LEARNING EVENT REPRESENTATIONS BY LOCALIZING TEMPORAL REVERSALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning has proved useful for time-series classification and forecasting, but common pretraining objectives do not explicitly require features that make individual events easy to recognize and locate. We introduce Temporal Reversal Discrimination (TRD), which trains an encoder to find a short segment reversed within an otherwise unchanged signal. A single linear auxiliary head learns this self-labeled task alongside the original pretraining objective and is discarded afterward, leaving inference unchanged. We evaluate TS2Vec, PatchTST and PMT on heart sounds and wearable operations using frozen features, small labeled training sets and separate labeled validation data. On CirCor heart sounds, linear classifiers fitted on 5% of training patients gain 14–47 points in mean onset macro-F_1 over pretraining without TRD. The benefit persists across the tested classifiers and labeled subsets. On OpenPack, linear classifiers trained on five labeled minutes per training person gain 2–18 points in mean macro-F_1, with mixed results at other settings. PatchTST and PMT also improve on seismic and ECG localization. In the CirCor controls, predicting the reversed positions produces better event features than assigning one reversal label to the window, even when the latter task is learned well. Comparisons with block shuffling show that the better manipulation depends on the encoder and dataset. On CirCor, all three encoders fine-tuned from TRD weights outperform their counterparts fine-tuned after original pretraining or trained from scratch. With the same separately labeled validation cohort, TRD fine-tuning on seven patient groups matches or exceeds the mean onset F_1 of training from scratch on 31 groups.

1 INTRODUCTION

Event-level analysis requires more than a label for an entire recording: seismic arrival times support earthquake location (Zhu & Beroza, 2019; Köhler et al., 2026), while cardiac-wave boundaries characterize physiological timing (Kalyakulina et al., 2020). Obtaining such temporal annotations can require expert inspection (Oliveira et al., 2022; Kalyakulina et al., 2020). Self-supervised time-series representations have proved useful for classification and forecasting (Yue et al., 2022; Nie et al., 2023). Cross-view discrimination and masked-signal reconstruction do not explicitly require semantic event identity and timing to be readily decoded from the learned features. We first keep each encoder fixed and train small classifiers on its features to test how readily event information can be recovered before fine-tuning the encoder (Section 3.1). Figure 1 illustrates the challenge: linear probes fitted on 5% of training patients fragment or confuse heart sounds when reading native TS2Vec (Yue et al., 2022) and PMT (Stangeland et al., 2026) features.

Temporal structure motivates a self-labeled localization task. A seismic arrival is followed by a decaying coda, and within a heartbeat the first heart sound is followed by the second. Reversing an onset–coda interval makes the coda build towards the arrival; reversing an interval containing both complete heart sounds places the second before the first and replays each from its original offset towards its onset. Temporal direction has supplied self-supervision through forward–backward discrimination of whole video clips, wearable-sensor windows and ECG segments (Wei et al., 2018; Saeed et al., 2019; Yuan et al., 2024; Zhang et al., 2023).

Code: https://anonymous.4open.science/r/TRD-4238

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.

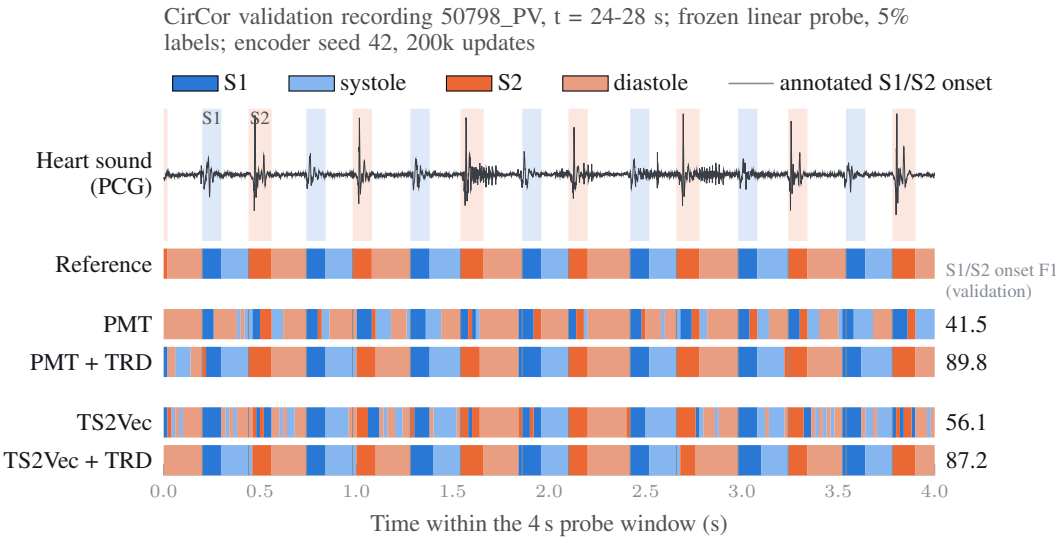


Figure 1: **Local reversal supervision improves low-label event readout.** Four seconds of a CirCor validation recording: reference heart-cycle states, and the states predicted by linear probes on frozen TS2Vec and PMT features, each pretrained with and without TRD. Probes are fitted on 5% of training patients. Native probes fragment or confuse the heart sounds; with TRD they track the reference. Values at right are typed-onset macro-F_1 at 50 ms over the whole validation split for encoder seed 42; Appendix E.5 describes how the excerpt was selected.

To make event identity and timing easier to recover, we introduce Temporal Reversal Discrimination (TRD), an auxiliary objective: reverse a short interval of an unlabeled window and train the encoder to identify the positions inside it. The manipulation labels every valid token without event annotations. One affine auxiliary head predicts these labels, and we add its binary cross-entropy loss to the original pretraining objective. The encoder architecture, original objective and its input views are unchanged; the auxiliary head is discarded after pretraining.

Existing time-series self-supervised encoders can support substantially better event prediction than their original pretraining achieves; TRD, a simple reversal-localization auxiliary, demonstrates this. Related temporal pretexts support localization through synthetic video boundaries (Xu et al., 2021), dense time-arrow features (Gallusser et al., 2023) and out-of-order frame localization (Dave et al., 2024); AnomalyBERT (Jeong et al., 2023) predicts corruption membership for anomaly scoring. TRD adds the same waveform-reversal localization task to contrastive and reconstruction-based objectives; its target identifies a manipulation, while evaluation tests real-event identity and timing in clean signals.

Across TS2Vec, PatchTST and PMT, TRD improves every primary paired comparison at 200k updates: mean gains are 14–47 onset macro-F_1 points with 5% of CirCor training patients and 2–18 macro-F_1 points with five labeled minutes per OpenPack training person, using linear classifiers (Table 1). We contribute:

- **A reversal-localization auxiliary:** one affine head predicts reversed-interval membership at each position alongside existing contrastive or reconstruction-based pretraining, without event annotations or changes to inference.
- **Replicated improvements to existing methods:** CirCor gains cover every encoder seed, classifier and training-label budget; gains remain under structured decoding, and fine-tuning all three encoders beats original-pretraining and scratch references at both budgets. With the same validation cohort, seven-group TRD fine-tuning matches or exceeds the mean onset F_1 of 31-group scratch training. PatchTST and PMT also improve seismic and ECG localization.
- **Controlled characterization:** dense targets outperform pooled targets on CirCor, including when pooled learning succeeds; preferred manipulations vary by method and dataset; CirCor gains persist across labeled subsets without validation-based classifier selection.

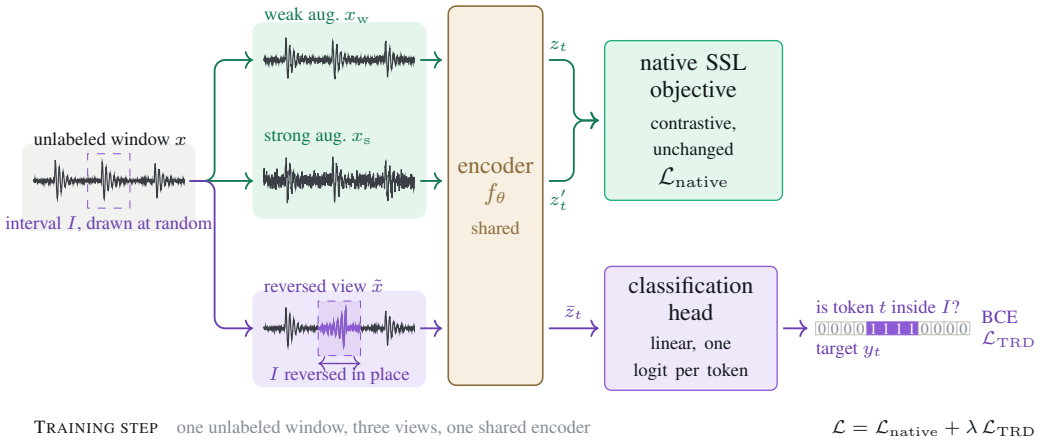


Figure 2: **One TRD training step.** An unlabeled window x yields the native training views, here PMT's x_w and x_s, and a third view $\tilde{x}$ in which one randomly placed interval I is reversed in place. All views pass through the same encoder f_θ. The native loss is unchanged, and a linear head on the tokens of $\tilde{x}$ is trained to mark the tokens inside I (Equations 1–4). The window is synthetic so that the reversal is visible: inside I the coda precedes its onset.

2 Local Reversal as Supervision

In Figure 2, a seismic onset followed by a decaying coda illustrates temporal asymmetry. In the synthetic window, reversing interval I places the coda before its onset, while leaving the surrounding signal unchanged. TRD uses the known location of this manipulation to supervise the encoder's token representations. The construction operates on time-series windows without requiring a particular event type; the seismic signal is an illustration. Our working hypothesis is that locating a reversed interval encourages features sensitive to event morphology and temporal context, such as onset–decay patterns and relations between neighbouring events. Learning these features may make event identity and timing easier to read from unmodified signals.

The task is *self-labeled*, because the targets follow from the manipulation; *dense*, because every valid token receives a target; and *local*, because most of the window remains unchanged context. Because the auxiliary head is a single affine map, training encourages the encoder to expose whether each token lies inside the reversed interval through a linear decision. We add this objective to a *host* (i.e., an encoder together with its native self-supervised objective and training procedure). Section 4.3 studies training without the native objective.

2.1 Formulation

Let $x \in \mathbb{R}^{L \times C}$ be a preprocessed window of L samples and C channels. At every training step and for every window, draw a start a uniformly from $\{0, \dots, L - s\}$ and let $I = [a, a + s)$, where the span s and the fade length k are fixed per dataset with $2k < s < L$. The reversed view $\tilde{x}$ reverses the samples of I jointly across channels and blends them with the original over k samples at each edge through

$$\tilde{x}_{a+j} = \alpha_j \, x_{a+s-1-j} + (1 - \alpha_j) \, x_{a+j}, \qquad j = 0, \dots, s-1, \tag{1}$$

where $\alpha_j \in [0, 1]$ equals one in the interior of I and ramps linearly to zero over the first and last k samples; outside I, $\tilde{x}_t = x_t$. The crossfade softens abrupt changes at the interval edges, and the faded samples count as part of I.

The shared encoder f_θ maps a view to one feature vector per position of its temporal grid. We call these positions *tokens*, $t = 1, \dots, T$: one sample for an encoder that keeps the input resolution, one patch for a patch-based encoder. The reversed view gives tokens $\bar{z}_t = f_\theta(\tilde{x})_t \in \mathbb{R}^D$, and the target marks each token's membership in I,

$$y_t = \mathbf{1}\big[\text{token } t \text{ lies in } I\big], \tag{2}$$

by exact sample membership at input resolution and an at-least-half-overlap rule on the encoder's temporal grid for patch-based encoders. Appendix B specifies each host's token coordinates, overlap

calculation and validity handling. A single affine head produces one logit per token, $u_t = \mathbf{w}^\top \bar{z}_t + b$. The auxiliary loss averages binary cross-entropy over the valid (unpadded) tokens $\mathcal{V}$ of the window, defined as

$$\mathcal{L}_{\text{TRD}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} H(u_t, y_t), \tag{3}$$

where H denotes binary cross-entropy with logits. The host computes its native loss $\mathcal{L}_{\text{native}}$ from the tokens of its own views, z_t and z'_t in Figure 2, exactly as the host would without TRD, and the total pretraining loss is

$$\mathcal{L} = \mathcal{L}_{\text{native}} + \lambda \mathcal{L}_{\text{TRD}}. \tag{4}$$

The principal native-versus-TRD comparisons use $\lambda = 1.5$ across hosts and datasets, with $\lambda = 0$ for the native control. Appendix G.1 examines the auxiliary weight, interval span and token-target convention. The head adds $D + 1$ parameters.

2.2 Adding TRD to a self-supervised recipe

The notation "+TRD" denotes adding the auxiliary objective to the same host. We use three distinct recipes: TS2Vec (Yue et al., 2022), a dilated-convolution encoder trained by hierarchical contrast over two overlapping crops; PatchTST (Nie et al., 2023), a patch transformer trained by masked-patch reconstruction; and PMT (Stangeland et al., 2026), a memory transformer trained contrastively on two augmented views. We chose these varied architectures and native objectives to test whether the addition is useful beyond a single learning recipe. The attachment is the same in each; Figure 2 illustrates the PMT case. The host builds its native views as usual, and TRD adds one further view, $\tilde{x}$, through the same encoder. The head reads the encoder's final tokens, before any objective-specific projection, and $\bar{z}_t$ never enters a contrastive negative pool or a reconstruction target. For TS2Vec, the auxiliary head uses the unpooled timestamp features, while native contrastive learning retains its temporal max-pooling hierarchy (Yue et al., 2022). Reversal is applied to the normalized waveform before any patching, so patch-based hosts see reversed samples, not reordered embeddings.

Span and fade are specified in physical units per dataset and shared across hosts (Table B.1). Appendices G and B give their selection basis, sensitivity and implementation details.

TRD encodes one additional view during pretraining and leaves inference unchanged. Across six CirCor/OpenPack profiles on one NVIDIA GH200 in FP32/TF32, its measured training-core update overhead is 3–29% (Appendix H).

3 Event-representation evaluation

We ask three questions: does TRD improve event prediction when the encoder is kept fixed; do the gains remain after sequence decoding or supervised fine-tuning; and how do position-wise labels and the choice of signal manipulation affect the results? This section tests usefulness; Section 4 compares the pretraining tasks.

3.1 Setup

We first keep each pretrained encoder fixed and train a classifier on its features. On CirCor, we also fine-tune the encoder and classifier together. CirCor and OpenPack are the main test datasets; CREW and LUDB provide supporting localization studies. Within each method and dataset, runs with and without TRD share initialization, pretraining data, original augmentations, optimizer and training schedule. Paired classifiers use the same training and selection procedures. Pretraining budgets are matched within comparisons; table captions give their endpoints and Appendix A gives schedules.

Principal datasets and tasks. The CirCor DigiScope dataset (Oliveira et al., 2022) contains heart-sound recordings. We predict heart-cycle states and evaluate the onsets of the first and second sounds (S1 and S2): a predicted onset must match both the sound type and the reference time within 50 ms, with F_1 averaged over the two types. On OpenPack (Yoshimura et al., 2024), a dataset of packaging work, we use left-wrist acceleration to recognize work operations and report macro-F_1 at 1 Hz over scored positions, excluding null operations and padding. Pretraining uses only the training partitions.

Table 1: **Held-out event prediction at 200k updates.** F_1 (%), mean ± SD over three encoder seeds; heads fit 31 CirCor patient groups or five minutes per OpenPack training person, with additional validation supervision (Section 3.1). Bold scores: highest mean per column within each of the three host blocks. Δ: +TRD minus Native (positive-seed count), bold when all seeds improve.

	CirCor (5%)		OpenPack (5 min)	
Recipe	Linear	Temporal	Linear	Temporal
TS2Vec, Native	57.58 ± 1.41	82.63 ± 1.31	34.07 ± 0.62	42.75 ± 1.09
with SoftCLT-t	57.17 ± 1.85	82.53 ± 0.98	35.60 ± 0.53	**43.48** ± 0.85
+TRD	**91.44** ± 0.47	**95.16** ± 0.08	**36.32** ± 2.01	39.48 ± 2.47
Δ (seeds)	**+33.86** (3/3)	**+12.54** (3/3)	**+2.25** (3/3)	−3.27 (1/3)
PatchTST, Native	41.63 ± 2.91	55.74 ± 2.62	48.07 ± 0.95	50.17 ± 1.39
+TRD	**55.86** ± 0.95	**80.17** ± 0.28	**53.39** ± 4.46	**51.95** ± 4.16
Δ (seeds)	**+14.23** (3/3)	**+24.43** (3/3)	**+5.32** (3/3)	+1.78 (1/3)
PMT, Native	45.01 ± 0.95	67.39 ± 0.52	30.38 ± 1.69	30.27 ± 0.74
+TRD	**92.33** ± 0.34	**95.44** ± 0.04	**48.54** ± 0.53	**47.17** ± 0.54
Δ (seeds)	**+47.32** (3/3)	**+28.06** (3/3)	**+18.15** (3/3)	**+16.90** (3/3)

Recordings from linked patient identities on CirCor and repeated person identities on OpenPack are kept together when partitioning training, validation and test (Appendix A).

Readouts and supervision. We train either a linear classifier or a small temporal convolutional classifier on each frozen encoder's features; the latter spans 180 ms on CirCor (Appendix A.2). We call the downstream classifier a readout and its labeled training-set size the fitting budget. Training uses 7 or 31 CirCor patient groups (1%, 5%), or one or five labeled minutes per OpenPack training person. We select the training epoch using another 130 labeled CirCor validation groups or three OpenPack validation people, then test the selected classifier without refitting. The primary comparison uses the linear classifier with the larger labeled training subset. A separate CirCor analysis uses five different labeled subsets and evaluates classifiers at a fixed final epoch, without validation selection (Appendix D.2).

Supporting localization. The Curated Regional Earthquake Waveforms (CREW) dataset (Aguilar Suarez & Beroza, 2024) provides three-component seismic records with P/S arrivals. We evaluate PatchTST and PMT features with downstream classifiers fitted on 74 or 147 training traces, fixed P/S thresholds of 0.5 and no calibration labels, using test pick-F_1 at ±3 s. On the Lobachevsky University Electrocardiography Database (LUDB) (Kalyakulina et al., 2020), all three hosts use downstream classifiers fitted on 8, 16, 32 or 64 beats, with 52 separate development beats for threshold calibration and P/T onset scoring at ±40 ms. Its annotation-placed one-second ECG inputs supply beat location; TS2Vec uses a sample-grid affine readout, while patch-based hosts use learned upsampling.

Replication. The main comparison uses one labeled training subset and fixed classifier initialization across three paired encoder seeds. We report sample standard deviations and counts of positive paired differences; Appendix A describes sampling for the repeated-subset and supporting studies.

