CONTRASTIVE TIME-INVARIANT LEARNING FOR TRANSFERABLE ECG REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning transferable electrocardiogram (ECG) representations requires models to preserve diagnostic information and remain stable under changes that do not affect the diagnosis. Contrastive learning can encourage this stability through positive-pair construction. However, we observe that small temporal shifts still alter model predictions and internal representations. To address this issue, we propose Contrastive Time-Invariant Learning (CTI), which forms positive pairs from two independently circularly shifted views of the same ECG. CTI jointly optimizes supervised classification and cross-view contrastive objectives to explicitly encourage representation consistency across temporal positions, without changing the encoder architecture or test-time inference. In alignment analyses under fixed shifts, CTI improves deep-layer representation consistency at corresponding waveform locations. In frozen-encoder cross-dataset evaluation on three external ECG datasets, CTI improves classification accuracy by 4.9–14.2 percentage points over a two-view cross-entropy baseline using the same shifted views. These results show that explicitly encoding temporal shifts as a contrastive positive-pair relation improves the transferability of ECG representations beyond temporal-shift augmentation alone.

1 INTRODUCTION

Learning generalizable and robust representations from raw electrocardiogram (ECG) waveforms is fundamental to automated ECG analysis. High-quality ECG representations should preserve diagnostically relevant morphology, rhythm, and temporal information. They should also provide reusable features for downstream classification and cross-dataset transfer. This requires the model to extract diagnostic information while reducing its sensitivity to changes that do not alter the diagnosis. Contrastive learning offers a direct way to learn such invariance by bringing the representations of positive views closer together. In this process, the choice of positive views determines which variations the training objective encourages the model to remain consistent across.

Existing ECG contrastive learning methods have extensively explored view construction using temporal crops, cross-lead relationships, and physiologically motivated signal transformations. Such view designs have achieved strong performance on downstream tasks. However, existing methods have largely overlooked a key property of ECG signals: different forms of temporal information carry markedly different diagnostic value. Waveform morphology, rhythm, and relative timing directly encode essential diagnostic information. In contrast, a signal's window-relative temporal position is typically a non-diagnostic source of variation. ECG representation learning therefore requires more than stronger temporal invariance. A useful representation should preserve diagnostically relevant temporal structure while avoiding unnecessary dependence on non-diagnostic window-relative temporal position.

However, stability to window-relative temporal position does not arise automatically from classification training. We observe that ECG models trained with standard classification objectives remain sensitive to small temporal shifts. Even a slight shift of the same ECG segment can change model predictions and internal representations. This observation shows that diagnostic labels alone do not guarantee the desired temporal stability. A classification loss mainly constrains the final prediction and does not explicitly require the same diagnostic content at different temporal positions to have

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.

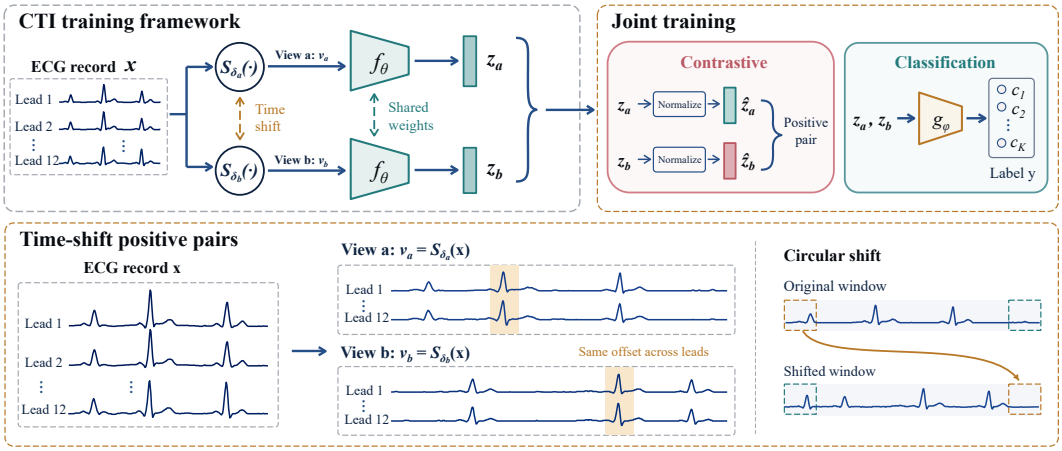


Figure 1: **Overview of Contrastive Time-Invariant Learning (CTI).** To reduce the dependence of ECG representations on window-relative temporal position, CTI forms positive pairs from two independently circularly shifted views of the same recording. A shared encoder is jointly trained with supervised classification and cross-view contrastive objectives. The lower panel illustrates synchronized shifts across leads and circular wrap-around.

consistent internal representations. A more direct constraint on this equivalence is therefore needed in representation space.

To address this issue, we propose Contrastive Time-Invariant Learning (CTI). CTI applies two independent small circular time shifts to the same ECG and treats the resulting views as a positive pair. Circular shifting changes the position of the signal within the input window without discarding original samples or introducing additional padding values. This provides a controlled way to construct views with different temporal positions. CTI jointly optimizes supervised classification and cross-view contrastive objectives over these two views. The classification objective provides supervision for diagnostic discrimination, while the contrastive objective directly encourages representation consistency across temporal positions. CTI changes only the training objective and requires no modification to the encoder architecture or test-time inference. Our experiments show that CTI achieves better deep-layer representation consistency after controlling for feature scale and aligning corresponding waveform locations. Cross-dataset transfer experiments also show gains from explicit cross-view contrastive constraints beyond temporal-shift augmentation alone.

Our main contributions are as follows:

1. **Characterizing temporal-shift sensitivity in ECG models.** We show that small temporal shifts can alter both the predictions and internal representations of ECG models. Through scale and alignment analyses, we further distinguish temporal displacement, changes in feature scale, and representation discrepancy.

2. **Contrastive Time-Invariant Learning.** We introduce CTI, which constructs positive pairs from independently circularly shifted views of the same ECG. CTI explicitly encourages invariance to the target temporal shift in representation space while leaving the encoder architecture and test-time inference unchanged.

3. **Cross-dataset validation and controlled analysis.** We evaluate frozen-encoder transfer on three external ECG datasets and further assess temporal-shift stability on HAR, Sleep-EDF, and Epilepsy. Controlled experiments on the training objective, boundary effects, and preprocessing are used to analyze the relationships among representation invariance, prediction stability, and transfer performance, as well as the conditions under which these effects hold.

2 Related Work

ECG Contrastive Representation Learning. Contrastive learning provides a framework for learning ECG representations by aligning related views of the same underlying physiological signal. Existing methods exploit temporal, multi-lead, and physiological structure to construct these views. CLOCS constructs positive pairs from temporal segments, different leads, and recordings associated with the same patient (Kiyasseh et al., 2021). 3KG introduces physiologically motivated spatial transformations in vectorcardiogram space and temporal masking, and evaluates how these view constructions affect downstream performance (Gopal et al., 2021). More recently, ACL-ECG incorporates anatomical priors into multi-lead contrastive learning and combines random resized cropping, cardiac-cycle masking, and zero-padded temporal shifting to construct training views (Liu et al., 2026). These studies demonstrate the value of ECG-specific view construction. However, improvements in downstream performance alone do not establish which input variations the learned representations become invariant to.

View Design, Positive Pairs, and Learned Invariance. The choice of contrastive views influences which information a representation retains. Tian et al. argue that effective views should reduce shared task-irrelevant information while preserving information needed for downstream tasks (Tian et al., 2020). Purushwalkam and Gupta show that the invariances learned through contrastive training depend on the augmentations and training data: aggressive cropping can promote occlusion invariance without ensuring viewpoint or category-instance invariance (Purushwalkam & Gupta, 2020). Xiao et al. further show that enforcing inappropriate invariance can discard information required by downstream tasks (Xiao et al., 2021). Together, these findings motivate evaluating the specific invariance induced by a view construction rather than assuming it follows from the training objective.

Related work in time-series representation learning also highlights the importance of training views and positive-pair quality. TS-TCC combines weakly and strongly augmented views with temporal and contextual contrastive objectives (Eldele et al., 2021). TS2Vec learns representations at multiple temporal scales through contextual views and hierarchical contrastive objectives (Yue et al., 2022). DBPM identifies noisy and faulty positive pairs that can impair representation learning and dynamically down-weights them based on their training histories (Lan et al., 2024). These methods motivate our focus on a specific positive-pair relationship: two views of the same ECG that differ only in window-relative temporal position.

