	1×10^{-4}	64	–	causal, 21+6 patches
MAE	Adam	1×10^{-4}	64	–	mask 0.4, linear decoder
Diffusion	Adam	1×10^{-4}	128	–	T=1000, cosine
DINO	AdamW	5×10^{-4}	128	$0.04 \rightarrow 0.1$	K=1024, EMA 0.9995
JEPA	SGD	3×10^{-3}	128	3×10^{-3}	pred. LR 5×10^{-3}, EMA 0.996
LeJEPA	AdamW	5×10^{-4}	128	1×10^{-2}	λ=0.008, 1024 proj.
SoftCLT	AdamW	2×10^{-4}	128	1×10^{-2}	τ_{temp}=2.0, λ=0.5

D.1 Main Results under Monash Pre-Training

This subsection gives the full scores underlying Table 2 and Figures 1 and 2. For each protocol, we report forecasting, classification, and anomaly detection in turn, and compare every pre-trained encoder against the other methods and baseline.

D.1.1 Linear probing

We report results for backbones pre-trained on *Monash* and evaluated with a linear head on the frozen encoder.

Forecasting. Table 10 shows that the transfer gain for forecasting is small. Under Monash, the random frozen encoder attains lower mean MSE in 40 of 49 objective–dataset comparisons and the lowest MSE on *ETTh1*, *Electricity*, and *Traffic*; the largest improvement is $\Delta = +0.016$ MSE (Diffusion on *ETTh2*). LeJEPA degrades on every benchmark ($\Delta = -0.413$ on *Traffic*), and SoftCLT trails the baseline on every dataset except *Weather*. Errors grow with the horizon for every objective, with a single exception (Diffusion on *ETTh2*), and no objective closes the gap at longer horizons (Tables 11 and 12).

Table 10: Linear-probing forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **Monash** dataset. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; lower is better and the best result per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	**0.379±0.001**	0.261±0.001	0.318±0.000	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	**0.435±0.000**	**0.298±0.000**
Signal-space predictive and reconstructive														
NTP	0.435±0.002	0.441±0.001	0.362±0.003	0.400±0.002	0.352±0.001	0.381±0.000	0.262±0.001	0.319±0.001	0.235±0.000	0.270±0.001	0.180±0.001	0.277±0.001	0.445±0.001	0.311±0.001
MAE	0.438±0.002	0.438±0.002	0.368±0.004	0.402±0.003	0.364±0.002	0.390±0.001	0.264±0.004	0.324±0.003	0.244±0.001	0.279±0.001	0.194±0.001	0.295±0.001	0.481±0.003	0.345±0.003
Diffusion	0.433±0.001	0.446±0.001	**0.334±0.001**	**0.385±0.001**	0.375±0.002	0.395±0.001	0.264±0.001	0.321±0.001	0.242±0.001	0.279±0.002	0.227±0.004	0.343±0.004	0.527±0.005	0.384±0.004
Latent-space alignment														
DINO	0.444±0.008	0.449±0.006	0.368±0.005	0.404±0.004	0.360±0.001	0.387±0.000	**0.254±0.002**	**0.311±0.001**	0.236±0.001	0.271±0.001	0.182±0.005	0.281±0.006	0.449±0.013	0.319±0.012
JEPA	0.433±0.003	0.438±0.002	0.363±0.005	0.402±0.004	0.364±0.001	0.384±0.002	0.270±0.003	0.325±0.002	0.245±0.002	0.278±0.002	0.178±0.000	0.275±0.001	0.442±0.000	0.311±0.001
LeJEPA	0.616±0.009	0.540±0.004	0.394±0.003	0.425±0.003	0.444±0.005	0.440±0.003	0.288±0.004	0.345±0.003	0.262±0.001	0.297±0.001	0.346±0.007	0.432±0.005	0.848±0.013	0.550±0.005
Contrastive														
SoftCLT	0.456±0.003	0.460±0.002	0.378±0.002	0.413±0.001	0.380±0.001	0.404±0.001	0.266±0.002	0.327±0.002	0.243±0.001	0.278±0.001	0.217±0.003	0.319±0.003	0.488±0.005	0.357±0.004

Forecasting Horizon-wise analysis. Tables 11 and 12 report the forecasting MSE at each prediction horizon. Error increases monotonically with the horizon for every objective and dataset, with a single exception (Diffusion on *ETTh2*). The leading model is stable across horizons but varies across datasets: the random frozen encoder is lowest at every horizon on *ETTh1*, *Electricity*, and *Traffic* (on *Electricity*, NTP and JEPA stay within 0.006–0.007 of it), NTP on *ETTm1*, DINO on *ETTm2*, NTP (tied with DINO at two horizons) on *Weather*, and Diffusion on *ETTh2* beyond the 96-step horizon. Consistent with our main findings, the gains from SSL do not grow with the horizon: NTP's margin over the random encoder shrinks from 0.016 to 0.008 on *Weather* and from 0.006 to 0.001 on *ETTm1* between the shortest and the longest horizon, so pre-training does not help more where forecasting is harder.

Table 11: Linear probing MSE across prediction horizons using an 8-layer backbone pre-trained on the Monash dataset. Lower is better; best per dataset and horizon in **bold**.

SSL Paradigm	ETTh1				ETTh2				ETTm1			
	96	192	336	720	96	192	336	720	96	192	336	720
Random frozen encoder	**0.373±.004**	**0.411±.004**	**0.429±.006**	**0.444±.008**	**0.281±.002**	0.345±.004	0.375±.006	0.401±.003	0.296±.001	0.333±.001	0.370±.003	0.424±.001
Signal-space predictive and reconstructive												
NTP	0.391±.002	0.428±.001	0.448±.001	0.475±.008	0.296±.004	0.361±.005	0.376±.003	0.416±.004	**0.290±.001**	**0.328±.001**	**0.365±.001**	**0.423±.002**
MAE	0.391±.002	0.426±.004	0.446±.003	0.488±.007	0.302±.006	0.370±.007	0.389±.008	0.412±.005	0.308±.002	0.343±.002	0.380±.004	0.427±.002
Diffusion	0.397±.001	0.432±.001	0.439±.002	0.463±.004	0.286±.001	**0.342±.001**	**0.322±.001**	**0.385±.002**	0.319±.003	0.356±.001	0.388±.003	0.438±.004
Latent-space alignment												
DINO	0.386±.001	0.423±.002	0.455±.019	0.511±.023	0.287±.004	0.367±.006	0.392±.009	0.427±.006	0.309±.001	0.337±.001	0.370±.001	0.425±.001
JEPA	0.388±.002	0.422±.002	0.443±.001	0.480±.006	0.298±.003	0.360±.004	0.380±.007	0.412±.006	0.305±.002	0.342±.001	0.376±.001	0.431±.001
LeJEPA	0.589±.011	0.612±.009	0.627±.008	0.638±.008	0.338±.005	0.391±.004	0.400±.003	0.448±.006	0.396±.005	0.423±.006	0.453±.005	0.504±.006
Contrastive												
SoftCLT	0.420±.002	0.449±.002	0.462±.002	0.494±.010	0.310±.003	0.371±.004	0.388±.003	0.442±.004	0.325±.001	0.359±.001	0.392±.002	0.444±.002

Table 12: Linear probing MSE across prediction horizons using an 8-layer backbone pre-trained on the Monash dataset. Lower is better; best per dataset and horizon in **bold**. MAE's Traffic column is averaged over three seeds; all other cells use five.

SSL Paradigm	ETTm2				Weather				Electricity				Traffic			
	96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720
Random frozen encoder	0.168±.001	0.226±.001	0.278±.001	0.372±.003	0.175±.001	0.218±.001	0.265±.000	0.333±.000	**0.144±.000**	**0.158±.000**	**0.174±.000**	**0.213±.000**	**0.412±.000**	**0.425±.000**	**0.438±.000**	**0.465±.000**
Signal-space predictive and reconstructive																
NTP	0.168±.001	0.228±.002	0.281±.002	0.370±.002	**0.159±.001**	**0.204±.001**	**0.253±.001**	**0.325±.000**	0.151±.001	0.166±.001	0.182±.001	0.220±.001	0.420±.001	0.436±.001	0.449±.001	0.474±.001
MAE	0.174±.004	0.230±.004	0.283±.006	0.370±.004	0.170±.002	0.213±.002	0.261±.001	0.331±.001	0.168±.001	0.180±.001	0.195±.001	0.232±.001	0.457±.003	0.471±.004	0.482±.004	0.510±.002
Diffusion	0.173±.001	0.229±.001	0.282±.003	0.370±.003	0.169±.001	0.212±.002	0.260±.002	0.326±.001	0.200±.003	0.210±.004	0.230±.004	0.269±.004	0.505±.005	0.515±.005	0.530±.005	0.558±.005
Latent-space alignment																
DINO	**0.164±.001**	**0.218±.001**	**0.270±.002**	**0.362±.004**	0.161±.002	**0.204±.002**	0.254±.001	**0.325±.001**	0.153±.005	0.168±.005	0.184±.005	0.223±.005	0.427±.013	0.440±.014	0.452±.014	0.477±.013
JEPA	0.179±.003	0.234±.003	0.287±.003	0.378±.003	0.172±.003	0.215±.003	0.263±.002	0.331±.002	0.150±.000	0.164±.000	0.180±.001	0.220±.001	0.418±.000	0.431±.000	0.445±.000	0.475±.001
LeJEPA	0.203±.003	0.254±.004	0.302±.005	0.397±.009	0.196±.001	0.234±.001	0.277±.001	0.341±.001	0.319±.007	0.331±.007	0.348±.007	0.387±.008	0.815±.013	0.834±.015	0.854±.015	0.883±.015
Contrastive																
SoftCLT	0.181±.002	0.233±.002	0.282±.002	0.370±.003	0.170±.001	0.212±.001	0.260±.001	0.330±.001	0.189±.003	0.203±.003	0.220±.003	0.259±.002	0.463±.005	0.477±.005	0.491±.005	0.520±.005

Classification. Table 13 reports the classification results. Three of the seven objectives outperform the random frozen encoder on a majority of datasets: NTP on 6 of 9, and Diffusion and LeJEPA on 5 of 9 each. The first two never attain the best accuracy on any dataset, although their gains can be large (NTP: $\Delta = +18.5$ on UWave and $+10.7$ on SpokenArabicDigits). LeJEPA, in contrast, attains the best accuracy on every dataset where it beats the baseline (*FaceDetection*, *Handwriting*, *SCP1*, *SpokenArabicDigits*, and *UWave*), including the largest gain in the table ($\Delta = +35.1$ on *UWave*, 71.69 vs. 36.56). These consistent gains for three of the seven objectives suggest that SSL pre-training is of value for classification. The remaining objectives are weaker: DINO and JEPA beat the baseline on only 3 of 9 datasets and SoftCLT on 4 of 9, with losses of up to $\Delta = -18.2$ (DINO on *JapaneseVowels*) and -14.4 (SoftCLT on *SCP1*). Overall, pre-training yields a benefit for classification: three of the seven objectives outperform the baseline on a majority of datasets, which suggests that gains are attainable with a suitable choice of objective.

Table 13: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash** dataset. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	**27.68±1.88**	54.49±0.54	4.31±0.54	66.63±1.33	**93.02±0.65**	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	24.94±0.19	57.76±0.60	6.26±2.08	66.83±3.16	89.62±0.99	74.75±1.22	52.89±3.57	88.29±0.40	55.06±1.83
MAE	24.56±0.46	54.20±0.91	5.06±1.30	62.05±1.39	86.27±2.59	75.70±1.97	51.66±2.65	81.51±2.38	34.56±2.93
Diffusion	24.41±1.12	55.72±0.49	6.73±1.15	64.59±3.37	90.59±0.40	68.33±4.37	52.33±0.65	84.56±1.44	50.12±3.56
Latent-space alignment									
DINO	27.30±0.94	51.77±0.44	10.94±0.57	72.20±0.53	74.81±0.76	72.42±1.65	54.11±2.26	72.56±2.45	30.50±1.60
JEPA	24.79±0.45	52.28±0.32	5.93±1.94	62.93±1.69	92.38±0.69	65.26±2.83	**54.89±2.73**	79.37±1.34	32.63±2.77
LeJEPA	26.09±1.24	**61.65±0.78**	**13.86±0.72**	62.54±1.73	86.98±1.09	**83.07±1.96**	48.89±1.65	**93.27±1.50**	**71.69±4.92**
Contrastive									
SoftCLT	27.23±1.09	51.90±0.58	7.93±0.90	**72.97±0.24**	75.89±1.58	62.32±2.11	53.11±1.03	74.93±3.96	48.63±6.92

Anomaly detection. Table 14 reports the anomaly-detection results. Pre-trained encoders outperform the random frozen encoder in 19 of 35 comparisons under $\mathrm{F1_{PA}}$, 33 of 35 under $\mathrm{F1_{Ev}}$, and 19 of 35 under $\mathrm{F1_{Af}}$. Across the 15 dataset–metric cells, NTP, DINO, and MAE win most often (13, 12, and 11 cells, respectively), and even the weakest objective by this count, LeJEPA, wins on 8 of 15. The largest gains occur on the spacecraft-telemetry datasets: on *MSL*, LeJEPA, JEPA, and Diffusion raise $\mathrm{F1_{PA}}$ from 15.57 to about 57 ($\Delta = +41$), and on *SMAP*, JEPA raises it from 7.98 to 34.83 ($\Delta = +26.9$); on *PSM*, DINO gains $\Delta = +23.7$ under $\mathrm{F1_{PA}}$ (71.23 vs. 47.55), and LeJEPA and DINO gain $\Delta = +25.5$ under $\mathrm{F1_{Ev}}$. The largest losses are those of JEPA on *PSM* ($\Delta = -21.7$ $\mathrm{F1_{PA}}$, -19.6 $\mathrm{F1_{Af}}$), of MAE on *PSM* ($\Delta = -11.2$ $\mathrm{F1_{PA}}$), and of JEPA and LeJEPA on *SWaT*, where both fall to about 3 $\mathrm{F1_{PA}}$ against 11.84 for the random encoder. Overall, pre-training yields a clear benefit for anomaly detection: every objective outperforms the random frozen encoder on a majority of dataset–metric cells, with *SWaT* as the only dataset where pre-training does not help.

Table 14: Linear-probing anomaly-detection F1 (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash** dataset. For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	28.58±0.34	13.11±0.09	29.45±0.19	55.02±9.27	29.64±4.71	49.55±2.94	22.68±0.61	6.19±0.10	18.39±0.42	61.75±0.34	20.00±0.43	**64.29±0.37**	11.50±0.01	**2.86±0.36**	9.54±0.16
MAE	29.36±0.16	**13.29±0.07**	29.96±0.20	36.31±5.19	19.13±3.30	41.34±2.72	27.05±0.85	**7.17±0.11**	21.45±0.64	61.09±1.27	19.05±0.76	63.34±1.25	7.72±4.80	1.28±0.53	10.87±2.39
Diffusion	56.53±0.29	12.26±0.07	48.34±0.26	43.78±1.52	34.64±4.27	41.58±1.73	25.23±3.65	6.01±0.49	20.06±2.50	52.03±0.26	30.24±0.35	55.57±0.12	8.59±2.15	1.39±0.69	9.39±1.99
Latent-space alignment															
DINO	25.01±1.13	12.16±0.30	27.18±0.68	**71.23±2.94**	41.91±3.03	**54.80±2.06**	20.43±2.62	5.63±0.70	16.70±1.83	**62.21±0.65**	17.50±0.65	64.01±0.90	11.68±0.10	2.01±0.36	11.50±2.07
JEPA	56.92±0.06	12.16±0.02	49.15±0.06	25.83±5.30	30.13±3.41	26.48±3.22	**34.83±0.41**	6.58±0.09	**27.31±0.58**	52.56±0.16	**31.15±0.12**	55.99±0.09	3.14±0.09	1.71±0.19	4.10±0.07
LeJEPA	**56.98±0.06**	12.19±0.04	**49.26±0.05**	47.20±0.16	**41.92±0.84**	42.67±0.27	34.63±0.83	6.46±0.16	27.24±0.52	52.69±0.04	31.01±0.06	56.06±0.08	3.07±0.08	0.97±0.12	4.09±0.05
Contrastive															
SoftCLT	30.30±0.05	13.18±0.02	30.77±0.05	44.77±2.79	38.39±1.06	39.82±2.12	29.54±0.34	**7.17±0.07**	23.42±0.27	52.02±0.20	30.43±0.16	55.84±0.07	11.32±0.35	2.10±1.09	12.97±0.71

D.1.2 MLP

We report results for backbones pre-trained on *Monash* and evaluated with an MLP head on the frozen encoder.