3.2 Frozen event prediction

Keeping the encoder fixed lets us ask whether reversal localization transfers to real-event prediction before event labels can reshape its features. At 200k updates, TRD improves the primary linear-classifier result in all 18 paired comparisons across three encoders, two main datasets and three seeds. CirCor gains extend to both classifiers and labeled training-set sizes; secondary OpenPack results vary across methods. Table 1 reports both classifiers at the larger labeled training-set size, with temporal SoftCLT (Lee et al., 2024) as a same-encoder TS2Vec comparison; Appendix D gives both sizes and additional references.

Heart-sound localization. On CirCor, every seed improves at both readouts and fitting budgets, including seven training patient groups (1%; Table D.1). Allowing the classifier to combine nearby features does not remove the benefit: TS2Vec rises from 82.63 to 95.16. A separate 50k validation diagnostic shows that the gain includes sound identity, not just onset timing: with true intervals supplied, TS2Vec/CirCor S1/S2 balanced accuracy rises from 89.50 to 95.78 with TRD (Appendix E.4).

Table 2: **Supporting localization across encoders.** Test F_1 (%), mean ± sample SD over three paired runs; Δ and s_Δ describe paired gains. CREW uses P/S matching at ±3 s, fixed thresholds and no calibration labels. LUDB uses P/T matching at ±40 ms within annotation-placed beats, with 52 additional calibration beats. Bold marks a higher TRD mean with three positive pairs. TS2Vec/CREW has no completed endpoint because full-length temporal contrast exceeds available GPU memory; head interfaces and all budgets are in Appendices D.7 and D.6.

Host	Native	+TRD	$\Delta \pm s_\Delta$	Positive
CREW: 20k updates, 147 fitting traces				
PatchTST	21.71 ± 0.27	$\mathbf{28.75} \pm 1.18$	$+7.04 \pm 1.44$	3/3
PMT	33.70 ± 1.81	$\mathbf{66.92} \pm 2.41$	$+33.21 \pm 0.61$	3/3
LUDB: 5k updates, 64 fitting beats				
TS2Vec	58.83 ± 9.11	54.01 ± 3.89	-4.82 ± 12.17	1/3
PatchTST	52.85 ± 9.82	$\mathbf{81.86} \pm 2.38$	$+29.01 \pm 10.53$	3/3
PMT	63.45 ± 4.39	$\mathbf{69.82} \pm 7.44$	$+6.38 \pm 5.01$	3/3

With TRD, all three hosts exceed both their native and untrained-encoder references on CirCor's primary linear readout (Appendix F). Native PatchTST and PMT score below their untrained references on CirCor, whereas native TS2Vec improves. On OpenPack, all three native recipes exceed their untrained references, and adding TRD improves the primary mean further.

Large gains from few labels could depend on which patients are labeled or where pretraining stops, so we vary both. CirCor gains persist across five labeled training subsets with classifiers evaluated at a fixed final epoch, without validation-based epoch selection (Appendix D.2). Selecting encoders on validation from initialization and checkpoints at 25k, 50k, 100k and 200k also preserves positive gains in every CirCor pair. Primary test gains under this policy are 31.64, 6.20 and 46.93 points for TS2Vec, PatchTST and PMT, respectively (Appendix D.3).

Wearable work-operation recognition. On OpenPack, we ask whether the benefit extends beyond physiological onsets to work operations recognized from wrist motion. PMT improves every seed at both readouts and fitting budgets, with primary and temporal gains of 18.15 and 16.90 points. Among the three principal recipes in Table 1, PatchTST+TRD attains the highest primary mean (53.39); its temporal gains are positive in only one seed per budget. TS2Vec improves on the primary readout and loses on the other three classifier/budget combinations (Table D.2). Appendix D.4 reports per-person results; Figure D.1 shows PMT's operation-level confusions.

At the larger fitting budget, TRD outperforms temporal SoftCLT (Lee et al., 2024) on TS2Vec across every CirCor seed and both readouts. On OpenPack, TRD has the higher linear mean while SoftCLT has the higher temporal mean (Table 1); Appendix C specifies the comparator. A supplementary PPT comparison is mixed on CirCor but improves every OpenPack seed at both readouts and fitting budgets (Appendix D.5).

Supporting seismic and ECG localization. TRD improves PatchTST and PMT in every paired seed at both CREW fitting budgets and all four LUDB budgets (Table 2; Appendices D.7 and D.6). Figure D.2 illustrates the PatchTST onset readout on individual beats and across the eligible LUDB test population. Transfer is not uniform: TS2Vec has negative mean LUDB gains at each budget, with mixed seed signs.

3.3 Structured readout and end-to-end adaptation

Downstream modeling that adds temporal structure or updates the encoder could close the gap; on CirCor, we evaluate structured decoding and supervised fine-tuning separately at the principal 200k pretraining endpoint.

Structured readout. TRD retains a positive advantage under shared structured decoding in every test seed on all three encoders. One transition/duration decoder is fitted from the same 31 training patient groups without additional calibration labels and uses whole-recording future context and backtracking. For TS2Vec, decoded F_1 rises from 90.97 to 96.29; decoded gains are 5.32 points for TS2Vec, 25.55 for PatchTST and 19.00 for PMT (Appendix E.2). The prior narrows the TS2Vec and PMT gaps but widens PatchTST's.

Table 3: **CirCor frozen readout and supervised adaptation.** Test typed-onset macro-F_1 at 50 ms (%), mean ± sample SD over three seeds. The columns compare frozen Native+TRD features with three end-to-end starting points; all use the same temporal-head architecture and 7 or 31 fitting groups (1%, 5%). All 130 validation groups support epoch selection. Bold marks the highest mean in each row. Fitting policies differ between frozen and adapted pipelines (Appendix E.3).

		Frozen encoder	End-to-end from		
Host	Groups	+TRD	Scratch	Native	+TRD
TS2Vec	7	92.73 ± 0.30	80.26 ± 1.44	87.49 ± 0.36	$\mathbf{92.75} \pm 0.77$
TS2Vec	31	95.16 ± 0.08	92.46 ± 0.24	93.72 ± 0.13	$\mathbf{95.94} \pm 0.26$
PatchTST	7	74.95 ± 0.46	70.44 ± 0.76	69.06 ± 1.01	$\mathbf{80.92} \pm 2.69$
PatchTST	31	80.17 ± 0.28	77.94 ± 0.44	78.84 ± 0.26	$\mathbf{87.34} \pm 2.23$
PMT	7	$\mathbf{94.06} \pm 0.94$	80.00 ± 0.64	79.37 ± 0.81	93.09 ± 0.15
PMT	31	$\mathbf{95.44} \pm 0.04$	90.07 ± 0.28	89.91 ± 0.12	94.64 ± 0.14

Table 4: **Supervision granularity and corruption choice.** CirCor test typed-onset macro-F_1 at 50 ms (%), mean ± sample SD over three 50k-update encoders. Linear heads fit 31 training patient groups (5%) and use validation-selected epochs within a 120-epoch budget. In (a), p is corruption probability and +TRD denotes added local dense reversal. The + conditions in (b) retain the native objective and use dense targets with $p = 1$ and the same local geometry. Bold marks column-best means in (a); in (b), it marks the higher auxiliary mean only when that operator beats the other in all three paired seeds. Bolding is not a significance test.

(a) Reversal supervision

Task	p	TS2Vec	PMT
Native	—	59.68 ± 0.95	45.43 ± 1.83
Whole / pooled	0.5	65.93 ± 0.79	64.66 ± 1.50
Local / pooled	0.5	59.25 ± 2.34	70.75 ± 5.37
Local / dense	0.5	73.80 ± 0.76	88.61 ± 0.28
+TRD	1	$\mathbf{85.73} \pm 2.08$	$\mathbf{90.38} \pm 0.56$

(b) Corruption choice across hosts

Host	Native	+Reversal	+Shuffle
TS2Vec	59.68 ± 0.95	85.73 ± 2.08	$\mathbf{94.90} \pm 0.32$
PatchTST	39.97 ± 0.93	$\mathbf{54.79} \pm 2.12$	44.18 ± 4.95
PMT	45.43 ± 1.83	90.38 ± 0.56	90.11 ± 0.28

End-to-end adaptation. On CirCor, fine-tuning from TRD weights outperforms both fine-tuning after original pretraining and training from scratch. Every paired run improves onset and state macro-F_1, across all three encoders and both labeled training-set sizes (Appendix E.3). Using seven training patient groups, TRD fine-tuning matches or exceeds the mean onset score of scratch training on 31 groups: 92.75 versus 92.46 for TS2Vec, 80.92 versus 77.94 for PatchTST, and 93.09 versus 90.07 for PMT. Under the evaluated procedures, a small temporal classifier on frozen TRD features also exceeds both reference means for all three encoders at both training-set sizes (Table 3).

4 Characterizing Reversal Supervision

We now examine the roles of the auxiliary targets, signal manipulation and original objective through separate matched comparisons (Appendix E.1).

4.1 Dense versus pooled targets

Position-wise targets distinguish TRD from earlier whole-input reversal pretexts, so we compare dense prediction (one target per position) with pooled prediction (one target per window). The pooled conditions use whole-window or local reversal; dense prediction uses local reversal. All three conditions corrupt each auxiliary input with probability $p = 0.5$; clean inputs receive negative targets. The two local tasks share the corruption distribution, while standard TRD reverses an interval in every auxiliary input ($p = 1$). Both pooled tasks use a mean-pooled affine binary auxiliary head.

With the same local-corruption distribution, dense prediction improves over local pooled prediction by 14.56 points for TS2Vec and 17.85 for PMT on the primary test score, with every paired seed positive—Table 4(a). Both dense conditions also outperform the whole-window pooled task, and $p = 1$ improves further on $p = 0.5$. Changing only the pooled task from global to local has host-

dependent effects. PMT learns both pooled pretexts well, yet dense supervision yields stronger event features; TS2Vec's local-pooled pretext remains near chance (Appendix E.1). Successful pooled pretext learning is therefore insufficient for the best event readout in this comparison.

4.2 Corruption choice across recipes

Position-wise labels do not determine which manipulation yields better event features. We compare reversal and block shuffling using the same interval geometry and auxiliary-loss weight. Shuffling changes the arrangement of blocks but preserves sample order inside each block. On CirCor, shuffle leads on TS2Vec, reversal on PatchTST, and PMT's primary means are close—Table 4(b). Across the larger-budget manipulation comparisons, reversal has the higher mean in three of four domains for each of PatchTST and PMT, with narrow, mixed-seed differences on PatchTST/CREW and PMT/CirCor; shuffle leads on TS2Vec in all three completed domains (Appendix E.1). The fixed readouts are linear on CirCor/OpenPack and downstream localization heads on CREW/LUDB; endpoints are 50k, 50k, 20k and 5k updates, respectively. On TS2Vec/CirCor, reversal also exceeds RMS-matched gain and external replacement on the primary readout, while independent targets reduce performance.

The pooled-target comparison supports dense supervision for the tested reversal task. At the manipulation-study endpoints, reversal and shuffle exceed the native mean in 30 and 40 of the 40 evaluated settings, respectively; these settings share encoders across readouts and budgets (Appendix E.1). Shuffle is more consistently beneficial in this matrix, while reversal has higher means in several patch-encoder settings. The findings separate two design choices: the granularity of supervision and the temporal manipulation supplying it.

4.3 Removing the native objective

Removing the original objective shows whether reversal localization can support useful event features on its own. At 200k updates, training with TRD alone gives PMT/CirCor performance close to training with both objectives. It is less reliable for TS2Vec/CirCor, where two of three seeds fail. Appendix F gives the complete comparisons and training dynamics.

5 Related work

Temporal self-supervision. Predictive and contrastive approaches include CPC, TS-TCC, TS2Vec and SoftCLT (van den Oord et al., 2018; Eldele et al., 2021; Yue et al., 2022; Lee et al., 2024); reconstruction approaches include masked-patch pretraining in PatchTST and past-to-current reconstruction in TimeSiam (Nie et al., 2023; Dong et al., 2024). Domain-specific event pretraining includes ATST-Frame for sound-event detection (Li et al., 2024), Beat-SSL for ECG, whose wave-segmentation comparison covers frozen TS2Vec on LUDB (Rizqyawan et al., 2026), and SeisLM for seismic event detection and phase picking (Liu et al., 2024). We instead measure what TRD adds to existing general-purpose time-series pretraining methods, rather than rank pretraining frameworks designed for one event domain. Temporal SoftCLT provides a same-encoder comparison; TimeDRL (Chang et al., 2024) and PPT supply additional numerical references under the configurations in Appendix C, with results in Appendix D.

Temporal order and time-arrow supervision. Shuffle and Learn (Misra et al., 2016) verifies video-frame ordering, SelfTime (Fan et al., 2020) learns inter-sample and intra-series temporal relations, and PPT (Kim et al., 2025) uses patch permutations with sequence-level order-consistency and contrastive objectives. Forward–backward video discrimination (Wei et al., 2018), activity-pretraining time flipping (Saeed et al., 2019), wearable arrow-of-time, permutation and time-warp prediction (Yuan et al., 2024), and ECG temporal/amplitude reversal detection (Zhang et al., 2023) assign whole-input transformation labels. The permutation task of Yuan et al. (2024) shuffles four chunks and predicts whether the input was transformed; our shuffle control instead permutes one interval and supervises its dense membership (Section 4.2).

Time-arrow supervision can also yield dense features: TAP (Gallusser et al., 2023) classifies consecutive microscopy image-pair order with a permutation-equivariant, spatially pooled head, producing pixel-level features for cell detection, segmentation and state classification; Chen et al. (2024) ex-

tend them to cell-event recognition. Xu et al. (2021) synthesize temporal video boundaries and classify their types to improve downstream localization. A close temporal-localization precedent is Dave et al. (2024): they localize out-of-order video frames through frame-wise binary prediction and combine temporal pretexts with contrastive learning; their frame-level formulations outperform clip-level counterparts. They motivate finer temporal targets partly by increasing pretext difficulty; our PMT comparison also shows a dense-target advantage when both local tasks attain high training accuracy (Section 4.1). JiTTER (Nam & Park, 2025) reconstructs block- and frame-shuffled audio sequences for sound-event detection; adding temporal flipping slightly lowers PSDS in its evaluated multitask reconstruction setup. TRD instead reverses samples within one contiguous waveform interval, jointly across channels before tokenization, and predicts interval membership without reconstructing the input. Unlike patch permutation, sample reversal can also change within-patch morphology.

Dense corruption discrimination. ELECTRA (Clark et al., 2020) identifies replaced text tokens, and TransEHR (Xu et al., 2023) uses value-replacement discrimination for clinical time series. AnomalyBERT (Jeong et al., 2023) predicts timestamp-level interval membership for replacement, length adjustment and peak-noise corruptions, deploying these predictions as anomaly scores. DiffCSE (Chuang et al., 2022) combines contrastive learning with replaced-token detection conditioned on a sentence embedding. CutPaste (Li et al., 2021) learns anomaly-detection features by classifying locally manipulated images; DRAEM (Zavrtanik et al., 2021) combines reconstruction and discrimination for anomaly localization. E-SSL (Dangovski et al., 2022) and AugSelf (Lee et al., 2021) add whole-input transformation-related prediction losses to existing representation learners; TRD instead supervises the location of a waveform manipulation to improve real-event prediction across time-series pretraining methods.

6 Discussion and Conclusion

Implications for event representations. Adding reversal localization improves all three pretraining methods in the primary event-prediction comparisons on CirCor and OpenPack. On CirCor, the gains remain after shared sequence decoding; they also remain under supervised fine-tuning at both labeled training-set sizes. They occur both where original pretraining helps and where it performs worse than initialization (Section 3.2). The PMT control shows why the prediction task matters: successfully assigning one reversal label to a window produces less useful event features than labeling positions within the same local corruption. Together, these findings support aligning pretraining supervision with the distinctions required for event prediction.

Interaction with the native objective. Corruption rankings vary across both recipes and datasets (Section 4.2). TS2Vec (Yue et al., 2022) combines timestamp-level contrast with losses on progressively max-pooled features. We hypothesize that shuffle's preservation of within-block patterns while changing their context may suit this multiscale objective; reversal also changes local morphology. The within-host reversals of the manipulation ranking on PatchTST and PMT indicate that architecture alone does not determine the preference; pooling has not been isolated as its cause.

Scope and further directions. Fine-tuning is evaluated on CirCor, and structured decoding uses one sequence-prior pipeline on that dataset. Temporal direction motivates the objective; its contribution remains to be separated from interval-insertion and crossfade cues (Appendix I).

Conclusion. Reverse a short interval of an unlabeled signal, and train the encoder to find it. This lightweight auxiliary improves the primary event predictions and provides better starting representations for CirCor fine-tuning, without adding computation at inference. The same encoder architectures can therefore support substantially better event prediction than their original pretraining achieves.

AI use statement

We used generative AI tools to assist with method implementation, development of data-preprocessing and analysis code, experimental design and critique, hypothesis refinement, result interpretation, and the formulation and derivation of a supporting mathematical argument. These tools also assisted with drafting and editing the manuscript and identifying related literature. Gen-

erative AI was not used to generate data. The authors reviewed and verified all AI-assisted code, derivations, text and claims and take responsibility for the final content of this work.

Ethics statement

This work studies self-supervised representation learning using existing research datasets. We identify no additional ethical concerns specific to the proposed method. The physiological experiments evaluate representation quality and do not establish suitability for clinical use.

Reproducibility statement

Section 2 defines the objective and its integration with the evaluated encoders. Appendices A and B describe data preparation, partitions, training settings and readout protocols. We report paired-seed results, label-draw robustness and controlled analyses in Appendices D and E, including the adaptation and selection protocol in Appendix E.3, with design sensitivity in Appendix G. The anonymous code repository is https://anonymous.4open.science/r/TRD-4238.

References

Albert Leonardo Aguilar Suarez and Greg Beroza. Curated regional earthquake waveforms (CREW) dataset. *Seismica*, 3(1), 2024. doi: 10.26443/seismica.v3i1.1049.

Ching Chang, Chiao-Tung Chan, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. TimeDRL: Disentangled representation learning for multivariate time-series. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 625–638. IEEE, 2024. doi: 10.1109/ICDE60146.2024.00054. URL https://doi.org/10.1109/ICDE60146.2024.00054.

Cangxiong Chen, Vinay P. Namboodiri, and Julia E. Sero. Self-supervised representation learning for cell event recognition through time arrow prediction. arXiv preprint arXiv:2411.03924, 2024. URL https://arxiv.org/abs/2411.03924.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Conf. North Amer. Ch. Assoc. Comput. Ling. (NAACL): Human Language Technologies*, pp. 4207–4218, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.311. URL https://aclanthology.org/2022.naacl-main.311/.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Inter. Conf. Learn. Represent. (ICLR)*, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *Inter. Conf. Learn. Represent. (ICLR)*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Ishan Rajendrakumar Dave, Simon Jenni, and Mubarak Shah. No more shortcuts: Realizing the potential of temporal self-supervision. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 38, pp. 1481–1491, 2024. doi: 10.1609/aaai.v38i2.27913. URL https://ojs.aaai.org/index.php/AAAI/article/view/27913. Issue 2.