Temporal-Shift Robustness and Invariance. Shift stability is not automatically guaranteed by standard neural network training. Studies in computer vision show that small translations or scale changes can substantially alter network predictions even when the semantic content of the input is largely preserved (Azulay & Weiss, 2019). Zhang further studies shift invariance from the perspective of sampling and anti-aliasing and improves model stability through architectural design (Zhang, 2019). These findings show that semantic consistency under an input transformation does not necessarily imply model invariance to that transformation.

Temporal transformations are also widely used in time-series representation learning. Temporal cropping, different temporal contexts, and other time-domain augmentations are commonly used to construct training views (Yue et al., 2022, Eldele et al., 2021). In ECG representation learning, ACL-ECG combines temporal shifting with random cropping and cardiac-cycle masking to construct contrastive views (Liu et al., 2026).

Our work further studies whether ECG models acquire representation invariance to non-diagnostic changes in window-relative temporal position, and how such invariance relates to prediction stability and cross-dataset transfer.

3 Method

Contrastive Time-Invariant Learning (CTI) aims to reduce the dependence of ECG representations on window-relative temporal position. It combines supervised classification with a contrastive objective over two shifted views of the same recording (Figure 1). Supervised classification provides constraints for diagnostic discrimination, while cross-view contrastive learning encourages representation consistency across temporal positions.

3.1 Problem Setup and Target Invariance

Let $x \in \mathbb{R}^{C\times T}$ denote an ECG segment and $y \in \{1, \ldots, K\}$ its diagnostic label. An encoder f_θ maps the input to a recording-level representation $z = f_\theta(x) \in \mathbb{R}^d$, and a linear classification head g_ϕ maps z to class logits. CTI applies supervised classification to z and a contrastive constraint directly to its normalized form $\hat{z}(x) = z/\max(\|z\|_2, \epsilon)$, where $\epsilon > 0$ ensures numerical stability.

Window-relative temporal position refers to the signal's sample-index position within a fixed input window. Let S_δ denote a temporal shift by δ samples. For shifts within the target range that preserve diagnostic information, CTI encourages $\hat{z}(x) \approx \hat{z}(S_\delta x)$ through shifted positive pairs, while supervised classification provides pressure for diagnostic discrimination.

3.2 Circular Time-Shift Positive-Pair Construction

To construct positive pairs with different window-relative temporal positions, we independently apply two circular shifts to the same ECG, making temporal position a directly controlled factor in view construction. We define a circular shift along the temporal axis as

$$[S_\delta(x)]_{c,t} = x_{c,(t-\delta) \bmod T}, \qquad c = 1, \ldots, C, \quad t \in \{0, \ldots, T-1\}. \tag{1}$$

C is the number of leads and T is the number of samples per lead. The indices c and t denote the lead and temporal position, respectively, and $\delta \in \mathbb{Z}$ is the shift offset in samples. The modulo operation wraps samples that cross one window boundary to the other end. The same offset is applied to every lead, preserving the relative timing across leads. Circular shifting preserves the original sample values and signal length while changing sample adjacency at the window boundary.

Let q_ρ denote the integer-offset sampling distribution, where ρ controls the shift range relative to the window length. Its nominal range is approximately $\pm\rho T$ samples, with $\rho = 0.05$ in our experiments. Its exact sampling and rounding rules are given in Appendix A.2. For the i-th recording and its diagnostic label (x_i, y_i), we independently sample two offsets $\delta_i^{(a)}, \delta_i^{(b)} \sim q_\rho$ and construct two views

$$v_i^{(a)} = S_{\delta_i^{(a)}}(x_i), \qquad v_i^{(b)} = S_{\delta_i^{(b)}}(x_i). \tag{2}$$

The superscripts a and b distinguish the two views. The two views form a positive pair and share the diagnostic label y_i. The contrastive objective aligns two views sampled from the shift distribution, with neither view fixed to the unshifted recording. The main CTI view generator introduces no additional cropping, resampling, amplitude perturbation, or spectral perturbation, keeping the targeted transformation separate from other augmentation factors.

3.3 Contrastive Time-Invariant Objective

To encourage consistent representations across temporal positions while learning diagnostic discrimination, we jointly optimize cross-view contrastive and supervised classification objectives. The contrastive term constrains the two shifted views of the same ECG, while the classification term provides supervision through their shared diagnostic label.

For a mini-batch of N recordings, let $z_i^{(r)} = f_\theta(v_i^{(r)})$ and let $\hat{z}_i^{(r)}$ denote its normalized representation. Here, $r \in \{a, b\}$ indexes the two shifted views. CTI uses a symmetric cross-view InfoNCE objective (van den Oord et al., 2018). For an anchor in view a, the positive is the same recording in view b, and the comparison set contains the N representations from view b:

$$\ell_i^{a\to b} = -\log \frac{\exp\Big((\hat{z}_i^{(a)})^\top \hat{z}_i^{(b)}/\tau\Big)}{\sum_{j=1}^{N} \exp\Big((\hat{z}_i^{(a)})^\top \hat{z}_j^{(b)}/\tau\Big)}. \tag{3}$$

$\tau > 0$ is the temperature parameter, and j indexes the recordings in the mini-batch. The term with $j = i$ corresponds to the positive, while the remaining $N - 1$ recordings serve as negatives. The reverse term $\ell_i^{b\to a}$ exchanges the roles of the two views. Same-view representations are excluded from the denominator. Different recordings remain negatives even when they share the same diagnostic label. Supervised classification is applied to both shifted views. The contrastive loss, classification

loss, and complete objective are

$$\mathcal{L}_{\text{con}} = \frac{1}{2N}\sum_{i=1}^{N}\left(\ell_i^{a\to b} + \ell_i^{b\to a}\right),$$
$$\mathcal{L}_{\text{CE}}^{\text{2view}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{r\in\{a,b\}}\text{CE}\Big(g_\phi(z_i^{(r)}), y_i\Big),$$
$$\mathcal{L}_{\text{CTI}} = \mathcal{L}_{\text{CE}}^{\text{2view}} + \lambda\mathcal{L}_{\text{con}}. \tag{4}$$

CE denotes the cross-entropy loss, and $\lambda \geq 0$ controls the weight of the contrastive loss. At inference, a single unshifted ECG is passed through the encoder and classification head, with no additional view generation or contrastive branch.

3.4 Measuring Temporal-Shift Sensitivity

To assess temporal-shift sensitivity, we separately measure changes in predictions and internal representations before and after shifting the input. For representation differences, we further apply scale normalization and positional alignment to distinguish the effects of feature scale and waveform displacement. All measurements are performed in evaluation mode.

Prediction stability. To quantify prediction changes under temporal shifts, we use the original prediction $\hat{y}(x) = \arg\max_u [g_\phi(f_\theta(x))]_u$ as the reference, where $u \in \{1, \ldots, K\}$ indexes the classes. For each recording, we independently sample k offsets $\delta_1, \ldots, \delta_k$ from q_ρ and define the prediction flip rate as

$$\text{FlipRate}(x) = \frac{1}{k}\sum_{j=1}^{k}\mathbf{1}\big\{\hat{y}(S_{\delta_j}x) \neq \hat{y}(x)\big\}. \tag{5}$$

We average this rate over test recordings, with lower values indicating greater prediction stability.

Representation sensitivity and scale. To analyze the effects of temporal shifts across encoder stages, we compare token-wise features of the original and shifted inputs. To account for feature scale across encoder depth, we compute raw and scale-normalized Euclidean distances.

Let $H_{\ell,m}(x)$ denote the feature of token m at encoder stage ℓ for input x. There are M patch tokens, with $m = 0$ denoting the classification token (CLS) and $m = 1, \ldots, M$ denoting the patch tokens. The mean token-wise Euclidean distance and its relative form are

$$D_\ell(x,\delta) = \frac{1}{M+1}\sum_{m=0}^{M}\left\|H_{\ell,m}(x) - H_{\ell,m}(S_\delta x)\right\|_2,$$
$$D_\ell^{\text{rel}}(x,\delta) = \frac{D_\ell(x,\delta)}{\max\{A_\ell(x), \epsilon\}}. \tag{6}$$

The relative form normalizes the distance by the mean token norm of the original input at the same stage, $A_\ell(x) = (M+1)^{-1}\sum_{m=0}^{M}\|H_{\ell,m}(x)\|_2$. Here, $\epsilon > 0$ ensures numerical stability. Distances are computed per recording and then averaged over the evaluation records. Raw and relative distances jointly characterize representation discrepancies across encoder depth and feature scale.

