TEMPORAL STRUCTURE IN CONTRASTIVE EEG REPRESENTATION LEARNING: A CONTROLLED STUDY FOR OBSTRUCTIVE SLEEP APNEA DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive self-supervised learning for physiological signals typically treats windows as independent apart from augmentation-defined positives, ignoring that windows from the same patient are temporally related. We introduce *Temporal NT-Xent*, which retains augmentation-defined positives and full-strength cross-patient negatives while continuously down-weighting the repulsive contribution of same-patient negatives according to their temporal distance. We evaluate this objective for EEG-based obstructive sleep apnea (OSA) detection along three complementary axes. First, on a canonical patient-held-out split — patients unseen during pretraining, validation, or test selection — Temporal NT-Xent produces a substantial downstream change relative to vanilla NT-Xent, raising AUROC from 0.5417 to 0.7667 and AUPRC from 0.4013 to 0.6673 (paired permutation p=0.0455 for AUROC). Second, the same frozen, 131K-parameter encoder transfers zero-shot to the independent SHHS-1 cohort without any apnea labels or target-domain fine-tuning, reaching AUROC $= 0.724$ (95% CI $[0.641, 0.801]$) against a single-channel source-to-target control of 0.718 — evidence that the learned representation carries signal beyond the source cohort, distinct from the in-domain question the canonical split answers. Third, a five-seed robustness analysis on the source cohort shows that the magnitude of the in-domain downstream gain is not seed-robust (mean ΔAUROC $= +0.0350 \pm 0.1600$, A1 exceeding A0 in only 2 of 5 seeds), a finding we report as a scientifically important characterization of small-cohort EEG SSL rather than a caveat to be minimized. Controlled comparisons against an explicit temporal-continuity objective, a SoftCLT-compatible soft instance-wise formulation, and three temporal kernels further show that the size and direction of the in-domain effect depend on how temporal information enters the objective. A representation-level diagnostic indicates the canonical gain is not explained by increased effective rank. Finally, the compact encoder retains 0.9991 ± 0.0004 cosine similarity to its outputs after INT8 post-training quantization (71% size reduction), a practical consequence of its small footprint that supports use in resource-constrained physiological monitoring. Together, these results characterize patient-level temporal structure as a meaningful inductive bias whose in-domain downstream benefit is stochastic in small cohorts, while the underlying representation demonstrates measurable cross-cohort transfer.

1 INTRODUCTION

EEG provides a dense, temporally structured measurement of physiological activity, with large volumes of unlabeled recordings but comparatively scarce clinical labels — an attractive regime for self-supervised learning (SSL). Contrastive objectives such as NT-Xent (Chen et al., 2020) have been applied to EEG and other biosignals (Franceschi et al., 2019; Yang et al., 2021), treating augmented views of a window as positives and all other windows in a batch as negatives, regardless of origin. This batch-agnostic negative structure is convenient but ignores that windows from the same patient are not independent: two temporally close windows may share physiological state even

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.

though they are not augmentation views of one another, while a window from a different patient carries no such relationship.

We introduce **Temporal NT-Xent**, which keeps the standard augmentation-defined positive pair and full-strength cross-patient negatives, and modulates only the weight of same-patient negatives as a continuous function of temporal distance (Eq. 3). The precise novelty is the objective construction: (i) the positive pair is an unmodified augmentation view, exactly as in vanilla NT-Xent; (ii) cross-patient negatives retain weight 1, exactly as in vanilla NT-Xent; (iii) same-patient windows are explicitly identified via patient/time metadata and receive a distance-dependent weight inside the softmax denominator, rather than being reassigned as positives or given a globally soft target. This differs from SoftCLT (Lee et al., 2024), which replaces the hard contrastive target with a continuous instance-wise (and, in its full form, timestamp-level) target for *every* pair; we compare against SoftCLT's architecture-compatible soft instance-wise component only, since its full temporal component assumes a hierarchical timestamp-level representation incompatible with our window-level encoder (Sec. 3.4). No prior EEG contrastive formulation, to our knowledge, isolates same-patient negative reweighting as the sole locus of temporal information while leaving the positive assignment and cross-patient negatives untouched — a design specific to windowed, patient-identified recordings, where same-patient windows are neither independent (as batch-agnostic NT-Xent assumes) nor guaranteed to share label-relevant state (as would justify treating them as positives).