Jiaxiang Dong, Haixu Wu, Yuxuan Wang, Yun-Zhong Qiu, Li Zhang, Jianmin Wang, and Mingsheng Long. TimeSiam: A pre-training framework for siamese time-series modeling. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Inter. Conf. Mach. Learn. (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11412–11436. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dong24e.html.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In Zhi-Hua Zhou (ed.), *Inter. Joint Conf. Artif. Intell. (IJCAI)*, pp. 2352–2359. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi: 10.24963/ijcai.2021/324. URL https://doi.org/10.24963/ijcai.2021/324. Main Track.

Haoyi Fan, Fengbin Zhang, and Yue Gao. Self-supervised time series representation learning by inter-intra relational reasoning. arXiv preprint arXiv:2011.13548, 2020. URL https://arxiv.org/abs/2011.13548.

Benjamin Gallusser, Max Stieber, and Martin Weigert. Self-supervised dense representation learning for live-cell microscopy with time arrow prediction. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Inter. Conf. Med. Imag. Comput.-Assist. Interv. (MICCAI)*, volume 14227 of *Lecture Notes in Computer Science*, pp. 537–547, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43993-3. doi: 10.1007/978-3-031-43993-3_52. URL https://doi.org/10.1007/978-3-031-43993-3_52.

Yungi Jeong, Eunseok Yang, Jung Hyun Ryu, Imseong Park, and Myungjoo Kang. AnomalyBERT: Self-supervised transformer for time series anomaly detection using data degradation scheme. In *Inter. Conf. Learn. Represent. Wksps. (ICLRW): Machine Learning for IoT: Datasets, Perception, and Understanding*, 2023. URL https://arxiv.org/abs/2305.04468.

A. I. Kalyakulina, I. I. Yusipov, V. A. Moskalenko, A. V. Nikolskiy, K. A. Kosonogov, G. V. Osipov, N. Yu. Zolotykh, and M. V. Ivanchenko. LUDB: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:186181–186190, 2020. doi: 10.1109/ACCESS.2020.3029211.

Jaeho Kim, Kwangryeol Park, Sukmin Yun, and Seulki Lee. PPT: Patch order do matters in time series pretext task. In *Inter. Conf. Learn. Represent. (ICLR)*, 2025. URL https://openreview.net/forum?id=7zwIEbSTDy.

Andreas Köhler, Tord Sture Stangeland, S. Mæland, and B. Dando. Adapting deep learning phase detectors for seismic array processing. *Journal of Geophysical Research: Machine Learning and Computation*, 3(3):e2026JH001249, 2026. doi: 10.1029/2026JH001249. URL https://doi.org/10.1029/2026JH001249.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Adv. Neural Inf. Process. Sys. (NeurIPS)*, volume 34, pp. 17710–17722, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/94130ea17023c4837f0dcdda95034b65-Abstract.html.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *Inter. Conf. Learn. Represent. (ICLR)*, 2024. URL https://openreview.net/forum?id=pAsQSWlDUf.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 9664–9674, June 2021.

Xian Li, Nian Shao, and Xiaofei Li. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 32:1336–1351, 2024. doi: 10.1109/TASLP.2024.3352248. URL https://doi.org/10.1109/TASLP.2024.3352248.

Tianlin Liu, Jannes Münchmeyer, Laura Laurenti, Chris Marone, Maarten V. de Hoop, and Ivan Dokmanić. SeisLM: a foundation model for seismic waveforms, 2024. URL https://arxiv.org/abs/2410.15765.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.),

European Conf. Comput. Vis. (ECCV), volume 9905 of *Lecture Notes in Computer Science*, pp. 527–544, Cham, 2016. Springer International Publishing. doi: 10.1007/978-3-319-46448-0_32. URL https://doi.org/10.1007/978-3-319-46448-0_32.

Hyeonuk Nam and Yong-Hwa Park. JiTTER: Jigsaw temporal transformer for event reconstruction for self-supervised sound event detection, 2025. URL https://arxiv.org/abs/2502.20857.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Inter. Conf. Learn. Represent. (ICLR)*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares, A. Elola, A. Bahrami Rad, R. Sameni, G. D. Clifford, and M. T. Coimbra. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE J. Biomed. Health Inform.*, 26(6):2524–2535, 2022. doi: 10.1109/JBHI.2021.3137048.

Muhammad Ilham Rizqyawan, Peter Macfarlane, Stathis Hadjidemetriou, and Fani Deligianni. Beat-SSL: Capturing local ECG morphology through heartbeat-level contrastive learning with soft targets. In *IEEE Inter. Sym. Biomed. Imag. (ISBI)*, 2026. doi: 10.1109/ISBI61048.2026.11515777. URL https://doi.org/10.1109/ISBI61048.2026.11515777.

Aaqib Saeed, Tanir Özçelebi, and Johan J. Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):61, June 2019. doi: 10.1145/3328932. URL https://doi.org/10.1145/3328932.

Tord Sture Stangeland, Andreas Köhler, Steffen Mæland, and Adín Ramírez Rivera. Progressive memory transformers: Memory-aware attention for time series. In *Adv. Neural Inf. Process. Sys. (NeurIPS)*, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. URL https://arxiv.org/abs/1807.03748.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 8052–8060, 2018. doi: 10.1109/CVPR.2018.00840. URL https://donglaiw.github.io/proj/learnAoT/index.html.

Gideon Weiss. Time-reversibility of linear stochastic processes. *Journal of Applied Probability*, 12(4):831–836, 1975. doi: 10.1017/S0021900200048804. URL https://doi.org/10.1017/S0021900200048804.

Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martínez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *IEEE Inter. Conf. Comput. Vis. (ICCV)*, pp. 7220–7230, October 2021.

Yanbo Xu, Shangqing Xu, Manav Ramprassad, Alexey Tumanov, and Chao Zhang. TransEHR: Self-supervised transformer for clinical time series data. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 623–635. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/xu23a.html.

Naoya Yoshimura, Jaime Morales, Takuya Maekawa, and Takahiro Hara. OpenPack: A large-scale dataset for recognizing packaging works in IoT-enabled logistic environments. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 90–97, 2024. doi: 10.1109/PerCom59722.2024.10494448.

Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digit. Med.*, 7(1):91, April 2024. doi: 10.1038/s41746-024-01062-3. URL https://www.nature.com/articles/s41746-024-01062-3.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 36(8):8980–8987, June 2022. doi: 10.1609/aaai.v36i8.20881. URL https://ojs.aaai.org/index.php/AAAI/article/view/20881.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *IEEE Inter. Conf. Comput. Vis. (ICCV)*, pp. 8330–8339, October 2021.

Wenrui Zhang, Shijia Geng, and Shenda Hong. A simple self-supervised ECG representation learning method via manipulated temporal–spatial reverse detection. *Biomed. Signal Process. Control*, 79:104194, 2023. doi: 10.1016/j.bspc.2022.104194. URL https://doi.org/10.1016/j.bspc.2022.104194.

Weiqiang Zhu and Gregory C. Beroza. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2019. doi: 10.1093/gji/ggy423. URL https://doi.org/10.1093/gji/ggy423.

Table A.1: Labels used to fit and select the readouts. Evaluation labels are excluded from these budgets.

Dataset	Head fitting	Additional selection/calibration
CirCor (principal)	7 or 31 patient groups	130 validation patient groups
CirCor (five draws)	7 or 31 patient groups	None (fixed-final heads)
CirCor (adaptation)	7 or 31 patient groups	130 validation patient groups
OpenPack	1 or 5 min per training person	3 validation people
CREW	74 or 147 traces	None
LUDB	8, 16, 32 or 64 beats	52 development beats

A Evaluation details

This appendix distinguishes the principal CirCor/OpenPack protocol from the supporting localization studies. Table A.1 records head-fitting and additional selection supervision; evaluation labels are separate from these budgets.

A.1 Datasets, partitions and preprocessing

CirCor. Inputs are one heart-sound channel at 4 kHz, demeaned and standardized independently per recording. The training, validation and test partitions contain 2,212/486/465 recordings from 610/130/132 groups, respectively; linked visits and recording sites remain together. The training source contains 13,627 four-second windows. Windows advance by their full length, with a final full-length tail window appended when necessary; that tail may overlap its predecessor. Window construction does not use event labels, and regions without annotation are excluded from supervised loss and scoring. Head-fitting budgets are 7 or 31 training patient groups.

OpenPack. Inputs are sixty-second windows of three-channel left-wrist acceleration, regularized and anti-aliased to 30 Hz, clipped at $\pm 3\,g$ and mapped to $[0, 1]$. Repeated participant identifiers are linked before partitioning: 70/14/20 training/validation/test sessions represent 10/3/3 people. The training source contains 2,196 windows, with the same final-full-tail rule as CirCor. Other sensors, timestamps and labels are not encoder inputs. Null operations and padding are excluded from supervised loss and scoring. Validation and test contain 28,502 and 34,361 supported scoring seconds; head fitting uses one or five labeled minutes per training person.

CREW. The matched CREW dataset, preprocessing and sampling protocol are described in Appendix D.7. The smaller head-fitting subset is nested within the larger under the common sampling policy.

LUDB. After preprocessing, LUDB contains 981/207/219 training/development/test beats from 137/29/29 represented subjects. Inputs are annotation-conditioned, one-second, twelve-lead ECG crops at 500 Hz, with independent per-trace demeaning and peak normalization. Lead-II P and T annotations define the target onsets; QRS is not predicted. Crop construction places the P onset 40–140 samples into the window; P-absent beats use a QRS-peak reference shifted by 102 samples. A neighbour guard and feasible-window checks exclude neighbouring beats. This preparation supplies beat-location information before the frozen-readout task.

Partition membership. The principal and standalone CirCor/OpenPack encoders use training-partition source pools with no linked-group overlap with validation or test. LUDB subject sets are disjoint after preprocessing. Five assigned subjects contribute no packaged beats; their partition assignments and rejected-beat identities are unavailable. For the LUDB duration analysis (Appendix F.3), exact fitting and calibration memberships are not verified as identical across recipes. CREW source, fitting and test populations are matched within each native/joint pair; Appendix D.7 reports their event and station overlap.

Fitted preprocessing statistics. CirCor waveform normalization is recording-specific, CREW/LUDB normalization is trace-specific, and OpenPack uses fixed resampling and clip-

ping rather than a fitted population scaler. PatchTST additionally fits channel standardization over its training loader after domain preprocessing; RevIN remains input-specific. CirCor/OpenPack frozen probes estimate feature mean and scale from all training-split features before choosing labeled groups, then apply those statistics to validation/test.

A.2 Readouts, selection and scoring

Principal CirCor/OpenPack head fitting. The linear head is one 1×1 convolution from features to state logits. The temporal head uses two width-five convolutions with 64 hidden channels and GELU, followed by an output layer; its receptive field on CirCor's 50 Hz aligned grid is 180 ms. Both use AdamW (learning rate 10^{-3}, weight decay 10^{-4}), batches of 64 windows and masked cross-entropy for up to 120 epochs. The epoch is selected by validation state macro-F_1. Label-draw seed 1729 and head-initialization seed 31415 are fixed across encoder seeds and matched arms. Saved heads and feature normalization are reused for test evaluation without refitting. Appendix D.2 describes the separate fixed-final-head protocol.

CREW head fitting. CREW uses final epoch-40 PatchTST and PMT readouts with fixed P/S thresholds of 0.5 and no calibration labels (Appendix D.7).

LUDB head fitting and calibration. The 5k-update three-host comparison fits frozen-feature heads for 40 epochs and evaluates the final head: TS2Vec uses an affine sample-grid readout, while PatchTST and PMT use learned upsampling. Each encoder seed uses one predetermined fitting draw shared by the native and joint arms, with nested budgets of 8, 16, 32 and 64 training beats. All heads use the same 52 additional development beats for threshold calibration, separate from head fitting. Evaluation localizes P/T onsets within annotation-placed beat crops at ± 40 ms on 219 test beats, with 180 P-eligible and 193 T-eligible targets. Appendix D.6 gives the paired results and the variation represented by their standard deviations.

From features to scores. CirCor features are aligned to 50 Hz with 200 frames per window; predictions are stitched into recordings, and each contiguous S1 or S2 argmax-state run contributes its start as a predicted onset. Matching is one-to-one to continuous-time reference onsets within 50 ms, excluding onsets within 100 ms of annotated-support boundaries on both sides; F_1 is averaged over sound types. OpenPack predictions are reconstructed in session time and scored on the support-aware 1 Hz grid, pooling supported frames across people. CREW uses macro P/S pick-F_1 at ± 3 s; LUDB uses P/T onset F_1 at ± 40 ms.

Structured readout and adaptation. Appendix E.2 evaluates a shared training-fitted transition/duration decoder on the principal three-host CirCor test readouts and a separate 50k TS2Vec validation head-fitting-budget check. The decoder operates offline over whole recordings with future context and backtracking. Appendix E.3 specifies three-host supervised adaptation at two fitting budgets, including the transferred learning-rate and validation epoch-selection policies.

Replication and fixed endpoints. The principal CirCor/OpenPack comparisons use encoder seeds 42, 7 and 123 at 200k updates with one common label draw and head-initialization seed; sample standard deviations describe these encoder outcomes. CREW source, head-fitting and test populations can vary across seeds and are matched within each native/joint pair. LUDB test results average three paired encoder outcomes with one predetermined fitting-label draw per seed, matched across native and joint; the four fitting budgets are nested within each draw. Supporting encoders use fixed 20k endpoints on CREW and 5k on LUDB; there is no completed full-length TS2Vec/CREW endpoint. The principal comparisons use fixed 200k endpoints; Appendix D.3 evaluates validation-selected encoder checkpoints.

Numerical validation of saved control readouts. Saved encoders and heads are checked on validation before test evaluation, with a per-metric tolerance of 0.001 on the 0–1 scale. Three reference evaluations use separately validation-calibrated probability bounds of 0.05 maximum absolute error, 0.002 mean absolute error and 0.01 changed-class fraction, with the same metric tolerance. Weights, selected head epochs, fitting memberships and training-feature normalization are unchanged, and

test scores are not used to calibrate these bounds. The largest observed metric discrepancy in these evaluations is 0.000216.

B IMPLEMENTATION DETAILS

This appendix specifies the feature taps, token targets and execution of the auxiliary in each evaluated host.

Common auxiliary. The corruption draws one start a per window per step, uniformly over $\{0, \dots, L - s\}$ for a fully valid window, and reverses the samples of $I = [a, a + s)$ jointly across channels with a linear crossfade of k samples at each edge (Eq. 1). The implementation requires $2k < s < L$. The head is one affine map from the D-dimensional token $\bar{z}_t$ to a logit ($D + 1$ parameters); the loss is the mean binary cross-entropy with logits over valid tokens (Eq. 3); $\lambda = 1.5$ in the principal native-versus-TRD comparisons and 0 in their native controls. The principal PMT and PatchTST comparisons use ordinary Torch random streams for corruption and encoder stochasticity. Joint TS2Vec saves and restores Python, NumPy, Torch CPU and initialized CUDA random states around the auxiliary pass; its independent-target control additionally uses a private target-label generator. PPT uses a private placement generator and forks auxiliary Torch stochasticity. Standalone passes advance ordinary forward random streams, while PPT retains its private placement generator. The matched comparisons share the declared native recipe; their realized stochastic trajectories can differ. Appendix E.1 describes the control operators, optimization and limits of stochastic matching.

Standalone training. With the native objective removed, each update takes one source window, builds one reversed view $\tilde{x}$ and encodes it once; only $\mathcal{L}_{\mathrm{TRD}}$ is minimized. Native weak/strong views, contrastive crops, masking and reconstruction, PPT permutations and TimeDRL prediction branches are bypassed; base preprocessing, encoder dropout and normalization remain. TS2Vec retains its post-update encoder averaging and the averaged-encoder feature interface used by the probes. This changes native-view generation, normalization exposure, random-number consumption and compute as well as the objective; the ablation compares training recipes. Standalone update timings and their native/joint references are in Appendix H.3.

PMT. The reversed view $\tilde{x}$ is batched with PMT's two native views x_{w} and x_{s} in one fused forward pass. Batches are 32 per view for the principal CirCor/OpenPack comparisons and 64 for CREW. Stem normalization statistics are kept separate per view. The auxiliary classifier reads the final contextual token stream after the memory/neighborhood stack and before native objective-specific projections. Frozen probes expose the same encoder features on clean inputs, followed by task alignment and, for CirCor/OpenPack, training-feature standardization. The tokens $\bar{z}_t$ of the reversed view do not enter the native contrastive negative pool; $\mathcal{L}_{\mathrm{native}}$ is computed from z_t and z'_t only. Targets are produced by the token-labelling helper `token_labels_from_spans`, which assigns each token the coordinate $t \cdot \mathrm{stride} + \lfloor \mathrm{patch}/2 \rfloor$ and marks it positive when its overlap with I reaches half the configured patch width; for odd patch widths the helper's interval width is $2\lfloor \mathrm{patch}/2 \rfloor$. The principal comparisons retain this token-coordinate convention. Appendix G.1 quantifies its difference from valid nominal convolution-support overlap and evaluates a separate aligned-target sensitivity on CirCor. Instance contrastive loss (ICL) matches sequence-summary (CLS) embeddings across the two native views, using other instances as negatives; progressive contrastive loss (PCL) contrasts corresponding memory states across views against negative states. The principal CirCor/OpenPack PMT recipes use ICL weight 0.5 and PCL weight 0.25, with HGCL disabled.

PatchTST. Reversal is applied to the instance-normalized context before patch construction, so the backbone sees reversed samples, not reordered patch embeddings. Targets are computed on the encoder's retained patch grid: a patch is positive when at least half of its valid samples lie in I, and fully invalid patches are excluded from the loss. The patch features of all channels are concatenated before the pointwise head. In the evaluated fused path the masked native view and the reversed view pass through the backbone in one batch (64 per view on CirCor/OpenPack), with BatchNorm handled per view so that the native view's running statistics are preserved. An unfused path (separate pass, no running-statistics update from the reversed view) exists and was used only in the cost study (Appendix H.2). The native objective is the released masked-patch-reconstruction recipe in FP32;

Table B.1: Physical scales in the matched comparisons. Fades are per edge; s/L is the corrupted sample fraction, rounded where necessary. The lower panel lists nominal patch width/stride in samples; 1/1 denotes sample-resolution features. The dash denotes no completed TS2Vec/CREW endpoint under the full-length protocol.

