FACETS: SELECTIVE CROSS-BAND SUPERVISION FOR ROBUST TIME-SERIES REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise in time-series self-supervised learning can corrupt not only observations but also the supervisory relation used to learn from them. When nuisance is predictable, reconstructable, or preserved across views, standard pretext objectives may consistently favor such nuisance patterns, causing them to become readily decodable from the learned representation. We call this failure mode *supervisory contamination*. FACETS treats robustness as a problem of supervision design: it makes selective agreement among frequency components of the same sequence the primary learning signal, down-weights weakly supported components via detached relative-regularity estimates, discounts batch-shared drift through directional rectification, and uses temporal prediction only as an auxiliary local constraint. We characterize when component agreement can improve the deployed original-view encoder by separating observed cross-band dependence from nuisance-mediated dependence and the component-to-deployment gap. A controlled diagnostic provides mechanistic evidence: prediction-only supervision makes injected interference highly decodable, whereas cross-band-only supervision sharply reduces nuisance decodability, at some cost in semantic accuracy. Across five biomedical and industrial datasets, FACETS improves 10%-label linear-evaluation accuracy by 2.07–4.44 percentage points over the strongest dataset-specific SSL baselines, achieves the best weighted F_1 on all five datasets, leads all six ECG transfer directions, and remains strongest under increasing Gaussian corruption. Together, these results show that robust self-supervised learning depends not merely on producing cleaner views, but on controlling which relations are trusted to organize the representation.

1 INTRODUCTION

Self-supervised learning (SSL) has become a central paradigm for learning reusable time-series representations when annotations are limited. Recent work has moved beyond task-specific encoders toward increasingly transferable representations, while emphasizing that temporal structure should be modeled across multiple scales and resolutions (Trirat et al., 2026; Wang et al., 2025). Yet most advances still focus on *how to encode an observation*. A more fundamental question receives less attention: *when the observation contains nuisance, which relation should an encoder be trained to preserve?*

This question matters because self-supervision is defined not only by the data, but by the relation from which the learning signal is constructed. Predictive and reconstructive objectives favor structure that can be recovered from context, whereas contrastive objectives depend on which samples or views are declared similar. Recent methods have therefore begun to treat relation construction itself as a modeling problem, for example by learning semantic similarity between temporal segments or adapting augmentations to construct more informative positive pairs (Xu et al., 2025; Liu et al., 2026). Under acquisition noise, however, relation quality becomes especially consequential. A nuisance pattern that is predictable, reconstructable, or repeatedly preserved across views can receive a consistent learning signal and become readily decodable from the learned representation. In this case, nuisance is not harmful merely because it is present in the input; it is harmful because the pretext relation repeatedly favors it.

We call this failure mode *supervisory contamination*. It exposes a distinction that is easy to overlook: cleaning an observation and improving the supervision constructed from that observation are not the

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.

same problem. Filtering may suppress visible corruption, but it does not determine which remaining statistical dependencies an objective will reinforce. Conversely, nuisance need not dominate a representation simply because it exists in the signal if the learning relation does not systematically reward it. Robust representation learning should therefore ask not only *what view should be presented to the encoder*, but also *what agreement should be trusted as supervision*. As illustrated in Figure 1, this shifts the focus from constructing cleaner views to constructing more reliable relations.

Frequency decomposition provides a natural alternative source of evidence. Task-relevant temporal structure often manifests across multiple scales, while perturbations can be unevenly distributed across frequency components (Liu et al., 2025). Instead of asking two independently transformed views to agree, one can therefore ask which structures recur across frequency components of the *same* observation. Importantly, however, cross-band agreement is not automatically semantic. Replacing augmented views with frequency components only moves the question of supervision; it does not remove it. Three challenges follow.

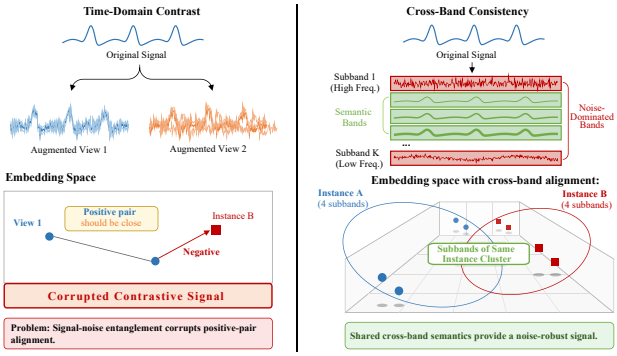


Figure 1: Motivation. Conventional two-view supervision may reinforce nuisance preserved across augmentations in the positive relation. Agreement among frequency components of the same observation offers candidate supervision, but component reliability and nuisance-mediated directions of agreement must be controlled.

C1—Relation reliability: which cross-band relations should be trusted? Frequency components differ substantially in structure and informativeness. A weakly structured or noise-dominated component should not shape the representation as strongly as one supported by stable within-band regularity. Cross-band learning therefore requires selective evidence rather than uniform alignment.

C2—Nuisance-mediated agreement: when does agreement reflect shared corruption rather than useful structure? Coherent baseline drift, periodic interference, or systematic acquisition bias may affect several frequency components simultaneously. High cross-band similarity can therefore arise from nuisance as well as semantics. Controlling only *which components* participate is insufficient; the objective must also control *which directions of agreement* are allowed to define similarity.

C3—Deployment transfer: does component-level supervision improve the representation that is actually used downstream? Frequency components are available during pretraining, whereas deployment consumes the representation of the original observation. Sharing encoder parameters provides a pathway for transfer, but it does not guarantee that useful component-level information reaches the deployed representation. Any account of cross-band learning must therefore distinguish useful agreement from nuisance-mediated dependence and from the component-to-deployment gap.

We address these challenges with FACETS, a framework built around *selective cross-band consistency*. The key departure is conceptual: frequency components are treated not merely as alternative inputs to fuse or signals to reconstruct, but as *candidate supervisory relations*. For **C1**, gated cross-band contrast is combined with detached relative-regularity estimates so that weakly supported components contribute less without allowing the probe objective to reshape the main encoder. For **C2**, directional rectification discounts similarity attributable to a shared mini-batch direction before cross-band relations are selected and scored. For **C3**, we explicitly analyze when component agreement can transfer task-relevant information to the original-view representation, separating observed cross-band dependence from nuisance-mediated dependence and the component-to-deployment gap. Frequency-guided temporal prediction is retained only as an auxiliary local constraint rather than the relation that organizes the global representation space. All decomposition, probe, prediction, and contrastive branches are training-only; deployment retains the original-view encoder alone.

Our contributions are threefold:

- **A supervision-centric view of robustness.** We formulate *supervisory contamination* as a testable failure mode in which nuisance becomes readily decodable not merely because it appears in the observation, but because the pretext relation consistently favors it. This reframes robust time-series SSL from solely constructing cleaner inputs toward constructing reliable supervisory relations.
- **Selective cross-band consistency.** We develop FACETS, which uses within-instance cross-band agreement as the primary source of self-supervision while explicitly controlling relation reliability and nuisance-mediated agreement through gated contrast, detached relative-regularity weighting, and directional rectification. Temporal prediction serves only as an auxiliary constraint, and downstream inference uses the original-view encoder alone.
- **Deployment-oriented analysis and mechanistic evidence.** We characterize when cross-band dependence can benefit the deployed representation by separating nuisance-mediated dependence from the component-to-deployment gap. We evaluate the framework across five biomedical and industrial datasets, six ECG transfer directions, multiple label budgets and corruption settings, and use a controlled nuisance-decoding diagnostic to directly probe the proposed failure mode.