We evaluate this objective along three axes that answer distinct scientific questions. **(1) In-domain transfer to unseen patients.** On a canonical patient-held-out split — train, validation, and test patients are disjoint, so the result reflects transfer to genuinely unseen patients rather than a random window-level split — Temporal NT-Xent raises AUROC from 0.5417 to 0.7667 and AUPRC from 0.4013 to 0.6673. This is the principal controlled demonstration that the objective can produce a substantially different, more label-informative downstream representation; it is not, by itself, evidence of robustness, which we assess next. **(2) Stochastic robustness of that in-domain effect.** Because the held-out cohort contains only 23 patients, we repeat pretraining and downstream evaluation across five seeds on the same fixed split. The mean AUROC difference falls to $+0.0350 \pm 0.1600$, with A1 exceeding A0 in only 2 of 5 seeds: the large canonical effect is real for that split but is not seed-robust, an important qualification of its magnitude rather than a refutation of the underlying hypothesis. **(3) Cross-cohort transfer.** We then ask a qualitatively different question — does the resulting frozen representation retain useful signal outside the source cohort entirely? Zero-shot evaluation on the independent SHHS-1 cohort, with no apnea labels or target-domain fine-tuning, gives AUROC $= 0.724$ (95% CI $[0.641, 0.801]$), against a single-channel source-to-target control of 0.718. This cross-cohort result is complementary evidence about the representation itself, and is not subject to the same within-cohort seed sensitivity as (1)-(2), since it uses the single canonical A1 encoder evaluated on entirely external data. Controlled comparisons against an explicit continuity objective, a SoftCLT-compatible formulation, and alternative temporal kernels (Sec. 5.3–5.4) further characterize *how* temporal information should enter the objective, and a compact-encoder quantization analysis (Sec. 5.6) characterizes the practical deployability of the representation studied throughout.

Contributions.

- We formulate Temporal NT-Xent, a subject-gated, distance-dependent weighting of same-patient negative interactions, $w_{ij} = 1 - \exp(-\lambda d_{ij})$, that preserves augmentation-defined positives and full-strength cross-patient negatives, and give an explicit mathematical contrast with SoftCLT's soft-target construction (Sec. 3.4).
- On a canonical patient-held-out split, we show a substantial in-domain downstream change (ΔAUROC $= +0.2250$, ΔAUPRC $= +0.2660$; Sec. 5.1), and quantify its stochastic sensitivity across five seeds (mean ΔAUROC $= +0.0350 \pm 0.1600$; Sec. 5.2) — two results that answer different questions and must be read together, not as contradictory evidence.
- We show that the resulting frozen encoder transfers zero-shot to the independent SHHS-1 cohort without target-domain labels or fine-tuning (AUROC $= 0.724$, 95% CI $[0.641, 0.801]$; Sec. 5.5), providing cross-cohort evidence for the learned representation that is distinct from, and not undermined by, the in-domain seed sensitivity.
- We provide a controlled comparison against an explicit temporal-continuity objective (PhysioCLR-style, AUROC $= 0.6183 \pm 0.1148$), a compatible SoftCLT formulation (AUROC

$\approx 0.5567 \pm 0.1451$), and three temporal kernels (aggregate AUROC 0.4900–0.6133), showing these mechanisms for injecting temporal information are not interchangeable (Sec. 5.3–5.4).

- We report a representation-level diagnostic (effective rank $\mathrm{erank}(A0)\approx 7.30$ vs. $\mathrm{erank}(A1)\approx 7.25 \pm 0.80$) showing the canonical gain is not explained by increased effective dimensionality, and characterize the practical deployability of the compact encoder used throughout via INT8 post-training quantization (0.9991 ± 0.0004 cosine similarity, 71% size reduction; Sec. 5.6).

2 Related Work

Contrastive self-supervised learning. SimCLR showed that augmentation-based positives with an NT-Xent objective learn useful representations without labels (Chen et al., 2020); similar recipes have been adapted to multivariate time series and physiological signals (Franceschi et al., 2019; Yang et al., 2021). These formulations treat every non-positive batch sample as an equally valid negative, an assumption continuous physiological recordings violate to some degree.

Temporal contrastive learning. TNC defines temporal neighborhoods under a local-stationarity assumption (Tonekaboni et al., 2021); TS-TCC combines temporal and contextual contrasting with weak/strong augmentation pairs (Eldele et al., 2021); SoftCLT replaces binary positive/negative assignment with soft instance-wise and temporal targets from a hierarchical, timestamp-level architecture (Lee et al., 2024). These methods establish that temporal relationships need not be binary, but each ties the temporal signal to a specific target construction or architecture. Our formulation differs in *where* temporal information enters the objective: the augmentation-defined positive and cross-patient negatives are untouched, and only the weight of same-patient negatives is modulated continuously by temporal distance, independent of any target-softening scheme.

EEG representation learning and OSA. Self-supervised EEG representation learning has been studied for sleep staging and broader biosignal transfer (Yang et al., 2021; 2023). Most evaluations report a single canonical split; we additionally quantify stochastic sensitivity across seeds on a fixed small held-out cohort, a common but under-reported regime in clinical EEG SSL.