Dataset	Window (s)	Rate (Hz)	Channels	Span (s)	Fade (ms)	s/L
CirCor	4	4,000	1	0.8	10	20%
OpenPack	60	30	3	16	200	26.7%
CREW	300	100	3	20	200	6.7%
LUDB	1	500	12	0.4	40	40%

Dataset	PMT	PatchTST	TS2Vec
CirCor	161/80	80/80	1/1
OpenPack	15/6	16/16	1/1
CREW	101/50	50/50	—
LUDB	11/5	10/5	1/1

the principal CirCor pretraining configuration uses a OneCycle schedule with peak learning rate 2×10^{-4}, as documented in Appendix C.

TS2Vec. Features are at stride one, so targets are exact half-open sample membership in I; NaN and padding positions are excluded from the loss, and the reversed span must fit inside a contiguous valid region of the input. TS2Vec's native views are two overlapping crops with its timestamp masking; because neither covers the whole window, the reversed view $\tilde{x}$ is built from the full window and encoded in its own pass, with the random-number state forked around the auxiliary. Batches are two windows on CirCor and eight on OpenPack. The auxiliary trains the live encoder, while frozen probes use the averaged encoder. There is no fused path for TS2Vec. The independent-target and order-preserving, RMS-matched gain controls share this module's interval placement, span and crossfade. Appendix E.1 specifies the control operators and optimization.

Supporting feature and readout interfaces. The supporting native/TRD/shuffle comparisons keep each host's feature grid and executed head fixed. TS2Vec retains all twelve LUDB input channels and sample-resolution features; its executed output is an affine map with 963 parameters, rather than the unused upsampling branch. CREW PatchTST supplies 600 tokens of dimension 384 to a 1,034,115-parameter localization head; LUDB PatchTST supplies 100 tokens of dimension 1,536 to a 16,522,755-parameter head. The patch-based heads use one U-Net upsampling stage, whereas the TS2Vec affine path already operates at sample resolution. These are within-host addition comparisons, not equal-capacity cross-host readout comparisons. Supporting TS2Vec/PatchTST runs and the OpenPack/CREW/LUDB shuffle comparisons use private corruption streams; the PMT CREW/LUDB native/TRD comparisons retain their specified ordinary-stream policy. All comparisons retain their prescribed source memberships, targets, auxiliary weights and selection policies; matched settings do not imply identical realized stochastic trajectories.

Corruption scales and token grids. Table B.1 gives the physical corruption scales for the principal matched comparisons and the corresponding encoder grids. The fraction s/L measures corrupted input samples, including the faded edges; patch-level positive-target fractions additionally depend on overlap, grid placement and validity. At the configured strides, CirCor span settings of 40/40/3,200 for PMT/PatchTST/TS2Vec all give 0.8 s; OpenPack settings of 80/30/480 give 16 s; the PMT CREW setting of 40 gives 20 s. The CREW comparison uses a 20-sample fade at 100 Hz.

Equal update counts do not equalize sample exposure or computation across hosts (Appendix H).

Inference path. Frozen-feature extraction for every evaluated profile calls neither the corruption nor the auxiliary head; the encoder and readout architecture of a +TRD run is identical to its control.

C BASELINE CONFIGURATIONS AND VERIFICATION

PatchTST. On CirCor, PatchTST uses masked-patch reconstruction in FP32, batches of 64, 16,000-sample contexts, patch width and stride 80, and mask ratio 0.4. The OneCycle schedule uses a peak learning rate of 2×10^{-4}, shared by the native and joint arms; the principal comparison evaluates both at 200k updates. At 200k updates, native PatchTST reaches 41.63 ± 2.91 test typed-onset F_1 at 50 ms with a linear head fitted on 31 training groups and selected on validation within a 120-epoch budget (Table D.1). In a one-seed ETTh1 source-task reproduction, test MSE differs from the released implementation by 0.0221%.

PPT. The PPT port passes trained-model source parity. On CirCor, its pretexts depend strongly on patch edges, its dense features compress, and its joint reversal detector remains near the position prior. The PPT/+TRD results therefore characterize the evaluated combination; the standalone condition removes the native pretexts and is analyzed separately in Appendix F.

TimeDRL. TimeDRL (Chang et al., 2024) separates timestamp-predictive and instance-contrastive objectives and is evaluated as an additional unpaired reference. Its CirCor/OpenPack features are assessed with the frozen readouts and fitting budgets of Appendix A. No TimeDRL+TRD result is included.

SoftCLT-t. SoftCLT-t replaces TS2Vec's temporal contrastive targets with soft targets on the same encoder while retaining the hard instance term. It changes the temporal target rather than adding a second complete objective and has no standalone arm.

Comparison scope. The native/+TRD comparisons evaluate the addition within each specified recipe. They do not establish the best achievable configuration of each published method.

D COMPLETE RESULTS AND ROBUSTNESS

D.1 CIRCOR AND OPENPACK AT ALL LABEL BUDGETS AND READOUTS

Tables D.1 and D.2 evaluate native and native+TRD TS2Vec, PatchTST, PMT and PPT, together with TS2Vec/SoftCLT-t and TimeDRL references, at 200k updates. Each recipe uses encoder seeds 42, 7 and 123, both frozen readouts and both head-fitting budgets. Validation-selected heads and feature normalization are applied to test data without refitting. Values are means and sample standard deviations over encoder seeds; the lower blocks give paired gains computed before rounding and the number of positive seed differences. SoftCLT-t uses the TS2Vec encoder with soft temporal contrastive targets and the unchanged hard instance term; TimeDRL is an additional native reference. PPT and its combination with TRD are discussed in Appendix D.5; Appendix C specifies the evaluated baselines. In each host block, bold scores mark the highest mean per column; bold gains are positive in every encoder seed. These are descriptive highlighting rules, not statistical-significance markers. The head-fitting budgets exclude the additional labeled validation groups used for head-epoch selection (Table A.1).

Variation across seeds. For PatchTST on OpenPack, the primary gain of $+5.32$ comprises seed differences of $+11.13$, $+0.87$ and $+3.97$. Its temporal/1-minute mean gain is $+0.017$ before rounding, and both temporal means are positive in only one seed. TS2Vec's OpenPack temporal/5-minute differences are $+0.42$, -4.38 and -5.85.

Temporal SoftCLT on the same encoder. At the larger CirCor budget, TS2Vec+TRD exceeds TS2Vec with SoftCLT-t by 34.27 points with the linear head and 12.64 with the temporal head, in every encoder seed. On OpenPack it exceeds TS2Vec with SoftCLT-t by 0.73 points on the primary readout, positive in one seed, and trails it on both temporal readouts in every seed.

Validation and test agreement. Mean-effect signs agree between validation and test for all 32 comparisons: four paired recipes, two datasets, two readouts and two budgets. The principal tables use the fixed 200k-update checkpoint for every arm.

Table D.1: CirCor held-out typed S1/S2 onset macro-F_1 at 50 ms (132 held-out patient groups). Label budgets are 7 (1%) and 31 (5%) labeled training patient groups. The lower block gives paired +TRD minus native gains from unrounded values, with the number of encoder seeds in which the gain is positive.

Arm	Linear / 1%	Linear / 5%	Temporal / 1%	Temporal / 5%
TS2Vec, native	52.09 ± 1.77	57.58 ± 1.41	77.40 ± 1.70	82.63 ± 1.31
with SoftCLT-t	51.67 ± 2.36	57.17 ± 1.85	77.76 ± 1.32	82.53 ± 0.98
+TRD	**88.16** ± 0.93	**91.44** ± 0.47	**92.73** ± 0.30	**95.16** ± 0.08
PatchTST, native	32.64 ± 4.34	41.63 ± 2.91	49.04 ± 2.95	55.74 ± 2.62
+TRD	**53.04** ± 2.12	**55.86** ± 0.95	**74.95** ± 0.46	**80.17** ± 0.28
PMT, native	39.64 ± 1.52	45.01 ± 0.95	62.76 ± 0.63	67.39 ± 0.52
+TRD	**89.55** ± 1.18	**92.33** ± 0.34	**94.06** ± 0.94	**95.44** ± 0.04
TimeDRL (reference)	37.85 ± 4.54	37.97 ± 5.36	57.33 ± 1.27	65.70 ± 0.89
Additional PPT comparison				
PPT	32.26 ± 8.06	31.88 ± 9.15	43.92 ± 11.52	46.14 ± 10.32
PPT +TRD	**34.80** ± 1.77	**35.08** ± 1.63	**50.01** ± 1.34	**52.27** ± 0.68
Δ TS2Vec	**+36.06** (3/3)	**+33.86** (3/3)	**+15.33** (3/3)	**+12.54** (3/3)
Δ PatchTST	**+20.40** (3/3)	**+14.23** (3/3)	**+25.91** (3/3)	**+24.43** (3/3)
Δ PMT	**+49.90** (3/3)	**+47.32** (3/3)	**+31.29** (3/3)	**+28.06** (3/3)
Δ PPT	+2.53 (1/3)	+3.20 (1/3)	+6.10 (2/3)	+6.13 (2/3)

Table D.2: OpenPack held-out support-aware operation macro-F_1 at 1 Hz (three held-out people, 34,361 supported seconds). Label budgets are one and five labeled minutes per training person. The lower block gives paired gains as in Table D.1.

Arm	Linear / 1 min	Linear / 5 min	Temporal / 1 min	Temporal / 5 min
TS2Vec, native	29.71 ± 0.53	34.07 ± 0.62	**33.50** ± 0.53	42.75 ± 1.09
with SoftCLT-t	**30.29** ± 0.29	35.60 ± 0.53	33.16 ± 0.44	**43.48** ± 0.85
+TRD	28.44 ± 1.06	**36.32** ± 2.01	26.97 ± 0.85	39.48 ± 2.47
PatchTST, native	31.29 ± 1.08	48.07 ± 0.95	33.79 ± 0.35	50.17 ± 1.39
+TRD	**34.97** ± 5.50	**53.39** ± 4.46	**33.81** ± 5.88	**51.95** ± 4.16
PMT, native	19.90 ± 0.72	30.38 ± 1.69	18.23 ± 1.53	30.27 ± 0.74
+TRD	**32.83** ± 1.37	**48.54** ± 0.53	**29.49** ± 0.48	**47.17** ± 0.54
TimeDRL (reference)	18.53 ± 0.20	24.84 ± 1.15	19.94 ± 0.45	27.39 ± 1.18
Additional PPT comparison				
PPT	14.86 ± 1.23	20.94 ± 0.88	15.98 ± 0.90	20.59 ± 0.21
PPT +TRD	**22.32** ± 1.57	**34.63** ± 2.14	**24.59** ± 1.53	**35.15** ± 3.02
Δ TS2Vec	−1.28 (0/3)	**+2.25** (3/3)	−6.53 (0/3)	−3.27 (1/3)
Δ PatchTST	**+3.68** (3/3)	**+5.32** (3/3)	+0.02 (1/3)	+1.78 (1/3)
Δ PMT	**+12.93** (3/3)	**+18.15** (3/3)	**+11.26** (3/3)	**+16.90** (3/3)
Δ PPT	**+7.46** (3/3)	**+13.69** (3/3)	**+8.61** (3/3)	**+14.56** (3/3)

D.2 Fitting-label draws with fixed-final heads

For each host, we cross three native/joint encoder-seed pairs at 200k updates with five predetermined fitting-label draws. The 7- and 31-patient-group subsets are nested within each draw and shared across recipes and encoder seeds, giving 15 paired evaluations for each readout and budget. Both frozen heads use fixed initialization and epoch 120, without validation-based epoch or threshold selection or early stopping; the principal tables instead use validation-selected heads. Feature normalization uses the complete training-split feature banks. Table D.3 gives the mean test outcomes; Table D.4 separates descriptive variation across encoder means and draw means.

Table D.3: CirCor test typed-onset macro-F_1 at 50 ms (%) for fixed-200k encoders, averaged over the crossed encoder-seed and fitting-draw design. Linear and temporal heads use epoch 120 without validation-based selection, with fitting sets of 7 (1%) or 31 (5%) patient groups. Positive counts are paired encoder/draw evaluations. +TRD retains the native objective.

Encoder	Head	Fit	Native	+TRD	Δ	Positive
PatchTST	Linear	1%	33.07	52.68	+19.60	15/15
PatchTST	Linear	5%	35.46	54.78	+19.32	15/15
PatchTST	Temporal	1%	47.30	72.71	+25.42	15/15
PatchTST	Temporal	5%	54.49	78.33	+23.83	15/15
PMT	Linear	1%	40.57	89.79	+49.22	15/15
PMT	Linear	5%	44.72	91.66	+46.93	15/15
PMT	Temporal	1%	58.87	91.61	+32.75	15/15
PMT	Temporal	5%	66.62	92.12	+25.50	15/15
TS2Vec	Linear	1%	50.58	88.56	+37.98	15/15
TS2Vec	Linear	5%	57.26	91.18	+33.92	15/15
TS2Vec	Temporal	1%	73.64	91.20	+17.55	15/15
TS2Vec	Temporal	5%	77.30	91.04	+13.73	15/15

Table D.4: Descriptive variation in the CirCor test paired gains (percentage points) from Table D.3: fixed-200k encoders, epoch-120 linear and temporal heads, and 7/31-patient-group fitting sets. s_{enc} is the sample SD across three encoder means after averaging over draws; s_{draw} is the sample SD across five draw means after averaging over encoders. These are distinct summaries, not standard errors or confidence intervals.

Encoder	Head	Fit	s_{enc}	s_{draw}
PatchTST	Linear	1%	6.22	4.00
PatchTST	Linear	5%	6.98	3.57
PatchTST	Temporal	1%	2.48	1.56
PatchTST	Temporal	5%	2.30	0.75
PMT	Linear	1%	1.22	1.80
PMT	Linear	5%	1.20	0.96
PMT	Temporal	1%	1.18	3.68
PMT	Temporal	5%	0.97	2.24
TS2Vec	Linear	1%	1.07	2.44
TS2Vec	Linear	5%	1.20	0.65
TS2Vec	Temporal	1%	1.34	2.63
TS2Vec	Temporal	5%	0.96	0.94

D.3 Validation-selected encoder checkpoints

Validation trajectories at 0, 25k, 50k, 100k and 200k show that joint PMT on OpenPack declines late and several native arms peak earlier. For each arm, encoder seed, readout and fitting budget, we select from steps $\{0, 25\mathrm{k}, 50\mathrm{k}, 100\mathrm{k}, 200\mathrm{k}\}$ using validation state macro-F_1, breaking ties in favour of the earlier step. Test evaluation uses the corresponding head, feature normalization and fitting membership without refitting. Tables D.5 and D.6 compare validation-selected checkpoints with the fixed-200k endpoint. Every CirCor paired difference remains positive. The primary linear gains are 31.64 points for TS2Vec, 6.20 for PatchTST and 46.93 for PMT, compared with 33.86, 14.23 and 47.32 at the fixed endpoint. OpenPack remains recipe-dependent: TS2Vec's primary selected-checkpoint gain is 0.53 points and positive in one seed, while its other three readout/budget combinations are negative. PMT improves throughout; PatchTST's temporal effects remain mixed across seeds.

D.4 OpenPack results by held-out person

At 200k updates, Table D.7 reports the primary linear readout separately for each test person, using five labeled minutes per training person. The pooled metric is computed from pooled predictions,

Table D.5: CirCor test results after per-arm, per-seed, per-readout, per-fitting-budget validation selection from encoder steps $\{0, 25\text{k}, 50\text{k}, 100\text{k}, 200\text{k}\}$. Linear and temporal heads fit 7 (1%) or 31 (5%) training patient groups. Selection uses state macro-F_1 with earlier-step ties; the selected heads and normalization are applied without refitting. The test outcome is typed S1/S2 onset macro-F_1 at 50 ms (%). Means over three encoders. +TRD retains the native objective.

Encoder	Head	Fit	Native	+TRD	Δ_{sel}	Δ_{200k}	Positive
PatchTST	Linear	1%	42.23	53.04	+10.81	+20.40	3/3
PatchTST	Linear	5%	49.66	55.86	+6.20	+14.23	3/3
PatchTST	Temporal	1%	56.96	75.27	+18.31	+25.91	3/3
PatchTST	Temporal	5%	64.89	80.17	+15.28	+24.43	3/3
PMT	Linear	1%	40.88	90.00	+49.12	+49.90	3/3
PMT	Linear	5%	45.40	92.33	+46.93	+47.32	3/3
PMT	Temporal	1%	62.32	94.26	+31.93	+31.29	3/3
PMT	Temporal	5%	68.38	95.44	+27.06	+28.06	3/3
TS2Vec	Linear	1%	53.96	88.16	+34.20	+36.06	3/3
TS2Vec	Linear	5%	59.81	91.44	+31.64	+33.86	3/3
TS2Vec	Temporal	1%	77.93	92.73	+14.80	+15.33	3/3
TS2Vec	Temporal	5%	83.17	95.16	+11.99	+12.54	3/3

Table D.6: OpenPack test results after per-arm, per-seed, per-readout, per-fitting-budget validation selection from encoder steps $\{0, 25\text{k}, 50\text{k}, 100\text{k}, 200\text{k}\}$. Linear and temporal heads fit one or five labeled minutes per training person. Selection uses state macro-F_1 with earlier-step ties; the selected heads and normalization are applied without refitting. The test outcome is operation macro-F_1 on scored 1 Hz positions (%). Means over three encoders. +TRD retains the native objective.

Encoder	Head	Fit	Native	+TRD	Δ_{sel}	Δ_{200k}	Positive
PatchTST	Linear	1 min	31.29	34.97	+3.68	+3.68	3/3
PatchTST	Linear	5 min	48.07	53.39	+5.32	+5.32	3/3
PatchTST	Temporal	1 min	33.79	33.81	+0.02	+0.02	1/3
PatchTST	Temporal	5 min	50.17	51.95	+1.78	+1.78	1/3
PMT	Linear	1 min	20.94	35.08	+14.15	+12.93	3/3
PMT	Linear	5 min	30.55	51.36	+20.81	+18.15	3/3
PMT	Temporal	1 min	19.49	32.86	+13.37	+11.26	3/3
PMT	Temporal	5 min	30.95	49.70	+18.75	+16.90	3/3
TS2Vec	Linear	1 min	30.39	28.49	−1.89	−1.28	0/3
TS2Vec	Linear	5 min	35.94	36.47	+0.53	+2.25	1/3
TS2Vec	Temporal	1 min	32.39	26.75	−5.64	−6.53	0/3
TS2Vec	Temporal	5 min	43.14	39.86	−3.28	−3.27	0/3

rather than by averaging the three participant F_1 scores. Mean gains are positive for all three people on every host at this readout, but three people do not establish population-level uncertainty.

Operation-level predictions. Figure D.1 shows the whole-test confusion matrices of PMT's primary operation readout for seed 42. With TRD, recall improves for all ten operations in this seed, rather than the gain being confined to one class. Across the three encoder seeds, mean operation recall ranges from 0.29–0.32 without and 0.48–0.50 with TRD; recall improves for ten operations in two seeds and nine in the third. The matrices summarize operation identity across scored seconds, complementing the per-person F_1 results in Table D.7.