2 Related Work

Self-supervised time-series representation learning. Time-series SSL spans contrastive, predictive/reconstructive, and knowledge-guided approaches (Zhang et al., 2024). Contrastive methods form positives at the instance, timestamp, or contextual level (Meng et al., 2023; Zhou et al., 2024; Eldele et al., 2023; Yue et al., 2022; Tian et al., 2025; Moakher et al., 2026); predictive and generative methods recover future, masked, or corrupted content (Zheng et al., 2021; Dong et al., 2023; Han et al., 2025); and co-training, distillation, and discrete representation learning introduce complementary supervision (Zhang et al., 2023a; Xu et al., 2024; Zheng et al., 2024). These methods mainly improve the encoder or the transformation pipeline. We instead focus on the reliability of the relation used as supervision when nuisance can survive view construction.

Frequency-aware and robust temporal learning. Frequency-aware SSL commonly treats the spectral domain as an additional view or a source of denoised targets. TF-C aligns temporal and frequency views (Zhang et al., 2022); CRT reconstructs masked temporal and spectral content (Zhang et al., 2023b); DECL selects denoisers to form contrastive views (Zhou et al., 2024); and recent methods use graph structure or content-aware spectral encoding for medical signals (Fan et al., 2025; Han et al., 2025). From an architectural perspective, TeCh replaces decentralized self-attention with a centralized core-token aggregation–redistribution module and adapts temporal and channel tokenization to medical time series (Guoqi et al., 2026). TeCh addresses how multichannel information is routed in a supervised classifier, whereas FACETS addresses which relation should organize an unlabeled representation under noise. TeCh is therefore treated as a supervised, label-budget-matched reference rather than an SSL baseline. FACETS otherwise differs from frequency-aware SSL in objective: frequency components are not outputs to reconstruct or alternative inputs to fuse, but candidate relations whose reliability is explicitly controlled.

3 Selective Cross-Band Self-Supervision

3.1 Problem Setup and Deployment Path

Let $x_i \in \mathbb{R}^{C\times T}$ denote the i-th multichannel time series, where C and T are the numbers of channels and time steps, respectively. We write the observation conceptually as $x_i = v_i + s_i$, where v_i contains task-relevant structure and s_i denotes nuisance. Let f_θ denote the shared encoder with parameters θ. The representation retained at deployment is

$$h_i = f_\theta(x_i) \in \mathbb{R}^d, \tag{1}$$

where d is the representation dimension.

During pretraining, we apply a five-level channel-wise db4 discrete wavelet transform (DWT). Let $\mathcal{W}_k$ denote the operator that reconstructs the k-th detail component to the original sequence length.

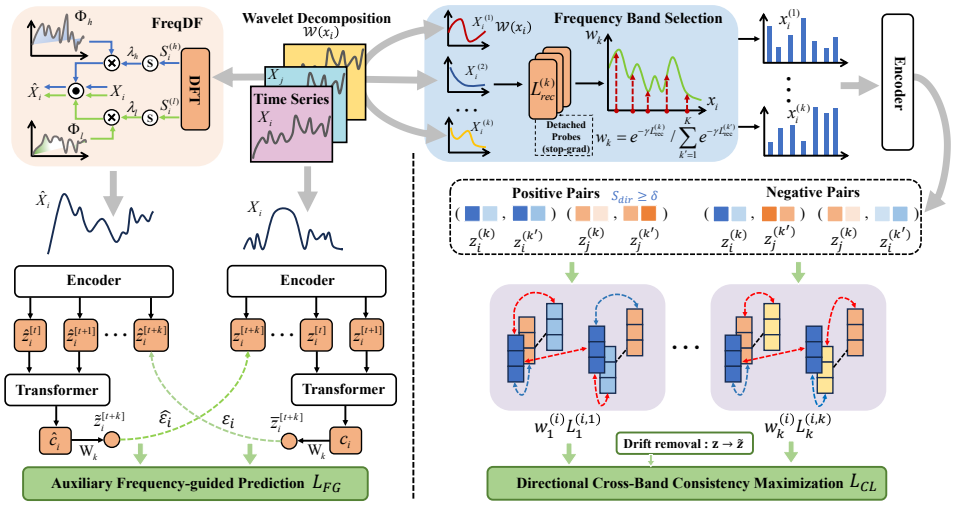


Figure 2: Overview of FACETS. Wavelet detail components define candidate cross-band relations, detached probes control their relative influence, and directional rectification discounts a batch-shared direction before pair construction and scoring. Frequency-guided temporal prediction supplies an auxiliary local constraint. Only the original-view encoder is retained downstream.

We define

$$x_i^{(k)} = \mathcal{W}_k(x_i), \qquad k \in \{1, \ldots, K\}, \tag{2}$$

with $K = 5$, corresponding to the reconstructed detail components $D_1, \ldots, D_5$. The level-five approximation component A_5 is excluded from the cross-band branch. Each detail component is encoded by the same encoder,

$$z_i^{(k)} = f_\theta(x_i^{(k)}) \in \mathbb{R}^d. \tag{3}$$

Thus, h_i denotes the representation of the original observation, whereas $z_i^{(k)}$ denotes the representation of its k-th wavelet detail component. The DWT branches are used only during pretraining: no component aggregation is required at inference. Figure 2 summarizes the pretraining architecture and the original-view deployment path.

3.2 Selective Cross-Band Consistency

Consider a mini-batch of B observations. For an anchor $z_i^{(k)}$, $i \in \{1, \ldots, B\}$ indexes the instance and $k \in \{1, \ldots, K\}$ indexes the wavelet component. Components from the same instance form candidate positives, whereas components from other instances serve as negatives.

Let $s(a, b)$ denote the similarity function used for pair construction and contrastive scoring. Unrectified variants use cosine similarity, whereas the full FACETS model uses the directionally rectified similarity introduced in Eq. (13). Because not every within-instance cross-band relation is reliable, we retain only candidate positives whose similarity exceeds a mini-batch-dependent threshold:

$$\mathcal{P}_i^{(k)} = \left\{ k' \neq k : s\left(z_i^{(k)}, z_i^{(k')}\right) \geq \delta \right\}, \tag{4}$$

where k' indexes a partner component and δ is the median within-instance component similarity in the current mini-batch.

For an anchor with a nonempty positive set $\mathcal{P}_i^{(k)}$, we define

$$q_{i,k}^{+} = \sum_{k' \in \mathcal{P}_i^{(k)}} \exp\Big(s(z_i^{(k)}, z_i^{(k')})/\tau\Big), \tag{5}$$

$$q_{i,k}^{-} = \sum_{j \neq i} \sum_{k'=1}^{K} \exp\Big(s(z_i^{(k)}, z_j^{(k')})/\tau\Big), \tag{6}$$

$$\ell_{i,k} = -\log \frac{q_{i,k}^{+}}{q_{i,k}^{+} + q_{i,k}^{-}}, \tag{7}$$

where j indexes another instance in the mini-batch and $\tau > 0$ is the contrastive temperature. This objective changes the supervisory relation: rather than enforcing agreement between two independently transformed views, it organizes the representation using selectively supported relations among frequency components of the same observation.