3 Temporal NT-Xent

3.1 Problem formulation

A recording from patient p is divided into windows $\mathcal{X}_p = \{x_{p,1}, \ldots, x_{p,T_p}\}$. Each window x_i carries a patient identifier s_i, a temporal coordinate t_i, and an augmented view x_i^+. An encoder f_θ maps a window to an embedding $z_i = f_\theta(x_i) \in \mathbb{R}^{128}$. For a pair of windows, define the temporal distance

$$d_{ij} = |t_i - t_j|. \tag{1}$$

Our hypothesis is that temporally adjacent windows may contain partially shared physiological state and therefore need not exert the same repulsive force as temporally distant windows — a hypothesis we test empirically rather than assume.

3.2 Vanilla NT-Xent

For anchor i with augmented positive $p(i)$ and similarity $\mathrm{sim}(u, v) = u^\top v/\|u\|\|v\|$, standard NT-Xent is

$$\mathcal{L}_{\mathrm{NT}} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \log \frac{\exp(\mathrm{sim}(z_i, z_{p(i)})/\tau)}{\exp(\mathrm{sim}(z_i, z_{p(i)})/\tau) + \sum_{j\in\mathcal{A}(i)} \exp(\mathrm{sim}(z_i, z_j)/\tau)}, \tag{2}$$

where $\mathcal{A}(i)$ is the set of negatives for anchor i (all other samples in the batch, excluding $p(i)$), and τ is temperature. Every negative receives equal weight.

3.3 TEMPORAL NT-XENT

We assign each negative $j \in \mathcal{A}(i)$ a weight w_{ij} that depends on same-patient membership and temporal distance:

$$w_{ij} = \begin{cases} 1 - \exp(-\lambda\, d_{ij}), & s_i = s_j, \\ 1, & s_i \neq s_j, \end{cases} \tag{3}$$

and define

$$\mathcal{L}_{\text{TempNTXent}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log \frac{\exp(\text{sim}(z_i, z_{p(i)})/\tau)}{\exp(\text{sim}(z_i, z_{p(i)})/\tau) + \sum_{j \in \mathcal{A}(i)} w_{ij} \exp(\text{sim}(z_i, z_j)/\tau)}. \tag{4}$$

We use $\tau = 0.5$ and $\lambda = 0.1$ throughout; the augmentation positive $p(i)$ is excluded from $\mathcal{A}(i)$. As $d_{ij} \to 0$, $w_{ij} \to 0$, so nearby same-patient windows exert little negative pressure; as d_{ij} grows, $w_{ij} \to 1$, recovering ordinary negative weighting. Temporal NT-Xent never re-labels a same-patient window as a positive — it only reweights an existing negative relationship inside the denominator of Eq. 4.

3.4 RELATIONSHIP TO SOFTCLT

SoftCLT (Lee et al., 2024) is the closest prior formulation because it also treats temporal relationships as non-binary. Its instance-wise component replaces the hard contrastive target with a continuous soft target for every pair,

$$\text{target}_{ij} \;\propto\; \exp\big(-\alpha \cdot \text{dist}(x_i, x_j)/\tau_I\big), \tag{5}$$

and its full temporal component operates on timestamp-level representations from a hierarchical TS2Vec-style architecture. Because our encoder (Sec. 4) produces one window-level embedding rather than a sequence of timestamp-level embeddings, this temporal component cannot be applied without changing the architecture; we therefore compare against the architecture-compatible soft instance-wise component only, using published defaults (τ_I=2, α=0.5, min–max normalized Euclidean distance).

Eq. 3–4 and Eq. 5 differ in construction, not just implementation: SoftCLT replaces the *target* for every pair with a continuous value; Temporal NT-Xent keeps the target hard (one positive, all else negative) and instead reweights the *loss contribution* of a specific, patient- and time-identified subset of negatives, inside the softmax denominator. Under the matched window-level protocol (Sec. 5.3), the compatible SoftCLT formulation does not reproduce the canonical A1 gain (AUROC $\approx 0.5567 \pm 0.1451$ vs. 0.5967 ± 0.1778 across five seeds), indicating the two mechanisms are not interchangeable in this setting.

4 EXPERIMENTAL SETUP

Dataset and split. We use the OpenNeuro ds008108 EEG sleep cohort (142 subjects: 56 OSA, 86 controls; AHI-based labels), with a fixed patient-level 60/20/20 split (`random_state=42`): 66 train, 22 validation, 23 held-out test subjects (test positive rate ≈ 0.348). SSL pretraining uses only training subjects; validation is used for checkpoint selection; test subjects are held out until final evaluation. No patient's windows appear in more than one partition.

Encoder. Each input is 20 EEG channels $\times$ 3,000 samples (30 s at 100 Hz). A log-magnitude STFT (n_{fft}=256, hop 32) yields a $20\times129\times94$ representation, processed by a compact STFTENCODER2D ($\approx$131K parameters, projection dim. 64) producing a 128-d embedding.