D.5 PPT AND ITS COMBINATION WITH TRD

The PPT pair tests reversal localization with an order-aware patch-pretext recipe. Native PPT and PPT+TRD use the same 200k endpoint, label budgets, frozen readouts and test splits as the principal comparisons (Tables D.1 and D.2).

On CirCor, the primary linear gain is +3.20 points, with seed differences of −3.97, +15.02 and −1.44; the positive mean is driven by one unusually weak native seed. The temporal gain at 5%

Table D.7: OpenPack test operation macro-F_1 (%) per held-out person at 200k updates, averaged across three encoder seeds. Linear heads fit five labeled minutes per training person. Positive counts give the number of encoder seeds with a gain. +TRD retains the native objective.

Encoder	Person	Native	+TRD	Δ	Positive
PatchTST	1	49.71	58.25	+8.54	3/3
PatchTST	2	40.27	44.36	+4.09	2/3
PatchTST	3	54.61	56.21	+1.59	2/3
PMT	1	35.81	49.58	+13.76	3/3
PMT	2	19.82	39.06	+19.24	3/3
PMT	3	37.04	59.72	+22.68	3/3
TS2Vec	1	35.09	35.97	+0.88	2/3
TS2Vec	2	29.73	32.28	+2.56	3/3
TS2Vec	3	38.62	41.29	+2.67	3/3

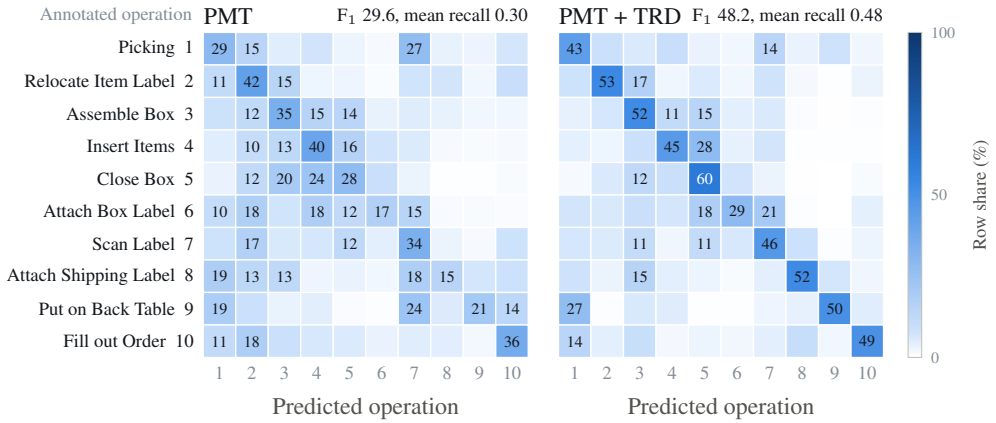


Figure D.1: **OpenPack operation discrimination from frozen PMT features.** Row-normalized test confusion (%) over 34,361 scored seconds from three held-out people, for PMT encoders after 200k updates (seed 42), without and with TRD. Linear heads are fitted on five labeled minutes per training person and selected on validation data. Rows give annotated operations and columns predictions; diagonal entries are class recalls, while off-diagonal entries show the distribution of errors. Numbers mark every diagonal cell and off-diagonal cells of at least 10%. Mean operation recall rises from 0.30 to 0.48.

labels is +6.13, positive in two seeds. On OpenPack, every seed improves at both budgets and readouts, with primary linear and temporal gains of +13.69 and +14.56 points.

On CirCor, the evaluated PPT pretexts are sensitive to patch edges and the dense features compress (Appendix C). The standalone PPT-configured results in Appendix F remove the native objectives and test a distinct training recipe.

D.6 LUDB: ECG ONSET LOCALIZATION

We evaluate all three hosts with and without TRD at 5k updates, using three paired encoder seeds (42, 7 and 123) and four nested fitting budgets. Inputs are annotation-placed, one-second, twelve-lead ECG beat crops; Appendix A.2 specifies calibration and scoring. Table D.8 gives all host/budget means and paired differences; Table D.9 provides the PMT seed-level detail. PatchTST and PMT improve in every seed and budget, whereas TS2Vec's means decrease with mixed paired signs. Variation includes encoder-seed and predetermined fitting-draw changes, matched within native/TRD/shuffle comparisons.

Onset-probability readout. Figure D.2 shows the PatchTST readouts behind Table D.8 at the 64-beat fitting budget. The native readout distributes onset probability across broader regions; with TRD, probability concentrates around the annotated onsets. Across the six onset-type/encoder-seed

Table D.8: **LUDB native-versus-TRD comparison at every fitting budget.** Test F_1 percentages, mean $\pm$ sample SD over three paired seeds; gains and their sample SDs are computed from paired differences before rounding. Encoders use 5k updates, final epoch-40 heads and P/T matching at ± 40 ms, with 52 additional calibration beats. Paired runs share each fitting draw; differences across hosts are not an equal-capacity head comparison.

Host	Fitting labels	Native	+TRD	$\Delta \pm s_\Delta$	Positive
TS2Vec	8	57.01 ± 11.14	52.32 ± 8.83	-4.69 ± 13.44	1/3
TS2Vec	16	55.50 ± 12.24	52.49 ± 2.24	-3.01 ± 14.17	2/3
TS2Vec	32	56.73 ± 9.45	54.99 ± 5.57	-1.74 ± 12.96	1/3
TS2Vec	64	58.83 ± 9.11	54.01 ± 3.89	-4.82 ± 12.17	1/3
PatchTST	8	45.71 ± 3.10	58.61 ± 5.82	$+12.90 \pm 3.14$	3/3
PatchTST	16	49.29 ± 6.70	71.24 ± 3.53	$+21.96 \pm 8.60$	3/3
PatchTST	32	55.49 ± 3.65	76.86 ± 1.72	$+21.37 \pm 5.09$	3/3
PatchTST	64	52.85 ± 9.82	81.86 ± 2.38	$+29.01 \pm 10.53$	3/3
PMT	8	36.79 ± 4.54	55.37 ± 5.51	$+18.58 \pm 7.80$	3/3
PMT	16	50.54 ± 5.72	61.33 ± 7.02	$+10.79 \pm 8.55$	3/3
PMT	32	51.42 ± 3.44	73.10 ± 2.28	$+21.67 \pm 5.70$	3/3
PMT	64	63.45 ± 4.39	69.82 ± 7.44	$+6.38 \pm 5.01$	3/3

Table D.9: LUDB PMT onset localization at 5k updates: test P/T onset F_1 at ± 40 ms (%) for every fitted seed/budget pair. Each encoder uses one predetermined fitting draw matched across Native and Native +TRD, with 52 separate development beats for threshold calibration and a 40-epoch final learned-upsampled head. Gains are rounded from the full-precision paired differences.

Training beats	Encoder seed	Native	+TRD	Δ
8	42	38.18	49.03	10.85
8	7	31.72	58.16	26.44
8	123	40.47	58.93	18.46
16	42	53.11	54.25	1.15
16	7	54.53	68.29	13.75
16	123	43.99	61.45	17.46
32	42	55.19	70.50	15.30
32	7	48.46	74.76	26.29
32	123	50.61	74.04	23.43
64	42	59.09	67.11	8.02
64	7	67.88	78.24	10.36
64	123	63.38	64.12	0.75

combinations, the median fraction of a beat's onset-probability mass within ± 40 ms is 27.5–43.5% without and 79.2–83.3% with TRD.

Example selection. Among the 16 test subjects with at least six beats annotated with both onsets, subject 9 is selected by minimizing the largest mid-rank-percentile distance from the median across its native score, TRD score and paired difference for seed 42. The excerpt comprises the consecutive three-beat group whose two scores are closest to the subject's scores, with earliest-order tie breaking. Test predictions determine the illustration only; model fitting, calibrated thresholds and whole-test scores are unchanged.

D.7 CREW: SEISMIC PHASE PICKING

We train PatchTST and PMT with and without TRD at fixed 20k endpoints using encoder seeds 42, 7 and 123. Native and TRD training share non-auxiliary configurations and source populations within each seed; PMT uses BF16 and PatchTST FP32. Inputs contain three seismic components (ENZ) over 300 s at 100 Hz, demeaned over time and peak-normalized across channels and time. PMT uses a 101-sample tokenizer kernel, stride 50 and batch size 64; PatchTST uses 50-sample patches and stride 50. Each training-source population contains 147,460 traces.

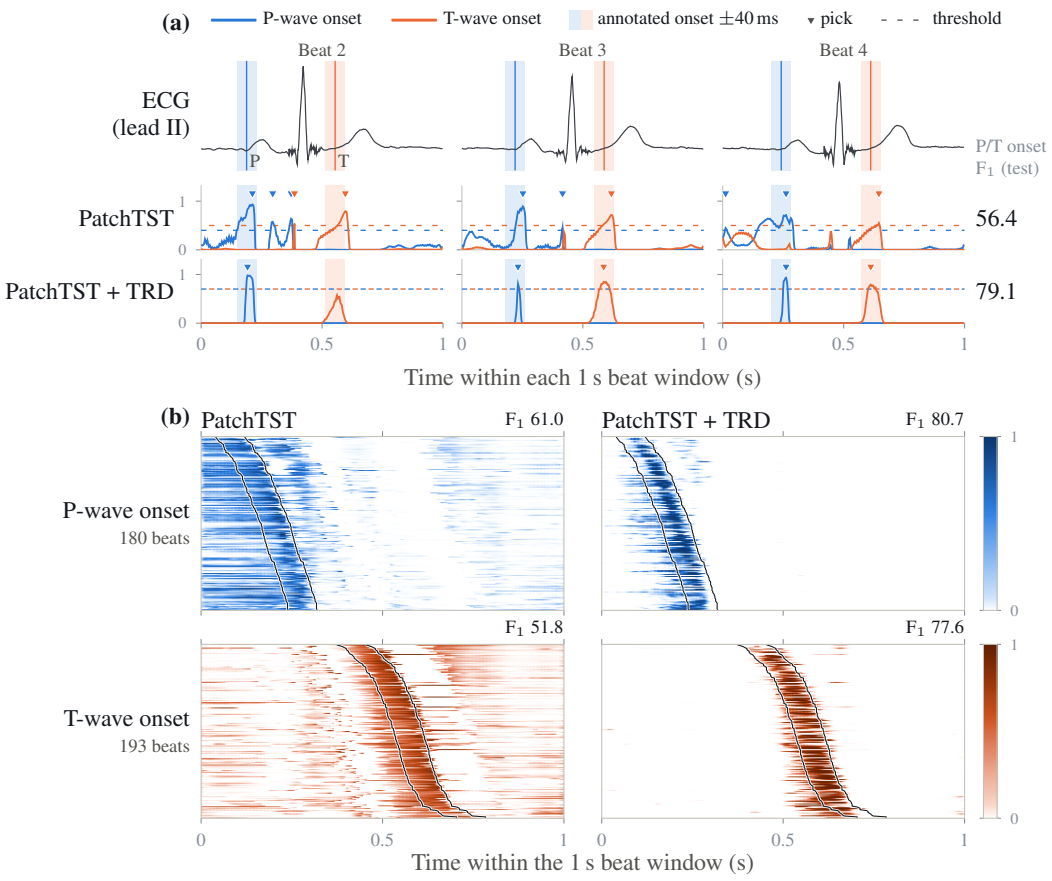


Figure D.2: **LUDB onset probabilities from frozen PatchTST features, without and with TRD.** Encoders use 5k updates (seed 42); heads are fitted on 64 training beats, with pick thresholds calibrated on 52 development beats. (a) Three consecutive beats from subject 9, each shown in its annotation-placed one-second input window; lead II is displayed, while the encoder receives all twelve leads. Curves give P- and T-wave onset probabilities, dashed lines calibrated thresholds, and triangles detected picks; shaded bands mark ±40 ms around the annotated onsets. (b) All 180 P-eligible and 193 T-eligible test beats, sorted by annotated onset time in the same order across both models; black curves mark the ±40 ms corridor. Probabilities use a common 0–1 scale. Scores describe the whole test split for this seed: macro onset F_1 in (a), per-type F_1 in (b).

Table D.10: **CREW native-versus-TRD comparison at every fitting budget.** Test F_1 percentages, mean $\pm$ sample SD over three paired seeds; gains and their sample SDs are computed from paired differences before rounding. Encoders use 20k updates, final epoch-40 heads and P/S matching at ± 3 s. TS2Vec has no completed full-length endpoint.

Host	Fitting labels	Native	+TRD	$\Delta \pm s_\Delta$	Positive
PatchTST	74	12.52 ± 2.15	26.66 ± 0.32	$+14.14 \pm 1.93$	3/3
PatchTST	147	21.71 ± 0.27	28.75 ± 1.18	$+7.04 \pm 1.44$	3/3
PMT	74	27.51 ± 3.22	65.69 ± 2.22	$+38.18 \pm 1.52$	3/3
PMT	147	33.70 ± 1.81	66.92 ± 2.41	$+33.21 \pm 0.61$	3/3

Table D.11: CREW test macro P/S pick-F_1 at ± 3 s (%) after 20k PMT pretraining updates and 40 head-fitting epochs, with fixed 0.5 thresholds and no calibration labels. Each Native/Native +TRD pair uses matched source, fitting and test populations; populations can vary across seeds. The two fitting budgets share each encoder pair. Gains are computed before rounding.

Training traces	Seed	Native	+TRD	Δ
74	42	27.98	67.06	+39.08
74	7	24.08	63.12	+39.04
74	123	30.47	66.89	+36.42
147	42	34.78	68.23	+33.46
147	7	31.62	64.14	+32.52
147	123	34.72	68.38	+33.66

Heads fit nested subsets of 74 and 147 training traces for 40 epochs; only the final head is evaluated, with P/S thresholds fixed at 0.5 and no calibration labels. Each seed is scored on 3,130 test traces using macro P/S pick-F_1 at ± 3 s. Source, fitting and test memberships are matched within each native/TRD/shuffle comparison and can vary across seeds, so the sample SD includes encoder and population variation. Both fitting budgets share each encoder; Table D.10 gives the complete host/budget summaries and Table D.11 gives PMT seed-level detail. No TS2Vec/CREW score is available: full-length native temporal contrast exceeded available GPU memory in all three conditions before a 20k endpoint was produced.

Partition structure. The training-source and test sets use the declared trace partitions but share station identities and, for two seeds, a small number of event identities (Table D.12). This evaluates phase picking under that split, rather than station-disjoint generalization.

E Controlled analyses

E.1 Supervision and corruption controls

Supervision granularity is evaluated on CirCor at 50k updates. Corruption comparisons use 50k prefixes of the 200k schedules on CirCor/OpenPack, 20k endpoints on CREW and 5k on LUDB, with seeds 42, 7 and 123. CirCor inputs are four-second windows at 4 kHz; local corruptions use a 0.8 s interval, a 10 ms fade per edge and auxiliary weight 1.5. On CirCor, linear and temporal heads use the same 7/31 training-patient-group masks and a 120-epoch fitting budget with validation-based selection. Head selection uses the validation cohort of 486 recordings from 130 patient groups; the test cohort contains 465 recordings from 132 patient groups. We report test typed-onset macro-F_1 at 50 ms, with linear/5% as the primary readout, using the same validation-selected heads and training-feature normalization without refitting. PMT follows the first 50k updates of its 200-epoch cosine schedule with 1,000-update ticks; TS2Vec uses a constant learning rate of 0.001. Random-stream isolation differs across conditions, so matched scientific settings and encoder seeds do not imply identical stochastic trajectories.

Supervision granularity. Whole-window reversal acts on the complete valid input without local fades. Whole-window and local pooled tasks average valid token features before a single affine

Table D.12: CREW training-source/test metadata overlap. Each seed uses 147,460 training-source traces and 3,130 test traces, with the same populations within a Native/Native +TRD pair. Shared counts refer to identities present in both source and test sets.

Seed	Test events	Shared events	Test stations	Shared stations
42	2666	0	617	578
7	2701	4	625	586
123	2697	2	638	604

Table E.1: CirCor test control results after 50k encoder updates: typed-onset macro-F_1 at 50 ms (%), primary linear readout fitted on 31 training patient groups (5%). Values are mean ± sample SD over three encoders. Heads use validation-selected epochs within a 120-epoch budget and are applied to test without refitting.

Encoder	Condition	Test
TS2Vec	Native	59.68 ± 0.95
TS2Vec	Local dense reversal p=1	85.73 ± 2.08
TS2Vec	Whole-window pooled p=0.5	65.93 ± 0.79
TS2Vec	Local pooled p=0.5	59.25 ± 2.34
TS2Vec	Local dense p=0.5	73.80 ± 0.76
TS2Vec	External replacement p=1	80.67 ± 1.84
TS2Vec	Four-block shuffle p=1	94.90 ± 0.32
TS2Vec	RMS-matched gain	73.82 ± 0.94
TS2Vec	Independent targets	54.03 ± 0.67
PMT	Native	45.43 ± 1.83
PMT	Local dense reversal p=1	90.38 ± 0.56
PMT	Whole-window pooled p=0.5	64.66 ± 1.50
PMT	Local pooled p=0.5	70.75 ± 5.37
PMT	Local dense p=0.5	88.61 ± 0.28

prediction; the target indicates whether the corresponding transformation was applied. The local dense task predicts interval membership, including the fades, at each valid token. The three matched tasks apply a transformation with probability $p = 0.5$; clean examples retain the auxiliary encoder pass and receive negative targets. Local pooled and dense tasks share occurrence, placement, source ordering, preprocessing, feature tap and head-initialization policy. Pooled BCE averages examples; dense BCE first averages valid tokens within each example and then averages examples, without prevalence reweighting. The standard dense reversal reference uses $p = 1$. Changing exposure or prediction granularity also changes target prevalence and semantics; these comparisons do not form a complete factorial.

Table E.1 reports test scores for the primary readout, and Table E.2 gives the paired test differences. Dense local supervision at matched $p = 0.5$ improves over pooled local supervision on both hosts in every seed.

Corruption choice. Gain changes amplitude without reordering samples; independent targets pair a reversed input with an independently sampled interval mask. Both use the affine membership head and auxiliary weight 1.5. On TS2Vec, replacement and four-block shuffle use the same local interval, outer fade, dense target, $p = 1$, auxiliary weight and readout policy. Replacement draws a segment from another eligible training recording, excluding the recipient recording and duplicate waveform identities, with a uniformly sampled valid start. This is an AnomalyBERT-inspired control, not a reproduction of the full method. Shuffle partitions the interval into four equal contiguous blocks and samples uniformly from the nine permutations that move every block. All channels share the permutation; sample orientation within each block is unchanged. Replacement and shuffle have no internal fades or added amplitude normalization and are not rescaled to match reversal's perturbation magnitude.