We next characterize when such component-level supervision can benefit the representation retained at deployment. Let $Z_i = (z_i^{(1)}, \ldots, z_i^{(K)})$ denote the collection of component representations, and let $I(\cdot;\cdot)$ and $I(\cdot;\cdot \mid \cdot)$ denote mutual information and conditional mutual information, respectively. Define

$$\Xi_i = \frac{1}{K(K-1)} \sum_{k \neq k'} I\Big(z_i^{(k)}; z_i^{(k')} \mid v_i\Big), \qquad \epsilon_{\mathrm{dep},i} = I(v_i; Z_i \mid h_i). \tag{8}$$

Here, Ξ_i captures cross-band dependence that remains after conditioning on the task-relevant structure v_i, and hence may be mediated by nuisance. The term $\epsilon_{\mathrm{dep},i}$ measures task-relevant information available in the component set Z_i but absent from the deployed representation h_i.

Proposition 1 (Cross-band transfer to the deployed representation)**.** *The original-view representation satisfies*

$$I(h_i; v_i) \geq \frac{1}{K(K-1)} \sum_{k \neq k'} I(z_i^{(k)}; z_i^{(k')}) - \Xi_i - \epsilon_{\mathrm{dep},i}. \tag{9}$$

The bound separates observed cross-band dependence from nuisance-mediated dependence and the component-to-deployment gap. It should be interpreted as an information decomposition rather than a task-performance guarantee: shared encoder parameters provide a pathway for transfer, but do not force either residual term to vanish. Appendix A provides the proof.

3.3 Controlling Component and Direction Reliability

Detached relative-regularity weighting. Relation gating determines which cross-band pairs are eligible to contribute, but the participating components need not be equally reliable. A weakly structured or noise-dominated component should not influence the objective as strongly as a component with more stable within-band structure. We therefore associate each component k with a lightweight reconstruction probe consisting of an encoder e_k and decoder d_k, and define

$$L_{i,\mathrm{rec}}^{(k)} = \Big\| x_i^{(k)} - d_k\Big(e_k(x_i^{(k)})\Big) \Big\|_2^2, \qquad w_i^{(k)} = \frac{\exp\Big(-\gamma\, \mathrm{sg}[L_{i,\mathrm{rec}}^{(k)}]\Big)}{\sum_{r=1}^{K} \exp\Big(-\gamma\, \mathrm{sg}[L_{i,\mathrm{rec}}^{(r)}]\Big)}. \tag{10}$$

Here, r is a dummy component index, $\gamma > 0$ controls the concentration of the relative weights, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The reconstruction error therefore affects the main objective only through a detached relative-regularity score and does not provide gradients to the shared encoder.

We do not interpret the probe loss as a calibrated estimate of semantics or signal-to-noise ratio (SNR). Under the explicit condition that expected probe error decreases with component SNR, the induced weighting yields a higher expected SNR than uniform weighting; Appendix A states the corresponding result and assumptions. The resulting cross-band objective is

$$\mathcal{L}_{\mathrm{CB}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} w_i^{(k)} \ell_{i,k}. \tag{11}$$

Directional rectification. Component-level reliability does not by itself prevent multiple bands from agreeing through the same nuisance factor. Coherent acquisition drift, periodic interference, or systematic sensor bias can induce a common direction across many component embeddings. We therefore estimate a mini-batch-shared direction and attenuate its contribution:

$$\bar{z} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K} z_i^{(k)}, \qquad u = \frac{\bar{z}}{\|\bar{z}\|_2 + \epsilon}, \qquad \widetilde{z}_i^{(k)} = z_i^{(k)} - \eta\big(z_i^{(k)\top} u\big)u. \tag{12}$$

Here, $\bar{z}$ is the mean component embedding in the mini-batch, u is its normalized direction, $\epsilon = 10^{-8}$ is a numerical stabilizer for near-zero batch-mean norms, and $\eta \geq 0$ controls the strength of projection removal.

For embeddings a and b, let $\widetilde{a}$ and $\widetilde{b}$ denote their rectified versions under the same batch direction u. We define

$$s_{\mathrm{dir}}(a,b) = \cos(\widetilde{a},\widetilde{b}) - \beta\,\mathrm{ReLU}\Big(\rho - \cos(\widetilde{a},\widetilde{b})\Big), \tag{13}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity, $\mathrm{ReLU}(t) = \max(0,t)$, ρ is the directional similarity margin, and $\beta \geq 0$ controls the penalty strength. In the full FACETS objective, we set $s = s_{\mathrm{dir}}$ in both the positive-set construction in Eq. (4) and the contrastive loss in Eq. (7).

Directional rectification affects only pair construction and contrastive scoring; the representation used downstream remains the original-view encoder output. The single-direction approximation is intended for coherent, approximately low-rank nuisance and is not assumed to remove arbitrary instance-specific corruption.

3.4 Auxiliary Local Prediction

Selective cross-band consistency determines the global representation geometry, but does not explicitly preserve short-range temporal predictability. We therefore introduce a frequency-guided view $\widehat{x}_i$ and use temporal prediction as an auxiliary local constraint.

Let $X_i = \mathfrak{F}(x_i)$ denote the discrete Fourier transform of x_i, $A_i[\ell] = |X_i[\ell]|$ the spectral amplitude at frequency-bin index $\ell \in \{1,\dots,L\}$, and $\bar{A}_i = L^{-1}\sum_{\ell=1}^{L} A_i[\ell]$ the instance-wise mean amplitude. Using the logistic sigmoid $\sigma(\cdot)$, we construct complementary soft masks

$$\Phi_l[\ell] = \sigma\big(\kappa(\bar{A}_i - A_i[\ell])\big), \qquad \Phi_h[\ell] = 1 - \Phi_l[\ell], \tag{14}$$

where Φ_l and Φ_h denote the complementary frequency-domain masks and $\kappa > 0$ controls their transition sharpness. We use $\kappa = 1.0$ in all reported experiments.

A lightweight gating network g takes the unfiltered time-domain sequence x_i as input and produces two bounded scalar gates,

$$(\lambda_l, \lambda_h) = g(x_i), \qquad \lambda_l, \lambda_h \in (0,1), \tag{15}$$

where λ_l and λ_h modulate the contributions of the two masks. The resulting frequency-guided view is

$$\widehat{x}_i = \mathfrak{F}^{-1}[(\lambda_l\Phi_l + \lambda_h\Phi_h) \odot X_i], \tag{16}$$

where $\mathfrak{F}^{-1}$ is the inverse Fourier transform and $\odot$ denotes element-wise multiplication.