Forecasting. Table 15 reports the forecasting results. Diffusion is the only objective with consistent gains: it outperforms the random frozen encoder on all five datasets under both MSE and MAE, and it attains the lowest MSE on four of five datasets and the lowest MAE on all five (tied with DINO on *ETTm2*). Its only loss in MSE is to DINO on *ETTm2*, by 0.001 (0.261 ± 0.001 vs. 0.260 ± 0.002), which is within seed variance. Since Diffusion beats the random encoder on only two of these five datasets under linear probing, this suggests that its representations encode forecasting-relevant information that is not linearly accessible and requires a non-linear head. The remaining objectives do not show consistent gains: four of seven improve on *Weather*, and DINO leads on *ETTm2*, but overall the pre-trained encoders outperform the control in only 11 of 35 comparisons, five of which are due to Diffusion.

Classification. Table 16 reports the classification results. The number of wins over the control drops relative to linear probing, from 29 to 25 of 63 objective–dataset pairs. This drop reflects a stronger control rather than weaker pre-trained encoders: with the MLP head, the random frozen encoder improves on eight of nine datasets, most strongly on *Heartbeat* (from 66.63 to 73.37). On that dataset, NTP, DINO, and SoftCLT exceed the control under linear probing but fall just below it with the MLP head (73.05, 72.56, and 73.29 vs. 73.37), and four of the eight lost wins are by less than 0.35 points, within seed variance. Though we do see that LeJEPA attains the best accuracy on four datasets, with gains of $\Delta = +41.2$ on *UWave* (82.97 vs. 41.81), $+15.2$ on *SpokenArabicDigits*, and $+14.7$ on *Handwriting*. This suggests that a non-linear head recovers from random features part of the class information that pre-training makes linearly accessible, which narrows the margin on the datasets where the advantage of pre-training was small.

Anomaly detection. Table 17 reports the anomaly-detection results. The MLP head supports better results for the pre-trained encoders: the wins under $\mathrm{F1_{PA}}$, $\mathrm{F1_{Ev}}$, and $\mathrm{F1_{Af}}$ rise to 26, 34, and 29 of 35, from 19, 33, and 19 under linear probing. The wins rise for six of the seven objectives and are

Table 15: Forecasting performance (MSE and MAE) using an MLP probe, reported as mean ± std. Datasets: **H1/H2** (ETTH1/2), **M1/M2** (ETTM1/2), **Wth** (Weather). Models are grouped by SSL paradigm; top results (lowest error) per metric and dataset are highlighted in **bold**.

SSL Paradigm	Loss	H1	H2	M1	M2	Wth
Random frozen encoder	MSE	0.425±.002	0.362±.002	0.354±.003	0.264±.002	0.233±.001
	MAE	0.434±.001	0.399±.001	0.385±.002	0.321±.001	0.270±.000
Signal-space predictive and reconstructive						
NTP	MSE	0.436±.004	0.369±.001	0.358±.002	0.264±.001	0.230±.000
	MAE	0.444±.002	0.402±.001	0.387±.001	0.322±.001	0.266±.000
MAE	MSE	0.444±.003	0.379±.001	0.352±.002	0.263±.026	0.232±.000
	MAE	0.444±.002	0.413±.001	0.384±.001	0.323±.018	0.266±.000
Diffusion	MSE	**0.420±.002**	**0.344±.001**	**0.351±.002**	0.261±.001	**0.228±.000**
	MAE	**0.432±.002**	**0.388±.000**	**0.382±.001**	0.318±.001	**0.264±.001**
Latent-space alignment						
DINO	MSE	0.447±.002	0.371±.003	0.361±.001	**0.260±.002**	0.229±.000
	MAE	0.455±.002	0.406±.002	0.390±.001	**0.318±.001**	0.265±.000
JEPA	MSE	0.436±.004	0.375±.005	0.362±.003	0.271±.002	0.233±.002
	MAE	0.442±.003	0.410±.002	0.390±.002	0.326±.001	0.268±.002
LeJEPA	MSE	0.583±.003	0.417±.003	0.395±.002	0.289±.002	0.246±.001
	MAE	0.529±.001	0.439±.002	0.414±.002	0.344±.001	0.283±.000
Contrastive						
SoftCLT	MSE	0.454±.008	0.374±.005	0.380±.005	0.282±.010	0.239±.002
	MAE	0.457±.007	0.411±.004	0.403±.005	0.336±.009	0.274±.003

Table 16: Classification accuracy (%) using an MLP probe on a frozen Monash-pretrained backbone. Top result per dataset is **bold**.

SSL Paradigm	Eth	Face	Hand	Hrt	Vwl	S1	S2	Dig	UW
Random frozen encoder	27.84±0.90	52.93±0.30	6.54±1.50	**73.37±0.20**	94.97±0.20	**78.23±1.00**	51.67±1.20	80.18±1.00	41.81±1.30
Signal-space predictive and reconstructive									
NTP	29.47±0.60	52.64±0.50	10.14±0.50	73.05±2.20	93.58±0.50	73.47±1.10	49.44±1.00	86.71±0.40	63.51±2.50
MAE	27.47±1.80	53.94±1.20	6.77±0.40	69.76±2.10	**95.95±0.80**	74.66±4.70	49.59±2.30	92.72±0.80	47.74±2.40
Diffusion	25.10±0.00	52.30±1.60	8.29±0.80	64.27±2.30	92.84±0.30	71.50±1.00	**55.69±2.20**	89.27±1.50	55.00±2.30
Latent-space alignment									
DINO	25.29±0.30	50.00±0.00	5.59±0.60	72.56±0.20	77.71±1.40	72.18±2.30	55.55±1.00	62.27±1.70	33.05±6.10
JEPA	25.10±0.30	51.84±0.40	7.30±1.80	64.15±2.70	93.51±0.50	70.14±2.20	54.72±1.10	83.13±1.10	46.25±1.40
LeJEPA	25.86±2.00	**62.15±0.80**	**21.24±0.40**	63.66±3.00	90.07±1.20	78.07±3.90	50.70±1.00	**95.37±0.60**	**82.97±0.80**
Contrastive									
SoftCLT	**29.85±2.10**	50.31±0.30	11.21±1.90	73.29±0.90	81.22±1.90	68.26±2.10	51.67±1.20	65.59±4.40	45.31±8.70

unchanged for NTP (13 of 15). For the random encoder the picture is mixed, as it gains on 8 of the 15 dataset–metric cells and drops on 7. This suggests that the MLP head is appropriate for this task, strengthening the pre-trained encoders relative to the baseline.

Table 17: Anomaly detection F1 (mean ± std) using an MLP probe on backbones pre-trained on Monash. For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation (Af) F1 (%). Top result per dataset and metric in **bold**.

SSL Paradigm	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	16.88±0.04	8.92±0.12	19.95±0.10	48.31±1.09	15.56±1.36	45.67±0.75	15.47±0.46	4.72±0.13	14.07±0.33	50.67±2.27	14.87±0.65	52.70±1.73	11.75±0.11	1.14±0.17	12.99±2.52
Signal-space predictive and reconstructive															
NTP	18.14±0.14	10.03±0.18	20.69±0.11	66.99±1.01	35.51±2.92	58.73±1.77	24.27±0.30	6.53±0.05	19.96±0.25	**64.49±0.27**	17.16±0.14	**66.55±0.21**	11.64±0.11	**2.24±0.48**	9.70±2.62
MAE	25.18±1.37	13.09±0.42	26.96±0.84	60.35±6.06	**43.09±2.49**	52.06±3.69	24.71±0.48	7.14±0.21	19.64±0.44	62.31±0.63	18.48±0.30	65.30±0.33	11.71±0.09	1.57±0.20	**13.84±0.30**
Diffusion	45.76±2.82	10.23±0.57	40.53±1.93	42.59±0.63	17.92±0.96	41.19±1.12	18.25±0.55	5.54±0.18	15.84±0.41	51.95±0.21	30.15±0.13	55.45±0.11	11.53±0.38	2.19±0.15	13.63±0.60
Latent-space alignment															
DINO	18.22±0.57	10.75±0.26	20.68±0.39	**70.55±4.04**	41.49±4.71	**58.85±3.24**	19.19±0.33	5.58±0.06	17.05±0.16	63.40±0.75	15.71±0.10	65.94±0.62	**11.78±0.04**	1.75±0.36	13.77±0.46
JEPA	**57.11±0.11**	12.24±0.05	**49.25±0.09**	42.48±2.00	37.63±1.98	38.90±1.40	**37.48±0.72**	**7.18±0.14**	**29.31±0.65**	52.54±0.15	**31.08±0.19**	56.08±0.06	3.10±0.12	1.59±0.13	4.08±0.07
LeJEPA	**57.11±0.12**	12.27±0.05	49.24±0.08	47.51±0.30	42.24±0.31	43.26±0.14	33.78±0.64	6.33±0.12	26.93±0.33	52.63±0.23	30.74±0.13	56.25±0.08	3.12±0.25	0.84±0.16	4.19±0.24
Contrastive															
SoftCLT	30.18±0.12	**13.20±0.10**	30.62±0.08	52.88±1.99	33.62±5.88	48.29±0.91	26.77±1.02	6.94±0.18	21.37±0.77	52.23±0.24	29.58±0.58	56.22±0.29	11.25±0.31	1.58±0.44	13.13±0.48

Summary. These results indicate that increasing the capacity of the probing head does not alter the task-level conclusions of our study: with an MLP head, pre-training remains clearly beneficial

for anomaly detection (26, 34, and 29 of 35 wins under $\mathrm{F1}_{\mathrm{PA}}$, $\mathrm{F1}_{\mathrm{Ev}}$, and $\mathrm{F1}_{\mathrm{Af}}$), selective for classification (25 of 63), and limited for forecasting (11 of 35). A more expressive head does not compensate for the limitations of the learned representations in forecasting, where only Diffusion shows consistent gains. It does, however, change the ranking between objectives: Diffusion moves from below the random encoder on three of five forecasting datasets under linear probing to above it on all five, and in classification the control benefits from the MLP head about as much as the pre-trained encoders do. Overall, the task-level behavior is driven primarily by the pre-trained representations rather than by the expressiveness of the downstream head, while the objective-level ranking depends on the head, which supports reporting linear and MLP probing side by side.

D.1.3 FINE-TUNING

We report results for backbones pre-trained on *Monash* and fine-tuned per dataset.

Forecasting. We report the fine-tuning results in Table 18. The gains for the pre-trained encoders are small and not consistent: they beat the baseline in only 3 of 35 comparisons under MSE, all on *Weather*. Relative to the frozen random encoder under linear probing, the from-scratch baseline itself improves on only three of five datasets (*ETTm1*, *ETTm2*, and *Weather*). It does, however, recover the datasets it lost under linear probing: it now attains the best result in 8 of the 10 dataset–metric cells, and the remaining two (*Weather*) are lost by at most 0.002, within seed variance. On *ETTh2*, this recovery is due to the pre-trained encoder rather than the baseline: the baseline is unchanged (0.350), while Diffusion, which led under linear probing (0.334), degrades to 0.354 after fine-tuning. LeJEPA degrades on every dataset (e.g., *ETTh1* MSE 0.505 vs. 0.415), and NTP is the second-weakest initialization (*ETTh1* 0.470, *ETTm2* 0.280), indicating that a forecasting pretext task does not transfer to downstream forecasting despite matching it. This gives empirical support to our finding that pre-training does not necessarily produce better representations for forecasting.

Table 18: Full fine-tuning forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) on backbones pre-trained on the **Monash** dataset. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; lower is better and the best result per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Supervised from scratch	**0.415±0.001**	**0.426±0.001**	**0.350±0.002**	**0.391±0.002**	**0.351±0.001**	**0.380±0.001**	**0.252±0.027**	**0.313±0.018**	0.230±0.002	0.265±0.001
Signal-space predictive and reconstructive										
NTP	0.470±0.008	0.462±0.004	0.385±0.001	0.412±0.000	0.356±0.002	0.385±0.001	0.280±0.001	0.332±0.001	0.232±0.002	0.267±0.001
MAE	0.438±0.004	0.438±0.003	0.368±0.002	0.403±0.002	0.354±0.005	0.383±0.001	0.261±0.000	0.322±0.000	0.234±0.003	0.268±0.001
Diffusion	0.441±0.018	0.450±0.008	0.354±0.004	0.397±0.003	0.359±0.003	0.389±0.002	0.268±0.001	0.324±0.000	0.229±0.001	0.266±0.001
Latent-space alignment										
DINO	0.460±0.022	0.456±0.012	0.376±0.007	0.405±0.005	0.356±0.001	0.386±0.001	0.260±0.001	0.317±0.001	0.230±0.000	0.266±0.000
JEPA	0.440±0.011	0.443±0.007	0.383±0.003	0.412±0.002	0.356±0.002	0.384±0.001	0.267±0.001	0.322±0.002	**0.228±0.000**	**0.263±0.001**
LeJEPA	0.505±0.007	0.490±0.004	0.405±0.006	0.432±0.003	0.374±0.004	0.403±0.003	0.281±0.002	0.337±0.003	0.232±0.001	0.269±0.001
Contrastive										
SoftCLT	0.436±0.002	0.449±0.001	0.373±0.002	0.414±0.003	0.361±0.004	0.389±0.002	0.261±0.001	0.320±0.001	**0.228±0.001**	0.264±0.001

Classification. Table 19 shows that full fine-tuning reshapes the linear-probing ranking. Pre-trained initializations outperform supervised training from scratch in 42 of 63 objective–dataset pairs, up from 29 under linear probing, and MAE is the only objective that wins on all nine datasets. DINO, which trailed under linear probing, achieves the top result on *FaceDetection* (66 vs. 50 for the baseline, $\Delta = +15$) and is strong on *SCP1* (86) and *UWave* (79). LeJEPA retains its linear-probing strengths on gesture data, leading *Handwriting* (25 vs. 5, $\Delta = +20$) and *UWave* (85 vs. 48, $\Delta = +37$, the largest gain in the table). MAE tops *JapaneseVowels* (97), *SCP1* (87), and *Heartbeat* (74). JEPA leads *SCP2* (57) and *SpokenArabicDigits* (98) but is otherwise among the weakest pre-trained initializations, falling below the baseline on *UWave* (45 vs. 48) and *Heartbeat* (61 vs. 69). SoftCLT leads *EthanolConcentration* (34) but with high variance (±11). The baseline remains competitive on *JapaneseVowels* (96) and *SpokenArabicDigits* (96), where the strongest methods are within roughly two points of it.