Table E.2: Paired primary control differences on CirCor test, in percentage points. Means and sample SDs are computed across three paired encoder-seed differences before rounding. The outcome is typed-onset macro-F_1 at 50 ms after 50k updates, with the linear readout fitted on 31 training patient groups. Test evaluation reuses the validation-selected heads; positive counts refer to the three encoder pairs.

Encoder	Treatment – reference	$\Delta \pm s_\Delta$	Positive
TS2Vec	Whole pooled – local pooled	6.68 ± 2.73	3/3
TS2Vec	Dense – pooled (local, $p = 0.5$)	14.56 ± 2.85	3/3
TS2Vec	Dense $p = 0.5$ – dense $p = 1$	-11.92 ± 2.11	0/3
PMT	Whole pooled – local pooled	-6.09 ± 6.68	1/3
PMT	Dense – pooled (local, $p = 0.5$)	17.85 ± 5.62	3/3
PMT	Dense $p = 0.5$ – dense $p = 1$	-1.77 ± 0.82	0/3
TS2Vec	Replacement – reversal	-5.06 ± 0.64	0/3
TS2Vec	Shuffle – reversal	9.18 ± 1.77	3/3
TS2Vec	Reversal – gain	11.91 ± 1.62	3/3
TS2Vec	Gain – native	14.14 ± 0.21	3/3
TS2Vec	Independent targets – native	-5.65 ± 1.16	0/3

Table E.3: Corruption supervision across readouts and fitting budgets on CirCor test. Typed S1/S2 onset macro-F_1 at 50 ms (%), mean $\pm$ sample SD over three TS2Vec encoder seeds after 50k updates. Linear and temporal heads fit 7 (1%) or 31 (5%) training patient groups and use validation-selected epochs within a 120-epoch budget. Test applies the same selected heads and training-feature normalization without refitting. Bold marks the highest mean among the six conditions in each column.

Condition	Linear / 1%	Linear / 5%	Temporal / 1%	Temporal / 5%
Native	53.76 ± 1.25	59.68 ± 0.95	78.13 ± 1.77	83.12 ± 1.76
Independent targets	50.68 ± 0.63	54.03 ± 0.67	74.86 ± 1.33	80.30 ± 1.45
Order-preserving gain	64.78 ± 1.31	73.82 ± 0.94	81.08 ± 1.21	88.08 ± 0.72
External replacement	75.92 ± 2.25	80.67 ± 1.84	86.26 ± 1.01	89.93 ± 0.38
Local reversal (+TRD)	79.82 ± 1.96	85.73 ± 2.08	90.17 ± 0.18	93.94 ± 0.28
Four-block shuffle	$\mathbf{92.11} \pm 0.61$	$\mathbf{94.90} \pm 0.32$	$\mathbf{94.03} \pm 0.27$	$\mathbf{96.46} \pm 0.10$

Let v be the original selected interval, w its taper and $r = (1 - w)v + w\ \mathrm{reverse}(v)$ the faded reversal. The order-preserving control returns $g = v + a(w \odot v)$, where

$$a = \frac{\|r - v\|_2}{\max(\|w \odot v\|_2, \varepsilon)}. \tag{E.1}$$

The norms include all samples and channels of the interval for one example, and ε is the smallest positive normal float64 value. The zero/zero case gives no increment; a nonzero numerator with a zero denominator or nonfinite values is rejected. Arithmetic before the norms retains input precision, norms use float64, and the gain coefficient is cast back to input precision.

Table E.3 reports all six corruption conditions at both readouts and fitting budgets on test.

Reversal exceeds gain in every test encoder/readout/budget comparison. At the primary readout, the paired reversal-minus-gain difference is 11.91 ± 1.62 points. Gain also exceeds native training in every test encoder/readout/budget comparison. Independent targets reduce the mean score in all four readout/budget settings, with eleven of the twelve test pairs negative. The four readout/budget outcomes reuse each encoder, so these twelve comparisons represent three pretraining replications. This control shows the cost of an unrelated target; its deficit is not an isolated estimate of the benefit of label alignment.

Four-block shuffle exceeds reversal by 9.18 ± 1.77 points on the primary test readout (Table E.2); its advantage also holds in every test seed at both fitting budgets with both readouts.

Corruption ranking across recipes and datasets. Tables E.4–E.7 report every completed native/reversal/shuffle mean at the prescribed readouts and fitting budgets. Both auxiliary conditions retain the native objective, use weight 1.5 and apply the manipulation with probability one. The

Table E.4: **CirCor operator comparison.** 50k prefixes of the 200k schedules; onset F_1 at 50 ms; 7/31 fitting patient groups and validation-selected heads. Values are test percentages, mean ± sample SD over three seeds. Both auxiliary conditions retain the native objective. Δ is Shuffle minus TRD, with paired SD and positive-seed count. Budgets/readouts share encoders.

Host	Head / labels	Native	+TRD	+Shuffle	$\Delta \pm s_\Delta$	Positive
TS2Vec	linear/1%	53.76 ± 1.25	79.82 ± 1.96	92.11 ± 0.61	$+12.29 \pm 1.47$	3/3
TS2Vec	linear/5%	59.68 ± 0.95	85.73 ± 2.08	94.90 ± 0.32	$+9.18 \pm 1.77$	3/3
TS2Vec	temporal/1%	78.13 ± 1.77	90.17 ± 0.18	94.03 ± 0.27	$+3.86 \pm 0.13$	3/3
TS2Vec	temporal/5%	83.12 ± 1.76	93.94 ± 0.28	96.46 ± 0.10	$+2.52 \pm 0.31$	3/3
PatchTST	linear/1%	30.32 ± 7.01	54.01 ± 0.46	37.12 ± 3.49	-16.90 ± 3.27	0/3
PatchTST	linear/5%	39.97 ± 0.93	54.79 ± 2.12	44.18 ± 4.95	-10.61 ± 6.67	0/3
PatchTST	temporal/1%	50.86 ± 1.62	75.30 ± 0.47	57.11 ± 0.23	-18.20 ± 0.37	0/3
PatchTST	temporal/5%	57.21 ± 0.39	79.03 ± 0.77	64.04 ± 0.68	-15.00 ± 0.91	0/3
PMT	linear/1%	40.76 ± 0.88	88.83 ± 0.64	86.78 ± 1.22	-2.05 ± 1.77	0/3
PMT	linear/5%	45.43 ± 1.83	90.38 ± 0.56	90.11 ± 0.28	-0.27 ± 0.64	1/3
PMT	temporal/1%	62.35 ± 1.35	93.98 ± 0.46	91.82 ± 0.13	-2.16 ± 0.39	0/3
PMT	temporal/5%	67.68 ± 0.58	94.97 ± 0.11	95.00 ± 0.16	$+0.03 \pm 0.05$	2/3

Table E.5: **OpenPack operator comparison.** 50k prefixes of the 200k schedules; operation macro-F_1 at 1 Hz; 1/5 fitting minutes per training person and validation-selected heads. Values are test percentages, mean ± sample SD over three seeds. Both auxiliary conditions retain the native objective. Δ is Shuffle minus TRD, with paired SD and positive-seed count. Budgets/readouts share encoders.

Host	Head / labels	Native	+TRD	+Shuffle	$\Delta \pm s_\Delta$	Positive
TS2Vec	linear/1 min	29.75 ± 0.53	27.50 ± 0.54	48.97 ± 2.68	$+21.47 \pm 2.39$	3/3
TS2Vec	linear/5 min	35.17 ± 0.39	35.38 ± 1.13	59.88 ± 1.59	$+24.51 \pm 1.35$	3/3
TS2Vec	temporal/1 min	32.82 ± 0.83	26.55 ± 0.85	48.50 ± 2.63	$+21.94 \pm 1.81$	3/3
TS2Vec	temporal/5 min	43.16 ± 0.37	39.04 ± 0.66	61.66 ± 1.93	$+22.62 \pm 1.55$	3/3
PatchTST	linear/1 min	22.08 ± 0.96	21.79 ± 0.98	33.31 ± 1.41	$+11.52 \pm 1.04$	3/3
PatchTST	linear/5 min	32.61 ± 1.30	35.34 ± 2.33	44.47 ± 2.65	$+9.13 \pm 2.63$	3/3
PatchTST	temporal/1 min	23.29 ± 0.86	20.31 ± 1.77	32.82 ± 2.11	$+12.50 \pm 2.30$	3/3
PatchTST	temporal/5 min	34.59 ± 0.83	33.12 ± 1.93	43.86 ± 2.11	$+10.74 \pm 2.17$	3/3
PMT	linear/1 min	20.29 ± 2.03	34.10 ± 0.63	30.63 ± 4.11	-3.47 ± 3.68	0/3
PMT	linear/5 min	31.15 ± 1.80	50.27 ± 0.50	45.78 ± 3.32	-4.48 ± 3.09	0/3
PMT	temporal/1 min	18.47 ± 0.13	31.40 ± 1.41	29.67 ± 3.41	-1.74 ± 3.03	1/3
PMT	temporal/5 min	29.68 ± 1.40	48.54 ± 0.98	46.53 ± 3.58	-2.01 ± 4.00	1/3

four-block derangement uses 0.2 s blocks within the 0.8 s CirCor interval with 10 ms outer fades; OpenPack uses 4 s blocks within 16 s and 200 ms outer fades. CREW uses four 5 s blocks within 20 s with 200 ms outer fades, and LUDB uses four 0.1 s blocks within 0.4 s with 40 ms outer fades. All channels share the permutation, sample order is preserved within blocks, and there are no internal fades or magnitude rescaling. CirCor/OpenPack heads use validation state-F_1 epoch selection; CREW/LUDB use final epoch-40 heads, with the fixed CREW thresholds or separate LUDB calibration described in Appendix A.2. There is no completed TS2Vec/CREW endpoint; all other host/dataset comparisons have three seeds. Relative to native training, shuffle improves all three seeds in 39 of the 40 settings; the exception is PatchTST/CirCor with the linear head at 5% fitting labels, where two seeds improve.

On CirCor, reversal leads on PatchTST at both readouts and budgets; PMT's larger-budget means are close and its smaller-budget comparisons favor reversal. On OpenPack, shuffle leads on TS2Vec and PatchTST, while PMT favors reversal on the primary readout in every seed. On CREW, PMT favors reversal in every seed at both budgets and PatchTST has small mixed-seed mean advantages for reversal. On LUDB, shuffle exceeds reversal in every seed/budget on TS2Vec and PMT; PatchTST's ranking varies with budget, favoring reversal in every seed at 64 fitting beats. These are observed rankings, not a validation-selected operator policy or a causal account of the difference.

Table E.6: **CREW operator comparison.** 20k encoders; P/S F_1 at ± 3 s; 74/147 fitting traces, fixed thresholds and final epoch-40 heads. Values are test percentages, mean $\pm$ sample SD over three seeds. Both auxiliary conditions retain the native objective. Δ is Shuffle minus TRD, with paired SD and positive-seed count. Budgets/readouts share encoders.

Host	Head / labels	Native	+TRD	+Shuffle	$\Delta \pm s_\Delta$	Positive
TS2Vec: no 20k native, reversal or shuffle endpoint (GPU memory).						
PatchTST	upsample/74	12.52 ± 2.15	26.66 ± 0.32	25.63 ± 3.04	-1.03 ± 2.76	1/3
PatchTST	upsample/147	21.71 ± 0.27	28.75 ± 1.18	27.71 ± 0.68	-1.05 ± 1.86	1/3
PMT	upsample/74	27.51 ± 3.22	65.69 ± 2.22	62.02 ± 1.79	-3.67 ± 1.64	0/3
PMT	upsample/147	33.70 ± 1.81	66.92 ± 2.41	64.12 ± 2.08	-2.79 ± 1.27	0/3

Table E.7: **LUDB operator comparison.** 5k encoders; P/T F_1 at ± 40 ms; 8/16/32/64 fitting beats and 52 calibration beats; final epoch-40 heads. TS2Vec uses an affine sample-grid readout, whereas patch-based hosts use learned upsampling. Values are test percentages, mean $\pm$ sample SD over three seeds. Both auxiliary conditions retain the native objective. Δ is Shuffle minus TRD, with paired SD and positive-seed count. Budgets/readouts share encoders.

Host	Head / labels	Native	+TRD	+Shuffle	$\Delta \pm s_\Delta$	Positive
TS2Vec	affine/8	57.01 ± 11.14	52.32 ± 8.83	73.45 ± 4.02	$+21.12 \pm 12.40$	3/3
TS2Vec	affine/16	55.50 ± 12.24	52.49 ± 2.24	71.74 ± 1.61	$+19.25 \pm 3.83$	3/3
TS2Vec	affine/32	56.73 ± 9.45	54.99 ± 5.57	74.79 ± 5.38	$+19.81 \pm 10.95$	3/3
TS2Vec	affine/64	58.83 ± 9.11	54.01 ± 3.89	72.44 ± 5.32	$+18.43 \pm 3.08$	3/3
PatchTST	upsample/8	45.71 ± 3.10	58.61 ± 5.82	59.01 ± 8.45	$+0.41 \pm 7.70$	2/3
PatchTST	upsample/16	49.29 ± 6.70	71.24 ± 3.53	69.97 ± 4.16	-1.27 ± 6.42	1/3
PatchTST	upsample/32	55.49 ± 3.65	76.86 ± 1.72	73.56 ± 5.57	-3.29 ± 6.08	1/3
PatchTST	upsample/64	52.85 ± 9.82	81.86 ± 2.38	77.96 ± 0.48	-3.90 ± 2.26	0/3
PMT	upsample/8	36.79 ± 4.54	55.37 ± 5.51	62.83 ± 7.36	$+7.46 \pm 2.38$	3/3
PMT	upsample/16	50.54 ± 5.72	61.33 ± 7.02	70.73 ± 7.02	$+9.40 \pm 4.25$	3/3
PMT	upsample/32	51.42 ± 3.44	73.10 ± 2.28	79.09 ± 1.98	$+5.99 \pm 3.17$	3/3
PMT	upsample/64	63.45 ± 4.39	69.82 ± 7.44	81.91 ± 1.69	$+12.09 \pm 9.13$	3/3

Pretext learning. Gain reaches mean training membership accuracy of 97.80%, against an 80% all-negative reference, with final BCE between 0.0559 and 0.0616 across seeds. Independent targets reach 80.00% training accuracy with BCE between 0.4780 and 0.4809, consistent with the nonuniform position prior of independently placed intervals. These are training pretext measurements, distinct from downstream onset F_1. Matched perturbation magnitude and loss weight do not equalize gradient scale, task difficulty or splice statistics.

TS2Vec's local-pooled auxiliary remains near chance, whereas whole-window orientation prediction learns well but yields smaller downstream gains than dense reversal supervision. PMT learns both pooled tasks well, yet dense local targets still give the stronger primary event readout. Replacement and shuffle also learn above the all-negative balanced-accuracy reference. Table E.8 summarizes BCE and balanced accuracy over the terminal training reporting interval. Balanced accuracy is calculated from aggregated positive and negative counts within each encoder before averaging across encoders, including runs with one-class batches. These statistics describe training, not held-out accuracy or whole-trajectory averages. The transformations differ in perturbation and splice statistics, so neither their pretext accuracy nor their downstream ordering identifies a unique directional cue.

E.2 Shared sequence decoding and head-fitting budget

The decoder is fitted once from the same 31 training patient groups used by the linear/5% heads: 121 recordings with 76,344 annotated frames on the 50 Hz grid. Transition counts use adjacent observed labels within continuous annotated support, and complete-duration fitting excludes censored runs at support and recording boundaries. Support is split before rasterization, including gaps shorter than one scoring-grid interval. The model uses smoothed transitions, log-normal duration hazards with an unbounded geometric tail, and a fixed unit prior weight. Its parameters are shared across hosts, conditions and encoder seeds, without additional calibration labels or per-host decoder tun-

Table E.8: CirCor auxiliary-task BCE and balanced accuracy over the terminal training reporting interval of 50k-update runs: three-encoder means. Balanced accuracy (BA) is derived from aggregated positive/negative counts within each encoder before averaging across encoders. These are training snapshots, not held-out pretext accuracy. Whole-window pooled tasks use $p = 0.5$; shuffle and replacement use $p = 1$.

Encoder	Auxiliary	BCE	BA (%)
PMT	Local dense, $p = 0.5$	0.0447	94.39
PMT	Local pooled, $p = 0.5$	0.1205	95.06
PMT	Whole-window pooled	0.0052	99.81
TS2Vec	Four-block shuffle	0.2296	81.70
TS2Vec	External replacement	0.2852	76.68
TS2Vec	Local dense, $p = 0.5$	0.0584	92.71
TS2Vec	Local pooled, $p = 0.5$	0.6979	49.28
TS2Vec	Whole-window pooled	0.0250	99.05

Table E.9: **Shared structured decoding on CirCor.** Test typed-onset macro-F_1 at 50 ms (%), mean $\pm$ sample SD over three 200k encoder seeds. Saved linear heads fit 31 training patient groups and use validation-based epoch selection. One transition/duration model fitted from those same training groups is shared across all conditions, without additional calibration; it uses whole-recording future context. Δ is +TRD minus Native with paired SD. +TRD retains the native objective. All paired gains are positive.

Host	Readout	Native	+TRD	Δ
TS2Vec	Raw	57.58 ± 1.41	91.44 ± 0.47	$+33.86 \pm 1.03$
TS2Vec	Decoded	90.97 ± 0.44	96.29 ± 0.11	$+5.32 \pm 0.35$
PatchTST	Raw	41.63 ± 2.91	55.86 ± 0.95	$+14.23 \pm 3.85$
PatchTST	Decoded	61.34 ± 6.90	86.89 ± 0.72	$+25.55 \pm 6.46$
PMT	Raw	45.01 ± 0.95	92.33 ± 0.34	$+47.32 \pm 0.96$
PMT	Decoded	77.00 ± 1.21	96.01 ± 0.15	$+19.00 \pm 1.26$

ing. Decoding uses whole-recording probabilities, future context and backtracking, without using evaluation annotation boundaries to partition the input. Scoring applies annotation support, the same 100 ms boundary exclusion and one-to-one typed-onset matching at 50 ms as for raw predictions.

Principal three-host test readouts. Table E.9 uses the principal 200k CirCor encoders and their saved linear/5% heads, whose epochs were selected on all 130 validation patient groups. The heads and feature statistics are reused without refitting on the 132-group, 465-recording test cohort, comprising 266,737 scored frames. The raw scores reproduce the principal results; all six raw/decoded TRD-minus-native contrasts are positive in every seed. The learned temporal prior narrows the TS2Vec and PMT gaps but widens the PatchTST gap. The independent-target, replacement and shuffle controls and the adapted encoders are outside this decoder comparison.