A Transformer predictor estimates the next encoder state for both the original sequence x_i and the frequency-guided view $\widehat{x}_i$. The corresponding losses are $\mathcal{L}_{\mathrm{pred}}^{\mathrm{orig}}$ and $\mathcal{L}_{\mathrm{pred}}^{\mathrm{filt}}$, giving

$$\mathcal{L}_{\mathrm{pred}} = \mathcal{L}_{\mathrm{pred}}^{\mathrm{orig}} + \mathcal{L}_{\mathrm{pred}}^{\mathrm{filt}}. \tag{17}$$

We use a one-step prediction horizon ($H = 1$); Appendix B.3 specifies the temporal targets and the prediction loss.

The complete pretraining objective is $\mathcal{L} = \mathcal{L}_{\mathrm{CB}} + \alpha\mathcal{L}_{\mathrm{pred}}$, where $\alpha \geq 0$ controls the contribution of temporal prediction. Selective cross-band consistency determines instance separation and alignment in the global representation space, whereas temporal prediction provides a local auxiliary constraint.

Table 1: Ten-percent-label evaluation over ten runs (%). TeCh† is trained end to end with the same label budget; the remaining methods use frozen linear evaluation.

Method	SleepEDF		FaultDiagnosis		CPSC18		PTB-XL		Georgia	
	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1
TF-C (NeurIPS'22)	62.07±.61	61.24±.54	76.82±.72	74.69±.71	38.46±.50	35.91±.64	51.21±.62	40.93±.37	48.97±.34	33.43±.38
TS2Vec (AAAI'22)	63.68±.56	62.75±.42	77.63±.49	77.15±.44	40.56±.59	37.47±.47	52.71±.59	44.37±.21	51.61±.43	36.42±.32
CRT (TNNLS'23)	62.25±.48	61.03±.38	75.82±.58	76.07±.32	39.25±.65	36.34±.34	52.32±.58	42.31±.35	49.92±.27	34.51±.15
SimMTM (NeurIPS'23)	63.84±.57	62.95±.43	78.39±.43	77.52±.39	40.74±.56	37.72±.32	52.63±.43	44.75±.39	51.34±.32	35.63±.24
TS-CoT (CIKM'23)	64.62±.59	63.98±.28	76.32±.44	75.96±.32	39.82±.78	37.43±.39	52.92±.54	43.14±.35	50.97±.31	34.85±.21
CA-TCC (TPAMI'23)	63.91±.48	63.37±.33	78.05±.43	77.39±.26	40.79±.54	38.29±.49	52.27±.32	44.91±.22	52.07±.27	36.26±.34
RCD-KD (NeurIPS'24)	67.02±.26	66.01±.18	80.83±.13	80.92±.14	42.92±.22	37.69±.34	52.82±.25	45.25±.24	54.66±.32	44.28±.24
DECL (IJCAI'24)	71.74±.43	70.96±.25	82.78±.38	82.17±.23	45.01±.32	41.65±.39	57.41±.36	49.32±.27	55.37±.26	40.35±.21
DU-IN (NeurIPS'24)	71.56±.39	70.88±.38	85.38±.11	83.31±.14	47.99±.26	45.19±.31	55.66±.36	53.34±.35	53.24±.24	44.37±.26
MedGNN (WWW'25)	68.70±.31	67.82±.28	80.67±.32	80.48±.35	48.27±.34	48.24±.32	58.13±.33	55.67±.32	50.46±.25	43.31±.24
CBD (AAAI'25)	71.24±.54	69.61±.39	79.41±.32	77.49±.44	48.37±.18	46.15±.38	54.81±.26	49.96±.28	56.14±.18	40.69±.52
TeCh† (ICLR'26)	72.10±.40	69.42±.45	79.81±2.09	79.72±2.06	46.18±.50	40.97±.62	51.55±.66	40.09±1.74	55.65±.15	40.15±.20
FACETS	**74.82±.34**	**74.39±.36**	**89.82±.27**	**88.62±.28**	**51.50±.26**	**50.57±.27**	**61.27±.26**	**56.06±.27**	**58.21±.33**	**49.89±.35**

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

We evaluate on five biomedical and industrial time-series datasets using the processed data and serialized partitions released with DECL (Zhou et al., 2024). The released preprocessing uses a 40/20/40 sample-level split rather than an entity-disjoint split. Self-supervised pretraining uses no labels and runs for 100 epochs. Linear evaluation freezes the encoder and trains a classifier with 10% or 30% of the training labels; fine-tuning updates the full network over label budgets from 1% to 100%; and cross-domain transfer uses 10% target labels. Accuracy and weighted F_1 are reported over ten runs. Dataset statistics, baseline protocols, architecture, optimization settings, and implementation details are provided in Appendix B.

4.2 MAIN REPRESENTATION RESULTS

Table 1 reports mean±standard deviation over ten runs with 10% of the training labels. Among the SSL methods, FACETS ranks first on every dataset and metric. Compared with the strongest external SSL baseline for each dataset, the accuracy gains are 3.08 points on SleepEDF, 4.44 on FaultDiagnosis, 3.13 on CPSC18, 3.14 on PTB-XL, and 2.07 on Georgia. The weighted-F_1 margins are especially large on FaultDiagnosis (+5.31) and Georgia (+5.52), so the improvement is not explained by majority-class accuracy alone.

The TeCh comparison asks a stricter but protocol-qualified question: can an SSL representation followed by a frozen linear classifier compete with a recent supervised medical-time-series architecture trained end to end under the same label budget? FACETS is higher than TeCh on all ten dataset–metric pairs. Accuracy gains range from +2.56 points on Georgia to +10.01 on FaultDiagnosis, while weighted-F_1 gains range from +4.97 on SleepEDF to +15.97 on PTB-XL. The larger weighted-F_1 gaps on CPSC18, PTB-XL, and Georgia indicate that the advantage is not limited to majority classes. This result strengthens the representation-quality claim, but it does not turn TeCh into an SSL baseline: the comparison is evidence about label efficiency, not an identical training objective. With 30% labels, FACETS remains best on nine of ten dataset–metric pairs among the SSL baselines; the only exception is PTB-XL weighted F_1, where it trails MedGNN by 0.25 points (Appendix C.1).

4.3 ABLATION AND MECHANISM ANALYSIS

The variants in Table 2 follow the definitions used in the implementation. `only FG` uses prediction on the filtered and original views; `only FG⁻` removes original-view prediction; `only CBCL` uses uniform cross-band contrast; `FS` adds detached relative-regularity weights; `FACETS*` adds the two-view prediction block but retains cosine similarity; and FACETS adds directional rectification.

The progression separates the roles of the components. `only FG⁻` is better than `only FG` on every dataset, showing that adding a prediction target can be harmful when prediction is the complete objective. `only CBCL` then improves over both prediction-only variants on all five datasets. Relative-regularity weighting adds 0.99–3.51 accuracy points over uniform CBCL. Adding prediction to the

Table 2: Ablation results under 10%-label linear evaluation. All entries are mean±standard deviation over ten runs (%).