Training. All objectives share encoder, augmentation (spectral subband masking, p=0.5), optimizer (AdamW, lr 10^{-3}, weight decay 10^{-4}), batch size 64, and budget (5 SSL epochs). Downstream, the encoder is frozen and a 16,897-parameter MLP is trained for 8 epochs with AdamW and binary focal loss; the checkpoint with highest validation AUROC is selected and the test set evaluated once. Patient-level predictions average window-level probabilities, $\hat{y}_p = \frac{1}{T_p} \sum_t \hat{y}_{p,t}$; all reported metrics are patient-level.

Conditions. A0: vanilla NT-Xent. A1: Temporal NT-Xent. A2: NT-Xent + an explicit temporal-continuity penalty between adjacent same-patient windows. A3 (PhysioCLR-style): $\mathcal{L}_{\text{temp}} =$

Table 1: Canonical (seed 42) patient-level test performance.

Method	AUROC	AUPRC
A0 — Vanilla NT-Xent	0.5417	0.4013
A1 — Temporal NT-Xent	**0.7667**	**0.6673**
A2 — NT-Xent + TempReg	0.5083	0.4778
A3 — PhysioCLR-style	0.7250	0.5126

Table 2: Five-seed aggregate performance for A0 vs. A1 (per-seed values in Appendix C).

Method	AUROC (mean ± SD)	AUPRC (mean ± SD)
A0	0.5617 ± 0.1008	0.4306 ± 0.0850
A1	0.5967 ± 0.1778	0.5204 ± 0.1401

$\frac{1}{|\mathcal{V}|}\sum_{(i,j)\in\mathcal{V}} \|z_i - z_j\|_2^2$, over adjacent same-patient pairs $\mathcal{V} = \{(i,j) : s_i{=}s_j,\ t_j{=}t_i{+}1\}$, identified via explicit metadata (verified invariant to batch permutation). We also evaluate the SoftCLT-compatible formulation (Eq. 5) and three kernels for w_{ij} (Sec. 5.4). A0/A1 are evaluated on five seeds $\{42, 123, 2025, 7, 13\}$ on the same fixed split, measuring stochastic variation rather than independent-cohort variation. Paired permutation tests use 10,000 within-patient label exchanges; full per-seed significance tables are in Appendix B.

5 Results

5.1 Canonical held-out evaluation

Temporal NT-Xent improves over vanilla NT-Xent by $\Delta\text{AUROC} = +0.2250$ and $\Delta\text{AUPRC} = +0.2660$ on the canonical split (paired permutation $p{=}0.0455$ for AUROC, $p{=}0.0704$ for AUPRC; 95% paired bootstrap CI for the AUROC difference: $[0.0500, 0.4333]$). This is a strong motivating result, but the held-out cohort contains only 23 patients, so we treat it as a canonical-split result rather than evidence of universal robustness, assessed next.

5.2 Multi-seed robustness

The mean paired AUROC difference is $+0.0350 \pm 0.1600$ (median -0.0167), with A1 exceeding A0 in only 2 of 5 seeds; for AUPRC the mean paired difference is $+0.0898 \pm 0.1568$, A1 higher in 3 of 5. Per-seed AUROC permutation p-values are 0.0455, 0.5751, 0.4964, 0.3552, 0.9217 (seeds 42, 123, 2025, 7, 13), and the sign of the effect changes across seeds (Appendix B). We read this as evidence of substantial stochastic sensitivity in the present small-cohort protocol: the canonical gain is real for that split but should not be extrapolated into a claim of consistent superiority.

5.3 Temporal mechanism comparison

Table 1 shows explicit continuity regularization (A2) does not reproduce A1's canonical gain (AUROC 0.5083), while the corrected PhysioCLR-style objective (A3) reaches a substantial single-split gain (0.7250). Across five seeds, A3 is itself highly variable: $\text{AUROC} = 0.6183 \pm 0.1148$ (range 0.5000–0.8083; full per-seed values in Appendix C), exceeding A1 in 2 of 5 seeds with no paired comparison reaching $p{<}0.05$. Alternative temporal mechanisms thus occupy a similarly variable regime rather than being uniformly weaker.

The SoftCLT-compatible formulation (Eq. 5) obtains, at seed 42, $\text{AUROC} = 0.5583$, $\text{AUPRC} = 0.5549$ ($\Delta\text{AUROC} = -0.2083$ vs. A1, $p{=}0.295$, not significant); across five seeds, $\text{AUROC} \approx 0.5567 \pm 0.1451$, beating A0 in 3/5 seeds and A1 in 2/5. Under the architecture-compatible comparison and identical protocol, SoftCLT does not reproduce the canonical A1 gain — an informative result about the transferability of instance-wise soft targets to this setting, not evidence that SoftCLT is broadly inferior.