Head-fitting budget on validation. The separate 50k TS2Vec analysis in Table E.10 compares 120- and 600-epoch fitting budgets for frozen native, gain and reversal encoders, using the same three seeds and linear/5% readout. The learning rate and other head settings are held fixed, with validation-based epoch selection and the shared decoder above.

With the 120-epoch-budget heads, the decoded reversal-minus-native differences are 2.32, 4.71 and 4.52 points for seeds 42, 7 and 123; reversal-minus-gain differences are 2.87, 3.17 and 3.70. The decoder reduces native typed predictions from a seed mean of 42,436 to 19,343 for 19,552 reference onsets, while also changing recall and timing. Its recovery describes the complete readout pipeline rather than measuring which information the encoder already contained. Five times the fitting budget adds approximately 1–1.5 raw onset-F_1 points per arm, leaving the large raw gap and smaller decoded gap in Table E.10. Selected epochs are 572–598, but validation state-score improvements between epochs 480 and 600 are only 0.027–0.085 points.

Table E.10: **Readout effects at two head-fitting budgets.** CirCor validation, 50k TS2Vec encoders, linear/5% heads; typed-onset F_1 at 50 ms. The 120-epoch columns give mean ± sample SD over three encoder seeds; the 600-epoch columns give means. The final row gives mean paired differences. Heads use 120- or 600-epoch fitting budgets with the same learning rate and validation-based epoch selection. Decoder parameters are shared across both budgets.

	120-epoch budget		600-epoch budget	
Pretraining	Raw	Decoded	Raw	Decoded
Native	57.97 ± 0.79	89.88 ± 0.74	59.42	89.81
Native + gain	72.60 ± 1.32	90.48 ± 0.38	73.63	90.79
Native +TRD	$\mathbf{81.56} \pm 2.66$	$\mathbf{93.73} \pm 0.76$	**82.78**	**93.96**
+TRD minus Native	+23.59	+3.85	+23.36	+4.15

E.3 End-to-end adaptation with limited fitting labels

We compare scratch, native-pretrained and Native+TRD starts for TS2Vec, PatchTST and PMT at 7 and 31 fitting patient groups (1%, 5%), using seeds 42, 7 and 123. Pretrained starts use the fixed 200k encoders; fitting memberships and head initialization are matched across starting conditions. The encoder and the two-width-five-convolution temporal head (64 hidden channels, GELU and output layer; initialization seed 31415) are trained together on clean inputs with masked state cross-entropy, without native SSL or TRD losses. Fresh AdamW uses batch 32, weight decay 10^{-4} and a 120-epoch limit; the 31-group budget contains 809 fitting windows and gives 26 updates per epoch.

Selection and held-out scoring. All reported adaptation runs use learning rate 10^{-3}. For TS2Vec and PMT at 31 fitting groups, a common rate per host is selected from $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$ by the equally weighted mean of best validation state-F_1 across the nine start/seed combinations, breaking ties toward the lower rate; both select 10^{-3}. PatchTST at 31 groups and all hosts at 7 groups use this rate as a fixed transferred setting, without an additional rate search. Every run selects the earliest epoch attaining its highest state macro-F_1 on all 130 validation patient groups; these selection labels are additional to the fitting groups. Starting-encoder feature statistics are fitted on all 13,627 training windows and kept fixed during adaptation, without validation/test statistics. TS2Vec adapts its averaged starting encoder without continuing self-supervised averaging. The complete selected encoder/head pair is scored on the 132-group, 465-recording test cohort with the unchanged protocol in Appendix A.2, without further updates, calibration, selection or decoding. The finite fitting budget does not establish convergence or an optimized supervised ceiling; a PMT scratch run at 31 groups selects epoch 120.

TRD-pretrained starts improve both onset and state macro-F_1 over native and scratch starts in every seed on every host at both fitting budgets. Table E.11 gives all absolute means and SDs; Table E.12 gives the paired onset gains and their SDs, computed before rounding.

Comparison with frozen features and across fitting budgets. The frozen temporal-readout means exceed both scratch and native-start adaptation means in all six host/budget settings (Table 3). Frozen fitting uses batch 64 and a fixed 10^{-3} rate; adaptation uses batch 32 and the rate policy above, with the same temporal-head architecture and validation cohort. These are comparisons of evaluated pipelines, not an intervention changing only encoder freezing. Seven-group TRD adaptation has mean onset F_1 of 92.75, 80.92 and 93.09 on TS2Vec, PatchTST and PMT, compared with 92.46, 77.94 and 90.07 for 31-group scratch training. This is a descriptive comparison of means, not an equivalence test or a claim that every seed wins across budgets. Both budgets use the same 130 additional validation groups, so this comparison concerns fitting-label efficiency, not a fourfold reduction in total annotation use.

E.4 Event identity with supplied location and altered context

Supplied true intervals at 50k. For the same 19,552 annotated events (9,686 S1 and 9,866 S2), we classify S1 versus S2 using the existing linear/5% probabilities inside each true interval. No head is refitted and encoder context is intact. The central-frame variant tests dependence on interval

Table E.11: **Complete CirCor adaptation outcomes.** Test typed-onset macro-F_1 at 50 ms and state macro-F_1 (%), mean ± sample SD over three seeds. Groups count fitting patients, not the additional 130 validation groups used for epoch selection. Native and TRD starts use 200k pretrained encoders.

Host	Groups	Starting point	Onset F_1	State F_1
TS2Vec	7	Scratch	80.26 ± 1.44	80.18 ± 0.30
TS2Vec	7	Native	87.49 ± 0.36	83.50 ± 0.31
TS2Vec	7	Native +TRD	92.75 ± 0.77	87.46 ± 0.46
TS2Vec	31	Scratch	92.46 ± 0.24	88.50 ± 0.26
TS2Vec	31	Native	93.72 ± 0.13	89.13 ± 0.14
TS2Vec	31	Native +TRD	95.94 ± 0.26	90.52 ± 0.06
PatchTST	7	Scratch	70.44 ± 0.76	66.77 ± 0.43
PatchTST	7	Native	69.06 ± 1.01	65.03 ± 0.65
PatchTST	7	Native +TRD	80.92 ± 2.69	74.89 ± 3.70
PatchTST	31	Scratch	77.94 ± 0.44	73.87 ± 0.02
PatchTST	31	Native	78.84 ± 0.26	73.69 ± 0.36
PatchTST	31	Native +TRD	87.34 ± 2.23	82.20 ± 3.98
PMT	7	Scratch	80.00 ± 0.64	77.52 ± 0.41
PMT	7	Native	79.37 ± 0.81	76.98 ± 0.44
PMT	7	Native +TRD	93.09 ± 0.15	87.59 ± 0.17
PMT	31	Scratch	90.07 ± 0.28	86.05 ± 0.09
PMT	31	Native	89.91 ± 0.12	85.60 ± 0.12
PMT	31	Native +TRD	94.64 ± 0.14	89.43 ± 0.03

Table E.12: **Paired CirCor adaptation gains.** Typed-onset macro-F_1 differences (percentage points), mean and sample SD of three seed-matched differences computed before rounding. +TRD denotes a Native+TRD pretrained start.

Host	Groups	Contrast	$\Delta \pm s_\Delta$	Positive
TS2Vec	7	+TRD − Native	$+5.26 \pm 1.08$	3/3
TS2Vec	7	+TRD − Scratch	$+12.49 \pm 1.78$	3/3
TS2Vec	7	Native − Scratch	$+7.23 \pm 1.17$	3/3
TS2Vec	31	+TRD − Native	$+2.23 \pm 0.33$	3/3
TS2Vec	31	+TRD − Scratch	$+3.48 \pm 0.39$	3/3
TS2Vec	31	Native − Scratch	$+1.25 \pm 0.11$	3/3
PatchTST	7	+TRD − Native	$+11.86 \pm 3.08$	3/3
PatchTST	7	+TRD − Scratch	$+10.48 \pm 1.98$	3/3
PatchTST	7	Native − Scratch	-1.38 ± 1.22	0/3
PatchTST	31	+TRD − Native	$+8.49 \pm 2.31$	3/3
PatchTST	31	+TRD − Scratch	$+9.40 \pm 2.59$	3/3
PatchTST	31	Native − Scratch	$+0.90 \pm 0.59$	3/3
PMT	7	+TRD − Native	$+13.72 \pm 0.70$	3/3
PMT	7	+TRD − Scratch	$+13.09 \pm 0.59$	3/3
PMT	7	Native − Scratch	-0.63 ± 0.35	0/3
PMT	31	+TRD − Native	$+4.73 \pm 0.26$	3/3
PMT	31	+TRD − Scratch	$+4.57 \pm 0.14$	3/3
PMT	31	Native − Scratch	-0.16 ± 0.39	1/3

averaging. Table E.13 shows that the identity advantage survives supplying event location; it does not decompose onset-F_1 into separate causal detection and identity contributions.

Context interventions at 25k. The context intervention uses 25k TS2Vec encoders and an annotation-assisted cohort of 3,011 training events from 28 patients within the 31-patient training mask, and 10,342 validation events from 114 patients. An unchanged 160 ms focal sound is centered in a four-second input, with 10 ms taper skirts on each side. A linear probe is fitted sepa-

Table E.13: **Event-identity readout with true locations supplied.** CirCor validation, 50k TS2Vec; mean ± sample SD over three encoder seeds. Identity is S1/S2 balanced accuracy (%), while untyped onset F_1 uses 50 ms matching. Bold marks column-best means. +TRD retains the native objective.

Condition	Interval-mean identity	Central-frame identity	Untyped onset F_1
Native	89.50 ± 0.66	83.26 ± 0.56	83.60 ± 0.77
Independent targets	85.97 ± 1.49	78.55 ± 1.78	78.64 ± 1.95
Gain	93.10 ± 0.30	90.48 ± 0.35	88.75 ± 0.66
+TRD	$\mathbf{95.78} \pm 0.68$	$\mathbf{94.29} \pm 1.10$	$\mathbf{91.37} \pm 0.74$

Table E.14: **Identity readout after changing context.** CirCor validation with 25k TS2Vec encoders; S1/S2 balanced accuracy (%), mean ± sample SD over three encoder seeds. Probes are fitted separately for each input condition on the annotation-assisted cohort described in the text. Bold marks row-best means. +TRD retains the native objective.

Input	Initialization	Native	Gain	+TRD
Natural context	81.83 ± 0.47	93.09 ± 1.11	96.27 ± 0.51	$\mathbf{98.38} \pm 0.36$
Original slots, no background	84.92 ± 0.48	97.31 ± 0.19	98.63 ± 0.17	$\mathbf{98.99} \pm 0.17$
Focal sound only	69.22 ± 0.28	76.06 ± 0.29	80.51 ± 0.78	$\mathbf{87.29} \pm 0.57$
Shuffled neighbours	81.19 ± 0.77	96.35 ± 0.14	$\mathbf{98.05} \pm 0.26$	97.41 ± 0.09

rately for each encoder seed, training arm and input condition under the same 120-epoch fitting and validation-selection policy. The comparison measures accessible identity information under each input distribution, rather than changing the inputs to one fixed classifier.

With only the focal sound, reversal exceeds native by 11.23 points and gain by 6.78, in every seed. After neighbouring packets are shuffled among their original slots, gain exceeds reversal by 0.64 points. Annotation-based reconstruction preserves slot timing, and local packets retain waveform shape and recording-level preprocessing effects. These findings show accessible local identity information and sensitivity to context, without identifying whether local or between-event supervision taught those features.

E.5 Teaser excerpt selection

Figure 1 uses 200k-update validation predictions from the linear heads fitted on 5% of CirCor training patient groups. The four arms are native PMT, PMT+TRD, native TS2Vec and TS2Vec+TRD; the figure shows encoder seed 42.

Of 486 validation recordings, 355 were eligible: each came from a standard auscultation position, contained at least eight eligible S1 and eight S2 onsets, and had an annotated run of at least six seconds containing a complete four-second probe window. For each eligible recording, the four record-level typed-onset F_1 scores and the two within-host TRD gains were converted to percentiles within this candidate set. The recording minimizing the largest absolute deviation of these six percentiles from the median was selected. This chose recording `50798_PV`. Among its complete four-second probe windows, the window whose per-arm scores were closest to the recording-level scores was selected: window 6, covering 24–28 s.

The right-hand values are the whole-validation typed-onset macro-F_1 scores at 50 ms for seed 42, rather than excerpt scores. Per-record scoring was checked against the saved whole-split event counts for all four arms and all three encoder seeds.

F Objective-removal ablation: complete matrix

F.1 Complete test endpoints

Tables F.1 and F.2 compare initialization, native training, joint training and TRD-only training on CirCor and OpenPack. Five encoder recipes are evaluated with linear and temporal readouts at both

Table F.1: **CirCor: complete objective-removal test matrix.** typed-onset macro-F_1 at 50 ms, mean $\pm$ sample SD over three encoder seeds. Bold marks the highest trained mean in each column within a recipe; initialization is a separate reference.

Training	Linear / 1%	Linear / 5%	Temporal / 1%	Temporal / 5%
TS2Vec				
Initialization	46.30 ± 0.73	47.22 ± 0.62	63.28 ± 1.62	69.91 ± 0.75
Native	52.09 ± 1.77	57.58 ± 1.41	77.40 ± 1.70	82.63 ± 1.31
Native +TRD	$\mathbf{88.16} \pm 0.93$	$\mathbf{91.44} \pm 0.47$	$\mathbf{92.73} \pm 0.30$	$\mathbf{95.16} \pm 0.08$
TRD-only	45.95 ± 37.72	46.42 ± 39.98	58.69 ± 29.05	62.54 ± 28.58
Masked-PatchTST				
Initialization	42.23 ± 3.40	49.66 ± 0.62	56.96 ± 0.33	64.90 ± 0.54
Native	32.64 ± 4.34	41.63 ± 2.91	49.04 ± 2.95	55.74 ± 2.62
Native +TRD	53.04 ± 2.12	$\mathbf{55.86} \pm 0.95$	$\mathbf{74.95} \pm 0.46$	$\mathbf{80.17} \pm 0.28$
TRD-only	$\mathbf{53.47} \pm 0.66$	54.66 ± 0.43	73.20 ± 0.56	79.94 ± 0.28
PMT				
Initialization	45.28 ± 0.90	48.47 ± 0.10	57.97 ± 1.21	65.46 ± 0.99
Native	39.64 ± 1.52	45.01 ± 0.95	62.76 ± 0.63	67.39 ± 0.52
Native +TRD	$\mathbf{89.55} \pm 1.18$	92.33 ± 0.34	$\mathbf{94.06} \pm 0.94$	95.44 ± 0.04
TRD-only	86.12 ± 1.13	$\mathbf{92.48} \pm 0.03$	91.19 ± 1.03	$\mathbf{95.58} \pm 0.09$
TimeDRL				
Initialization	24.66 ± 0.60	24.11 ± 0.34	50.82 ± 0.46	57.41 ± 0.41
Native	$\mathbf{37.85} \pm 4.54$	$\mathbf{37.97} \pm 5.36$	$\mathbf{57.33} \pm 1.27$	$\mathbf{65.70} \pm 0.89$
TRD-only	25.92 ± 1.23	25.97 ± 1.59	48.46 ± 0.85	55.32 ± 0.43
PPT-configured				
Initialization	31.92 ± 0.33	30.72 ± 1.09	52.14 ± 0.83	59.97 ± 0.38
Native	32.26 ± 8.06	31.88 ± 9.15	43.92 ± 11.52	46.14 ± 10.32
Native +TRD	34.80 ± 1.77	35.08 ± 1.63	50.01 ± 1.34	52.27 ± 0.68
TRD-only	$\mathbf{43.85} \pm 1.43$	$\mathbf{48.02} \pm 2.73$	$\mathbf{66.96} \pm 1.08$	$\mathbf{74.08} \pm 0.75$

fitting budgets and three encoder seeds. Initialization is the untrained encoder of the standalone recipe with a fitted readout; trained conditions use 200k updates. Native training uses the recipe's self-supervised objective; joint training adds TRD, while standalone training removes the native objective together with its views and branches. TimeDRL has no joint condition, and PPT-configured and masked-PatchTST are two configurations of the same encoder family. Validation-selected heads, training-feature normalization and label draws are applied to test without additional fitting or checkpoint selection. Head-fitting and additional validation-supervision budgets follow Table A.1. The objective-removal comparison uses test cohorts previously evaluated for native and joint training.

Seed dispersion and readout dependence. PMT/CirCor standalone primary scores are 92.51, 92.45 and 92.49, with mean 92.48 and sample SD 0.03. Its mean difference from joint is 0.15 points, with mixed paired signs; the larger-budget temporal means are also close, 95.58 standalone and 95.44 joint. TS2Vec/CirCor instead has primary standalone scores of 26.22, 92.47 and 20.57, so its mean of 46.42 obscures one strong and two weak outcomes. All three joint runs score above 91 on that readout.

On OpenPack's primary readout, TS2Vec standalone exceeds native by 2.00 points in every seed, whereas its temporal/5-minute readout is 2.98 points below native in every seed. Joint PMT and masked-PatchTST have primary means 2.06 and 5.24 points above standalone, respectively, with mixed paired signs for both comparisons. All five standalone recipes improve over their own primary-readout initialization in every OpenPack seed.

Additional recipes. The PPT-configured encoder improves over joint training on both larger-budget readouts in every seed. Linear and temporal gains are 12.94 and 21.81 points on CirCor, and 12.25 and 10.88 on OpenPack. OpenPack primary standalone scores are 51.18, 39.95 and 49.51; this dispersion remains part of the result. Removing native objectives also removes the associated views and branches, so the experiment identifies a useful standalone recipe rather than an isolated effect

Table F.2: **OpenPack: complete objective-removal test matrix.** operation macro-F_1 at 1 Hz, mean $\pm$ sample SD over three encoder seeds. Bold marks the highest trained mean in each column within a recipe; initialization is a separate reference.