Variant	SleepEDF		FaultDiagnosis		CPSC18		PTB-XL		Georgia	
	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1	Acc.	WF_1
only FG	70.96±.50	69.15±.47	84.44±.47	83.53±.46	42.35±.39	39.64±.40	52.72±.35	46.31±.32	52.77±.39	43.82±.36
only FG^{-}	71.33±.41	70.64±.43	85.02±.36	84.85±.33	46.25±.37	43.73±.36	54.39±.31	50.33±.30	53.63±.32	44.18±.34
only CBCL	71.68±.27	71.19±.32	85.38±.38	85.05±.39	47.23±.36	45.72±.38	55.45±.37	52.37±.35	54.29±.22	44.82±.26
FS	72.67±.29	71.78±.22	87.07±.26	86.24±.27	50.74±.37	48.47±.37	57.01±.36	53.11±.30	55.41±.24	45.96±.26
FACETS*	73.34±.31	72.76±.30	88.71±.38	87.35±.36	51.03±.29	49.79±.24	60.03±.26	55.60±.34	57.29±.36	48.00±.35
FACETS	**74.82±.34**	**74.39±.36**	**89.82±.27**	**88.62±.28**	**51.50±.26**	**50.57±.27**	**61.27±.26**	**56.06±.27**	**58.21±.33**	**49.89±.35**

weighted cross-band model is beneficial, particularly on PTB-XL (+3.02) and Georgia (+1.88), even though prediction alone is weaker. Directional rectification gives a final gain on every dataset. Thus, cross-band contrast supplies the main instance-level relation, weighting controls the contribution of each band, prediction preserves local dynamics, and rectification changes how pairs are selected and scored. Because all downstream evaluations retain only the original-view encoder, these gains also show that component-level pretraining supervision benefits the representation used at deployment, without requiring component aggregation at inference.

Supervisory-contamination diagnostic. We conduct a controlled mechanism test on FaultDiagnosis by injecting four class-independent nuisance groups during self-supervised pretraining. After freezing the encoder, separate linear probes predict semantic class and nuisance identity (chance: 25%). Prediction-only `only FG` makes nuisance highly decodable (96.6%), whereas prediction-free `FS` reduces it to 27.6% but loses substantial semantic information. FACETS recovers semantic decodability to 86.4% while nuisance remains lower than under `only FG` (71.8%). This single-seed diagnostic therefore supports a semantic–nuisance trade-off, rather than nuisance invariance; full protocol details are given in Appendix D.2.

4.4 Label Efficiency and Cross-Domain Generalization

Figure 3 reports fine-tuning from 1% to 100% labels against representative baselines. FACETS is highest at every displayed label ratio on SleepEDF, FaultDiagnosis, and PTB-XL. The largest gaps occur at 1%–5%, but the advantage remains at 100%, indicating that the initialization is useful beyond the most label-scarce regime. Representative baselines are shown for readability; an extended comparison is provided in Appendix C.2.

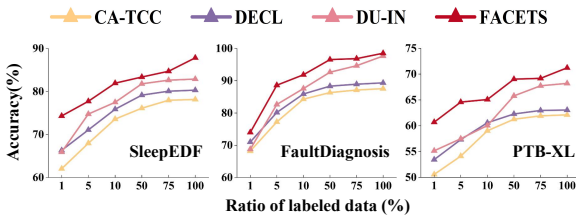


Figure 3: In-domain fine-tuning accuracy from 1% to 100% labels on SleepEDF, FaultDiagnosis, and PTB-XL.

For cross-domain transfer, we pretrain on one ECG cohort and fine-tune on another using 10% target labels. FACETS achieves the highest accuracy in all six directions in Table 3. Relative to the strongest alternative in each direction, the gains are +7.77 (P→C), +6.04 (G→C), +0.53 (C→P), +0.96 (G→P), +0.88 (C→G), and +0.24 (P→G). The mean advantage over supervised TeCh is 3.08 points across the six directions (58.67 versus 55.59), but the directional margins vary from 0.24 to 7.77 points. The larger gains when CPSC18 is the target, together with the narrow margins for transfer to Georgia, show that cross-band pretraining improves transfer without making it source–target invariant.

Table 3: Cross-domain accuracy (%) with 10% target labels. C: CPSC18, P: PTB-XL, G: Georgia.

Method	P→C	G→C	C→P	G→P	C→G	P→G
DECL	46.23	44.25	60.21	59.03	56.41	55.69
RCD-KD	46.42	43.38	56.50	57.49	56.06	54.05
DU-IN	46.13	45.06	61.48	60.43	55.83	54.19
CBD	44.78	43.35	57.33	58.37	53.49	52.42
MedGNN	45.23	45.27	60.64	59.58	55.89	54.88
TeCh†	48.50	52.23	59.97	60.83	55.87	56.16
FACETS	**56.27**	**58.27**	**62.01**	**61.79**	**57.29**	**56.40**

4.5 Robustness

Figure 4 evaluates additive zero-mean Gaussian noise with standard deviation from 0 to 0.9 under the 10%-label linear-evaluation protocol. The encoder is frozen, and only a linear classifier is trained on 10% of the clean training labels. Gaussian corruption is then applied only to the raw held-out test inputs, with no corruption-time updating or retraining. Every method degrades as corruption grows, but FACETS remains the highest curve throughout the evaluated range. The Georgia waveform case study and complete baseline-wander and impulse-noise results are reported in Appendix C.3. These results therefore characterize test-time corruption robustness rather than robustness to noisy pretraining or noisy probe fitting.

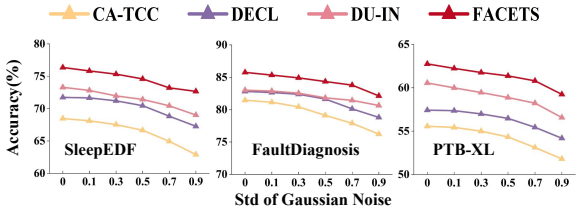


Figure 4: 10%-label linear-evaluation accuracy under additive zero-mean Gaussian noise. The frozen encoder and linear classifier are trained on clean data; corruption is applied only to held-out test inputs.

The broader stress tests identify a failure boundary: smooth baseline wander is comparatively benign, whereas strong impulse noise can overwrite short discriminative events across several bands simultaneously. Sensitivity results are in Appendix C.4. Computational profiling is reported in Appendix C.5. The additional pretraining cost arises from multi-subband processing and auxiliary prediction and probe modules, all of which are removed at deployment so that inference uses only the shared encoder.

5 Discussion

The results support a supervision-centric view of robustness. Frequency decomposition is useful not merely because it produces cleaner inputs, but because it provides alternative supervisory relations whose reliability can be controlled. The nuisance-decoding diagnostic illustrates this distinction: prediction-only supervision makes injected nuisance highly decodable, whereas prediction-free cross-band supervision reduces nuisance decodability at a cost in semantic information. FACETS lies between these extremes, consistent with using cross-band consistency as the primary relation and prediction as an auxiliary constraint. Supervisory contamination therefore concerns whether the learning objective repeatedly makes nuisance useful to encode, rather than whether nuisance is simply present in the observation.

Several limitations remain. Relative regularity is only a proxy for component reliability, and reconstructable nuisance may violate the assumed probe–SNR relation. Directional rectification models shared corruption with a single batch direction and may be insufficient for heterogeneous or class-aligned nuisance; a fixed db4 decomposition may likewise be suboptimal when spectral structure varies substantially across samples or channels. Shared encoder parameters also do not guarantee complete transfer from component-level supervision to the deployed representation. Finally, the nuisance-decoding diagnostic is single-seed, and the DECL benchmark uses sample-level rather than entity-disjoint splits. Appendix E discusses these failure modes in more detail.