To compare representations at corresponding waveform locations, we reindex the shifted patch tokens using the known temporal offset and apply the same scale normalization to compute the aligned relative distance $D_\ell^{\text{align,rel}}$. This operation applies to shifts by integer multiples of the patch length, with the full definitions provided in Appendix B.1.

4 Experiments

We analyze prediction and representation changes under temporal shifts and test whether CTI improves cross-dataset transfer beyond shift augmentation. Controlled experiments examine training constraints, while non-ECG tasks assess classification and shift stability.

4.1 Experimental Setup

Datasets. We train on the five-class PTB-XL source dataset (Wagner et al., 2020, Strodthoff et al., 2021) and evaluate frozen-encoder transfer on three external ECG datasets: Chapman (Zheng et al., 2020), CPSC2018 (Liu et al., 2018), and Georgia (Perez Alday et al., 2020). Each ECG contains 12 leads and 1,000 samples at 100 Hz. Each target dataset retains its own label set.

Comparison methods. We compare CTI with supervised and self-supervised representation-learning methods. CE denotes supervised training with cross-entropy on the original inputs. CE-2V uses the same two independently shifted views as CTI but only the cross-entropy objective. Other supervised methods include SupCon (Khosla et al., 2020), a supervised BYOL variant (BYOL-Sup) (Grill et al., 2020), SAM (Foret et al., 2021), Manifold Mixup (Verma et al., 2019), and R-Drop (Liang et al., 2021). Self-supervised methods include SimCLR (Chen et al., 2020a), MoCo v2 (Chen et al., 2020b), BYOL (Grill et al., 2020), VICReg (Bardes et al., 2022), Barlow Twins (Zbontar et al., 2021), and SimSiam (Chen & He, 2021). We also compare JS-Consistency, Phase-guided (Demirel & Holz, 2025), and ShiFT variants (Shamba et al., 2026) to examine prediction-consistency constraints and different approaches to temporal shifting. In addition, we add the CTI constraint to existing representation-learning methods, including CLOCS (Kiyasseh et al., 2021), 3KG (Gopal et al., 2021), ST-MEM (Na et al., 2024), ACL-ECG (Liu et al., 2026), and TF-C (Zhang et al., 2022), and compare their performance before and after adding the constraint.

Implementation and evaluation. To control for backbone differences, the main method comparisons use a common Transformer encoder. For the main ECG transfer experiments, we train for 60 epochs and use the final-epoch encoder. CTI uses $\lambda = 0.3$ and $\tau = 0.07$. We use AdamW (Loshchilov & Hutter, 2019) for the main encoder and linear-classifier training. During transfer, the encoder remains frozen, and a new linear classifier is trained using target-dataset labels. We primarily report Accuracy and Macro-F1, and use FlipRate to evaluate prediction stability. The main experiments report the mean and sample standard deviation across five random seeds. Classification and stability metrics are expressed as percentages, and performance changes are expressed in percentage points. Data processing, model configurations, and supplementary evaluation details are provided in Appendix A.

4.2 Temporal-Shift Sensitivity

To assess the sensitivity of ECG models to changes in window-relative temporal position, we analyze prediction and representation changes under small temporal shifts. In our diagnostic analysis, a CE model trained without shift augmentation has a prediction FlipRate of $20.62 \pm 1.36\%$ under random circular shifts. Training with diagnostic labels alone can therefore leave predictions dependent on window-relative temporal position, motivating explicit consistency constraints between shifted views.

As shown in Figure 2a–b, the raw Euclidean distance between unaligned representations increases with network depth, whereas the scale-normalized relative distance can decrease, indicating that raw-distance growth partly reflects changes in feature scale. Shifts also displace waveform locations, so we control feature scale and align corresponding waveform locations to compare representations.

Under a fixed full-patch shift, the aligned relative discrepancy increases from the embedding stage to the final encoder block for both CE and the augmentation-matched CE-2V control, whereas CTI maintains a lower discrepancy that decreases slightly (Figure 2c). At the final encoder block, CTI achieves an aligned relative discrepancy of 0.0174, approximately 43% lower than CE-2V's 0.0303. Thus, explicit cross-view contrastive constraints further improve deep-layer consistency at corresponding waveform locations beyond matched shift augmentation. To examine whether this consistency extends to larger displacements, we evaluate full-patch shifts of 50–250 samples. Across these magnitudes, CTI achieves a final-block CLS cosine distance approximately 62–63% lower than CE-2V (Figure 2d), extending this advantage to recording-level representations under the larger displacements evaluated. Appendix B.1 provides the complete comparison across distance metrics.

4.3 Cross-Dataset Transfer

To evaluate cross-dataset transfer of the learned representations, we freeze source-trained encoders and fit linear classifiers on three external ECG datasets. CTI achieves the highest mean Accuracy and Macro-F1 on all three datasets in Table 1, outperforming the supervised and self-supervised methods compared. Compared with the augmentation-matched CE-2V baseline, CTI improves accuracy by

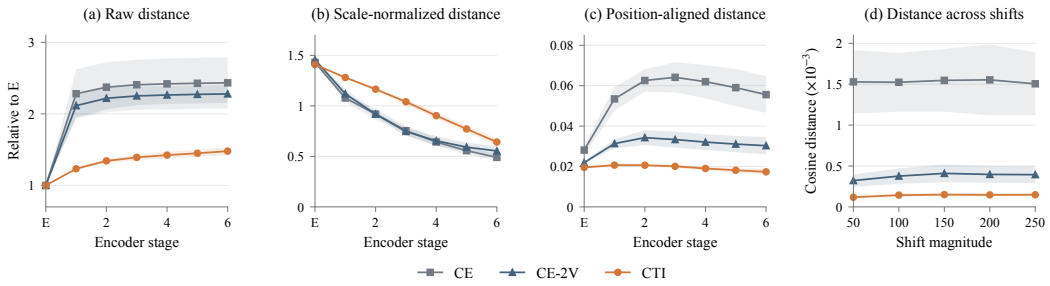


Figure 2: **Temporal-shift sensitivity of ECG representations.** (a) Raw distance relative to the embedding stage (E). (b–c) Scale-normalized distances before and after position alignment. (d) Cosine distance between classification-token representations at the final encoder block across full-patch shift magnitudes. Curves and shading show the mean and standard deviation across five seeds.