Table 19: Full fine-tuning classification accuracy (%, mean ± std over five seeds) on backbones pre-trained on the **Monash** dataset. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Supervised from scratch	27.15±1.81	50.40±0.31	4.54±0.81	68.78±1.11	96.38±0.50	78.77±0.90	52.44±2.29	95.94±0.39	48.31±1.29
Signal-space predictive and reconstructive									
NTP	26.23±2.09	52.31±0.50	18.06±0.77	74.02±2.99	95.54±0.73	86.35±1.98	50.84±1.15	97.03±0.36	72.66±3.73
MAE	29.09±1.27	50.46±0.51	11.06±1.86	**74.39±2.79**	**97.03±0.69**	**87.20±2.43**	52.78±1.57	97.45±0.25	78.59±1.72
Diffusion	24.81±2.72	56.61±6.36	5.27±1.46	62.80±0.63	92.77±0.72	82.25±2.54	55.70±1.14	97.86±0.19	76.56±1.29
Latent-space alignment									
DINO	26.52±0.78	**65.74±0.55**	10.26±0.84	72.56±0.40	83.44±1.33	86.09±1.57	56.25±0.99	96.84±0.18	79.14±1.69
JEPA	25.00±0.17	51.39±0.97	7.65±1.65	61.10±4.72	93.04±1.05	71.50±3.42	**57.36±1.82**	**98.30±0.24**	44.53±3.58
LeJEPA	25.10±0.27	51.74±1.22	**25.09±1.03**	63.90±2.13	93.17±0.80	78.24±2.56	50.69±2.46	97.11±0.19	**85.39±0.81**
Contrastive									
SoftCLT	**34.32±10.59**	54.12±5.77	11.56±2.84	72.68±0.34	88.31±1.43	72.02±1.34	54.02±2.20	89.21±6.59	54.53±12.61

Anomaly detection. We report the fine-tuning results in Table 20. The pre-trained encoders beat the from-scratch baseline in 21, 24, and 24 of 35 comparisons under $\mathrm{F1}_{\mathrm{PA}}$, $\mathrm{F1}_{\mathrm{Ev}}$, and $\mathrm{F1}_{\mathrm{Af}}$, against 19, 33, and 19 over the random frozen encoder under linear probing, and six of the seven objectives win on a majority of the 15 dataset–metric cells. Fine-tuning preserves the linear-probing advantage of JEPA and LeJEPA on spacecraft telemetry: they lead *MSL* on all three metrics (PA 57.3, Ev 12.4, Af 49.5 vs. 49.4/10.2/44.2 for the baseline), and JEPA leads *SMAP* on PA (36.2) and Af (28.2). On *PSM* the picture changes from linear probing: Diffusion (PA 48.3, Af 45.5) and LeJEPA (PA 48.1, Ev 43.3, Af 43.6) are far ahead of both the baseline (PA 28.3±10.4) and JEPA (PA 20.8), and LeJEPA's event-wise F1 is more than 15 points above any other objective. On *SMD*, NTP and SoftCLT lead the point-adjusted and affiliation metrics (NTP PA 54.2, SoftCLT Af 58.1), whereas the event-wise metric favors JEPA (31.2), LeJEPA (30.6), and Diffusion (30.2), all of which exceed the baseline (20.8) by roughly 10 points. *SWaT* remains an outlier: all other objectives lie within one point of the baseline on PA, while JEPA (7.2) and LeJEPA (8.9) fall below it. DINO is the only objective that degrades under fine-tuning, dropping below the baseline under $\mathrm{F1}_{\mathrm{PA}}$ on every dataset (e.g., *MSL* PA 31.0±6.5), with seed variance several times larger than that of the other objectives. Overall, pre-training remains beneficial for anomaly detection under full fine-tuning.

Table 20: Full fine-tuning anomaly-detection F1 (%, mean ± std over five seeds) on backbones pre-trained on the **Monash** dataset. For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Supervised from scratch	49.44±0.34	10.21±0.12	44.17±0.36	28.29±10.44	25.56±7.59	23.39±9.33	29.21±0.41	6.22±0.11	24.61±0.30	50.25±0.18	20.84±0.30	53.13±0.84	**11.62±0.11**	1.40±0.34	13.54±0.64
Signal-space predictive and reconstructive															
NTP	51.05±2.27	11.03±0.48	44.99±1.65	24.06±2.02	27.76±2.24	25.36±0.81	31.40±1.31	6.48±0.15	25.66±0.92	**54.16±0.70**	21.91±0.14	57.45±0.34	11.59±0.06	1.40±0.24	13.40±0.40
MAE	50.62±1.63	10.75±0.29	44.68±1.13	20.35±0.97	26.23±0.76	23.01±0.61	30.20±0.32	5.55±0.09	25.18±0.31	52.51±0.79	21.65±0.33	56.12±0.42	**11.62±0.11**	1.26±0.05	13.51±0.47
Diffusion	54.01±1.19	11.64±0.25	46.05±1.14	**48.30±0.29**	25.01±0.78	**45.46±0.05**	34.96±0.33	7.32±0.05	27.80±0.30	51.48±0.46	30.24±0.31	54.63±0.47	11.38±0.36	1.19±0.04	13.02±0.86
Latent-space alignment															
DINO	30.99±6.51	7.61±0.80	32.70±4.43	18.28±8.64	20.34±7.38	19.66±8.13	24.36±6.19	**8.66±1.25**	22.15±3.72	46.20±3.51	19.45±1.41	49.96±2.95	11.59±0.13	1.44±0.50	**13.81±0.37**
JEPA	**57.32±0.08**	**12.38±0.02**	**49.48±0.08**	20.80±0.90	26.71±0.62	23.86±0.39	**36.20±0.37**	6.85±0.06	**28.20±0.25**	52.70±0.00	**31.17±0.01**	56.04±0.00	7.18±0.00	0.75±0.14	8.68±0.03
LeJEPA	57.28±0.13	12.35±0.11	49.44±0.10	48.06±0.75	**43.34±1.61**	43.61±0.69	31.95±1.33	6.50±0.18	25.46±0.85	52.19±0.00	30.55±0.06	56.16±0.01	8.88±1.98	0.78±0.18	10.73±2.19
Contrastive															
SoftCLT	53.76±2.41	11.67±0.43	46.66±1.73	32.46±3.25	25.37±1.53	32.28±1.58	33.72±1.33	6.49±0.15	27.37±0.92	53.93±0.35	25.71±1.17	**58.07±0.62**	11.17±0.35	**1.82±0.85**	11.21±1.69

Summary. Under full fine-tuning, no single objective dominates across tasks, and the results highlight a clear distinction between tasks. In anomaly detection, JEPA, LeJEPA, and Diffusion take the top score on most dataset–metric pairs (9 of 15), while DINO becomes unstable and falls below the baseline under $\mathrm{F1}_{\mathrm{PA}}$. In classification, the ranking is dataset-specific: MAE, LeJEPA, JEPA, DINO, and SoftCLT each lead on at least one benchmark, MAE is the only objective that beats the baseline on all nine datasets, and the gap to the baseline is largest on the gesture datasets. For these two tasks, pre-training provides a strong initialization that can be effectively exploited through full fine-tuning. In contrast, for forecasting, the benefits of pre-training remain limited even when all parameters are optimized: the supervised-from-scratch baseline achieves the lowest MSE and MAE on four of the five datasets, and no pre-trained model improves on it except by a marginal

0.001–0.002 on *Weather*. This suggests that current SSL objectives do not capture the inductive biases required for accurate temporal extrapolation.

D.2 Pre-Training Source and Scale

This section gives the extended results for Table 3.

D.2.1 Overview

Figure 7 summarizes performance under Monash, Monash-sized synthetic, Large Synthetic, In-domain and Hybrid pre-training. Because Monash contains approximately 680K samples, whereas Large Synthetic and Hybrid contain 2.5M samples, we use the matched-scale comparisons to separate data source from corpus size. Throughout these experiments, the effect of data composition varies by task. At matched scale, Monash-sized synthetic pre-training is often competitive with Monash pre-training, although the two are not uniformly equivalent. At the larger 2.5M-sample scale, Large Synthetic and Hybrid pre-training also produce similar results for many objectives, suggesting limited benefit from adding the evaluated real-data component in those settings. For example, LeJEPA on *UWave* obtains an accuracy of 63.3% with Large Synthetic pre-training, 71.2% with Monash-sized synthetic pre-training, and 71.69% with Monash pre-training. Overall, these results support synthetic data as a useful pre-training source within the evaluated regime, but do not establish that transfer is independent of data origin: the objective–dataset exceptions above show that source and scale interact.

D.2.2 Extended Forecasting Results (Linear Probing)

In-domain pre-training In-domain pre-training (Table **??**) lowers MSE relative to Monash in 34 of 49 objective–dataset comparisons, and lowers it relative to all other schemes for NTP, MAE, and SoftCLT on nearly all datasets. It is the only regime in which several objectives match or beat the random encoder, led by NTP on Traffic (0.410 vs. 0.435) and ETTm1 (0.341 vs. 0.356), and the benefit is broadest on Weather, where six of seven objectives outperform random. Still, the random encoder remains better in 34 of the 49 comparisons, with one tie; JEPA does not benefit on any ETT dataset; and LeJEPA's ETTh1 MSE falls from 0.616 to 0.486 while the random encoder stays at 0.414.

Table 21: Linear-probing forecasting MSE and MAE (mean $\pm$ std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained **in-domain** (on each target dataset). Lower is better; best per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	0.379±0.001	0.261±0.001	0.318±0.000	0.248±0.000	0.281±0.001	0.172±0.000	0.267±0.000	0.435±0.000	0.298±0.000
Signal-space predictive and reconstructive														
NTP	0.426±0.002	0.434±0.001	0.351±0.001	0.392±0.001	**0.341±0.004**	**0.375±0.002**	0.264±0.001	0.320±0.001	**0.235±0.001**	**0.269±0.001**	**0.165±0.000**	**0.258±0.000**	**0.410±0.001**	**0.280±0.000**
MAE	0.424±0.004	0.436±0.003	0.349±0.002	0.389±0.001	0.356±0.003	0.386±0.002	**0.256±0.001**	0.317±0.001	0.240±0.001	0.275±0.001	0.182±0.002	0.280±0.002	0.447±0.003	0.326±0.005
Diffusion	0.429±0.002	0.440±0.002	**0.336±0.001**	**0.385±0.001**	0.362±0.001	0.391±0.001	0.264±0.001	0.321±0.001	0.237±0.000	0.274±0.000	0.198±0.001	0.308±0.001	0.458±0.001	0.338±0.002
Latent-space alignment														
DINO	0.418±0.002	0.429±0.002	0.358±0.001	0.396±0.001	0.352±0.001	0.383±0.001	0.263±0.002	0.317±0.002	0.238±0.001	0.274±0.001	0.176±0.003	0.273±0.005	0.437±0.001	0.310±0.001
JEPA	0.437±0.002	0.437±0.001	0.364±0.002	0.405±0.001	0.370±0.001	0.388±0.000	0.275±0.001	0.333±0.000	0.237±0.000	0.271±0.000	0.188±0.003	0.287±0.003	0.453±0.003	0.326±0.003
LeJEPA	0.486±0.004	0.477±0.003	0.384±0.003	0.415±0.002	0.377±0.003	0.401±0.002	0.279±0.001	0.337±0.001	0.253±0.001	0.287±0.000	0.733±0.007	0.701±0.004	0.691±0.086	0.505±0.063
Contrastive														
SoftCLT	0.438±0.006	0.444±0.004	0.354±0.002	0.392±0.002	0.365±0.002	0.388±0.001	0.260±0.003	**0.314±0.002**	0.243±0.001	0.276±0.001	0.217±0.012	0.325±0.014	0.454±0.003	0.324±0.004

Impact of Pre-training Data Composition. Figures 3–6 and Tables 10,22— 24 report the effect of pre-training data composition on forecasting performance across datasets. Overall, the effect of data source is dataset-dependent and generally modest.

For ETTh1 and ETTh2, Synthetic and Hybrid pre-training often provide slight improvements over Monash for several methods, although the relative differences remain small. Among signal-space objectives, Diffusion is particularly strong on ETTh2, while *Latent-space alignment* (LeJEPA) consistently underperforms on ETTh1.

On the ETTm datasets, the ranking between methods remains largely stable across data regimes. On ETTm1, *Signal-space predictive and reconstructive* methods (NTP) are consistently among the strongest performers, while on ETTm2, *Latent-space alignment* (DINO) achieves the best results across all regimes, with only minor variation between Monash, Synthetic, and Hybrid pre-training.

Weather and Electricity exhibit similarly limited sensitivity to data composition. On these datasets, *Latent-space alignment* (JEPA) and *Signal-space predictive and reconstructive* methods (NTP) are generally competitive, while LeJEPA consistently underperforms, particularly on Electricity.

Traffic exhibits the largest sensitivity to data composition. In particular, LeJEPA degrades substantially under Synthetic and Hybrid pre-training, JEPA performs better under Monash (0.442) than under Synthetic or Hybrid (0.456/0.458), whereas NTP improves slightly under Synthetic and Hybrid (0.434 vs. 0.443).

Overall, these results reinforce that data composition alone does not provide a uniform improvement in forecasting performance. Instead, its effect depends strongly on the interaction between dataset characteristics and the SSL objective, with no single pre-training regime consistently dominating across benchmarks.

Table 22: Linear-probing forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **large synthetic** corpus (2.5M samples). The first row is a randomly initialized frozen encoder baseline; the remaining rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives. Lower is better and the best result per dataset and metric (including the baseline) is in **bold**; ties at the reported precision are both bolded.