Training	Linear / 1 min	Linear / 5 min	Temporal / 1 min	Temporal / 5 min
TS2Vec				
Initialization	16.94 ± 0.74	19.94 ± 0.88	19.86 ± 0.24	31.53 ± 1.17
Native	$\mathbf{29.71} \pm 0.53$	34.07 ± 0.62	$\mathbf{33.50} \pm 0.53$	$\mathbf{42.75} \pm 1.09$
Native +TRD	28.44 ± 1.06	$\mathbf{36.32} \pm 2.01$	26.97 ± 0.85	39.48 ± 2.47
TRD-only	26.91 ± 0.76	36.07 ± 0.75	26.12 ± 0.45	39.77 ± 1.31
Masked-PatchTST				
Initialization	18.67 ± 0.47	25.11 ± 0.10	21.00 ± 0.45	27.40 ± 0.32
Native	31.29 ± 1.08	48.07 ± 0.95	33.79 ± 0.35	50.17 ± 1.39
Native +TRD	$\mathbf{34.97} \pm 5.50$	$\mathbf{53.39} \pm 4.46$	$\mathbf{33.81} \pm 5.88$	$\mathbf{51.95} \pm 4.16$
TRD-only	31.63 ± 2.74	48.15 ± 3.57	27.63 ± 3.07	44.00 ± 3.17
PMT				
Initialization	14.95 ± 0.42	17.80 ± 0.63	15.10 ± 0.24	22.87 ± 0.11
Native	19.90 ± 0.72	30.38 ± 1.69	18.23 ± 1.53	30.27 ± 0.74
Native +TRD	$\mathbf{32.83} \pm 1.37$	$\mathbf{48.54} \pm 0.53$	29.49 ± 0.48	$\mathbf{47.17} \pm 0.54$
TRD-only	31.53 ± 5.08	46.47 ± 3.13	$\mathbf{30.08} \pm 3.44$	46.49 ± 3.21
TimeDRL				
Initialization	17.63 ± 0.28	22.86 ± 0.20	18.90 ± 0.14	24.78 ± 0.03
Native	18.53 ± 0.20	24.84 ± 1.15	19.94 ± 0.45	27.39 ± 1.18
TRD-only	$\mathbf{19.19} \pm 0.42$	$\mathbf{26.91} \pm 0.57$	$\mathbf{20.34} \pm 0.56$	$\mathbf{30.03} \pm 0.10$
PPT-configured				
Initialization	15.64 ± 0.05	19.58 ± 0.12	16.41 ± 0.24	23.48 ± 0.16
Native	14.86 ± 1.23	20.94 ± 0.88	15.98 ± 0.90	20.59 ± 0.21
Native +TRD	22.32 ± 1.57	34.63 ± 2.14	24.59 ± 1.53	35.15 ± 3.02
TRD-only	$\mathbf{30.18} \pm 4.78$	$\mathbf{46.88} \pm 6.06$	$\mathbf{29.51} \pm 5.01$	$\mathbf{46.03} \pm 5.61$

Table F.3: LUDB standalone development results, P/T onset F_1 at ± 40 ms: initialization $\rightarrow$ 200k, three-encoder means. Counts are labeled beats, not patients.

Recipe	$\sim$8 beats	$\sim$16	$\sim$32	$\sim$64
PMT	$50.71 \rightarrow 35.35$	$52.98 \rightarrow 38.32$	$50.43 \rightarrow 42.33$	$56.14 \rightarrow 45.50$
TS2Vec	$0.00 \rightarrow 39.49$	$0.00 \rightarrow 38.17$	$0.00 \rightarrow 44.84$	$0.00 \rightarrow 43.96$
Masked-PatchTST	$49.61 \rightarrow 36.63$	$52.75 \rightarrow 48.14$	$56.48 \rightarrow 58.96$	$59.50 \rightarrow 63.59$
TimeDRL	$62.84 \rightarrow 62.68$	$65.47 \rightarrow 72.72$	$73.89 \rightarrow 79.26$	$77.08 \rightarrow 83.36$
PPT-configured	$35.66 \rightarrow 31.39$	$39.98 \rightarrow 40.66$	$33.21 \rightarrow 46.69$	$34.98 \rightarrow 62.32$

of a single loss gradient. TimeDRL/CirCor has a small primary improvement over initialization but lower temporal scores; on OpenPack it improves over native on both larger-budget readouts.

Numerical verification. Validation replay uses the source numerical backend and a maximum per-metric tolerance of 0.001 on the 0–1 scale. Verification coverage differs across encoders: some evaluations include probability checks, while others were verified only at the metric level. Nine evaluations use probability tolerances calibrated on validation; their maximum metric discrepancy is 0.0670 percentage points. Test scores were not used to select replay tolerances, model parameters, head epochs or repeated attempts. These checks establish numerical reproducibility to the stated tolerances, not equality of all model probabilities or an explanation of pretraining outcomes.

F.2 LUDB STANDALONE DEVELOPMENT ENDPOINTS

Table F.3 gives initialization and 200k development outcomes at every label budget. TimeDRL improves at its three larger budgets in every seed, while its approximately eight-beat mean is unchanged to within 0.2 points. A matched native/joint LUDB comparison at 200k updates is unavailable.

Table F.4: CirCor validation trajectories, linear/5% typed-onset F_1 at 50 ms. Native and +TRD columns are three-seed means; TRD-only columns are individual seeds. Dashes mark unmeasured Native/+TRD checkpoints.

Updates	Native	+TRD	TRD-only 42	TRD-only 7	TRD-only 123
0	45.16	45.15	44.58	45.94	44.94
5k	—	—	30.74	30.09	31.86
25k	58.67	78.22	29.71	29.32	30.93
50k	57.97	81.56	27.80	46.38	31.11
100k	56.95	85.30	31.03	82.16	28.32
200k	56.24	88.10	24.47	89.32	20.06

Table F.5: LUDB development trajectories at approximately 64 labeled beats, standalone training; P/T onset F_1 at ± 40 ms, three-encoder means. Columns evaluate the stated checkpoints under each recipe's training schedule.

Recipe	Initialization	5k	25k	50k	100k	200k
PMT	56.14	73.45	59.66	56.45	43.05	45.50
TS2Vec	0.00	22.97	25.71	28.71	40.48	43.96
Masked-PatchTST	59.50	74.98	72.68	73.41	70.22	63.59
TimeDRL	77.08	81.90	83.50	82.49	87.10	83.36
PPT-configured	34.98	76.70	69.07	61.87	63.64	62.32

F.3 Standalone training dynamics

TS2Vec/CirCor. Table F.4 separates the consistently useful joint recipe from the seed-sensitive standalone trajectories. The eventual successful standalone seed remains below initialization at 25k and improves later; the two weak seeds do not show comparable recovery. Their failure is therefore not simply a late-training decline.

We evaluate the 200k encoders on a common 64-patient validation bank to characterize the standalone failures. For the tested reversal placement, failed standalone seeds 42 and 123 have membership BCE values of 0.48092 and 0.48062, respectively. Both are near the position prior's 0.48069, with zero positive recall. Their logits change little when the interval moves and resemble the position-prior curve even on an all-zero input; their live features also lose variation across waveforms. The successful seed has BCE 0.05993 and positive recall 94.25% on the same placement, while all three joint encoders discriminate the corruption strongly. These pretext measurements are separate from the downstream onset scores in the table. They characterize a waveform-insensitive solution in the failed models, but do not isolate why optimization reaches it. Pretraining corruption placements are unavailable, and initialization, source order, dropout and corruption sampling vary together.

LUDB and OpenPack duration effects. Table F.5 reports the larger-budget LUDB trajectories for all five standalone recipes. PMT's decline from 5k to 200k occurs in every seed at every LUDB label budget, while final training reversal accuracy is approximately 99.98% and BCE 0.00055. Thus successful training-pretext discrimination and useful frozen event readout need not improve together. On OpenPack, standalone PMT's primary validation score falls from 54.75 at 100k to 49.80 at 200k in every seed.

G Design sensitivity

We examine the auxiliary weight, reversal span, token-target convention, and encoder stem and patch width, with formulation checks described below. Each experiment specifies its dataset, endpoint and replication.

Table G.1: PMT/CirCor weight and span sensitivity after 50k encoder updates: validation typed-onset macro-F_1 at 50 ms from linear heads fitted on 31 training patient groups (5%) and selected on validation. Values are mean ± sample SD of paired gains over three encoder seeds, in percentage points. The central setting is $\lambda = 1.5$, span 0.8 s. The six settings form a one-factor-at-a-time union, not a Cartesian grid; all other settings and the 10 ms fade are fixed. Span changes also change positive-target prevalence.

λ	Span (s)	$\Delta_{\text{native}} \pm s$	$\Delta_{\text{central}} \pm s$	Positive vs native
0.5	0.8	41.83 ± 1.43	-3.40 ± 1.03	3/3
1.5	0.8	45.23 ± 1.82	Reference	3/3
4.5	0.8	46.25 ± 2.34	1.02 ± 0.61	3/3
1.5	0.2	40.54 ± 2.65	-4.69 ± 1.08	3/3
1.5	0.4	45.50 ± 1.32	0.27 ± 0.66	3/3
1.5	1.6	38.18 ± 2.54	-7.05 ± 1.03	3/3

Table G.2: Exhaustive label-grid comparison with valid nominal convolution support. Disagreement is reported across all valid targets and across the union of positions positive under either convention, enumerating every valid interval placement for each dataset.

Dataset	Patch/stride	Placements	All (%)	Union-positive (%)
CirCor	161/80	12801	0.992188	4.841418
OpenPack	15/6	1321	0.720918	2.670069
CREW	101/50	28001	0.329851	4.829448
LUDB	11/5	301	1.790698	4.388893

G.1 Weight, span and target-coordinate sensitivity

For PMT/CirCor, the default uses auxiliary weight 1.5 and a 0.8 s span. We vary the weight at this span and the span at this weight, using the matched 50k schedule and a 10 ms fade. The settings form a six-setting one-factor-at-a-time union, not a Cartesian grid. All six improve the primary validation readout over native training in every encoder seed, with mean gains of 38.18–46.25 points (Table G.1). The corruption scale affects the gain magnitude and changes the positive-target fraction.

The implementation appendix specifies the token-coordinate convention used in the principal comparisons. We enumerate every valid interval start to compare that convention with a target based on valid nominal convolution support: a token is positive when at least half of its valid nominal support overlaps the half-open corruption interval. The encoder's token-validity mask continues to determine loss support. Table G.2 measures label disagreement; nominal convolution support is distinct from the final contextual receptive field. Downstream sensitivity is evaluated on CirCor with three encoder seeds: support-aligned targets change the primary test score by -0.31 ± 0.39 points, where uncertainty is the sample SD of paired changes. This is small relative to the native-to-joint gain in the tested configuration.

Boundary regression. A boundary variant adds a clipped signed-distance-to-splice Smooth-L1 term inside the auxiliary weight. On LUDB development data at 5k encoder updates, it improves the ±20 ms endpoint at the two smallest budgets in every seed but has mixed effects elsewhere.

Span selection. The 800 ms CirCor span was selected using a PMT validation comparison with 3 seeds, 4,260 encoder updates and 30-epoch heads, then shared across every evaluated CirCor host.

Stem and patch-width sensitivity. On CirCor with one seed, 10k updates and 60-epoch heads, the linear/5% gain was +35.44, +30.14, +30.32 and +28.76 across a standard and an affine stem at 40 and 80 ms patch widths. All four configurations give positive gains in this one-seed comparison.

Table H.1: **Training-core time and peak allocated memory.** Native versus Native +TRD on one GH200. Time is milliseconds per optimizer update; overhead and its repeat range are percentages. PMT and PatchTST use fused views; TS2Vec uses a separate auxiliary pass.

Host	Batch	Native (ms)	+TRD (ms)	Overhead (%)	Repeat range (%)
CirCor					
PMT	32	95.83	98.52	+2.8	[−5.9, +10.5]
TS2Vec	2	99.15	123.40	+24.5	[+11.5, +41.1]
PatchTST	64	14.40	18.50	+28.5	[+15.4, +44.9]
OpenPack					
PMT	32	97.18	107.51	+10.6	[−0.2, +19.1]
TS2Vec	8	35.93	45.97	+27.9	[+23.0, +37.5]
PatchTST	64	14.32	17.80	+24.3	[+15.3, +35.5]

Dataset	Host	Native peak (GiB)	+TRD peak (GiB)
CirCor	PMT	2.37	3.18
	TS2Vec	31.50	31.50
	PatchTST	1.30	2.49
OpenPack	PMT	3.36	4.58
	TS2Vec	2.08	2.08
	PatchTST	1.41	2.74

H Computational cost and execution details

H.1 Measurement protocol and native–joint comparison

Measurements use one NVIDIA GH200 (120 GB), PyTorch 2.10, FP32 with TF32 and cuDNN benchmarking, and one numerical CPU thread. A training-core update includes corruption, all encoder views, losses, the backward pass, gradient clipping, the optimizer step and, for TS2Vec, encoder averaging. It excludes data loading, host-side augmentation, logging, checkpointing and downstream probes. Each profile has four repeats, each comprising five windows of 20 updates; all measured windows are included. Table H.1 reports profile-level update times and overheads, with the overhead range across repeats. Peak memory is the maximum allocated-memory peak over measured windows after optimizer initialization. Batches and native recipes differ across hosts, so the comparisons are native versus joint within each profile.

The auxiliary head adds 129–385 parameters in the measured hosts; most additional work comes from encoding the extra view and retaining its activations. Reversing a short interval does not shorten that encoder input. The profile-level overheads in Table H.1 range from 2.8% to 28.5%; individual repeat ranges are wider. Peak allocated memory increases for PMT and PatchTST. TS2Vec's peak is unchanged because allocations from its separate passes do not peak together, although the auxiliary still allocates memory.

H.2 Execution-mode sensitivity

The execution-mode benchmark compares fused and unfused execution against its corresponding measured native baselines (Table H.2). Each overhead uses the native baseline measured for that benchmark; denominators are not shared with Table H.1. For PMT, the measured unfused path includes compiler dispatch: the separate auxiliary batch of 32 falls back to eager execution after clean views are compiled at batch 64. The unfused overhead therefore describes that execution path, not a version with both batch shapes compiled. PatchTST fusion reduces update time by 27.1% on CirCor and 21.5% on OpenPack relative to its unfused path, while increasing peak allocated memory by 0.37 and 0.30 GiB, respectively. The evaluated TS2Vec implementation has no fused path.

H.3 Standalone training

Standalone training keeps one corrupted encoder pass and removes the native objective's views and branches. The objective-removal benchmark measures native, joint and standalone training with four

Table H.2: **Execution-mode sensitivity.** Update-time overhead relative to the native baseline measured in the execution-mode benchmark, in percent. The PMT unfused path includes the eager fallback described in the text.

Dataset	Host	Fused overhead (%)	Unfused overhead (%)
CirCor	PMT	+9.9	+160
OpenPack	PMT	+4.8	+155
CirCor	PatchTST	+23.2	+69
OpenPack	PatchTST	+26.4	+61

Table H.3: **Standalone training-core cost.** Milliseconds per optimizer update in the objective-removal benchmark, averaged over four paired repeats. Savings are percentages relative to the named reference. Batches and encoder configurations remain recipe-specific. +TRD retains the native objective; TRD-only removes it.

		Update time (ms)			Savings (%)	
Recipe	Batch	Native	+TRD	TRD-only	vs Native	vs +TRD
CirCor						
PMT	32	101.05	102.92	71.12	+29.6	+30.9
TS2Vec	2	98.84	112.88	9.58	+90.3	+91.5
Masked-PatchTST	64	14.59	18.41	15.45	−5.9	+16.1
TimeDRL	32	13.23	—	8.66	+34.5	—
PPT-configured	32	38.39	49.51	13.98	+63.6	+71.8
OpenPack						
PMT	32	102.46	105.48	73.14	+28.6	+30.7
TS2Vec	8	38.59	45.72	9.57	+75.2	+79.1
Masked-PatchTST	64	15.34	19.18	15.01	+2.1	+21.7
TimeDRL	32	12.72	—	7.89	+38.0	—
PPT-configured	32	54.53	60.44	13.52	+75.2	+77.6

paired repeats under the measurement protocol above (Table H.3). Savings use the corresponding native or joint time measured within this benchmark. Joint execution follows the efficacy path: fused PMT and masked-PatchTST, and separate-pass TS2Vec and PPT. TimeDRL has no joint reference. Savings are relative to the indicated reference; negative savings indicate a slower standalone update.

Standalone PMT reduces update time by 30.9% on CirCor and 30.7% on OpenPack relative to fused joint training. Its peak allocated memory is 1.11 and 1.54 GiB, respectively, compared with 3.19 and 4.58 GiB for joint training in the objective-removal benchmark. The corresponding CirCor primary test scores are close (Section 4.3); on OpenPack, standalone has a lower mean endpoint. The large TS2Vec savings coexist with two failed standalone CirCor seeds. Masked-PatchTST saves time relative to joint training but has little advantage over its already single-view native objective: the reported savings are -5.9% on CirCor and $+2.1\%$ on OpenPack.

H.4 Whole-job runtime and estimated training cost

Whole-job wall times include loading, probes, startup and other work excluded from the training-core timings. Mean native-to-joint times over three encoder seeds are 7.04 to 7.97 h for CirCor PMT, 22.45 to 24.06 h for CirCor TS2Vec, and 2.38 to 2.70 h for CirCor PatchTST. The corresponding OpenPack pairs are 8.25 to 8.19 h, 14.84 to 13.62 h, and 3.17 to 3.27 h. These are end-to-end observations, not controlled measurements of the auxiliary's update overhead.

Estimated training cost. For each validation checkpoint, training-core time is estimated as its update count multiplied by the measured per-update time. At the smaller of a pair's estimated 200k-update budgets, the best available validation score with TRD exceeds native for all three CirCor hosts and for PMT on OpenPack, but not for TS2Vec on OpenPack. This comparison uses checkpoints from each recipe's training schedule and validation-based selection; it is not an equal-compute retraining experiment.

H.5 Inference and structured decoding

The corruption and auxiliary head are absent from frozen-feature extraction in every evaluated profile. The encoder and downstream readout architectures are unchanged between native and joint training. For the 50k TS2Vec validation diagnostic (Appendix E.2), structured decoding takes approximately 0.3 CPU seconds for 486 CirCor recordings, or 0.03 ms per second of audio, excluding encoder work. This decoder timing concerns the offline readout diagnostic, separately from the cost of adding TRD during pretraining.

I Time reversibility and what a local reversal leaves detectable

A stationary process is time-reversible when its finite-dimensional distributions are unchanged by reversing the time index. Every stationary scalar Gaussian process is reversible: its law is fixed by its constant mean and an autocovariance symmetric in the lag (Weiss, 1975). Thus an unblended reversed interval has the same marginal distribution as the original interval, but its relation to unchanged context can differ.

For example, consider a stationary, unit-variance Gaussian AR(1) process with $0 < \rho < 1$ and $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$. Reversing $I = [a, a+s)$ without a fade places X_{a+s-1} next to X_{a-1}, changing the covariance across that splice from ρ to ρ^s. Crossfading softens the abrupt splice but can introduce a local variance cue:

$$\begin{aligned}&\mathrm{Var}[\alpha_j X_{a+s-1-j} + (1-\alpha_j) X_{a+j}] \\ &\quad = 1 - 2\alpha_j(1-\alpha_j)\left(1 - \rho^{|s-1-2j|}\right).\end{aligned} \tag{I.1}$$

This is below one when $0 < \alpha_j < 1$ and the two mixed indices are distinct. The faded samples can therefore have lower variance than their surroundings even though the source process is reversible.

These examples show why source reversibility alone does not make local-reversal membership unpredictable, or imply that the auxiliary cannot benefit a downstream task. They do not establish which cue the trained encoders use. Section 6 therefore distinguishes temporal direction as a motivation from an identified cause of the gains.