Table 1: **Cross-dataset ECG transfer across representation-learning methods.**

		Chapman		CPSC2018		Georgia	
Method	Labels	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Supervised methods							
CE	Yes	59.07 ± 0.44	43.17 ± 0.56	46.20 ± 0.76	26.36 ± 1.41	48.49 ± 1.93	25.60 ± 2.44
CE-2V	Yes	64.12 ± 1.14	49.04 ± 1.12	49.25 ± 0.89	31.89 ± 1.53	51.51 ± 0.52	28.96 ± 1.33
JS-Consistency	Yes	66.49 ± 1.15	50.53 ± 1.36	50.78 ± 1.02	32.65 ± 1.48	49.70 ± 2.40	28.00 ± 1.72
Phase-guided	Yes	64.34 ± 1.17	49.29 ± 1.47	48.73 ± 1.10	31.12 ± 0.91	51.61 ± 0.80	29.32 ± 1.66
ShiFT-Sup	Yes	74.07 ± 0.36	58.53 ± 0.61	52.57 ± 0.65	34.75 ± 0.63	54.68 ± 1.41	38.06 ± 1.76
SupCon	Yes	64.41 ± 0.99	48.63 ± 0.81	49.39 ± 0.95	27.77 ± 1.97	51.14 ± 0.92	28.24 ± 1.84
BYOL-Sup	Yes	69.63 ± 0.84	54.37 ± 1.15	51.79 ± 1.10	34.85 ± 1.56	53.60 ± 1.05	34.71 ± 2.60
SAM	Yes	60.75 ± 0.91	44.96 ± 1.08	45.44 ± 0.73	25.58 ± 0.74	49.28 ± 1.12	25.67 ± 1.43
Manifold Mixup	Yes	58.19 ± 1.77	42.71 ± 1.58	44.88 ± 2.27	24.30 ± 2.26	50.59 ± 0.74	25.67 ± 1.11
R-Drop	Yes	58.38 ± 1.23	42.03 ± 1.38	45.15 ± 0.44	21.52 ± 2.59	47.42 ± 1.15	23.93 ± 0.93
Self-supervised methods							
SimCLR	No	66.89 ± 0.46	50.68 ± 0.42	43.41 ± 0.91	26.01 ± 1.50	51.82 ± 1.69	34.38 ± 1.84
MoCo v2	No	70.87 ± 0.74	53.84 ± 0.58	44.39 ± 1.18	26.25 ± 2.57	53.03 ± 1.51	39.33 ± 1.46
BYOL	No	69.78 ± 0.68	52.58 ± 0.65	40.99 ± 1.70	18.45 ± 1.00	49.12 ± 0.83	32.08 ± 0.83
VICReg	No	66.67 ± 2.60	50.33 ± 2.57	43.66 ± 0.97	25.62 ± 0.94	48.75 ± 1.85	28.96 ± 3.05
Barlow Twins	No	41.41 ± 1.16	19.31 ± 3.50	34.08 ± 0.17	5.70 ± 0.06	36.20 ± 2.63	15.57 ± 2.82
SimSiam	No	43.35 ± 3.17	20.10 ± 4.97	35.10 ± 1.12	10.15 ± 1.21	43.41 ± 3.17	20.85 ± 1.60
ShiFT	No	68.93 ± 0.42	52.42 ± 0.31	43.91 ± 0.47	24.05 ± 0.89	52.16 ± 0.91	36.54 ± 1.85
Our method and label-free reference							
CTI-SSL	No	69.98 ± 0.84	53.70 ± 0.70	44.03 ± 1.40	30.13 ± 1.44	51.22 ± 0.57	34.98 ± 1.14
CTI (Ours)	Yes	**78.33 ± 0.29**	**62.17 ± 0.31**	**56.07 ± 0.77**	**41.69 ± 0.96**	**56.44 ± 1.28**	**43.28 ± 2.73**

14.21, 6.82, and 4.93 percentage points on Chapman, CPSC2018, and Georgia, respectively. The corresponding Macro-F1 gains are 13.13, 9.80, and 14.32 percentage points. CTI improves overall classification performance while also achieving better performance averaged equally across classes. With the same shifted views, CTI thus provides additional transfer gains through explicit cross-view contrastive constraints beyond shift augmentation.

To examine whether this constraint can complement other representation-learning objectives, we add the CTI constraint to existing methods and compare each with its original version (Table 2). CE, CE-2V, SupCon, and BYOL-Sup improve across all six transfer metrics, with SupCon achieving a mean gain of 12.45 percentage points. ECG-specific methods also benefit. CLOCS and 3KG

Table 2: **Effect of adding the CTI constraint on cross-dataset ECG transfer.**

	Chapman		CPSC2018		Georgia		
Method	Δ Accuracy	Δ Macro-F1	Δ Accuracy	Δ Macro-F1	Δ Accuracy	Δ Macro-F1	Mean Δ
CE	+13.56	+13.59	+3.72	+9.21	+5.90	+14.39	+10.06
CE-2V	+12.54	+11.47	+5.11	+9.69	+5.19	**+15.57**	+9.93
SupCon	**+15.05**	**+14.89**	**+8.81**	**+16.05**	+5.82	+14.11	**+12.45**
BYOL-Sup	+8.75	+7.86	+4.28	+7.33	+3.77	+12.29	+7.38
SimCLR	+3.71	+3.56	−0.21	+3.77	+0.24	+1.22	+2.05
CLOCS (CMSMLC)	+3.66	+3.47	+3.47	+6.04	**+6.24**	+11.07	+5.66
3KG	+0.46	+0.37	+1.24	+1.93	+2.04	+2.38	+1.40
ST-MEM	+2.61	+1.38	+3.29	+8.68	+4.59	+7.98	+4.76
ACL-ECG	−0.10	−0.22	+1.94	+3.14	+1.70	+2.41	+1.48
TF-C	+0.43	+0.40	+0.58	+3.49	+0.52	+0.49	+0.99

Δ: change after adding CTI, in percentage points. Mean Δ: average over the six dataset–metric entries. Gray denotes decreases, bold marks column maxima, and dashes indicate unavailable results.

improve on all three datasets, while ST-MEM and ACL-ECG improve both Accuracy and Macro-F1 on CPSC2018 and Georgia. CTI thus extends to other training frameworks, using temporal-shift positive-pair constraints to further improve transfer performance in existing supervised, contrastive, and masked representation-learning methods.

To test whether the transfer gains extend to full fine-tuning, we jointly update the pretrained encoder and a new classifier on each target dataset. Compared with CE-2V, CTI improves accuracy by 5.35, 5.29, and 4.14 percentage points on Chapman, CPSC2018, and Georgia, respectively, with corresponding Macro-F1 gains of 5.84, 7.45, and 10.56 percentage points (Appendix B.1, Table 6). CTI's transfer advantage therefore persists when the encoder is further optimized on target data, extending beyond frozen-representation linear evaluation.

4.4 Ablations and Controlled Analysis

To examine how training constraints relate to CTI's transfer gains, we compare output-consistency training and ablate the representation constraints. JS-Consistency reduces source-domain FlipRate to 4.90%, below CTI's 13.51%, yet CTI achieves higher transfer accuracy on all three external ECG datasets (Figure 3a). Relative to CE-2V, CTI also transfers better despite a slightly higher FlipRate. CTI's transfer advantage does not require a lower source-domain FlipRate. Compared with constraining source-domain output consistency alone, CTI's representation-level contrastive constraint may be more conducive to learning diagnostic features that can be reused across datasets.

Component ablations reveal distinct roles for positive-pair consistency and the representation distribution. Adding only a feature-alignment constraint that brings positive-pair representations closer yields lower transfer performance than CE-2V. Adding a distributional constraint alongside alignment further improves transfer and brings performance closer to full CTI, which achieves the highest accuracy on all three target datasets in this component study (Figure 3b, Appendix B.2). This suggests that CTI's transfer gains may arise from two complementary effects: reducing representation discrepancies between temporal positions of the same recording while preserving distinguishability between different recordings. Transfer gains also hold across training shift magnitudes, zero-padded and reflection-padded training views, and Transformers with different positional encodings and a 1D CNN (Appendices B.1–B.3), showing that they are not confined to the default shift magnitude, circular wrapping, or a single encoder.

4.5 Evaluation Beyond ECG

To test whether CTI also improves temporal-shift stability beyond ECG, we train models separately for human activity recognition (HAR), sleep staging (Sleep-EDF), and epilepsy recognition (Epilepsy), and evaluate classification performance and temporal-shift stability under different boundary op-

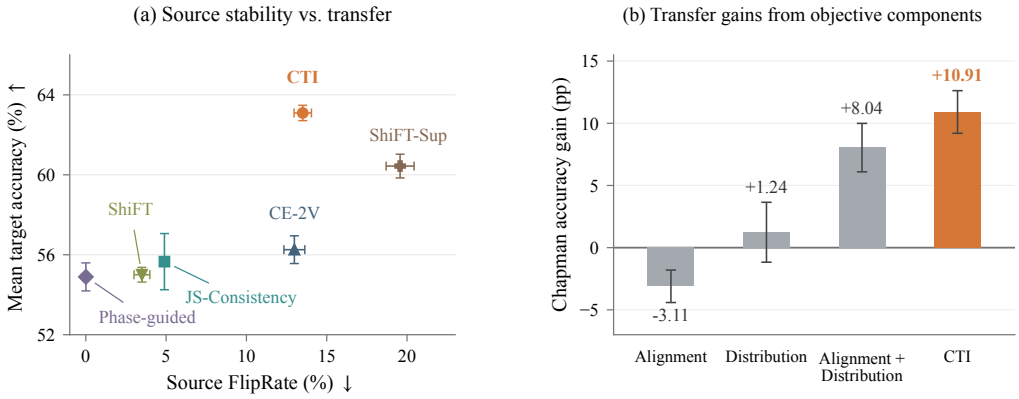


Figure 3: **Training constraints and transfer performance.** (a) Source FlipRate versus mean accuracy across three ECG targets. (b) Chapman accuracy gains over CE-2V with different objective components. Error bars show standard deviations across seeds.