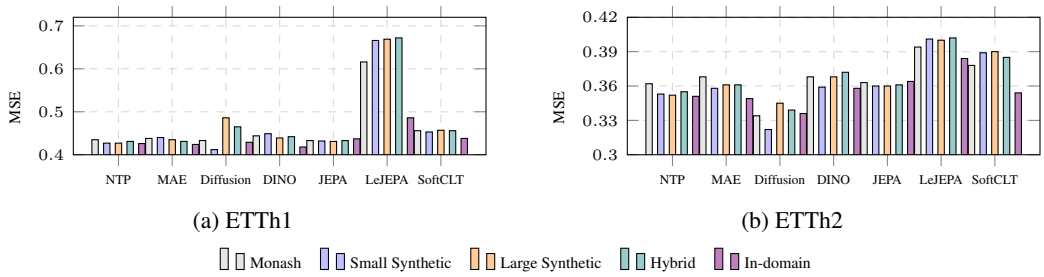
Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	**0.393±0.001**	0.356±0.001	**0.379±0.001**	0.261±0.001	0.318±0.000	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	0.435±0.000	**0.298±0.000**
Signal-space predictive and reconstructive														
NTP	0.427±0.002	0.435±0.002	0.352±0.001	0.394±0.001	**0.350±0.001**	**0.379±0.001**	0.260±0.001	0.318±0.001	0.237±0.000	**0.271±0.000**	**0.172±0.000**	0.268±0.000	**0.434±0.000**	0.299±0.001
MAE	0.435±0.003	0.446±0.002	0.361±0.002	0.396±0.002	0.362±0.023	0.389±0.014	0.269±0.003	0.327±0.002	0.242±0.011	0.280±0.008	0.191±0.003	0.293±0.004	0.465±0.003	0.338±0.004
Diffusion	0.486±0.009	0.482±0.007	**0.345±0.002**	0.396±0.003	0.390±0.004	0.409±0.005	0.275±0.003	0.331±0.003	0.245±0.004	0.283±0.003	0.327±0.018	0.439±0.017	0.647±0.023	0.463±0.013
Latent-space alignment														
DINO	0.439±0.008	0.448±0.005	0.368±0.006	0.404±0.004	0.359±0.002	0.386±0.001	**0.256±0.001**	**0.314±0.000**	**0.236±0.001**	0.273±0.001	0.185±0.003	0.286±0.004	0.452±0.006	0.323±0.007
JEPA	0.431±0.002	0.439±0.004	0.360±0.004	0.404±0.007	0.369±0.003	0.389±0.004	0.280±0.008	0.336±0.008	0.255±0.008	0.288±0.008	0.195±0.002	0.294±0.002	0.456±0.003	0.327±0.001
LeJEPA	0.669±0.008	0.561±0.004	0.400±0.004	0.428±0.003	0.460±0.005	0.452±0.003	0.287±0.002	0.345±0.002	0.265±0.002	0.300±0.001	0.433±0.018	0.505±0.015	0.925±0.027	0.585±0.014
Contrastive														
SoftCLT	0.457±0.002	0.463±0.002	0.390±0.002	0.422±0.003	0.395±0.002	0.412±0.001	0.275±0.002	0.332±0.002	0.246±0.001	0.282±0.001	0.223±0.003	0.321±0.003	0.524±0.004	0.390±0.004

Table 23: Linear-probing forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **hybrid** mixture of Monash and synthetic data (2.5M samples). The first row is a randomly initialized frozen encoder baseline; the remaining rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives. Lower is better and the best result per dataset and metric (including the baseline) is in **bold**; ties at the reported precision are both bolded.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	**0.379±0.001**	0.261±0.001	**0.318±0.000**	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	0.435±0.000	**0.298±0.000**
Signal-space predictive and reconstructive														
NTP	0.431±0.001	0.438±0.002	0.355±0.001	0.396±0.001	**0.350±0.001**	**0.379±0.001**	**0.260±0.000**	**0.318±0.001**	**0.235±0.000**	**0.270±0.000**	0.173±0.000	0.269±0.000	**0.434±0.001**	0.300±0.001
MAE	0.431±0.003	0.441±0.002	0.361±0.003	0.398±0.002	0.369±0.003	0.394±0.002	0.265±0.002	0.324±0.001	0.241±0.001	0.277±0.002	0.186±0.001	0.285±0.002	0.463±0.011	0.333±0.011
Diffusion	0.465±0.006	0.470±0.004	**0.339±0.002**	**0.390±0.002**	0.401±0.012	0.417±0.007	0.272±0.002	0.329±0.002	0.251±0.003	0.290±0.003	0.307±0.019	0.424±0.017	0.606±0.018	0.449±0.014
Latent-space alignment														
DINO	0.442±0.005	0.449±0.004	0.372±0.003	0.407±0.001	0.353±0.001	0.384±0.001	0.265±0.001	0.321±0.001	**0.235±0.001**	0.271±0.001	0.184±0.001	0.284±0.002	0.453±0.004	0.327±0.004
JEPA	0.433±0.003	0.440±0.002	0.361±0.005	0.404±0.004	0.370±0.003	0.390±0.002	0.278±0.006	0.334±0.005	0.255±0.007	0.288±0.007	0.196±0.003	0.295±0.003	0.458±0.004	0.327±0.003
LeJEPA	0.672±0.006	0.562±0.003	0.402±0.006	0.430±0.003	0.466±0.002	0.455±0.001	0.292±0.001	0.350±0.001	0.264±0.002	0.299±0.002	0.434±0.008	0.507±0.006	0.896±0.008	0.573±0.007
Contrastive														
SoftCLT	0.456±0.003	0.462±0.003	0.385±0.001	0.418±0.001	0.386±0.004	0.406±0.003	0.266±0.001	0.325±0.003	0.245±0.001	0.280±0.001	0.215±0.001	0.315±0.001	0.507±0.003	0.377±0.001

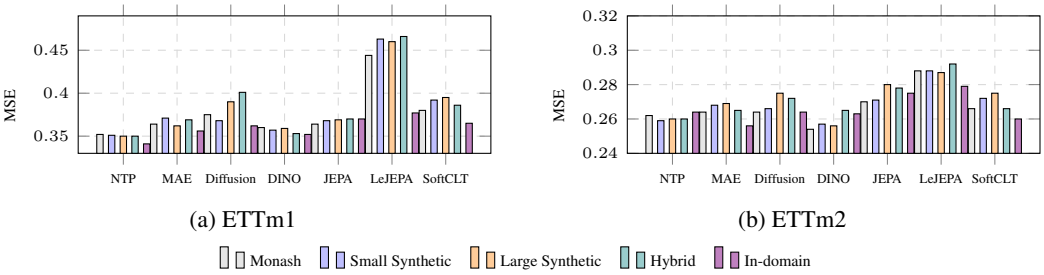
Table 24: Linear-probing forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **Monash-sized synthetic** corpus (matched scale, per-model configuration). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; lower is better and the best result per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	0.414±0.002	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	**0.379±0.001**	0.261±0.001	0.318±0.000	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	0.435±0.000	**0.298±0.000**
Signal-space predictive and reconstructive														
NTP	0.427±0.001	0.436±0.000	0.353±0.001	0.394±0.001	**0.351±0.001**	0.380±0.001	0.259±0.001	0.318±0.001	**0.237±0.001**	**0.271±0.000**	**0.172±0.000**	0.268±0.000	**0.433±0.000**	0.299±0.000
MAE	0.440±0.002	0.450±0.002	0.358±0.002	0.395±0.001	0.371±0.002	0.395±0.001	0.268±0.001	0.325±0.001	0.241±0.001	0.276±0.001	0.192±0.003	0.297±0.006	0.461±0.004	0.336±0.002
Diffusion	**0.412±0.001**	0.430±0.001	**0.322±0.001**	**0.374±0.001**	0.368±0.001	0.388±0.001	0.266±0.002	0.321±0.001	0.246±0.000	0.279±0.000	0.181±0.001	0.286±0.001	0.455±0.002	0.333±0.002
Latent-space alignment														
DINO	0.449±0.011	0.452±0.007	0.359±0.001	0.396±0.001	0.357±0.001	0.386±0.001	**0.257±0.001**	**0.315±0.001**	**0.237±0.000**	0.273±0.001	0.179±0.001	0.279±0.002	0.443±0.003	0.314±0.004
JEPA	0.432±0.004	0.436±0.001	0.360±0.008	0.401±0.005	0.368±0.002	0.386±0.002	0.271±0.003	0.327±0.002	0.241±0.003	0.274±0.002	0.186±0.002	0.284±0.002	0.447±0.002	0.319±0.003
LeJEPA	0.666±0.010	0.560±0.004	0.401±0.003	0.430±0.002	0.463±0.012	0.453±0.006	0.288±0.003	0.346±0.003	0.267±0.001	0.302±0.001	0.432±0.028	0.502±0.021	0.861±0.025	0.550±0.013
Contrastive														
SoftCLT	0.453±0.001	0.460±0.001	0.389±0.002	0.423±0.001	0.392±0.001	0.411±0.001	0.272±0.002	0.330±0.001	0.244±0.001	0.281±0.000	0.219±0.001	0.318±0.000	0.512±0.002	0.382±0.002



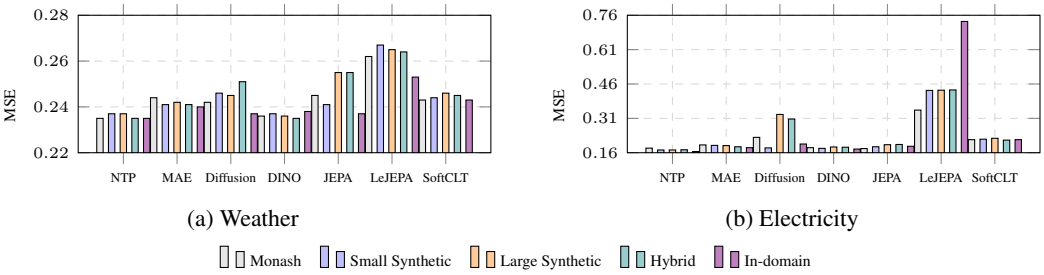
(a) ETTh1 (b) ETTh2

Figure 3: Impact of pre-training data composition on forecasting performance (MSE, linear probing, 8-layer backbone) for the ETTh datasets. Bars: Monash (680K), Small Synthetic (680K, matched scale), Large Synthetic (2.5M), Hybrid (2.5M), and In-domain. Lower is better.



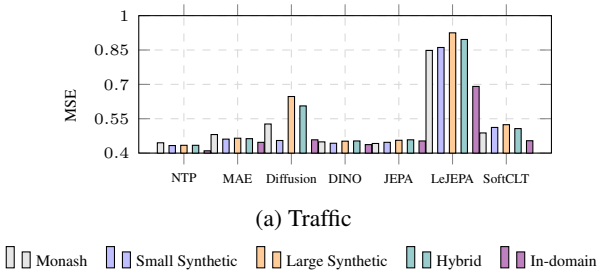
(a) ETTm1 (b) ETTm2

Figure 4: Impact of pre-training data composition on forecasting performance (MSE, linear probing, 8-layer backbone) for the ETTm datasets. Lower is better.



(a) Weather (b) Electricity

Figure 5: Impact of pre-training data composition on forecasting performance (MSE, linear probing, 8-layer backbone) for Weather and Electricity. Lower is better.



(a) Traffic

Figure 6: Impact of pre-training data composition on forecasting performance (MSE, linear probing, 8-layer backbone) for Traffic. Lower is better.

D.2.3 Extended Classification Results (Linear Probing)

Tables 13, 25, 26, and 27 report classification performance under the Monash large Synthetic, Hybrid pre-training, and Monash-sized synthetic (Monash sized) regimes, respectively. Overall, SSL methods present substantial gains over the non-pretrained baseline on several datasets, and the leading objective on each dataset is largely stable across pre-training sources.

Across all four sources, LeJEPA is the strongest frozen classifier. It leads *Handwriting*, *SpokenArabicDigits*, and *UWave* under every source, often by large margins over both the baseline and the next objective (Handwriting: 13.0, $\Delta = +8.7$ under large Synthetic, 16.4, $\Delta = +12.1$ under Monash-sized synthetic and 12.7, $\Delta = +8.4$ under Hybrid), and it leads *FaceDetection* under Monash and Hybrid and *SCP1* under Monash and large Synthetic. On *FaceDetection* and *SCP1* the margin to NTP or MAE is within about one point under every source, so those two leads are not robust to source.

MAE and NTP lead the remaining datasets. MAE leads *Heartbeat* under every source (73.0–75.6 vs. 66.6 for random features) and *EthanolConcentration* under Monash and Monash-sized synthetic, though every Ethanol result lies within the random encoder's seed variance. NTP leads *JapaneseVowels* under the three synthetic-containing sources (95.5–96.2 vs. 93.0), the one dataset without a gain under Monash pre-training.

The effect of pre-training source is present but secondary to the choice of objective. Source moves NTP most: its *UWave* accuracy rises from 55.0 under Monash to 61.6 and 62.8 under the synthetic sources and its *FaceDetection* accuracy from 55.0 to 59.1–60.9. LeJEPA's *UWave* accuracy is 71.69 under Monash and 71.2 under matched synthetic but 63.3 and 58.4 under large Synthetic and Hybrid, and MAE's falls from 40.0 to 28.1–31.8 under the synthetic sources. These shifts change absolute scores; they do not change which objective leads *Handwriting*, *SpokenArabicDigits*, *UWave*, or *Heartbeat*.

DINO and JEPA, the other latent-space alignment objectives, do not share LeJEPA's advantage. DINO is second on *Handwriting* under every source (10.0–12.0) and leads *SCP2* under Monash (56.0), but falls 13–17 points below random features on *JapaneseVowels* under every source. JEPA leads *SCP2* under Monash-sized synthetic and Hybrid (55.8, 54.5) and is otherwise at or below the random encoder on most datasets. LeJEPA's classification advantage is therefore a property of that objective rather than of latent-space alignment.

Seed variance is low for most objectives. LeJEPA's standard deviations range from 0.4 to 4.9 points, with the largest on *UWave* and *SCP1*; Diffusion and MAE reach 5–6 points on *JapaneseVowels* and *SpokenArabicDigits*; SoftCLT is the least stable, with 9–15 points on *JapaneseVowels*, *SpokenArabicDigits*, and *UWave* under the synthetic sources.

Overall, classification benefits substantially from SSL, and the objective that leads each dataset is stable across sources: LeJEPA on the handwriting and gesture benchmarks, MAE on *Heartbeat*, and NTP on *JapaneseVowels* where any gain appears. Pre-training source refines absolute accuracy but does not alter these leaders.

Table 25: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **large synthetic** corpus (2.5M samples). Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	27.68±1.88	54.49±0.54	4.31±0.54	66.63±1.33	93.02±0.65	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	27.80±1.80	**60.90±0.70**	4.30±0.70	70.50±4.10	**96.20±0.60**	75.50±1.70	53.40±2.30	85.30±0.50	62.80±1.20
MAE	26.50±3.70	55.60±0.50	4.50±0.60	**74.80±0.70**	82.20±6.10	78.00±1.50	52.90±4.00	81.30±2.10	28.60±2.70
Diffusion	24.40±0.80	56.30±0.80	7.70±1.80	65.20±2.60	90.30±0.50	66.70±3.60	52.20±1.30	84.20±1.40	56.70±2.80
Latent-space alignment									
DINO	26.80±1.10	53.40±0.60	12.00±1.30	71.80±0.60	79.80±0.80	71.70±1.50	53.60±1.90	82.70±1.60	44.00±7.70
JEPA	24.90±0.30	52.50±0.30	4.20±0.50	63.00±3.70	89.00±3.50	66.70±1.90	53.60±2.30	71.90±3.50	27.90±3.50
LeJEPA	25.10±0.50	60.80±0.80	**13.00±1.10**	66.40±1.60	82.50±4.70	**83.10±1.30**	47.10±1.80	**90.80±1.50**	**63.30±4.80**
Contrastive									
SoftCLT	**28.75±2.79**	58.60±0.66	5.34±0.60	72.30±0.19	62.06±3.92	75.70±4.87	**54.22±1.25**	67.15±2.66	49.81±3.14

Table 26: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **hybrid** mixture of Monash and synthetic data (2.5M samples). Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	27.68±1.88	54.49±0.54	4.31±0.54	66.63±1.33	93.02±0.65	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	**28.44±1.40**	59.13±0.60	5.98±1.20	70.64±1.60	**95.52±0.50**	74.81±1.80	52.78±2.50	84.57±1.00	54.44±3.10
MAE	28.06±2.50	55.83±0.50	5.48±1.00	**75.61±1.10**	84.05±2.00	**81.91±2.00**	51.44±1.90	84.17±0.80	31.75±2.50
Diffusion	24.90±0.20	56.30±0.70	7.10±1.60	63.40±2.50	91.20±0.70	66.60±2.10	51.30±1.30	84.70±1.60	56.10±5.90
Latent-space alignment									
DINO	27.80±1.10	54.00±0.80	11.60±0.90	71.70±0.70	80.30±0.80	68.40±1.50	52.70±1.70	83.50±1.60	44.80±6.00
JEPA	24.94±0.20	51.71±0.50	6.07±1.90	63.71±1.90	91.89±0.70	66.48±3.20	**54.45±3.00**	75.94±5.30	32.56±6.60
LeJEPA	24.87±0.90	**60.85±0.40**	**12.66±1.90**	64.97±1.00	85.08±2.70	81.50±4.90	47.00±2.50	**92.45±0.60**	**58.44±3.10**
Contrastive									
SoftCLT	26.62±0.99	56.91±1.26	6.31±0.52	74.34±1.37	74.60±4.83	76.66±6.92	54.11±2.93	84.69±3.32	58.81±9.16

Table 27: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash-sized synthetic** corpus (matched scale, per-model configuration). Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	27.68±1.88	54.49±0.54	4.31±0.54	66.63±1.33	93.02±0.65	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	27.84±1.37	**60.65±0.93**	4.80±1.20	70.83±2.57	**96.11±0.47**	75.56±1.73	53.55±2.85	86.05±0.75	61.56±2.38
MAE	**28.14±1.83**	53.55±1.24	4.30±0.51	**73.46±2.26**	93.89±2.03	**83.96±3.39**	50.33±2.97	76.75±2.27	28.06±2.30
Diffusion	25.02±0.56	56.90±0.88	9.41±1.20	65.85±1.54	91.35±0.85	68.13±2.67	52.11±2.29	88.55±1.22	55.13±2.34
Latent-space alignment									
DINO	27.38±0.93	52.87±0.67	11.62±1.34	71.51±1.26	75.89±0.65	73.72±1.43	52.78±0.99	77.97±1.64	36.25±5.13
JEPA	25.10±0.64	52.06±0.33	6.85±1.05	63.51±2.53	91.24±0.37	68.94±2.79	**55.78±2.64**	79.92±0.41	42.50±1.47
LeJEPA	26.08±2.12	59.67±0.61	**16.38±0.69**	60.68±2.73	81.46±2.47	82.12±1.38	46.44±1.15	**91.11±1.58**	**71.19±2.18**
Contrastive									
SoftCLT	27.91±3.82	52.04±1.27	6.16±1.04	73.17±0.81	64.90±14.72	61.09±3.74	55.00±2.14	64.06±12.81	33.56±14.72

D.2.4 Extended Anomaly Detection Results (Linear Probing)

Tables 14,29 , 28, and 30 provide the full anomaly-detection results under Monash-sized synthetic, large synthetic, and hybrid pre-training, respectively. Overall, anomaly detection benefits more clearly from SSL than forecasting, but the leading objective depends on the dataset rather than on the objective family.