6 Conclusion

We presented FACETS, a time-series self-supervised learning framework that treats robustness as a problem of controlling which cross-band relations are trusted as supervision. Across in-domain evaluation, cross-domain transfer, and corruption tests, the results support selective cross-band consistency as the primary supervisory relation with temporal prediction as an auxiliary constraint. The nuisance-decoding diagnostic further shows that robustness depends not merely on whether nuisance is present, but on whether the learning objective repeatedly makes it useful to encode.

AI USE STATEMENT

Generative AI tools assisted with language editing and exposition. They were not used to generate datasets, implement the method, or produce experimental results. The authors verified the resulting text against the implementation, experimental records, and cited literature, and take responsibility for all claims and content.

ETHICS STATEMENT

This work uses previously released biomedical and industrial time-series benchmarks and does not introduce new human-subject data collection. The learned representations are research prototypes and should not be interpreted as clinical diagnostic systems or safety-certified industrial monitors. Potential deployment requires entity-disjoint validation, subgroup analysis where applicable, and evaluation under acquisition conditions representative of the intended use.

REPRODUCIBILITY STATEMENT

Section 4 reports the benchmark protocol, comparison setting, repeated-run protocol, main representation results, ablations, transfer, and Gaussian-noise evaluation. Appendix B provides dataset statistics and implementation details; Appendix C provides the 30%-label results, extended fine-tuning comparisons, corruption tests, sensitivity analysis, and computational profiling; Appendix D provides exact ablation definitions and the nuisance-decoding diagnostic; Appendix A gives complete proofs and assumptions; and Appendix E discusses interpretation and failure modes.

REFERENCES

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Self-supervised contrastive representation learning for semi-supervised time-series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15604–15618, 2023.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *Proceedings of the ACM on Web Conference 2025*, pp. 5054–5064, 2025.

Yu Guoqi, Juncheng Wang, Chen Yang, Jing Qin, Angelica Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108935–108955, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b12604df7a807fb10eecf58833535a9d-Paper-Conference.pdf.

Yudong Han, Haocong Wang, Yupeng Hu, Yongshun Gong, Xuemeng Song, and Weili Guan. Content-aware balanced spectrum encoding in masked modeling for time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17059–17067, 2025.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. 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.

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM society European conference*, volume 3, 2016.

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.

Shikang Liu, Xiren Zhou, and Huanhuan Chen. Multiscale temporal dynamic learning for time series classification. *IEEE Transactions on Knowledge and Data Engineering*, 37(6):3543–3555, 2025.

Yusen Liu, Zhichen Lai, Hua Lu, Xu Cheng, Tianqing Zhu, Xiufeng Liu, and Huan Huo. Saurl-ts: A self-adaptive framework for unsupervised time series representation learning. *Pattern Recognition*, 175:113129, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113129. URL https://www.sciencedirect.com/science/article/pii/S0031320326000920.

Qianwen Meng, Hangwei Qian, Yong Liu, Lizhen Cui, Yonghui Xu, and Zhiqi Shen. Mhccl: masked hierarchical cluster-wise contrastive learning for multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9153–9161, 2023.

Yessin Moakher, Youssef Attia El Hili, and Vasilii Feofanov. Utica: Multi-objective self-distllation foundation model pretraining for time series classification. *arXiv preprint arXiv:2603.01348*, 2026.

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.

Tian Tian, Chunyan Miao, and Hangwei Qian. Frera: a frequency-refined augmentation for contrastive learning on time series classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2835–2846, 2025.

Patara Trirat, Yooju Shin, Junhyeok Kang, Youngeun Nam, Jihye Na, Minyoung Bae, Joeun Kim, Byunghyun Kim, and Jae-Gil Lee. Universal time-series representation learning: A survey. *ACM Comput. Surv.*, 58(12), June 2026. ISSN 0360-0300.

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.

Shiyu Wang, Jiawei LI, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16980–17016, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2b187165e28fdfdc0ffb34d1bfff2b0c-Paper-Conference.pdf.

Maxwell Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik Raghuram, James Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53188–53206, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/83eb339ed42297658fa24b5cec939285-Paper-Conference.pdf.

Qing Xu, Min Wu, Xiaoli Li, Kezhi Mao, and Zhenghua Chen. Reinforced cross-domain knowledge distillation on time series data. *Advances in Neural Information Processing Systems*, 37:47217–47241, 2024.

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.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE transactions on pattern analysis and machine intelligence*, 46(10):6775–6794, 2024.

Weiqi Zhang, Jianfeng Zhang, Jia Li, and Fugee Tsung. A co-training approach for noisy time series learning. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 3308–3318, 2023a.

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16129–16138, 2023b.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-in: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. *Advances in Neural Information Processing Systems*, 37: 79996–80033, 2024.

Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12220–12233, 2021.

Shuang Zhou, Daochen Zha, Xiao Shen, Xiao Huang, Rui Zhang, and Korris Chung. Denoising-aware contrastive learning for noisy time series. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5644–5652. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

A Theoretical Details

A.1 Proof of Cross-Band Transfer

For a fixed ordered pair $k \neq k'$, the chain rule gives

$$I(z_i^{(k)}; z_i^{(k')}) = I(z_i^{(k)}; v_i) + I(z_i^{(k)}; z_i^{(k')} \mid v_i) - I(z_i^{(k)}; v_i \mid z_i^{(k')}) \tag{18}$$

$$\leq I(z_i^{(k)}; v_i) + I(z_i^{(k)}; z_i^{(k')} \mid v_i). \tag{19}$$

Because $z_i^{(k)}$ is a coordinate of Z_i, data processing gives $I(v_i; Z_i) \geq I(v_i; z_i^{(k)})$. Combining this inequality with Eq. (19) and averaging over the $K(K-1)$ ordered pairs yields

$$I(v_i; Z_i) \geq \frac{1}{K(K-1)} \sum_{k \neq k'} I(z_i^{(k)}; z_i^{(k')}) - \Xi_i. \tag{20}$$

Relating the component set to the deployed representation,

$$I(v_i; h_i) = I(v_i; Z_i, h_i) - I(v_i; Z_i \mid h_i) \tag{21}$$

$$= I(v_i; Z_i) + I(v_i; h_i \mid Z_i) - \epsilon_{\mathrm{dep},i} \tag{22}$$

$$\geq I(v_i; Z_i) - \epsilon_{\mathrm{dep},i}. \tag{23}$$

Substitution proves Proposition 1.

A.2 Conditional Expected-SNR Result

Let $r_k = \mathrm{SNR}_k$ and $\bar{L}_k = \mathbb{E}[L_{\mathrm{rec}}^{(k)} \mid r_k]$. The next result describes a population tendency rather than claiming that realized reconstruction loss universally estimates SNR.