Table 3: **Classification performance and temporal-shift stability beyond ECG.**

Dataset	Method	Accuracy ↑	Macro-F1 ↑	FlipRate ↓		
				Circular	Zero	Reflect
HAR	CE	92.88 ± 0.41	92.80 ± 0.42	3.65 ± 0.22	3.82 ± 0.28	3.50 ± 0.24
	CTI (Ours)	$\mathbf{94.78 \pm 0.27}$	$\mathbf{94.75 \pm 0.29}$	$\mathbf{1.49 \pm 0.20}$	$\mathbf{1.50 \pm 0.19}$	$\mathbf{1.50 \pm 0.21}$
Sleep-EDF	CE	75.74 ± 1.20	66.41 ± 1.62	15.94 ± 1.46	16.37 ± 1.50	16.11 ± 1.42
	CTI (Ours)	$\mathbf{79.14 \pm 0.38}$	$\mathbf{72.53 \pm 1.25}$	$\mathbf{8.45 \pm 0.40}$	$\mathbf{9.18 \pm 0.44}$	$\mathbf{9.27 \pm 0.48}$
Epilepsy	CE	97.92 ± 0.14	96.74 ± 0.22	1.77 ± 0.21	1.65 ± 0.18	1.69 ± 0.17
	CTI (Ours)	$\mathbf{98.83 \pm 0.40}$	$\mathbf{98.18 \pm 0.61}$	$\mathbf{0.58 \pm 0.13}$	$\mathbf{0.58 \pm 0.13}$	$\mathbf{0.60 \pm 0.11}$

erations. As shown in Table 3, CTI improves both accuracy and Macro-F1 over CE on all three tasks, with accuracy gains of 1.90, 3.40, and 0.91 percentage points, respectively. CTI thus improves classification and prediction stability across different types of time-series signals beyond ECG.

Although training uses only circular shifts, CTI also maintains lower prediction FlipRates under zero-padded and reflection-padded shifts. Across the three boundary operations, CTI reduces FlipRate relative to CE by approximately 57–61% on HAR, 42–47% on Sleep-EDF, and 64–67% on Epilepsy. These results demonstrate that CTI alleviates prediction sensitivity to changes in temporal position, with gains extending to boundary operations not used during training.

5 Conclusion

We introduced Contrastive Time-Invariant Learning (CTI), which jointly optimizes classification and contrastive objectives to encourage consistent representations of two independently shifted views of the same ECG. With matched shifted views and training budgets, CTI improves frozen-encoder transfer accuracy by 4.9–14.2 percentage points over CE-2V across three external ECG datasets and achieves better deep-layer representation consistency in position-aligned analyses. Its transfer advantage does not require a lower source-domain FlipRate. Adding this constraint to a range of existing representation-learning methods also improves cross-dataset classification performance. These results show that positive-pair construction not only determines which training views a model encounters, but can also specify the changes across which representations should remain consistent. Explicit cross-view contrastive constraints on temporal shifts thus improve ECG representation transferability beyond shift augmentation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing and polishing the writing of this manuscript (sentence-level phrasing and consistency checks), and we did not use generative AI tools for research ideation, experimental design, or the core methodology of Contrastive Time-Invariant Learning. All AI-assisted text was reviewed and verified by the authors against the underlying experimental results, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes previously released, publicly available, de-identified datasets (PTB-XL, Chapman, CPSC2018, Georgia, CODE-15%, HAR, Sleep-EDF, and Epilepsy). It does not involve new human-subject data collection. Our results are retrospective and should not be read as evidence of deployment readiness: classification errors remain, and a representation with improved controlled-shift consistency can still fail under demographic, device, lead-quality, or workflow shifts that we do not test. Any clinical use of models built on this work would require prospective validation, calibration monitoring, and human oversight, and the model should be positioned as a decision-support aid used alongside clinician judgment, not as a standalone diagnostic tool.

REPRODUCIBILITY STATEMENT

The training objective, view construction, and sensitivity assessment are described in Section 3. Encoder details are given in Appendix A.1. Section 4.1 summarizes the main evaluation, with data partitions, baseline adaptations, evaluation procedures, and statistical families detailed in Appendix A. Supplementary diagnostic and component studies are identified separately with their respective protocols. Code and experiment configurations will be made publicly available.

REFERENCES

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b. URL https://arxiv.org/abs/2003.04297.

Berken Utku Demirel and Christian Holz. Shifting the paradigm: A diffeomorphism between time series data manifolds for achieving shift-invariancy in deep learning. In *International Conference on Learning Representations*, 2025.

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 *International Joint Conference on Artificial Intelligence*, pp. 2352–2359, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3KG: contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 156–167. PMLR, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Bob Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. CLOCS: contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 2021.

Xiang Lan, Hanshu Yan, Shenda Hong, and Mengling Feng. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. In *International Conference on Learning Representations*, 2024.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905, 2021.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Wenhan Liu, Zhijing Wu, and Zhaohui Yuan. ACL-ECG: Anatomy-aware contrastive learning for multi-lead electrocardiograms. *Sensors*, 26(3):1080, 2026. doi: 10.3390/s26031080.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, 2024.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12): 124003, 2020.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: invariances, augmentations and dataset biases. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3407–3418, 2020.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013. URL `https://doi.org/10.24432/C54S4K`. Data set.

Antônio H. Ribeiro, Gabriela M. M. Paixão, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. CODE-15%: a large scale annotated dataset of 12-lead ECGs. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.4916206. Data set, version 1.0.0.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Abdul-Kazeem Shamba, Kerstin Bach, and Gavin Taylor. Learning by shifting: Temporal view construction for time series contrastive learning. *arXiv preprint arXiv:2606.21957*, 2026. doi: 10.48550/arXiv.2606.21957.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 2019.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Qiuyi Wu and Ernest Fokoue. Epileptic seizure recognition. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5G308. Data set.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334. PMLR, 2019.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3988–4003, 2022.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020.

A Experimental Details

A.1 Datasets and Evaluation Protocol

Data and task construction. We train encoders on PTB-XL and evaluate frozen-representation transfer on Chapman, CPSC2018, and Georgia. Each dataset defines a separate single-label classification task with its own classes. PTB-XL uses five diagnostic superclasses: NORM, MI, STTC, CD, and HYP. For records with multiple supported diagnostic classes, we select the class associated with the highest diagnostic confidence, resolving ties by the original code order. Chapman and CPSC2018 map specified diagnostic codes to five and nine classes, respectively. When multiple mapped classes occur in a record, we select the first matching class in its diagnosis-code list. Georgia uses the first diagnosis code per record and retains the five most frequent codes as classes. These procedures convert the original annotations into the single-label tasks evaluated here.

Signal processing and filtering. Inputs contain 12 leads and 1,000 samples at 100 Hz. For Chapman and CPSC2018, we resample the signals, retain the first 1,000 samples or zero-pad shorter records, and standardize each lead within each record. For Georgia, we retain 12-lead recordings with an original length of 5,000 samples, resample them to 1,000 samples, and preserve physical amplitudes. Each retained recording contributes one input sample, without sliding-window expansion. Filtering excludes records that cannot be read, have incompatible input formats, fail the dataset-specific label-selection rules, or contain non-finite values after processing. Table 4 reports the samples used after label selection and signal processing.

Data splits. PTB-XL follows the official folds, using folds 1–8, 9, and 10 for training, validation, and testing, respectively, with no patients shared across splits. Chapman and CPSC2018 use random record-level splits of 80%, 10%, and 10%. Georgia uses random record-level splits of 70%, 15%, and 15%. All three external splits use random seed 42 and are shared across methods.

Table 4: Processed datasets used in the main comparison. Counts denote training, validation, and test records after loading and filtering.