Across all four sources, JEPA, LeJEPA, and Diffusion lead the spacecraft-telemetry datasets. All three reach 56.5–57.0 $\mathrm{F1_{PA}}$ on *MSL* under every source ($\Delta{=}{+}41$ over the random frozen encoder), and JEPA and LeJEPA reach 33.8–35.5 on *SMAP* ($\Delta{=}{+}26$ to $+28$), 6–10 points above the next objective. On the service-monitoring datasets DINO leads instead: it attains the highest $\mathrm{F1_{PA}}$ on *PSM* under every source (63.9–72.7) and on *SMD* under every source (61.9–63.5), where NTP is within 0.3–1.6 points. JEPA, the strongest objective on *MSL* and *SMAP*, falls below the random frozen encoder on *PSM* under every source (19.3–25.8 vs. 47.6), so leadership on one dataset does not carry to another within the same family.

The effect of pre-training source is meaningful but not uniform. NTP and MAE improve on *PSM* under the synthetic-containing sources: NTP rises from 55.0 under Monash to 58.1, 63.6, and 67.7 under matched synthetic, hybrid, and large synthetic, and MAE from 36.3 to 49.5, 47.4, and 64.0, with seed standard deviations of 5–12 points. JEPA, LeJEPA, and Diffusion move by less than one point on *MSL* and *SMAP* across sources. SoftCLT is unstable under synthetic sources (*PSM* 33.8±14.4 and *SMD* 40.0±11.6 at matched scale). *SWaT* is a consistent outlier: no objective exceeds

the random frozen encoder on PA or affiliation F1 under any source, and JEPA and LeJEPA collapse to $\approx$3 $\mathrm{F1_{PA}}$ throughout.

Taken together, the objective that leads each dataset is stable across sources: JEPA and LeJEPA on *MSL* and *SMAP*, DINO on *PSM* and *SMD* under point-adjusted F1, and JEPA under event-wise F1 on *SMD* (31.2–31.3 under every source). Source changes absolute scores, most for NTP and MAE on *PSM*, but does not produce a universally superior pre-training regime or change which objective leads.

Anomaly detection benefits from pre-training on four of the five benchmarks (Table 14). On the spacecraft-telemetry datasets JEPA and LeJEPA lead: on MSL they raise affiliation F1 from 19 to 49 and point-adjusted F1 from 16 to 57 (Δ=+41), matched by Diffusion, and on SMAP from 8 to 27 and 8 to 35, 6–8 points above MAE, the strongest signal-space objective; seed standard deviations are below one point. On the service-monitoring datasets NTP and DINO lead instead: DINO raises PSM affiliation F1 from 46 to 55 and point-adjusted F1 from 48 to 71, NTP and DINO gain 6 on SMD under both, while JEPA falls below the baseline on PSM and SoftCLT on PSM and SMD. Event-wise F1 confirms the benefit but changes the leader: it rises from 7 to 13 on MSL (MAE, NTP), 16 to 42 on PSM (LeJEPA, DINO), and 18 to 31 on SMD (JEPA, LeJEPA). SWaT is the exception: under point-adjusted and affiliation F1 no objective improves on the random frozen encoder, and the event-wise gain (NTP 2.9 vs. 1.1) is within two points.

The transfer gains persist under full fine-tuning (Tables 20 and 19), but the ranking does not. JEPA and LeJEPA lead *MSL* at $\mathrm{F1_{PA}}$ 57.3 vs. 49.4 from scratch ($\Delta = +7.9$), Diffusion and LeJEPA lead *PSM* at 48.3 vs. 28.3 ($\Delta = +20.0$), and LeJEPA reaches 85.4 on *UWave* vs. 48.3 ($\Delta = +37$) and 25.1 on *Handwriting* vs. 4.5 ($\Delta = +20$), the two yielding the largest transfer gain in the table. Yet DINO, below the random frozen encoder on *FaceDetection* under linear probing (52.0 vs. 54.5), leads it after fine-tuning (65.7 vs. 50.4), and Diffusion, below it on *PSM* (43.8 vs. 47.6), leads there too (48.3); conversely NTP gains 7 $\mathrm{F1_{PA}}$ on *PSM* frozen but falls to 24.1 vs. 28.3 adapted, and DINO becomes unstable under anomaly-detection fine-tuning, with seed variance an order of magnitude above the other objectives (*MSL* PA 31.0$\pm$6.5) and results below the baseline on every dataset except *SWaT*.

Sensitivity to the detection metric Which objective leads depends on the F1 criterion (Table 14). On *SMD*, DINO attains the highest point-adjusted F1 (62.2) while JEPA attains the highest event-wise F1 (31.2 vs. 17.5 for DINO); on *PSM*, DINO leads point-adjusted and affiliation F1 (71.2, 54.8) while LeJEPA matches it event-wise (41.9 vs. 41.9); on *MSL*, the point-adjusted leaders JEPA and LeJEPA ($\approx$57) trail MAE and NTP event-wise (12.2 vs. 13.3 and 13.1). The same reversals hold under synthetic and hybrid pre-training (Appendix D.2.4) and under full fine-tuning, where NTP leads *SMD* point-adjusted (54.2) and SoftCLT affiliation (58.1) but JEPA, LeJEPA, and Diffusion lead event-wise (31.2, 30.6, 30.2 vs. 20.8 from scratch). Event-wise and affiliation scoring preserve the benefit of pre-training on *MSL*, *PSM*, *SMAP*, and *SMD* but change the leading objective on each of them.

Table 28: Linear-probing anomaly-detection F1 (%, mean $\pm$ std over five seeds) using an 8-layer backbone pre-trained on the **large synthetic** corpus (2.5M samples). For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

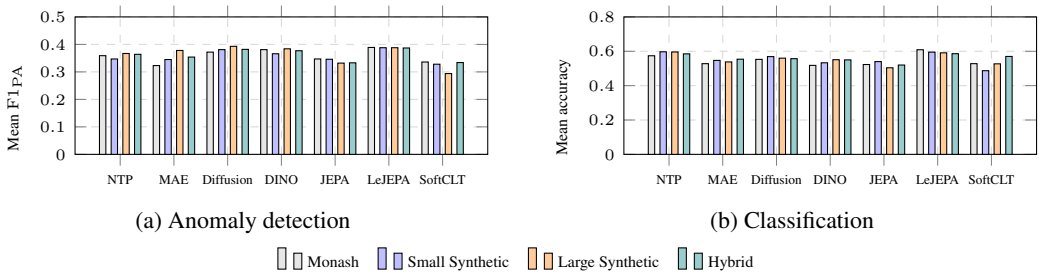
Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	24.66±0.51	11.54±0.17	27.10±0.33	67.66±1.17	41.39±3.77	51.03±2.36	17.42±0.77	4.73±0.14	14.53±0.53	61.84±0.66	17.25±0.68	64.17±0.60	11.70±0.10	1.59±0.52	12.14±2.64
MAE	29.32±0.27	**12.95±0.04**	29.95±0.20	63.95±4.78	32.87±4.45	54.44±1.98	25.90±0.71	6.44±0.15	20.80±0.47	57.96±1.84	17.02±0.79	59.90±1.31	11.70±0.10	1.66±0.30	12.89±1.83
Diffusion	56.66±0.01	12.27±0.01	48.50±0.02	43.78±2.74	40.73±1.27	40.32±2.66	**36.37±0.43**	6.76±0.05	**28.19±0.35**	52.42±0.27	30.24±0.31	55.93±0.14	7.19±1.10	0.86±0.16	8.32±1.06
Latent-space alignment															
DINO	25.51±1.61	12.03±0.54	27.15±1.01	**72.66±2.25**	40.52±4.92	**58.06±2.18**	18.72±2.27	5.50±0.73	15.34±1.54	**63.48±1.32**	16.71±0.40	**65.38±0.68**	11.74±0.06	1.45±0.11	12.97±1.46
JEPA	56.95±0.09	12.20±0.04	49.25±0.10	19.33±0.84	25.96±0.61	22.76±0.27	33.76±0.42	6.29±0.09	25.74±0.50	52.65±0.13	**31.25±0.05**	56.07±0.09	3.21±0.00	**1.90±0.10**	4.12±0.02
LeJEPA	**57.03±0.07**	12.21±0.03	**49.26±0.05**	47.15±0.23	**41.61±0.42**	42.69±0.13	34.10±0.53	6.35±0.12	27.06±0.29	52.55±0.16	30.95±0.04	55.98±0.09	3.10±0.07	0.94±0.06	4.11±0.07
Contrastive															
SoftCLT	25.99±4.82	12.08±1.66	26.91±4.26	37.12±13.89	31.65±7.94	36.76±10.46	28.16±1.21	**7.06±0.10**	22.37±1.00	45.81±5.15	22.18±6.04	48.98±4.59	10.01±1.66	1.39±0.38	12.02±1.87

Table 29: Linear-probing anomaly-detection F1 (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **hybrid** mixture of Monash and synthetic data (2.5M samples). For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	26.34±0.43	12.30±0.10	28.03±0.31	63.57±2.92	31.96±2.27	57.45±1.90	19.29±0.42	5.26±0.11	15.87±0.32	61.33±0.91	18.82±0.86	63.71±0.32	11.67±0.12	1.72±0.35	12.01±2.64
MAE	29.16±0.35	12.81±0.25	29.76±0.29	47.44±11.65	24.78±4.10	47.09±6.98	27.32±2.63	6.83±0.49	21.53±1.83	61.62±1.25	17.71±0.62	63.93±1.01	11.51±0.15	1.72±0.31	10.55±2.19
Diffusion	56.66±0.00	12.26±0.01	48.55±0.08	44.64±2.25	26.41±3.09	43.03±2.74	27.48±2.17	5.85±0.19	22.64±1.58	52.27±0.27	29.85±0.28	55.94±0.20	9.70±2.40	1.20±0.36	11.55±2.67
Latent-space alignment															
DINO	23.52±0.74	12.41±0.56	25.83±0.52	**72.56±1.06**	**42.51±4.53**	**58.88±2.57**	18.94±1.23	5.61±0.51	15.52±0.83	**61.94±0.51**	16.53±0.34	**64.07±0.39**	11.68±0.13	1.67±0.10	11.56±1.81
JEPA	**56.99±0.05**	12.19±0.04	**49.24±0.04**	19.96±2.25	26.42±1.56	23.07±0.99	33.96±0.36	6.33±0.09	25.92±0.60	52.56±0.16	**31.16±0.12**	55.98±0.11	3.20±0.04	**1.80±0.14**	4.13±0.03
LeJEPA	56.97±0.05	12.17±0.02	49.21±0.07	46.52±0.96	42.42±0.38	41.88±1.11	**34.50±0.46**	6.42±0.10	**27.11±0.38**	52.69±0.03	31.07±0.09	56.02±0.07	3.06±0.07	1.16±0.18	4.06±0.06
Contrastive															
SoftCLT	25.56±6.11	**13.00±0.18**	26.54±5.27	50.34±1.48	38.26±2.09	46.18±2.33	30.56±0.34	**7.45±0.19**	24.03±0.26	49.17±3.40	27.48±0.73	52.46±4.04	11.13±0.52	1.06±0.14	13.11±0.89

Table 30: Linear-probing anomaly-detection F1 (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash-sized synthetic** corpus (matched scale, per-model configuration). For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	25.01±0.61	11.72±0.28	27.26±0.37	58.10±8.56	29.27±7.40	51.90±4.65	17.07±0.82	4.68±0.24	14.30±0.62	61.53±1.59	17.33±0.82	63.70±0.93	11.81±0.02	**1.71±0.23**	13.32±1.25
MAE	28.98±0.72	12.83±0.24	29.72±0.42	49.52±11.79	23.64±2.97	47.03±3.98	24.45±1.76	6.24±0.34	19.63±1.27	59.73±1.40	17.53±0.60	61.46±1.38	9.70±3.86	1.36±0.48	11.18±2.15
Diffusion	56.66±0.00	12.23±0.03	48.59±0.06	43.41±2.83	34.18±4.56	41.17±2.88	30.84±0.70	6.30±0.07	24.94±0.56	52.33±0.52	29.81±0.35	55.91±0.31	7.35±0.25	0.93±0.15	8.93±0.45
Latent-space alignment															
DINO	25.56±0.80	11.96±0.34	27.39±0.39	**63.91±7.12**	33.67±2.49	**55.15±5.22**	18.60±1.55	5.39±0.51	15.31±1.10	**63.06±0.51**	16.24±0.17	**65.39±0.45**	11.75±0.09	1.36±0.20	13.20±1.64
JEPA	56.94±0.07	12.16±0.03	49.13±0.06	24.60±3.98	29.40±2.78	25.68±2.26	**35.50±0.63**	6.70±0.14	**27.87±0.50**	52.63±0.12	**31.19±0.05**	56.03±0.03	3.21±0.01	1.61±0.16	4.16±0.02
LeJEPA	**56.97±0.04**	12.20±0.02	**49.22±0.07**	46.87±0.23	**41.52±0.59**	42.46±0.30	34.47±0.56	6.43±0.12	27.15±0.31	52.49±0.14	30.93±0.09	55.94±0.03	3.07±0.09	0.91±0.06	4.09±0.07
Contrastive															
SoftCLT	30.37±0.07	**13.20±0.04**	30.80±0.04	40.20±1.29	35.59±0.84	37.81±0.76	29.66±0.41	**7.29±0.06**	23.50±0.33	52.47±0.34	30.49±0.20	56.39±0.12	11.22±0.37	1.42±0.44	13.24±0.46



(a) Anomaly detection (b) Classification

Figure 7: Impact of pre-training corpus under linear probing (8-layer backbone). Bars: Monash (680K), Small Synthetic (680K, matched scale), Large Synthetic (2.5M), and Hybrid (2.5M). (a) $\mathrm{F1_{PA}}$ averaged over the five anomaly-detection datasets; higher is better. (b) Accuracy averaged over the nine UEA datasets; higher is better. The corresponding forecasting results are shown per dataset in Figures 3–6.