Proposition 2 (Conditional expected-SNR advantage)**.** *Suppose $\bar{L}_k = g(r_k)$ for a strictly decreasing function g, and define*

$$\bar{w}_k = \frac{\exp[-\gamma g(r_k)]}{\sum_{q=1}^{K} \exp[-\gamma g(r_q)]}, \qquad \gamma > 0. \tag{24}$$

If the r_k are not all equal, then

$$\sum_{k=1}^{K} \bar{w}_k r_k > \frac{1}{K} \sum_{k=1}^{K} r_k. \tag{25}$$

Proof. Because g is strictly decreasing, $\bar{w}_k$ is strictly increasing in r_k. Reindex so that $r_1 \le \cdots \le r_K$, which also gives $\bar{w}_1 \le \cdots \le \bar{w}_K$. Since $\sum_k \bar{w}_k = 1$,

$$\sum_{k=1}^{K} \bar{w}_k r_k - \frac{1}{K}\sum_{k=1}^{K} r_k = \frac{1}{K} \sum_{1 \le p < q \le K} (\bar{w}_q - \bar{w}_p)(r_q - r_p). \tag{26}$$

Every summand is nonnegative, and at least one is positive when the SNRs are not all equal. The desired strict inequality follows. □

The premise can fail for regular nuisance such as smooth drift or narrowband periodic interference. This is why Eq. (10) is described as relative-regularity weighting rather than SNR estimation.

A.3 Probe–Encoder Gradient Decoupling

Lemma 1. *Let θ denote the main-encoder parameters. With $w_k = \mathrm{softmax}_k[-\gamma\,\mathrm{sg}(L_{\mathrm{rec}}^{(k)})]$, $\partial w_k/\partial\theta = 0$ through the probe-reconstruction pathway.*

Proof. The stop-gradient operator is the identity in the forward pass and has zero derivative in the main optimization graph. Therefore the probe loss cannot update the encoder through the weight computation. □

Detachment removes the direct reconstruction gradient but not the effect of the weights. For $\mathcal{L}_{\mathrm{CB}} = \sum_k w_k \mathcal{L}_k$, the encoder gradient is $\nabla_\theta \mathcal{L}_{\mathrm{CB}} = \sum_k w_k \nabla_\theta \mathcal{L}_k$; an inaccurate ranking can still over- or under-weight a component. Thus, stop-gradient separates the two optimization paths without guaranteeing that the band ranking is correct.

B Experimental Details

B.1 Datasets and Split Scope

We use the processed datasets and serialized partitions released with DECL (Zhou et al., 2024), without additional repartitioning. SleepEDF contains full-night EEG recordings from 82 individuals and is represented as single-channel, 100-Hz, 30-s epochs from five sleep stages (Kemp et al., 2000). FaultDiagnosis contains vibration measurements under healthy, inner-race-fault, and outer-race-fault conditions (Lessmeier et al., 2016). CPSC18 uses 12-lead ECG records resampled to 200 Hz and standardized to 10 s, retaining nine single-label classes (Liu et al., 2018). PTB-XL uses the corresponding 200-Hz, 10-s representation and five diagnostic superclasses (Wagner et al., 2020). Georgia follows the same sampling rate and duration; after excluding classes with fewer than 100 samples, six classes remain (Perez Alday et al., 2020). Dataset statistics are summarized in Table 4.

The released preprocessing uses a 40/20/40 train/validation/test split over processed sample indices and does not partition by subject, patient, recording, or acquisition run. Category labels remain attached to the serialized samples because the same interface supports self-supervised pretraining and downstream evaluation, but no SSL loss, positive-pair gate, probe weight, or checkpoint decision reads these labels. We use the final pretraining epoch rather than selecting a checkpoint using validation labels.

Under the primary 10%-label protocol, the downstream classifier uses 10% of the training partition, corresponding to 4% of all processed samples. Partitions and downstream label subsets are fixed across the ten repetitions, while training randomness is resampled. Consequently, the reported standard deviations quantify optimization and initialization variability under fixed data splits and label subsets, rather than variation over alternative partitions.

B.2 Evaluation Protocols and Baselines

Self-supervised pretraining runs for 100 epochs and uses the final encoder checkpoint. For linear evaluation, the pretrained encoder is frozen and a linear classifier is trained using either 10% or 30% of the labels in the training partition. Fine-tuning instead updates the full network using 1%, 5%,

Table 4: Processed datasets from the DECL benchmark. The test split contains the same number of samples as the training split.

Dataset	Train	Valid	Length	Channels	Classes
SleepEDF	16,923	8,462	3,000	1	5
FaultDiagnosis	5,456	2,728	5,120	1	3
CPSC18	13,754	6,877	2,000	12	9
PTB-XL	6,509	3,254	2,000	12	5
Georgia	6,334	3,167	2,000	12	6

10%, 50%, 75%, or 100% of the training labels. Cross-domain transfer first pretrains on one ECG cohort and then fine-tunes on another using 10% of the target training labels. We report accuracy and weighted F_1 over ten runs with fixed data partitions and label subsets and resampled training randomness.

We compare with TF-C, TS2Vec, CRT, SimMTM, TS-CoT, CA-TCC, RCD-KD, DECL, DU-IN, MedGNN, and CBD (Zhang et al., 2022; Yue et al., 2022; Zhang et al., 2023b; Dong et al., 2023; Zhang et al., 2023a; Eldele et al., 2023; Xu et al., 2024; Zhou et al., 2024; Zheng et al., 2024; Fan et al., 2025; Han et al., 2025). TeCh (Guoqi et al., 2026) is included as a supervised, label-budget-matched reference. Unlike the SSL baselines, which use self-supervised pretraining followed by frozen linear evaluation in the main 10%- and 30%-label comparisons, TeCh is trained end to end with the corresponding labeled fraction. We therefore report it separately as a supervised reference rather than treating it as an SSL baseline. All methods use the same released data partition and downstream label budget within each comparison.

B.3 Architecture and Optimization

Shared encoder and predictor. The shared encoder consists of three 1-D convolutional blocks with 64, 128, and 256 feature maps. Each block uses kernel size 8, batch normalization, ReLU, max pooling with kernel size 2, and dropout 0.2, followed by a projection to a 128-dimensional representation. The temporal prediction module contains four Transformer blocks with four attention heads per block and a feed-forward hidden dimension of 100.

Wavelet components. A level-five channel-wise db4 decomposition produces one approximation coefficient set cA_5 and five detail coefficient sets $cD_5, \ldots, cD_1$. Each detail component is reconstructed separately to the original sequence length, yielding $D_1, \ldots, D_5$, which serve as the $K = 5$ candidate subbands. The approximation component A_5 is excluded from the cross-band branch and is not processed by the shared encoder.

Relative-regularity probes. Each subband probe is a two-layer 1-D convolutional encoder–decoder with a linear bottleneck. Probe reconstruction errors are updated once per epoch and enter the main cross-band objective only through the detached relative-regularity weights defined in Eq. (10).

Frequency-guided view. For each sequence, FreqDF constructs the complementary Fourier-domain masks defined in Eq. (14). The lightweight gating network takes the unfiltered time-domain sequence x_i as input and returns the bounded scalars $(\lambda_l, \lambda_h) = g(x_i)$. These instance-dependent gates modulate the complementary masks before the inverse Fourier transform in Eq. (16). We use $\kappa = 1.0$ in all reported experiments.