Dataset	Input $C \times T$	Classes	Train	Validation	Test
PTB-XL	12×1000	5	17,084	2,146	2,158
Chapman	12×1000	5	33,563	4,192	4,197
CPSC2018	12×1000	9	8,264	1,033	1,033
Georgia	12×1000	5	3,557	761	763
HAR	9×128	6	5,881	1,471	2,947
Sleep-EDF	1×3000	5	27,425	8,078	6,805
Epilepsy	1×178	2	7,360	1,840	2,300

Encoder. The common encoder has six Transformer layers with hidden dimension $d = 256$, eight attention heads, and dropout 0.1. We use CLS after the final LayerNorm. Patch lengths are 50, 8, 150, and 9 samples for ECG, HAR, Sleep-EDF, and Epilepsy, respectively.

Training and evaluation. The main experiments follow the training setup in Section 4.1. With the encoder frozen, we train a new linear classifier for 30 epochs and select the classifier with the highest validation accuracy. CE-2V, JS-Consistency, and CTI use the same shift distribution. We assess performance differences between methods using two-sided t-tests paired by random seed, applying Holm correction (Holm, 1979) to multiple comparisons within each family.

A.2 Temporal-Shift Construction and Baselines

Shift sampling. Temporal offsets are sampled as

$$\delta = \mathrm{trunc}(\rho T(2u-1)), \quad u \sim \mathrm{Unif}[0,1), \quad \rho = 0.05, \tag{7}$$

Truncation rounds toward zero, giving a nonuniform integer-offset distribution. Offsets are sampled independently across recordings and views and shared across leads.

Classification controls. CE trains on unshifted recordings using diagnostic labels. CE-2V averages the classification losses of two independently shifted views. CTI uses these same views and classification losses, with the additional symmetric cross-view contrastive objective defined in Section 3. SAM, Manifold Mixup, and R-Drop use inputs without temporal shifting for supervised training. SAM computes the update gradient after perturbing the model parameters. Manifold Mixup mixes features and labels at the input or an intermediate encoder layer. R-Drop performs two forward passes on the same input with independent dropout and encourages consistency between their output distributions.

Contrastive and negative-free baselines. SimCLR uses views of the same recording as positives and a full two-view contrastive denominator that includes same-view negatives. In the implemented SupCon baseline, class labels additionally define positives across recordings, and the supervised contrastive loss is combined with averaged two-view cross-entropy. BYOL-Sup retains the shifted-view classification objective and replaces the cross-view contrastive term with a predictor-based stop-gradient objective without explicit negatives. CTI-SSL uses only CTI's circular-shift positive pairs and cross-view contrastive objective, without diagnostic-label supervision.

Additional self-supervised baselines. MoCo v2, BYOL, SimSiam, VICReg, and Barlow Twins use the common Transformer encoder while retaining their method-specific projection heads, prediction heads, and target-network mechanisms where applicable. These five methods use two independently zero-padded shifted views, the same number of training epochs as the main experiments, and common optimizer settings. Method-specific loss parameters are fixed before training and are not tuned on the test set. Frozen evaluation uses encoder representations without the auxiliary training heads.

Prediction consistency. JS-Consistency adds Jensen–Shannon divergence between the predicted class distributions of the two shifted views to their averaged classification loss. Its weight is calibrated on training data to match the initial encoder-gradient norm of CTI's weighted contrastive term.

Phase normalization. Phase-guided normalization adapts the guidance network and frequency-domain transformation of Demirel & Holz (2025) to the common encoder. It uses Fourier magnitudes to guide input phase normalization, retaining the original guidance regularizer. The transformation remains active during inference and frozen-encoder transfer.

ShiFT variants. ShiFT uses the overlapping prefix/suffix views and NT-Xent objective of Shamba et al. (2026), with a projection head and the common encoder in place of InceptionTime. Its two views overlap by 0.5 and have length approximately $2T/3$. ShiFT trains without source labels, while ShiFT-Sup adds averaged view cross-entropy with contrastive weight 0.3. Unlike CTI's full-window circular shifts, these views change input length and edge content. During evaluation, we discard the projector and extract CLS features from full-length inputs.

B Additional ECG Analyses

B.1 Representation Analysis and Transfer

Analysis setting. To assess feature scale and temporal correspondence, we compare distances before and after alignment. Figure 2a–c uses encoders selected on the source validation set, a common set of 512 source test records, and a fixed 50-sample shift. Panel (d) uses the same checkpoints and records to evaluate CLS cosine distance before the final LayerNorm under shifts of 50, 100, 150, 200, and 250 samples (1–5 patches). For each seed, we average over records and positive and negative shift directions before computing the mean and sample standard deviation across seeds.

Alignment controls. Temporal shifts change waveform locations within the token sequence, so comparisons at the same token index can include differences caused by waveform displacement. To compare representations at corresponding waveform locations, we reindex the shifted patch tokens using the known temporal offset.

Suppose the input is partitioned into $M = T/P$ equal-length, non-overlapping patches of P samples without additional padding. A shift $\delta = qP$, $q \in \mathbb{Z}$, induces an exact circular correspondence between patches, where q is the integer offset in patch units. We define the aligned features as

$$\widetilde{H}_{\ell,m}(S_\delta x) = \begin{cases} H_{\ell,0}(S_\delta x), & m = 0, \\ H_{\ell,1+((m-1+q) \bmod M)}(S_\delta x), & m = 1, \dots, M. \end{cases} \tag{8}$$

Patch tokens are compared at corresponding waveform locations, while CLS retains its original index and is included in the distance calculation. The aligned mean distance and its relative form are

$$D_\ell^{\text{align}}(x,\delta) = \frac{1}{M+1}\sum_{m=0}^{M}\left\|H_{\ell,m}(x) - \widetilde{H}_{\ell,m}(S_\delta x)\right\|_2, \qquad D_\ell^{\text{align,rel}}(x,\delta) = \frac{D_\ell^{\text{align}}(x,\delta)}{\max\{A_\ell(x),\epsilon\}}. \tag{9}$$

The relative form uses the same scale normalization as Eq. 6.

Reindexing establishes positional correspondence between tokens, while the effects of temporal position on their features remain reflected in the aligned distance.

Additional distance measures. To account for both views, symmetric normalization uses

$$D_\ell^{\text{sym}}(x,\delta) = \frac{2D_\ell(x,\delta)}{\max\{A_\ell(x),\epsilon\} + \max\{A_\ell(S_\delta x),\epsilon\}}. \tag{10}$$

Alignment changes only the numerator. Cosine distance is $1 - \cos$ similarity, averaged over patch tokens or computed for CLS.

Scale and alignment results. Figure 2a shows raw distances relative to each seed's embedding-stage mean, while panels (b) and (c) aggregate per-record relative distances. Raw distances increase with depth, whereas scale-normalized distances decrease. After alignment, CTI's relative discrepancy decreases from embedding to the final encoder block, while those of CE and CE-2V increase.

Table 5 further shows that, after controlling feature scale and aligning corresponding waveform locations, CTI achieves the lowest relative, symmetrically normalized, and patch cosine distances, together with the lowest CLS cosine distance. CTI thus exhibits greater consistency in both local representations at corresponding waveform locations and the global recording representation. Unaligned distances rank methods differently because comparisons at the same token index conflate waveform displacement with representation differences at corresponding locations.

Table 5: **Representation distances at the final encoder block.** Fixed one-patch shift. Distances are dimensionless, with scaling indicated where used.

Metric	CE	CE-2V	CTI (Ours)
Unaligned relative distance	0.4898 ± 0.0315	0.5538 ± 0.0493	0.6418 ± 0.0326
Unaligned symmetric distance	0.4897 ± 0.0314	0.5539 ± 0.0492	0.6418 ± 0.0325
Unaligned patch cosine distance	0.1232 ± 0.0086	0.1470 ± 0.0274	0.2361 ± 0.0242
Aligned relative distance	0.0556 ± 0.0090	0.0303 ± 0.0043	0.0174 ± 0.0013
Aligned symmetric distance	0.0556 ± 0.0090	0.0303 ± 0.0043	0.0174 ± 0.0013
Aligned patch cosine distance ($\times 10^{-3}$)	1.9674 ± 0.6405	0.6344 ± 0.1716	0.1699 ± 0.0271
CLS cosine distance ($\times 10^{-3}$)	1.5131 ± 0.3978	0.3327 ± 0.0692	0.1189 ± 0.0153

Class structure on target datasets. We further examine the class structure of final-epoch frozen encoder representations on three external ECG datasets. As shown in Figure 4, CTI achieves higher mean silhouette scores (Rousseeuw, 1987) than CE and CE-2V on all three datasets and reduces the prevalence of strongly negative per-record scores. These results indicate better overall class separation in the target domains, with a clearer relationship between within-class similarity and separation from other classes.