Table 3: Five-seed comparison across temporal mechanisms.

Method	AUROC (mean ± SD)	AUPRC (mean ± SD)	Beats A1 (seeds)
A1 — Temporal NT-Xent	0.5967 ± 0.1778	0.5204 ± 0.1401	—
A3 — PhysioCLR-style	0.6183 ± 0.1148	0.5474 ± 0.1039	2/5
SoftCLT-compatible	0.5567 ± 0.1451	—	2/5

Table 4: Temporal kernel ablation, five-seed aggregate.

Kernel	AUROC (mean ± SD)	AUPRC (mean ± SD)
Exponential	0.6133 ± 0.1071	0.5620 ± 0.0668
Linear	0.4900 ± 0.1596	0.4794 ± 0.1580
Hard cutoff	0.5467 ± 0.1473	0.4542 ± 0.1014

5.4 Temporal kernel ablation

We compare three kernels for w_{ij} in Eq. 3: exponential $1 - \exp(-\lambda d)$, linear $\min(1, \alpha d)$, and hard cutoff $\mathbf{1}[d > c]$, each over five seeds (per-seed values in Appendix C).

The exponential kernel attains the highest mean AUROC, but no kernel dominates uniformly (e.g. at seed 123 the linear kernel reaches 0.7667 vs. 0.6583 for exponential; Appendix C). The exact mapping from temporal distance to negative weight materially affects downstream behavior, supporting temporal structure as a specific, parameterization-sensitive inductive bias rather than a generic scalar correction.

5.5 External cross-cohort transfer: SHHS-1

The canonical split and multi-seed analysis (Sec. 5.1–5.2) both ask an *in-domain* question: given the source cohort, how does the choice of contrastive objective affect transfer to patients held out from that same cohort, and how stable is that effect across stochastic reruns? SHHS-1 (Sleep Heart Health Study) lets us ask a different, complementary question: does the resulting *frozen* representation carry signal that generalizes under a genuine cross-cohort distribution shift — different recording sites, equipment, and population — without any adaptation to the new domain?

We take the single canonical A1 encoder (seed 42, Sec. 5.1), freeze it, and evaluate it zero-shot on SHHS-1: no apnea labels from SHHS are used at any point, and no parameters are fine-tuned on SHHS data. This gives AUROC $= 0.724$ (95% CI $[0.641, 0.801]$), compared against a source-to-target single-channel control of 0.718 that isolates the effect of using a single, more portable EEG channel under the same transfer protocol (Table 5). The two numbers are close, so this result does not establish that Temporal NT-Xent's representation is superior to the control channel configuration; rather, both confirm that the encoder's learned features retain discriminative signal for OSA-related structure outside the cohort on which they were trained. Because this evaluation uses a single fixed encoder rather than repeated stochastic reruns, it is not subject to the same source of variance as the multi-seed analysis in Sec. 5.2 — it is a different kind of evidence (transfer breadth) rather than a more stable estimate of the same in-domain effect, and the two should not be read as contradicting one another.

5.6 Deployment: quantization of the compact encoder

Both the canonical in-domain evaluation and the SHHS transfer result above rely on the same ≈131K-parameter encoder, whose small footprint makes post-training compression practically relevant for wearable or resource-constrained physiological monitoring. We apply post-training W8A8 INT8 quantization (source-training calibration, no target-domain data), which reduces model size from 0.54 MB to 0.16 MB (-71%) with AUROC/AUPRC on the canonical split changing by only $-0.0056/-0.0084$ (Table 6). Encoder output cosine similarity between FP32 and INT8 is 0.9991 ± 0.0004 (logit-level Pearson/Spearman: 0.9996/0.9991), indicating the encoder's numerical function is highly preserved under compression. We report this as a practical consequence of the

Table 5: Zero-shot cross-cohort transfer to SHHS-1 using the frozen canonical A1 encoder (no target-domain labels or fine-tuning).

Configuration	AUROC	95% CI
A1 encoder $\rightarrow$ SHHS-1 (zero-shot)	0.724	$[0.641, 0.801]$
Single-channel source-to-target control	0.718	—

Table 6: INT8 post-training quantization of the A1 encoder.

Metric	FP32	INT8	Difference
Model size	0.54 MB	0.16 MB	-71%
AUROC	0.7667	0.7611	-0.0056
AUPRC	0.6673	0.6589	-0.0084

compact architecture used throughout the paper, not as evidence for the temporal mechanism itself, and we do not claim measured on-device latency or clinical deployment readiness.