E TRAINING DURATION, DEPTH, AND CONVERGENCE

In this section we provide results on training duration, transformer depth, and convergence.

E.1 EXTENDED RESULTS FOR 40 EPOCH EVALUATION

To test whether the per-model results are limited by pre-training budget, we double the schedule from 20 to 40 epochs while holding every other hyper-parameter at its native per-model value: the same 8-layer encoder, 336-step context, patch length, optimiser, and learning rate, so the only change is twice as many parameter updates. All backbones are pre-trained on Monash and evaluated by linear probing over 5 seeds. We run the same experiment with a 24-layered encoder on single seed. Tables 31, 32, 33, and 34 report classification, anomaly detection, forecasting, and 24 layer classification. We also report results of 24 layered transformer with 40 epochs

Results. Doubling the pre-training budget leaves the 20-epoch conclusions intact. Rankings are unchanged across all three task families: LeJEPA still leads the gesture and morphology benchmarks, signal-space objectives the periodic ones, JEPA and LeJEPA the spacecraft-telemetry anomaly datasets, and the Random frozen encoder baseline still attains the lowest forecasting MSE on ETTh1, Electricity, and Traffic. Most cells move by less than one accuracy point or 0.005 MSE. SoftCLT is the exception. SoftCLT improves forecasting on every dataset (ETTh1 $0.456 \rightarrow 0.442$, Electricity $0.217 \rightarrow 0.195$, Traffic $0.488 \rightarrow 0.431$), but remains above the random encoder on six of seven, matching it only on Traffic (0.431 vs. 0.435). The additional updates otherwise amplify the failure modes already visible at 20 epochs: JEPA collapses on JapaneseVowels, SpokenArabicDigits, and UWave with seed standard deviations of 8–22 points, and LeJEPA's forecasting degradation deepens on every dataset (Traffic $0.848 \rightarrow 0.940$), consistent with the isotropy induced by SIGReg accumulating with training. The forecasting transfer gain does not emerge with more training, so the gap to the random encoder is not a symptom of under-training, in line with the convergence analysis in Appendix E.2.

Table 31: Linear-probing forecasting error (MSE and MAE, mean ± std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **Monash** dataset for **40 epochs**. Lower is better; the best result per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	**0.379±0.001**	**0.261±0.001**	0.318±0.000	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	0.435±0.000	0.298±0.000
Signal-space predictive and reconstructive														
NTP	0.436±0.002	0.441±0.001	0.362±0.001	0.400±0.001	**0.351±0.000**	0.381±0.000	0.262±0.001	0.319±0.001	**0.235±0.000**	**0.270±0.001**	0.179±0.000	0.277±0.001	0.444±0.001	0.311±0.001
MAE	0.451±0.004	0.443±0.002	0.362±0.002	0.397±0.003	0.366±0.001	0.392±0.000	0.263±0.002	0.324±0.002	0.244±0.002	0.280±0.002	0.192±0.001	0.294±0.002	0.474±0.004	0.342±0.006
Diffusion	0.436±0.004	0.448±0.004	0.334±0.001	0.385±0.002	0.379±0.004	0.397±0.003	0.264±0.001	0.321±0.001	0.243±0.002	0.280±0.002	0.235±0.015	0.350±0.016	0.550±0.036	0.404±0.024
Latent-space alignment														
DINO	0.445±0.005	0.448±0.004	0.367±0.001	0.401±0.001	0.356±0.001	0.386±0.001	0.263±0.004	**0.318±0.003**	0.235±0.000	0.271±0.000	0.183±0.001	0.283±0.001	0.451±0.002	0.326±0.002
JEPA	0.437±0.005	0.442±0.003	0.361±0.004	0.404±0.005	0.368±0.005	0.389±0.004	0.273±0.006	0.328±0.005	0.253±0.009	0.286±0.009	0.189±0.009	0.287±0.010	0.457±0.018	0.323±0.013
LeJEPA	0.654±0.006	0.555±0.003	0.397±0.002	0.428±0.002	0.480±0.007	0.458±0.004	0.295±0.001	0.353±0.001	0.267±0.001	0.302±0.001	0.402±0.010	0.473±0.008	0.957±0.030	0.601±0.014
Contrastive														
SoftCLT	0.442±0.014	0.448±0.012	0.372±0.013	0.408±0.009	0.377±0.012	0.400±0.012	0.263±0.007	0.322±0.008	0.244±0.001	0.279±0.001	0.195±0.025	0.292±0.028	**0.431±0.000**	**0.293±0.000**

Table 32: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on **Monash** for **40 epochs**. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	27.68±1.88	54.49±0.54	4.31±0.54	66.63±1.33	93.02±0.65	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	27.20±1.50	54.90±0.30	6.20±1.40	70.80±1.80	**93.50±0.90**	76.10±1.30	53.70±2.10	86.50±1.20	57.60±2.60
MAE	25.70±1.59	52.82±0.93	5.98±1.12	**74.44±0.79**	92.00±1.06	**83.89±2.89**	51.56±2.80	73.26±2.27	26.00±6.48
Diffusion	25.20±1.30	57.70±0.50	9.50±0.60	64.10±2.00	90.60±1.40	67.50±1.70	51.30±1.70	88.50±0.30	56.70±1.60
Latent-space alignment									
DINO	26.60±1.40	53.40±1.50	11.60±1.20	72.40±0.40	76.70±1.40	68.70±1.10	53.80±1.50	75.70±2.00	41.20±7.50
JEPA	24.90±0.20	50.10±0.50	3.80±0.00	66.80±1.70	38.90±22.00	61.20±4.10	**54.60±2.50**	52.90±13.60	20.10±7.90
LeJEPA	25.02±1.51	**60.60±0.87**	**15.83±1.72**	62.05±1.76	83.73±1.59	81.77±2.10	48.00±2.07	**91.87±0.98**	**76.00±2.56**
Contrastive									
SoftCLT	**28.20±1.30**	52.70±1.00	7.50±1.70	72.70±0.80	69.50±7.00	64.40±3.80	52.60±1.90	58.20±12.30	41.90±14.80

Table 33: Linear-probing anomaly-detection F1 (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on **Monash** for **40 epochs**. For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	28.97±0.53	13.08±0.16	29.71±0.30	52.94±7.71	31.48±9.15	48.71±1.80	22.78±0.48	6.20±0.09	18.48±0.32	**61.97±0.34**	20.02±0.22	**64.58±0.41**	11.55±0.11	**2.57±0.36**	10.40±1.72
MAE	29.36±0.10	**13.32±0.09**	29.93±0.05	30.11±3.76	15.92±2.40	38.77±1.88	27.38±0.94	**7.25±0.08**	21.55±0.70	59.85±1.71	19.55±0.63	62.75±1.51	3.85±3.94	0.82±0.27	9.69±2.30
Diffusion	56.67±0.01	12.24±0.02	48.62±0.06	45.51±1.16	23.52±2.35	43.82±1.59	25.39±1.54	6.06±0.41	20.67±0.84	52.33±0.23	29.83±0.19	56.19±0.32	11.18±1.08	1.71±0.36	13.15±1.11
Latent-space alignment															
DINO	25.10±0.99	12.31±0.50	27.24±0.65	**66.04±8.28**	35.16±7.38	**54.59±4.01**	21.04±2.74	5.82±0.82	17.11±1.98	60.65±4.61	19.02±3.81	62.49±4.04	11.71±0.09	1.70±0.46	12.30±2.14
JEPA	**57.01±0.07**	12.18±0.05	**49.20±0.03**	21.98±2.64	27.71±1.73	24.24±1.38	33.55±0.28	6.26±0.05	25.67±0.19	52.63±0.13	**31.22±0.04**	56.03±0.05	3.22±0.00	1.99±0.16	4.13±0.02
LeJEPA	56.96±0.08	12.18±0.03	49.20±0.06	47.06±1.11	**42.15±0.53**	42.02±0.96	**34.19±1.36**	6.35±0.28	**26.97±0.94**	52.67±0.01	31.00±0.09	55.98±0.01	3.11±0.09	1.12±0.20	4.10±0.07
Contrastive															
SoftCLT	30.28±0.18	13.17±0.08	30.73±0.09	45.52±3.55	37.97±3.55	40.32±2.43	29.70±0.39	7.18±0.09	23.53±0.33	52.21±0.21	30.60±0.14	55.98±0.16	11.30±0.31	1.93±0.59	13.06±0.42

Table 34: Linear-probing classification accuracy (%, single seed) using a **24-layer** backbone pre-trained on **Monash** for **40 epochs**. Higher is better; best per dataset (among completed objectives) in **bold**. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary).

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	25.10	53.21	3.76	67.32	82.70	75.43	51.67	54.21	31.87
Signal-space predictive and reconstructive									
NTP	24.71	59.48	7.88	67.80	89.73	74.74	52.22	88.40	53.12
MAE	25.10	51.50	3.76	71.22	73.24	56.31	49.44	55.66	26.25
Diffusion	25.10	56.58	10.71	65.85	**91.62**	73.04	53.89	89.72	55.63
Latent-space alignment									
DINO	27.38	51.08	11.76	70.73	72.43	73.04	**54.44**	65.58	30.00
JEPA	25.10	52.16	3.76	61.46	90.27	68.26	53.33	74.17	42.19
LeJEPA	26.62	**61.83**	**15.53**	65.37	73.78	**78.50**	47.22	**90.72**	**76.25**
Contrastive									
SoftCLT	**28.52**	50.26	4.47	**73.66**	52.16	63.14	53.89	34.79	25.31

E.2 Depth Scaling

Downstream performance across depths Figure 8 shows forecasting performance as a function of Transformer depth. Across objectives, MSE is approximately flat between 2 and 8 layers and often increases at greater depths; the shallowest configuration is frequently within approximately 0.01 MSE of the best result. The rise beyond 8 layers is particularly visible for MAE on *ETTh1* (0.438 at N=8 vs. 0.554 at N=24) and SoftCLT on *ETTm1* (0.375 at N=8 vs. 0.692 at N=12); these are single-seed sweep values and differ from the five-seed the main tables. The broad conclusion that SSL provides limited forecasting benefit remains stable across depths. Although the deeper models train stably under the fixed budget (Appendix E.2), this experiment is not compute-normalized and does not isolate whether degradation at greater depth arises from overfitting, optimization efficiency, or objective–capacity interactions.

Implications for Depth Scaling. These results support our main finding that the absence of forecasting gains beyond eight layers is not caused by insufficient training. Even at 24 layers, models converge within the allocated budget, suggesting that the lack of improvement at greater depth reflects the learned representations rather than optimization constraints. While we do not perform compute-normalized scaling (e.g., matched optimization steps), the consistent convergence behavior across depths indicates that additional training would likely yield only marginal improvements.

Finally, deeper models incur substantially higher computational cost. For example, 24-layer models trained on Synthetic or Hybrid data require up to 28 hours per run (Appendix C.5), reinforcing the practical relevance of the depth result.

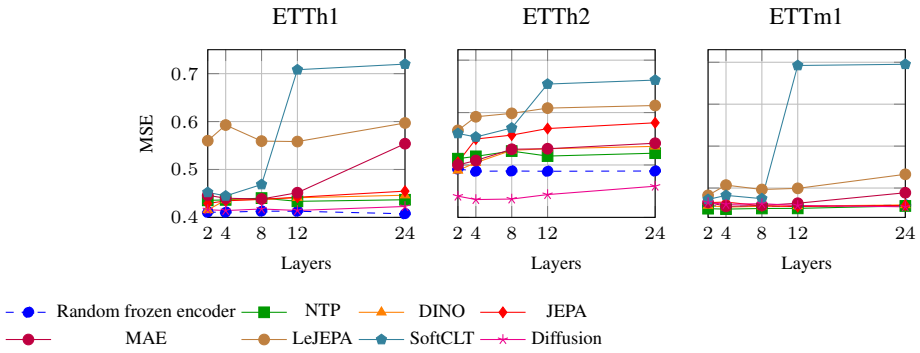


Figure 8: Forecasting performance (MSE) as a function of backbone depth.

E.3 Convergence Analysis

Convergence Under Fixed Pre-training Budget. All models are pre-trained for a fixed budget of 20 epochs across all backbone depths, including the deepest 24-layer configurations. Empirically, we observe stable convergence behavior across all evaluated SSL paradigms. Training loss decreases consistently in early epochs and plateaus thereafter, with no evidence of instability or divergence even for deep models.

Model selection is based on the best validation checkpoint rather than the final epoch. Across methods, peak validation performance typically occurs before the full budget is exhausted, most commonly around epochs 15–16. This indicates that the fixed 20-epoch budget is sufficient for convergence and does not disadvantage deeper models.

Training Dynamics for Deep Models. Figure 9 presents representative training curves for 8-layer models pre-trained on the Synthetic dataset for all seven paradigms. All methods exhibit rapid initial improvement followed by saturation. DINO converges quickly within the first few epochs, MAE shows a gradual but consistent decrease, and NTP stabilizes early with minor fluctuations. Improvements beyond epoch 15 are marginal across all methods.

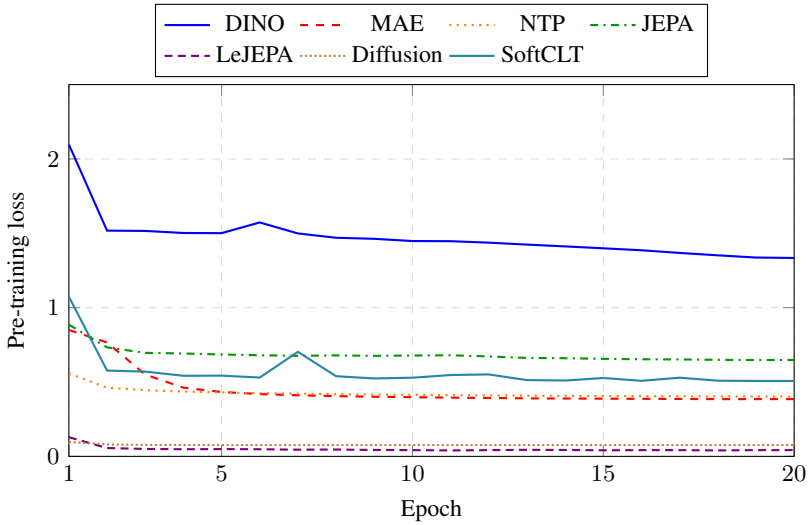


Figure 9: Pre-training loss vs. epoch for all seven objectives, pre-trained on **hybrid** data (Monash + synthetic, 2.5M samples, single seed, 20 epochs). Each curve is the model's own raw per-epoch training loss; because the objectives differ (DINO/JEPA/LeJEPA latent-space, MAE/Diffusion reconstruction, NTP prediction, SoftCLT contrastive), absolute values are not directly comparable across methods.

F CONTROLLED ABLATIONS

F.1 AUGMENTATIONS ABLATION STUDY LEJEPA

Setup. To isolate the impact of augmentation design on LeJEPA, we evaluate several families of transformations beyond the default DWT-based pipeline. Each augmentation is applied symmetrically to construct dual views and is parameterized by a single magnitude controlling the strength of the transformation. We consider the following families:

- **Galilean & Special-Relativistic:** global temporal rescaling and local Lorentz-style transformations of the form $\gamma(x - v \cdot t)$.
- **Coordinate Transformations:** global polar warps and local rotations in the value–time plane.
- **Hyperbolic Transformations:** global hyperbolic warps and local Möbius shifts on the Poincaré disk.
- **DWT (baseline):** wavelet-based augmentations with structured perturbations in the frequency domain.