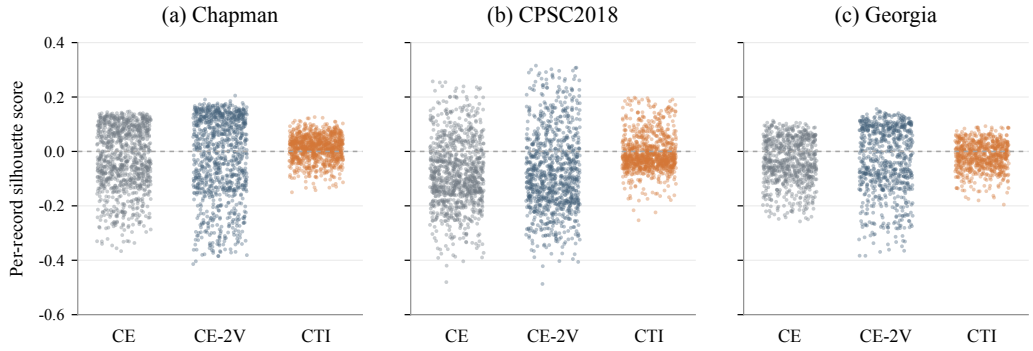


Figure 4: **Class structure of frozen representations on external ECG datasets.** Each point shows a record's Euclidean silhouette score computed using ground-truth classes in the original 256-dimensional representation space, averaged over five training seeds. Higher scores indicate better separation from other classes.

Full fine-tuning. Starting from the final source-training encoders, we initialize new classification heads and update all encoder layers on each target dataset for 30 epochs, selecting checkpoints by validation accuracy. Fine-tuning uses unshifted inputs and cross-entropy without an additional contrastive loss. Table 6 reports results on the original test sets. CTI improves both metrics in all five paired seeds on every target dataset. The gains remain positive in every seed after excluding test waveforms that exactly duplicate training or validation waveforms. CTI's transfer gains thus extend beyond frozen evaluation to target-domain encoder fine-tuning.

Table 6: **Full fine-tuning on external ECG datasets.**

Dataset	Method	Accuracy ↑	Macro-F1 ↑
Chapman	CE-2V	81.11 ± 0.41	64.41 ± 0.86
	CTI (Ours)	$\mathbf{86.45 \pm 0.42}$	$\mathbf{70.25 \pm 0.77}$
CPSC2018	CE-2V	54.77 ± 0.50	38.65 ± 0.74
	CTI (Ours)	$\mathbf{60.06 \pm 1.15}$	$\mathbf{46.09 \pm 1.27}$
Georgia	CE-2V	56.17 ± 0.94	41.50 ± 2.33
	CTI (Ours)	$\mathbf{60.31 \pm 1.11}$	$\mathbf{52.06 \pm 1.78}$

Encoder architectures. To test whether CTI's transfer gains depend on a particular encoder, we compare CTI with a matched CE-2V baseline using Transformers with absolute or relative positional encoding and a 1D CNN. CTI improves mean transfer accuracy for all three encoder configurations on each target dataset, with gains of 2.44–14.21 percentage points (Table 7). These results show that CTI's transfer gains extend across the evaluated positional encodings and to a convolutional encoder.

Architecture implementation. For Table 7, the relative-position Transformer retains the common encoder configuration but fixes absolute position embeddings to zero. Each layer and attention head learns an additive attention-logit bias indexed by signed patch distance, with offsets from -19 to 19 and separate entries for CLS-to-CLS, CLS-to-patch, and patch-to-CLS interactions. All biases are initialized to zero. The 1D CNN is a ResNet1D with a 64-channel, kernel-15, stride-2 stem followed by kernel-3, stride-2 max pooling. Three residual stages have widths 64, 128, and 256, each containing two blocks with two kernel-3 convolutions, BatchNorm, and ReLU. The first blocks of stages two and three downsample with stride 2 and use projected shortcuts. Global average pooling followed by a linear layer produces the 256-dimensional recording representation. All architectures use the main data splits and training setup, comparing CE-2V and CTI over five paired seeds.

Table 7: **Cross-dataset transfer gains of CTI across encoder architectures.**

Encoder	Chapman	CPSC2018	Georgia
Transformer (absolute position)	+14.21	+6.82	+4.93
Transformer (relative position)	+13.85	+4.69	+4.80
1D CNN	+4.31	+8.11	+2.44

B.2 Objective and View-Construction Ablations

Objective components. To analyze the roles of positive-pair alignment and the distribution of representations across recordings in transfer, we add feature-alignment and distributional constraints separately and jointly while retaining the two-view classification objective, and compare these alternatives with full CTI. Encoders are frozen after 60 epochs, and linear classifiers are fitted with Adam for 20 epochs using the main seed set.

Alignment minimizes squared distance between normalized paired CLS features. For each view, the distributional constraint takes the logarithm of the mean of $\exp(-2\|\hat{z}_i - \hat{z}_j\|_2^2)$ over distinct recordings, then averages the resulting values across the two views (Wang & Isola, 2020). Regularizer weights are calibrated on training data to match the initial encoder-gradient norm of CTI's weighted contrastive term. This alternative objective is not an exact finite-batch decomposition of InfoNCE.

Table 8: **Objective components.** Transfer accuracy under different representation objectives. All rows include two-view classification.

Objective	Chapman	CPSC2018	Georgia
CE-2V	67.18 ±1.38	51.62 ±1.88	52.24 ±1.91
+ Alignment	64.07 ±0.86	50.28 ±0.44	49.88 ±1.25
+ Distribution	68.42 ±1.30	52.24 ±1.92	53.37 ±0.49
+ Alignment and distribution	75.23 ±1.71	54.13 ±0.61	53.32 ±2.46
CTI (Ours)	78.09 ±0.75	56.52 ±0.22	57.69 ±0.70

Table 8 extends the component comparison in Figure 3b to all three targets. Adding only feature alignment yields lower accuracy than CE-2V on every dataset. Combining alignment with the distributional constraint improves transfer, but full CTI still achieves the highest means. These results show that bringing positive-pair representations closer alone is insufficient to improve transfer, whereas jointly constraining positive-pair consistency and the representation structure across recordings is more conducive to learning transferable representations.

Shared versus independent offsets. Table 9 compares shared and independent offsets with and without the contrastive term. Both settings retain two forward passes and independent dropout, while only independent offsets vary the temporal position between views.

Table 9: **Effect of temporal diversity between training views.** Transfer accuracy.

Objective	Offsets	Chapman	CPSC2018	Georgia
CE-2V	Shared	65.67 ±0.69	50.26 ±1.59	51.25 ±0.64
CE-2V	Independent	67.18 ±1.38	51.62 ±1.88	52.24 ±1.91
CE + contrastive	Shared	73.40 ±0.56	53.96 ±0.93	53.74 ±0.71
CTI (Ours)	Independent	78.09 ±0.75	56.52 ±0.22	57.69 ±0.70

With shared offsets, adding the contrastive objective still improves transfer accuracy. Independent offsets further yield higher mean accuracy on all three target datasets, indicating an additional benefit from constructing positive pairs at different temporal positions. CTI's transfer advantage thus benefits from both the contrastive objective and temporal-position differences between views. Relative displacement and augmentation diversity change together in this comparison.

Effect of training shift magnitude. To assess sensitivity to the training shift range, we compare CE-2V and CTI at $\rho \in \{0.025, 0.05, 0.10\}$ under matched training settings (Table 10). CTI achieves higher accuracy on all three target datasets at every range. CTI's transfer gains thus persist across the default shift magnitude and the smaller and larger ranges evaluated.