6 Representation Analysis

An earlier internal pipeline inadvertently fixed the SSL random seed during checkpoint generation, duplicating encoder checkpoints across nominal seeds; we discard those checkpoints for representation-level claims and use only correctly regenerated, independent checkpoints below. Using the corrected checkpoints, effective rank is $\mathrm{erank}(A0) \approx 7.30$ and $\mathrm{erank}(A1) \approx 7.25 \pm 0.80$: A1's effective rank is not higher than A0's, and its spectrum is, if anything, marginally sharper. Temporal NT-Xent therefore does not mitigate dimensional collapse relative to vanilla NT-Xent and can attain a lower effective rank. This rules out a tempting explanation — that Temporal NT-Xent works by expanding the representation into more usable dimensions — without offering a replacement causal account. We treat representation geometry strictly as a diagnostic: the canonical downstream improvement is not explained by increased effective dimensionality or improved spectral conditioning, and effective rank does not capture whatever fine-grained same-patient structure the temporal weighting reshapes.

7 Discussion

What does temporal structure change? Temporal NT-Xent does not collapse temporally related windows into positives; it changes their contribution to the contrastive denominator (Eq. 4) — a middle ground between treating every non-positive window as equally unrelated and declaring temporal neighbors to be positives. The kernel ablation (Sec. 5.4) shows this middle ground is itself sensitive to the chosen mapping.

Why is the downstream effect seed-sensitive? The held-out set contains only 23 patients, so a few patient-level ranking changes can move AUROC noticeably. Additional stochasticity sources include SSL initialization, downstream classifier initialization, validation-based checkpoint selection, and patient-level probability aggregation. Representation-level statistics and patient-level AUROC measure different properties: a global embedding change does not guarantee a corresponding ranking change for a specific classifier on a small cohort. The correct reading is not that temporal structure is ineffective, but that its downstream benefit is not reliably estimated from a single stochastic run here.

What do SoftCLT and the kernel comparison add? If temporal information alone were sufficient, we would expect broadly similar behavior across formulations that use it. Instead, A1, A3, the SoftCLT-compatible objective, and the three kernels all produce distinct performance profiles — evidence that temporal structure is a specific inductive bias whose effect depends on formulation, which is precisely the empirical characterization this paper provides.

How should the SHHS result be read alongside the seed analysis? Sec. 5.2 and Sec. 5.5 are sometimes conflated, but they answer different questions. The five-seed analysis asks whether a *specific training run's* in-domain advantage over vanilla NT-Xent reproduces under re-initialization on the source cohort; it does not. The SHHS evaluation asks whether *one fixed, already-trained* representation carries transferable signal on entirely external data; it does. A representation can be simultaneously sensitive to training stochasticity in its downstream classification margin over a baseline, and still encode genuinely transferable structure once trained — the SHHS result is evidence for the latter and says nothing about the former. We therefore do not present SHHS as proof that Temporal NT-Xent is statistically superior to competing objectives, nor as resolving the seed sensitivity documented above; it is complementary evidence about what the learned representation retains.

8 LIMITATIONS

Small cohort and fixed split. The processed cohort has 142 subjects, only 23 held-out test patients, limiting statistical power; five seeds measure stochastic variation on *one* held-out cohort, not variation across independent cohorts. True patient-level cross-validation would require repeating SSL pretraining inside every fold to avoid subject leakage (downstream-only CV would let the encoder see validation/test subjects during pretraining) and was computationally prohibitive at this scale; five-seed evaluation is a stochastic robustness analysis, not a substitute for independent cohorts. **Seed sensitivity.** The canonical improvement is not replicated across all five seeds (mean $\Delta\text{AUROC} = +0.0350 \pm 0.1600$); we do not claim robust, universal downstream superiority. **Preprocessing lineage.** Raw EDF artifacts available in our environment were incomplete simulation stubs, so a corrected preprocessing pipeline could not be rerun from raw recordings; results use the existing processed cohort, and residual preprocessing effects cannot be independently quantified. **Representation interpretation.** Representation geometry is descriptive; we do not claim effective rank or spectral structure causally determines downstream performance. **Clinical scope.** This is a representation-learning benchmark, not a clinical validation study.

9 BROADER IMPACT

This work is a research artifact, not a clinical diagnostic system. It may reduce reliance on manually labeled physiological data and clarifies how to evaluate temporal inductive biases in small cohorts. The principal risk is overinterpreting single-split results as clinical evidence; we avoid claims of diagnostic equivalence or deployment readiness.