Results. Across the forecasting benchmarks, DWT-based augmentations consistently provide the most stable and competitive performance.

LeJEPA For **forecasting**, DWT achieves on par with hyperbolic, (lowest MSE on ETTM1, lowest MAE on ETTM2, and ETTH1), on par with hyperbolic, and passing the others. while on **classification**, LeJEPA wins 5 out of 9, winning them by a large margin, while on the other 4, DWT is highly competitive.

Discussion. These results indicate that the effectiveness of augmentations depends strongly on the alignment between the induced invariances and the underlying signal structure. DWT-based augmentations preserve temporal locality while introducing structured perturbations in the frequency domain, which appears well-suited for forecasting.

In contrast, the alternative transformations introduce more global or nonlinear distortions that are less aligned with the statistical properties of real-world time series. While these transformations are theoretically motivated, they do not translate into effective inductive biases in practice.

Conclusion. Overall, the ablation confirms that the choice of augmentation is critical for LeJEPA. DWT-based augmentations provide a robust and well-aligned inductive bias, while alternative transformation families fail to consistently improve performance. This justifies our design choice of using wavelet-based augmentations as the default across all experiments.

Table 35: LeJEPA Forecasting performance (MSE and MAE) across augmentation strategies on Monash. Lower is better; best results per dataset and metric are highlighted in **bold**.

Augmentation	Loss	ETTH1	ETTH2	ETTM1	ETTM2	Weather
Galilean & Relativistic	MSE	0.682	**0.391**	0.465	0.295	0.262
	MAE	0.572	**0.426**	0.448	0.353	0.298
Coordinate / Polar	MSE	0.642	0.398	0.455	0.294	**0.261**
	MAE	0.552	0.430	0.446	0.351	**0.296**
DWT	MSE	0.611	0.399	**0.444**	0.296	0.263
	MAE	**0.537**	0.427	0.439	**0.349**	0.298
Hyperbolic	MSE	**0.606**	0.406	0.448	**0.293**	0.263
	MAE	0.538	0.441	**0.437**	0.350	0.299

Table 36: Linear-probing classification accuracy (%, single seed, 123) across augmentation strategies for a **LeJEPA** backbone pre-trained on Monash. Higher is better; best per dataset in **bold**. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary).

Augmentation	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Galilean & Relativistic	**28.14**	62.54	16.12	60.98	73.78	81.57	48.89	91.86	71.56
Coordinate / Polar	24.33	62.54	**16.47**	60.00	84.05	80.89	45.00	92.09	74.06
DWT	25.10	60.64	13.41	**64.88**	**87.57**	**84.98**	47.22	**92.68**	**76.56**
Hyperbolic	22.05	**65.07**	11.76	61.46	67.03	73.38	**50.56**	77.81	70.63

F.2 Results for Same Parameter Evaluation

We isolate the effect of the pre-training *objective* from confounds of capacity and optimisation by re-training every model under an identical budget: an 8-layer encoder with a 4-layer predictor (when needed), embedding dimension 128, patch size 16, learning rate 1×10^{-4}, batch size 64, and 20 pre-training epochs, with a fixed context window. All models are pre-trained on Monash and evaluated by linear probing with these settings held constant, so any difference in downstream performance reflects the SSL objective alone. NTP and MAE already use this configuration in the main experiments, so their rows are unchanged from the Monash tables. Tables 37, 38, and 39 give the per-dataset results.

Overall Trends. Under a matched budget the paradigm ranking is strongly task-dependent. Predictive *Latent-space alignment* (JEPA, LeJEPA) dominates anomaly detection and gesture/telemetry classification, *Signal-space predictive and reconstructive* objectives are strongest on periodic and speech signals, and the *random frozen encoder* baseline is essentially unbeaten on forecasting. Invariance-based DINO is competitive on point-adjusted anomaly metrics but collapses on shift-sensitive datasets, and LeJEPA – the strongest classifier – is the weakest forecaster, consistent with the precision–invariance hypothesis of Section 5.

Forecasting. Table 37 shows that, under a matched budget, pre-training provides no consistent forecasting benefit. The Random frozen encoder baseline attains the lowest MAE on ETTh1, ETTm1, Electricity, and Traffic and the lowest MSE on ETTh1 and Electricity; the remaining wins are marginal and split between Diffusion (ETTh2, MSE 0.329), DINO (ETTm2, MSE 0.257), and NTP/JEPA (Weather MSE 0.235, Traffic MSE 0.433). Most pre-trained models land within 0.01–0.02 MSE of the baseline. LeJEPA, the strongest classifier, is by far the weakest forecaster (ETTh1 MSE 0.618, Traffic 0.822), consistent with the near-isotropic embedding geometry induced by SIGReg (Appendix A.5).

Table 37: Linear-probing forecasting error (MSE and MAE, mean $\pm$ std over five seeds, averaged over horizons $\{96, 192, 336, 720\}$) using an 8-layer backbone pre-trained on the **Monash** dataset under an **equal pre-training budget** (identical LR, batch size, epochs, and context across all objectives). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; lower is better and the best result per dataset and metric is in **bold**.

Objective	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Random frozen encoder	**0.414±0.002**	**0.426±0.002**	0.350±0.002	0.393±0.001	0.356±0.001	**0.379±0.001**	0.261±0.001	0.318±0.000	0.248±0.000	0.281±0.001	**0.172±0.000**	**0.267±0.000**	0.435±0.000	**0.298±0.000**
Signal-space predictive and reconstructive														
NTP	0.435±0.002	0.441±0.001	0.362±0.003	0.400±0.002	**0.352±0.001**	0.381±0.000	0.262±0.001	0.319±0.001	0.235±0.000	0.270±0.001	0.180±0.001	0.277±0.001	0.445±0.001	0.311±0.001
MAE	0.438±0.002	0.438±0.002	0.368±0.004	0.402±0.003	0.364±0.002	0.390±0.001	0.264±0.004	0.324±0.003	0.244±0.001	0.279±0.001	0.194±0.001	0.295±0.001	0.481±0.003	0.345±0.003
Diffusion	0.433±0.002	0.446±0.001	**0.333±0.002**	**0.384±0.002**	0.375±0.003	0.395±0.002	0.263±0.001	0.320±0.001	0.242±0.001	0.280±0.001	0.226±0.002	0.342±0.003	0.523±0.007	0.383±0.004
Latent-space alignment														
DINO	0.437±0.006	0.443±0.003	0.363±0.003	0.396±0.003	0.356±0.002	0.385±0.001	**0.257±0.005**	**0.313±0.003**	0.236±0.001	0.270±0.001	0.179±0.003	0.278±0.004	0.443±0.005	0.316±0.008
JEPA	0.440±0.001	0.441±0.001	0.365±0.003	0.406±0.002	0.365±0.001	0.386±0.001	0.275±0.002	0.331±0.001	**0.235±0.000**	**0.269±0.000**	0.174±0.000	0.268±0.000	**0.433±0.001**	0.305±0.001
LeJEPA	0.618±0.016	0.541±0.008	0.388±0.003	0.421±0.002	0.439±0.010	0.436±0.006	0.282±0.002	0.340±0.001	0.258±0.001	0.294±0.001	0.342±0.008	0.430±0.007	0.822±0.021	0.540±0.010
Contrastive														
SoftCLT	0.458±0.005	0.457±0.002	0.359±0.013	0.404±0.006	0.380±0.001	0.404±0.002	0.265±0.001	0.323±0.002	0.244±0.002	0.279±0.001	0.216±0.003	0.318±0.002	0.489±0.002	0.359±0.002

Classification. Table 38 shows that with matched capacity LeJEPA is the strongest encoder, taking the top accuracy on five of nine datasets – FaceDetection (61.8), Handwriting (15.2), SCP1 (80.0), SpokenArabicDigits (93.1), and UWaveGestureLibrary (78.8) – with the largest margins on the gesture benchmarks (UWave $+42$ over the baseline). The remaining datasets are led by SoftCLT on

Heartbeat (73.2) and JEPA on SCP2 (57.1), while the random frozen encoder itself attains the highest accuracy on EthanolConcentration (27.7) and JapaneseVowels (93.0). Under a matched budget, no signal-space objective leads any dataset.

Table 38: Linear-probing classification accuracy (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash** dataset under an **equal pre-training budget** (identical LR, batch size, epochs, and context across all objectives). Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), UWave (UWaveGestureLibrary). Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset is in **bold**.

Objective	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave
Random frozen encoder	**27.68±1.88**	54.49±0.54	4.31±0.54	66.63±1.33	93.02±0.65	76.72±2.86	50.00±2.60	77.61±0.96	36.56±2.78
Signal-space predictive and reconstructive									
NTP	24.94±0.19	57.76±0.60	6.26±2.08	66.83±3.16	89.62±0.99	74.75±1.22	52.89±3.57	88.29±0.40	55.06±1.83
MAE	24.56±0.46	54.20±0.91	5.06±1.30	62.05±1.39	86.27±2.59	75.70±1.97	51.66±2.65	81.51±2.38	34.56±2.93
Diffusion	24.94±0.79	56.11±0.24	7.50±0.45	65.17±2.48	91.13±0.55	67.44±4.07	52.55±1.09	85.80±1.15	52.00±2.79
Latent-space alignment									
DINO	25.94±0.37	51.44±0.18	8.80±0.73	72.68±0.53	78.00±0.78	69.42±2.61	54.22±0.75	75.79±2.63	32.44±3.41
JEPA	24.87±0.46	52.90±0.37	8.97±0.36	60.00±2.86	91.40±0.52	70.85±2.28	**57.11±2.39**	83.45±0.36	46.63±1.93
LeJEPA	25.70±1.89	**61.78±0.30**	**15.20±0.55**	61.85±1.52	88.38±1.04	**80.00±1.47**	49.55±2.29	**93.08±0.29**	**78.81±2.07**
Contrastive									
SoftCLT	26.54±1.19	52.51±0.68	7.18±0.45	**73.17±0.62**	70.43±1.17	64.99±1.75	53.11±1.60	87.42±1.23	55.00±5.81

Anomaly Detection. Table 39 reports point-adjusted (PA), event-wise (Ev), and affiliation (Af) F1 (Sec B.2.3). Predictive Latent-space alignment leads the spacecraft-telemetry benchmarks: LeJEPA tops MSL (PA 57.0, Af 49.3) and JEPA tops SMAP (PA 37.1, Af 29.3), both far above the Random frozen encoder baseline (MSL PA 15.6, SMAP PA 8.0). On PSM and SMD the Signal-space predictive and reconstructive and invariance objectives take over the point metrics – DINO leads PSM (PA 55.6, Af 49.9) and NTP leads SMD (PA 61.8, Af 64.3) – while the event-wise metric consistently favours the latent-alignment and contrastive models (LeJEPA PSM-Ev 42.4; JEPA/LeJEPA/SoftCLT SMD-Ev ≈31), which the point-adjusted score under-credits. Two consistent failure modes appear: SWaT is degenerate for every method (all PA within a point of the baseline, and JEPA/LeJEPA collapse to ≈3), and MAE is unstable on PSM (PA 32.6±4.8) and SWaT (5.8±4.8).

Table 39: Linear-probing anomaly-detection F1 (%, mean ± std over five seeds) using an 8-layer backbone pre-trained on the **Monash** dataset under an **equal pre-training budget** (identical LR, batch size, epochs, and context across all objectives). For each dataset we report point-adjusted (PA), event-wise (Ev), and affiliation-based (Af) F1. Rows are grouped into signal-space predictive and reconstructive, latent-space alignment, and contrastive objectives; the best result per dataset and metric is in **bold**.

Objective	MSL			PSM			SMAP			SMD			SWaT		
	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af	PA	Ev	Af
Random frozen encoder	15.57±0.27	7.32±0.06	18.66±0.14	47.55±2.10	16.40±1.23	46.12±1.74	7.98±0.14	2.47±0.04	8.43±0.09	56.08±1.38	17.55±0.36	58.25±0.79	**11.84±0.01**	1.09±0.06	**14.23±0.16**
Signal-space predictive and reconstructive															
NTP	28.58±0.34	13.11±0.09	29.45±0.19	55.02±9.27	29.64±4.71	49.55±2.94	22.68±0.61	6.19±0.10	18.39±0.42	**61.75±0.34**	20.00±0.43	**64.29±0.37**	11.50±0.01	**2.86±0.36**	9.54±0.16
MAE	29.36±0.16	**13.29±0.07**	29.96±0.20	36.31±5.19	19.13±3.30	41.34±2.72	27.05±0.85	7.17±0.11	21.45±0.64	61.09±1.27	19.05±0.76	63.34±1.25	7.72±4.80	1.28±0.53	10.87±2.39
Diffusion	56.66±0.01	12.26±0.02	48.56±0.08	46.35±0.67	21.57±1.95	45.02±1.78	24.38±0.59	5.94±0.10	20.06±0.48	52.35±0.21	29.81±0.27	56.08±0.37	9.25±1.95	1.52±0.61	10.59±1.44
Latent-space alignment															
DINO	18.75±0.64	9.29±0.51	21.26±0.60	**55.56±7.97**	31.26±6.64	**49.88±5.75**	13.38±0.66	3.75±0.17	11.89±0.50	59.58±0.66	16.51±0.72	61.49±0.86	11.83±0.01	1.22±0.13	14.16±0.21
JEPA	56.84±0.04	12.10±0.03	49.10±0.09	43.31±0.38	37.48±0.27	39.23±0.32	**37.13±0.13**	**7.09±0.03**	**29.28±0.08**	52.64±0.06	30.95±0.12	55.99±0.12	3.02±0.03	1.10±0.06	4.04±0.02
LeJEPA	**57.03±0.05**	12.22±0.04	**49.26±0.06**	47.38±0.58	**42.41±0.81**	42.81±0.32	34.24±0.29	6.38±0.04	26.99±0.21	52.74±0.14	**30.99±0.06**	56.06±0.14	3.18±0.07	1.01±0.06	4.13±0.05
Contrastive															
SoftCLT	30.32±0.09	13.18±0.03	30.73±0.07	48.17±0.85	34.50±2.19	42.01±1.73	28.98±0.28	7.06±0.05	23.00±0.21	52.58±0.35	30.70±0.19	56.25±0.16	11.37±0.08	2.33±1.70	12.19±2.20

Summary. With capacity, optimisation, and context held identical across all models, the objective alone produces a clean task-dependent split: predictive *Latent-space alignment* gives the best initialisation for anomaly detection and classification (LeJEPA in particular), *Signal-space predictive and reconstructive* objectives excel on periodic/speech signals, while random init wins on forecasting.

F.3 LeJEPA ablations

Sensitivity to SIGReg Strength. Removing SIGReg produces representation collapse in the evaluated setting: at $\lambda = 0$, the reported per-dimension embedding standard deviation is 0.000, and mean classification accuracy falls to 31.26%, compared with 54.11% for the random frozen encoder.

All evaluated positive weights avoid this collapse. Across $\lambda \in \{0.002, 0.008, 0.02, 0.05\}$, effective rank ranges from 107.7 to 114.5, and per-dimension embedding standard deviation ranges from 0.964 to 0.983. Mean classification accuracy is 60.43%, 61.00%, 60.24%, and 59.84%, respectively. Each configuration exceeds the random encoder's mean accuracy, although individual datasets remain sensitive to the regularization weight. Increasing λ does not produce a monotonic improvement in classification performance.