Table 10: **Effect of training shift magnitude on cross-dataset ECG transfer.**

ρ	Method	Chapman	CPSC2018	Georgia	Mean
0.025	CE-2V	65.25 ± 1.38	49.91 ± 0.67	51.51 ± 0.76	55.56 ± 0.35
	CTI (Ours)	$\mathbf{78.41 \pm 0.62}$	$\mathbf{55.72 \pm 0.83}$	$\mathbf{56.36 \pm 1.08}$	$\mathbf{63.50 \pm 0.74}$
0.05	CE-2V	64.12 ± 1.14	49.25 ± 0.89	51.51 ± 0.52	54.96 ± 0.48
	CTI (Ours)	$\mathbf{78.33 \pm 0.29}$	$\mathbf{56.07 \pm 0.77}$	$\mathbf{56.44 \pm 1.28}$	$\mathbf{63.61 \pm 0.67}$
0.10	CE-2V	64.83 ± 0.72	49.66 ± 1.52	50.48 ± 1.13	54.99 ± 0.59
	CTI (Ours)	$\mathbf{78.34 \pm 0.63}$	$\mathbf{55.91 \pm 0.75}$	$\mathbf{56.88 \pm 0.94}$	$\mathbf{63.71 \pm 0.55}$

Contrastive hyperparameters. We evaluate the nine combinations of $\lambda \in \{0.1, 0.3, 1.0\}$ and $\tau \in \{0.03, 0.07, 0.20\}$ with a matched CE-2V control. Across this grid, mean accuracy over the three targets ranges from 59.67% to 63.61%, exceeding the control's 54.96% by 4.71–8.65 percentage points. CTI improves mean accuracy and Macro-F1 in all nine evaluated settings (Figure 5), showing that its transfer gains are not confined to a single loss weight or temperature.

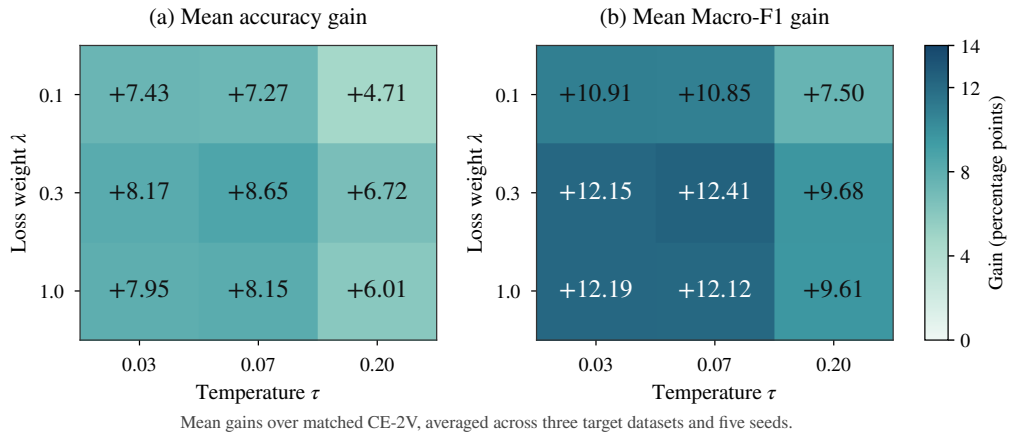


Figure 5: **Sensitivity to contrastive loss weight and temperature.** Mean gains over the matched CE-2V baseline in (a) accuracy and (b) Macro-F1, in percentage points.

B.3 Boundary Effects

Boundary handling at test time. To test whether CTI trained with circular shifts remains stable under other boundary treatments, we apply circular, zero-padded, and reflection-padded shifts at test time and compare prediction FlipRates. The three evaluations share the same temporal offsets and differ only in boundary handling. Table 11 also reports classification accuracy on unshifted inputs.

Table 11: **Source ECG classification and boundary sensitivity on PTB-XL.**

Method	Accuracy ↑	Circular FlipRate ↓	Zero FlipRate ↓	Reflect FlipRate ↓
CE	65.07 ± 1.03	20.62 ± 1.36	21.02 ± 1.34	20.83 ± 1.34
CTI (Ours)	67.61 ± 0.90	11.54 ± 1.15	11.61 ± 1.20	11.56 ± 1.12

Compared with CE, CTI reduces FlipRate by 9.08, 9.41, and 9.27 percentage points under circular, zero-padded, and reflection-padded shifts, respectively, while improving clean-input accuracy from

65.07% to 67.61%. These results demonstrate that CTI alleviates prediction sensitivity to changes in temporal position, maintaining stability gains of similar magnitude even under zero and reflection padding, neither of which is used during training.

Boundary handling during training. To further test whether the transfer gains depend on circular wrapping, we train CE-2V and CTI with circular, zero-padded, or reflection-padded shifts, making matched comparisons within each boundary treatment. CTI improves both accuracy and Macro-F1 on all three target datasets under all three training settings (Table 12). Thus, even with shifted views that do not wrap the signal across window boundaries, explicit cross-view contrastive constraints still provide additional transfer gains.

Table 12: **Transfer gains under different training boundary treatments.**

	Chapman		CPSC2018		Georgia	
Training boundary	Δ Accuracy	Δ Macro-F1	Δ Accuracy	Δ Macro-F1	Δ Accuracy	Δ Macro-F1
Circular	+14.21	+13.13	+6.82	+9.80	+4.93	+14.32
Zero padding	+14.77	+13.07	+5.92	+10.20	+5.69	+16.73
Reflection padding	+13.62	+12.72	+5.85	+9.24	+4.80	+14.30

B.4 Patient-Disjoint Multi-Label Transfer

To evaluate CTI under patient-disjoint splits and multi-label diagnosis, we further assess frozen-encoder transfer on CODE-15% (Ribeiro et al., 2021), retaining its six original diagnostic labels and allowing multiple labels per recording. Compared with CE-2V, CTI improves Macro-AUROC, Macro-AUPRC, and Macro-F1 by 4.55, 8.85, and 8.30 percentage points, respectively (Table 13), with gains in all three metrics for all five seeds. CTI's transfer advantage thus extends to multi-label ECG classification with no patients shared across target-data splits.

Data processing and splits. Original recordings are sampled at 400 Hz and stored as 4096×12 arrays. We extract the central 4,000 samples, downsample to 100 Hz using polyphase filtering, and standardize each lead within each recording to obtain 12×1000 inputs, retaining the original zero padding. After excluding non-finite or all-zero waveforms and groups of identical waveforms with conflicting labels, we construct approximately 80%/10%/10% patient-group splits containing 275,867 training, 34,342 validation, and 34,490 test recordings. The test set contains 23,375 patients. Neither patients nor exact duplicate waveforms are shared across splits.

Classifier and thresholds. We freeze the encoders trained on PTB-XL and fit a new six-output linear classifier using CODE-15% training labels with binary cross-entropy, selecting the classifier by validation Macro-AUROC. For Macro-F1, a separate threshold for each label is selected to maximize validation F1 over 0.01–0.99 in increments of 0.01, choosing the larger threshold in a tie. The test set is used for neither model nor threshold selection.

Table 13: **Patient-disjoint multi-label transfer on CODE-15%.**

Method	Macro-AUROC ↑	Macro-AUPRC ↑	Macro-F1 ↑
CE	84.96 ± 0.59	26.25 ± 0.67	31.09 ± 0.63
CE-2V	89.29 ± 0.42	32.30 ± 0.69	37.72 ± 0.66
CTI (Ours)	$\mathbf{93.84 \pm 0.29}$	$\mathbf{41.16 \pm 0.78}$	$\mathbf{46.02 \pm 0.96}$
CTI − CE-2V	$+4.55$	$+8.85$	$+8.30$

C EVALUATION BEYOND ECG

Tasks and preprocessing. Models are trained separately on HAR (Reyes-Ortiz et al., 2013), Sleep-EDF (Kemp et al., 2000), and Epilepsy (Wu & Fokoue, 2017). Dataset sizes and encoder configurations are given in Appendix A.1. HAR uses the official subject-separated test set and a window-level training/validation split. Sleep-EDF-20 uses 12/4/4 training/validation/test subjects and 30-second Fpz-Cz EEG windows at 100 Hz. Epilepsy uses a random window split. Non-ECG preprocessing follows TS-TCC (Eldele et al., 2021), with channel standardization fitted only on training data. For Epilepsy, we apply channel standardization to raw signals rather than time-index-specific MinMax scaling, to avoid introducing additional temporal-position dependence through preprocessing.

Evaluation settings. Table 3 uses the validation-selected model and its jointly trained classification head for each task. Accuracy and Macro-F1 are evaluated on unshifted inputs. Circular, zero-padded, and reflection-padded FlipRates use identical offsets, changing only the boundary treatment.