10 CONCLUSION

We introduced Temporal NT-Xent, which incorporates patient-level temporal relationships into the negative term of a contrastive objective (Eq. 4) while retaining augmentation-defined positives and full-strength cross-patient negatives. On a canonical patient-held-out OSA evaluation, this produces a substantial in-domain change relative to vanilla NT-Xent (AUROC $0.5417 \rightarrow 0.7667$, AUPRC $0.4013 \rightarrow 0.6673$), and the same frozen, compact encoder transfers zero-shot to the independent SHHS-1 cohort without target-domain labels or fine-tuning (AUROC $= 0.724$, 95% CI $[0.641, 0.801]$) — complementary evidence that the learned representation carries signal beyond the source cohort. A five-seed analysis on the source cohort shows that the magnitude of the in-domain downstream gain is stochastically sensitive (mean $\Delta\text{AUROC} = +0.0350 \pm 0.1600$), and controlled comparisons with explicit continuity regularization, a SoftCLT-compatible objective, and alternative temporal kernels show the in-domain effect depends on the precise formulation through which temporal information enters the objective. The compact encoder additionally retains high numerical fidelity under INT8 quantization (71% size reduction, 0.9991 ± 0.0004 cosine similarity), a practical consequence of the architecture used throughout. Read together, these results indicate that temporal structure can materially alter EEG contrastive representation learning and yields a representation with measurable cross-cohort transfer, while the magnitude of its in-domain downstream advantage over vanilla NT-Xent is stochastic in this small-cohort regime — a conclusion that motivates evaluating temporal physiological SSL objectives through controlled objective comparisons, cross-cohort transfer, and multi-seed, patient-level analysis rather than a single downstream score.

AI Assistance Disclosure

AI-assisted tools, including large language models, were used during the preparation of this manuscript for drafting and revising text, language editing, structural refinement, code-related assistance, and organization of experimental analyses. The authors reviewed and verified the generated material and retain responsibility for the experimental methodology, implementation, results, interpretations, and final claims.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

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 *IJCAI*, 2021.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*, 2019.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *ICLR*, 2024.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *ICLR*, 2021.

Chaoqi Yang, Danica Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *NeurIPS*, 2023.

A Reproducibility and additional experimental detail

All models are trained with the protocol in Sec. 4. Unit tests verify the A3 temporal-pair construction is invariant to batch permutation and correctly excludes cross-patient or non-adjacent pairs. SoftCLT hyperparameters (τ_I=2, α=0.5) follow the original paper's defaults, applied with min–max normalized Euclidean distance and no external data, SHHS labels, or test-subject information.

B Per-seed significance tests

Table 7: Per-seed A0 vs. A1 AUROC/AUPRC and paired permutation p-values (10,000 within-patient label exchanges).

Seed	A0 AUROC	A1 AUROC	ΔAUROC	p(AUROC)	A0 AUPRC	A1 AUPRC
42	0.5417	0.7667	+0.2250	0.0455	0.4013	0.6673
123	0.7167	0.5917	−0.1250	0.5751	0.5794	0.5022
2025	0.4417	0.3500	−0.0917	0.4964	0.3662	0.3850
7	0.5833	0.7667	+0.1833	0.3552	0.4141	0.6621
13	0.5250	0.5083	−0.0167	0.9217	0.3921	0.3855

C PER-SEED RESULTS: A3 AND KERNEL ABLATION

Table 8 gives the per-seed values underlying the A3 (PhysioCLR-style) aggregate reported in Sec. 5.3, and Table 9 gives the per-seed values underlying the kernel-ablation aggregate reported in Sec. 5.4. Both are provided for exact reproducibility of the mean/SD figures quoted in the main text.

Table 8: A3 (PhysioCLR-style) per-seed AUROC/AUPRC.

Seed	A3 AUROC	A3 AUPRC
42	0.5917	0.5608
123	0.5000	0.4015
2025	0.5750	0.5159
7	0.6167	0.5691
13	0.8083	0.6898

Table 9: Per-seed AUROC for the temporal kernel ablation.

Seed	Exponential	Linear	Hard cutoff
42	0.7667	0.5250	0.3750
123	0.6583	0.7667	0.6750
2025	0.5333	0.3167	0.3750
7	0.6500	0.4917	0.7250
13	0.4583	0.3500	0.5833

D REPRESENTATION ANALYSIS

This appendix gives the methodology and full results underlying the representation-level diagnostic summarized in Sec. 5.1–5.2 and the Representation Analysis section of the main text.

Representation extraction protocol. Embeddings are extracted from the frozen STFTENCODER2D (Sec. 4) on the held-out test split of the canonical patient-level partition, producing one 128-dimensional embedding per window; patient-level embedding sets are pooled window-wise prior to spectral analysis. As noted in the main text, an earlier internal pipeline inadvertently fixed the SSL random seed during checkpoint generation, so encoder checkpoints nominally associated with different seeds were in fact duplicates of a single training run. Those checkpoints are invalid for any representation-level (as opposed to downstream-prediction-level) claim and are excluded here; all effective-rank statistics below use only correctly, independently regenerated checkpoints. This is a stricter validity requirement than the downstream classification results in Sec. 5.2, which do not depend on checkpoint identity across seeds in the same way and were audited separately.