The forecasting outcome differs: every evaluated positive weight yields higher MSE than the random encoder on all five forecasting datasets in the table. Thus, preventing collapse and obtaining classification gains do not by themselves establish a positive forecasting transfer gain.

F.4 Representation Analyses

F.4.1 Embedding Analysis

We provide a qualitative analysis of the learned representations using a t-SNE visualization on the *UWaveGestureLibrary* dataset (Figure 10). Each point corresponds to an embedded time series instance, colored by its class label.

Since this dataset is used exclusively for classification, the visualization reflects the degree of *class separability* induced by each SSL paradigm. Clear differences emerge between methods. DINO produces well-separated and compact clusters, indicating strong discriminative structure in the embedding space. LeJEPA also yields highly separable clusters, though with slightly more dispersion, suggesting that its invariance constraints preserve class-level structure while allowing some intra-class variability. JEPA forms coherent but less sharply separated clusters, reflecting a softer organization of the latent space.

In contrast, *Signal-space predictive and reconstructive* methods exhibit weaker class separation. MAE produces partially clustered representations with noticeable overlap between classes, while NTP shows more entangled embeddings with limited global organization. Diffusion lies between these extremes, forming moderately structured but less compact clusters compared to alignment-based approaches.

Overall, the visualization is consistent with the classification results. LeJEPA's separable clusters translate directly into top linear-probing accuracy. DINO's compact clusters indicate that class structure is present in the embedding, yet its linear-probing accuracy on this dataset is low; since t-SNE is non-linear, this structure is not linearly accessible from the frozen representation, which is consistent with DINO's recovery under full fine-tuning (Section 4).

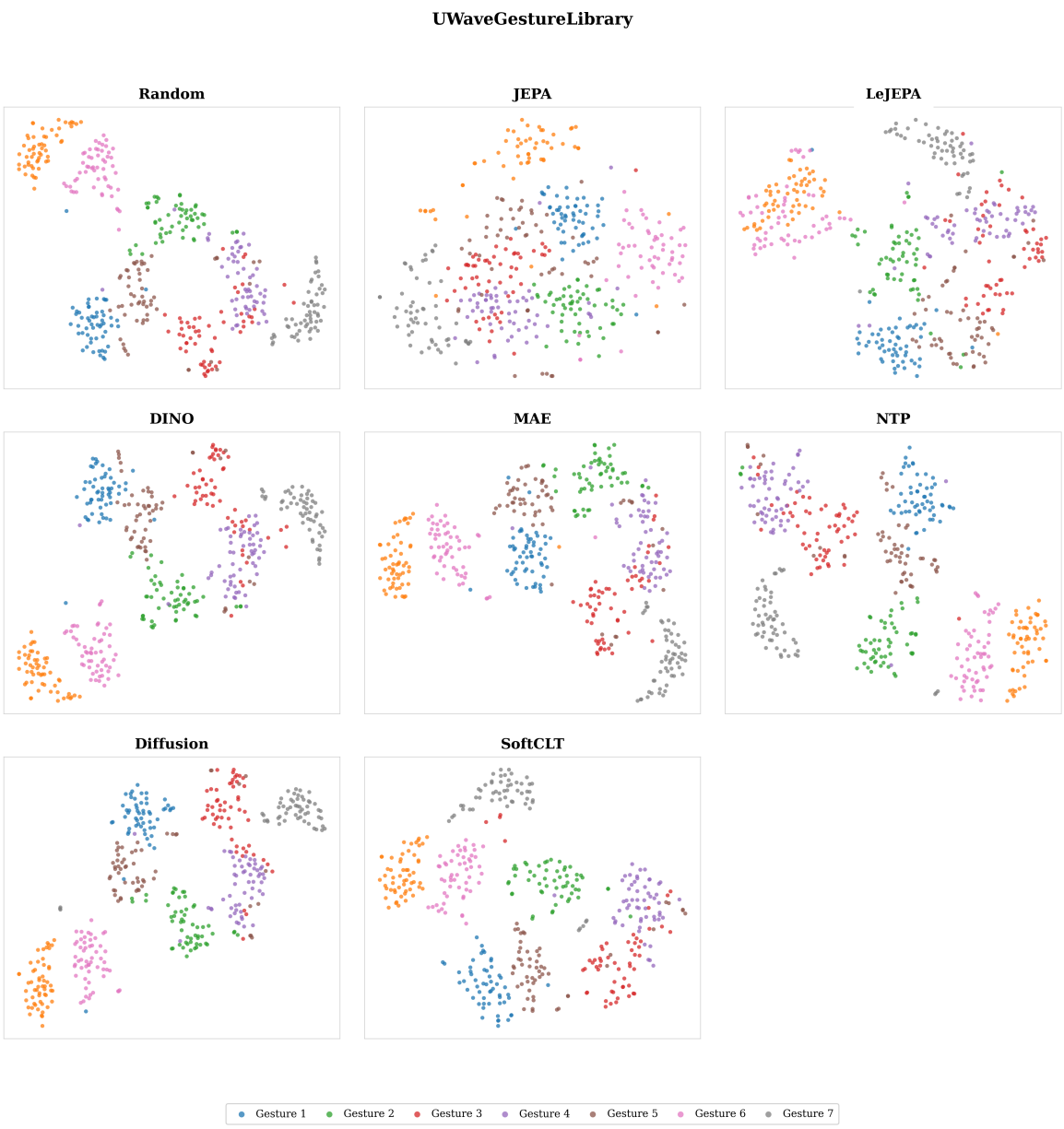


Figure 10: Embedding Analysis

F.4.2 Temporal Latent Drift

To probe how the learned representations respond to temporal perturbations, we measure the displacement of frozen encoder outputs under two families of input perturbation. For each objective we load the Monash-pre-trained backbone (seed 1337, context 336) and embed 300 non-overlapping windows from each of Weather and ETTh1 (z-scored per channel, stride 48). Each window is perturbed either by a circular time-shift of $k \in \{1, 2, 4, 8, 16, 32\}$ steps or by zeroing a random $p \in \{10\%, 20\%\}$ of its timesteps, and re-embedded. Drift is the cosine distance $D = 1 - \cos(\mathbf{z}, \mathbf{z}_\varepsilon)$ between the clean and perturbed embeddings, averaged over windows and datasets. We mean-pool the patch tokens before computing D; concatenating tokens instead would register a whole-patch shift as a large change purely because tokens move to different slots, which conflates re-ordering with a change in content. Results are from a single pre-training seed and carry no downstream label; they describe representation geometry only.

Table 40: Temporal latent drift (cosine distance, lower means the representation moved less) of frozen Monash-pre-trained encoders under time-shift and time-masking, averaged over Weather and ETTh1. Shifts of one full patch (16 steps) map every patch onto a neighbouring slot; sub-patch shifts change patch boundaries.

Family	Model	Shift 4	Shift 16	Mask 20%
Signal-space predictive and reconstructive	NTP	0.023	0.030	0.018
	MAE	0.030	0.013	0.017
	Diffusion	0.008	0.001	0.053
Latent-space alignment	DINO	0.188	0.383	0.078
	JEPA	0.673	0.005	0.479
	LeJEPA	0.062	0.237	0.157
Contrastive	SoftCLT	0.048	0.005	0.064

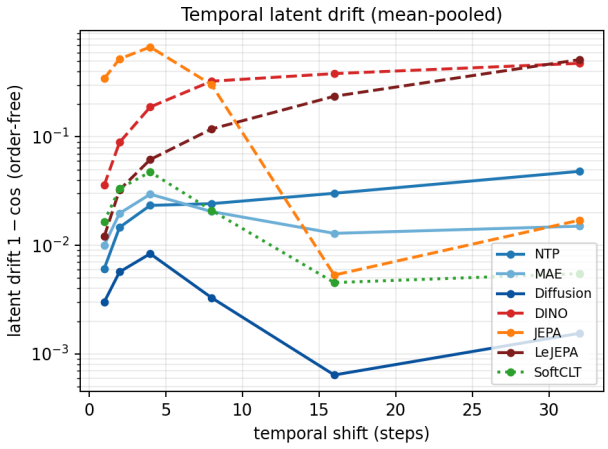


Figure 11: Temporal latent drift as a function of shift (log scale, mean-pooled representation). Curves that dip at multiples of the patch length (16, 32) indicate representations determined by patch content irrespective of position.

Findings. Two patterns emerge (Table 40, Figure 11). First, whole-patch shifts separate the objectives into two groups. MAE, Diffusion, JEPA, and SoftCLT return to near-zero drift at $k{=}16$: after pooling, their representation is largely a function of which patch contents are present, not where they occur. NTP does not ($0.023 \rightarrow 0.030$), consistent with its causal mask making token representations position-dependent by construction. DINO and LeJEPA show no such dip and drift monotonically with shift, so their tokens depend on absolute position and the pooled embedding describes the window as a whole. This is a direct geometric counterpart of the distinction drawn in Section 5: the reconstructive and denoising objectives yield local, content-based representations, while the two view-based objectives yield holistic ones. Second, latent-space alignment encoders are *more* displaced by local masking than signal-space predictive and reconstructive ones (LeJEPA 0.16, JEPA 0.48 vs. NTP and MAE $\approx$0.02). We caution against reading this as a lack of robustness: cosine drift in a near-isotropic embedding (Appendix A.5) registers any change, whereas in an embedding dominated by a few high-variance directions most input changes are invisible to cosine distance even when a linear head can exploit them. The probe therefore characterises what the representations encode, not how much fine-grained information they retain; it does not by itself establish the precision side of the precision–invariance hypothesis.

F.4.3 LeJEPA Analysis

Embedding Geometry. Figure 12 examines mean-pooled embeddings in $\mathbb{R}^D$, whereas SIGReg operates on flattened patch embeddings in $\mathbb{R}^{LD}$. The diagnostics therefore characterize geometry after pooling. The displayed random projections approximately follow a standard normal distribution, and

the covariance spectrum indicates high effective rank and approximately isotropic structure. The PCA visualization also suggests that class-related structure remains present. Downstream classification results provide the quantitative evidence that these representations retain useful discriminative information.

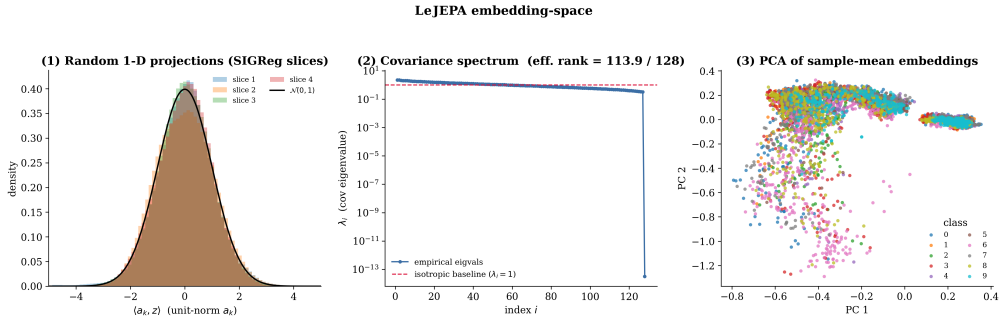


Figure 12: Geometry of LeJEPA's mean-pooled embeddings on SpokenArabicDigits. Left: distributions along sampled random directions compared with a standard normal distribution. Middle: covariance spectrum, with effective rank approximately 114 out of 128 dimensions. Right: PCA visualization colored by class, showing class-related structure within the approximately isotropic representation.

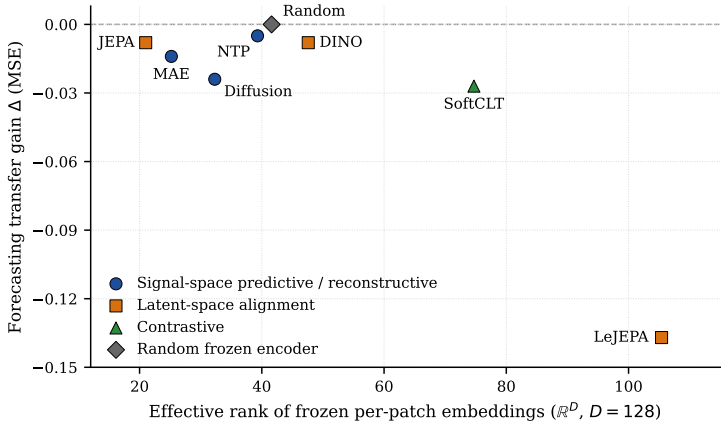


Figure 13: Forecasting transfer gain Δ (Eq. 8; MSE, linear probing, Monash pre-training), averaged over the seven datasets in Table 10, against the effective rank of frozen per-patch embeddings in $\mathbb{R}^D$. Negative transfers indicate degradation. These ranks are computed from per-patch embeddings, whereas Figure 12 examines mean-pooled embeddings. LeJEPA exhibits high effective rank and the largest negative forecasting transfer among the evaluated objectives.

Association with Forecasting Performance. Figure 13 associates LeJEPA's high effective rank with a negative forecasting transfer. The regularization sweep reinforces the coexistence of useful classification representations and poor forecasting performance: all evaluated positive weights avoid collapse and improve mean classification accuracy over random features, while their forecasting MSE remains above the control.

These observations motivate the hypothesis that the geometry encouraged by SIGReg interacts differently with classification and forecasting. They do not establish isotropy as the cause of the forecasting degradation. Removing SIGReg produces a collapsed representation, so that ablation does not provide a usable non-isotropic control. Comparisons with DINO also change the alignment loss, teacher dynamics, and other objective components. Isolating the mechanism requires interventions that vary representation geometry while preserving non-collapsed solutions and otherwise comparable training conditions.

Table 41: Sensitivity of LeJEPA to the SIGReg weight λ. Left: effective rank of the frozen per-patch embedding covariance and per-dimension embedding standard deviation ($D = 128$). Middle: linear-probe forecasting MSE averaged over horizons $\{96, 192, 336, 720\}$. Right: linear-probe classification accuracy (%), with the unweighted mean across nine datasets. The $\lambda = 0$ configuration removes SIGReg and exhibits representation collapse. A dash denotes an unreported result. Datasets: Ethanol (EthanolConcentration), FaceDet (FaceDetection), Handwrit (Handwriting), Heartbt (Heartbeat), JapVowel (JapaneseVowels), SCP1/SCP2 (SelfRegulationSCP1/2), Digits (SpokenArabicDigits), and UWave (UWaveGestureLibrary).

	Isotropy		**Forecasting MSE**					**Classification accuracy (%)**								
λ	Rank	Std	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Ethanol	FaceDet	Handwrit	Heartbt	JapVowel	SCP1	SCP2	Digits	UWave	Mean
0	20.5	0.000	0.718	0.410	0.699	0.341	–	24.71	50.00	3.76	72.20	8.38	49.83	50.00	10.00	12.50	31.26
0.002	107.7	0.972	0.665	0.390	0.486	0.306	0.266	24.33	59.96	13.88	61.95	88.11	83.28	46.11	91.91	74.38	60.43
0.008	109.2	0.964	0.616	0.394	0.444	0.288	0.262	26.09	61.65	13.86	62.54	86.98	83.07	48.89	93.27	71.69	60.89
0.02	114.2	0.981	0.619	0.390	0.451	0.291	0.264	25.86	62.68	16.24	61.95	85.68	80.55	45.56	92.68	70.94	60.24
0.05	114.5	0.983	0.636	0.395	0.456	0.294	0.261	23.95	61.27	17.06	62.44	80.81	80.89	43.89	91.04	77.19	59.84
Random frozen encoder	–	–	0.414	0.350	0.356	0.261	0.248	27.68	54.49	4.31	66.63	93.02	76.72	50.00	77.61	36.56	54.11