Effective-rank methodology. For a set of N embeddings stacked as $Z \in \mathbb{R}^{N\times 128}$, let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{128} \geq 0$ be the singular values of Z (equivalently, the square roots of the eigenvalues of the empirical covariance $Z^\top Z$). The singular-value spectrum is normalized to a probability distribution,

$$p_k = \frac{\sigma_k}{\sum_{l=1}^{128} \sigma_l}, \qquad k = 1, \ldots, 128, \tag{6}$$

and the effective rank is defined via the Shannon entropy of this distribution (Roy & Vetterli, 2007):

$$\mathrm{erank}(Z) = \exp\left(-\sum_{k=1}^{128} p_k \log p_k \right). \tag{7}$$

$\mathrm{erank}(Z)$ ranges from 1 (all variance concentrated in a single direction, maximal collapse) to 128 (variance spread uniformly across all directions). This entropy-based definition is the one used throughout the paper; a participation-ratio estimate was not computed in the underlying analysis and is not reported.

Results. Table 10 reports the effective rank of the A0 and A1 embedding spaces on the held-out test split, using the corrected checkpoints described above. Consistent with Sec. 5.1–5.2 of the main text, $\mathrm{erank}(A1)$ is not higher than $\mathrm{erank}(A0)$.

Table 10: Effective rank (Eq. 7) of held-out test embeddings, corrected checkpoints only.

Method	erank
A0 — Vanilla NT-Xent	≈ 7.30
A1 — Temporal NT-Xent	7.25 ± 0.80

Interpretation. This analysis is supplementary and diagnostic: it characterizes one geometric property of the embedding space and does not, by itself, establish or explain downstream classification performance. The valid effective-rank results do not support a claim that Temporal NT-Xent improves representation geometry, increases effective dimensionality, or mitigates dimensional collapse relative to vanilla NT-Xent; if anything, $A1$'s spectrum is marginally sharper. We do not resurrect or report any effective-rank figures from the invalidated hardcoded-seed checkpoints described above. As stated in the main text, the mechanism behind the canonical downstream gain should not be attributed to a simple expansion of effective dimensionality.

E Edge Deployment and Quantization Details

This appendix expands on the deployment characterization in Sec. 5.6.

Model artifacts. The quantized artifact is the same ≈131K-parameter STFTEncoder2D used throughout the paper (Sec. 4), evaluated as an FP32 checkpoint (0.54 MB) and a quantized INT8 checkpoint (0.16 MB). We do not report an export-format-specific (e.g. ONNX runtime) benchmark, as this was not part of the validated experimental artifacts.

Quantization procedure. Quantization is post-training W8A8 (8-bit weights, 8-bit activations), with calibration performed on the source-training split only; no target-domain (SHHS) data is used at any point. This is a static, calibration-based quantization procedure, as distinct from dynamic (calibration-free, activation-range-at-runtime) quantization; we report only the static W8A8 configuration, which is what the underlying implementation performs.

Accuracy and representation preservation. Table 11 reproduces the canonical-split accuracy comparison from Table 6 alongside the representation-fidelity statistics reported in the main text: encoder-output cosine similarity between FP32 and INT8 of 0.9991 ± 0.0004, and logit-level Pearson/Spearman correlation of 0.9996/0.9991 between FP32 and INT8 outputs. Prediction flip-rate at the classification threshold was not separately computed in the underlying analysis and is not reported here.

Table 11: Full accuracy and representation-fidelity comparison for INT8 post-training quantization of the A1 encoder (canonical split).

Metric	Value
Model size (FP32 → INT8)	0.54 MB → 0.16 MB (-71%)
AUROC (FP32 → INT8)	$0.7667 \rightarrow 0.7611$ (-0.0056)
AUPRC (FP32 → INT8)	$0.6673 \rightarrow 0.6589$ (-0.0084)
Encoder output cosine similarity (FP32 vs. INT8)	0.9991 ± 0.0004
Logit-level Pearson correlation	0.9996
Logit-level Spearman correlation	0.9991

Runtime. The validated artifacts available for this study measure numerical and predictive fidelity under quantization, not on-device latency, throughput, or memory-bandwidth behavior; no hardware-specific runtime benchmark was collected, and none is reported here. This is consistent with the scope stated in Sec. 5.6: the quantization result characterizes representation and accuracy preservation under compression, not measured on-device performance.

Scientific interpretation. These results support edge-deployment *feasibility* in the specific sense that the compact encoder's numerical function and canonical-split accuracy are highly preserved after an off-the-shelf static INT8 quantization pass. They do not constitute a real-device runtime evaluation and should not be read as evidence of clinical deployment readiness; both remain open questions outside the scope of the present artifacts.