Temporal prediction objective. Let $h_i^{[t]}$ and $\widehat{h}_i^{[t]}$ denote temporally indexed encoder states of the original and frequency-guided views, respectively. Conditioned on the states through time t, the Transformer produces predictions $\widetilde{h}_i^{[t+j]}$ and $\widetilde{\widehat{h}}_i^{[t+j]}$ over a horizon H. The per-instance losses are

$$\mathcal{E}_i^{\text{orig}} = \frac{1}{H}\sum_{j=1}^{H}\left\| h_i^{[t+j]} - \widetilde{h}_i^{[t+j]} \right\|_2^2, \qquad \mathcal{E}_i^{\text{filt}} = \frac{1}{H}\sum_{j=1}^{H}\left\| \widehat{h}_i^{[t+j]} - \widetilde{\widehat{h}}_i^{[t+j]} \right\|_2^2. \tag{27}$$

Their batch-averaged losses are

$$\mathcal{L}_{\text{pred}}^{\text{orig}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{E}_i^{\text{orig}}, \qquad \mathcal{L}_{\text{pred}}^{\text{filt}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{E}_i^{\text{filt}}. \tag{28}$$

The two views are predicted independently. We use $H = 1$ in all reported experiments. Accordingly, `only FG` optimizes both prediction terms, `only FG`$^-$ retains only $\mathcal{L}_{\text{pred}}^{\text{filt}}$, and `FACETS`* and FACETS combine both prediction terms with their cross-band objectives.

Optimization. Unless varied in the sensitivity analysis, we use $\tau = 0.2$, $\gamma = 0.1$, $\alpha = 0.5$, $\eta = 1.0$, $\rho = 0.5$, and $\beta = 0.5$. Directional normalization uses $\epsilon = 10^{-8}$. Training uses batch size 128 and Adam with initial learning rate 10^{-4} and weight decay 5×10^{-4} on one NVIDIA GeForce RTX 4090 GPU. A single run-level seed initializes NumPy and PyTorch randomness, including parameter initialization and data-loader order.

For the computational profile reported in Table 7, parameter counts include only trainable modules participating in the self-supervised pretraining optimizer. Modules excluded from that optimizer are not included in the reported pretraining parameter count.

C Supplementary Experimental Results

C.1 Thirty-Percent-Label Evaluation

Table 5 reports the 30%-label linear-evaluation results over ten runs, following the same frozen-encoder protocol as the primary 10%-label evaluation.

Table 5: Thirty-percent-label evaluation over ten runs (%). Venue and publication year are shown in parentheses. TeCh† is an end-to-end supervised reference; the remaining methods use frozen linear evaluation.

Accuracy

Method	SleepEDF	FaultDiag.	CPSC18	PTB-XL	Georgia
TF-C (NeurIPS'22)	68.23±.67	79.03±.76	41.34±.62	52.31±.69	51.82±.46
TS2Vec (AAAI'22)	68.91±.68	80.58±.62	42.36±.65	53.52±.46	53.18±.52
CRT (TNNLS'23)	67.93±.54	78.92±.74	41.05±.74	53.48±.54	51.83±.34
SimMTM (NeurIPS'23)	69.89±.63	82.10±.57	42.96±.62	53.75±.45	53.56±.33
TS-CoT (CIKM'23)	68.49±.65	80.61±.54	41.28±.79	53.04±.51	52.53±.41
CA-TCC (TPAMI'23)	70.56±.49	82.15±.58	42.64±.57	53.98±.37	53.49±.30
RCD-KD (NeurIPS'24)	73.61±.35	85.02±.48	52.62±.32	58.89±.33	53.83±.28
DECL (IJCAI'24)	77.21±.47	87.94±.39	48.03±.36	59.73±.45	57.82±.37
DU-IN (NeurIPS'24)	74.35±.19	86.67±.38	55.02±.61	61.40±.25	56.63±.27
MedGNN (WWW'25)	73.87±.36	81.54±.35	51.11±.47	60.76±.31	52.38±.27
CBD (AAAI'25)	74.72±.44	84.32±.41	49.76±.34	60.42±.32	52.01±.38
TeCh† (ICLR'26)	76.04±.33	83.74±1.48	50.77±.38	54.70±.55	55.79±.12
FACETS	**78.78±.30**	**92.21±.26**	**58.11±.29**	**63.01±.30**	**59.20±.37**

Weighted F_1

Method	SleepEDF	FaultDiag.	CPSC18	PTB-XL	Georgia
TF-C (NeurIPS'22)	67.85±.56	78.41±.61	37.96±.73	43.85±.43	36.73±.49
TS2Vec (AAAI'22)	68.47±.37	80.22±.49	39.72±.47	45.12±.32	39.34±.36
CRT (TNNLS'23)	67.16±.40	78.50±.46	38.85±.45	44.09±.38	37.02±.23
SimMTM (NeurIPS'23)	68.72±.53	81.74±.44	39.87±.43	45.81±.31	39.83±.41
TS-CoT (CIKM'23)	68.12±.34	79.72±.42	39.45±.54	44.13±.24	37.84±.45
CA-TCC (TPAMI'23)	69.73±.36	81.73±.37	40.81±.46	46.27±.26	40.16±.39
RCD-KD (NeurIPS'24)	71.54±.36	82.72±.55	52.68±.33	53.19±.38	46.25±.21
DECL (IJCAI'24)	76.75±.35	87.52±.28	44.90±.43	51.14±.32	44.71±.32
DU-IN (NeurIPS'24)	74.59±.14	86.66±.40	54.72±.64	57.80±.24	52.48±.22
MedGNN (WWW'25)	72.61±.33	81.42±.36	51.12±.46	**58.65±.33**	47.15±.25
CBD (AAAI'25)	73.50±.48	83.79±.44	48.71±.26	56.47±.45	42.79±.39
TeCh† (ICLR'26)	74.10±.32	83.75±1.48	46.93±.49	46.70±.86	40.10±.12
FACETS	**78.43±.34**	**92.16±.26**	**55.76±.27**	58.40±.28	**53.71±.38**

Across the ten dataset–metric pairs covered by the SSL baselines, FACETS obtains the best result on nine. The sole exception is PTB-XL weighted F_1: MedGNN reaches 58.65, whereas FACETS reaches 58.40, a difference of 0.25 percentage points. Against supervised TeCh, FACETS is higher on

all ten pairs: the accuracy margins range from +2.74 points on SleepEDF to +8.47 on FaultDiagnosis, and the weighted-F_1 margins range from +4.33 on SleepEDF to +13.61 on Georgia. The Georgia result is particularly informative: the accuracy margin is only +3.41 points, whereas the weighted-F_1 margin is +13.61, indicating a substantially larger improvement on class-balanced performance than on aggregate accuracy. The 30%-label comparison therefore reinforces the label-efficiency claim without erasing the protocol distinction between supervised training and frozen linear evaluation.

C.2 Extended In-Domain Fine-Tuning Comparisons

Figure 5 extends the fine-tuning comparison to additional baselines on SleepEDF, FaultDiagnosis, and PTB-XL.

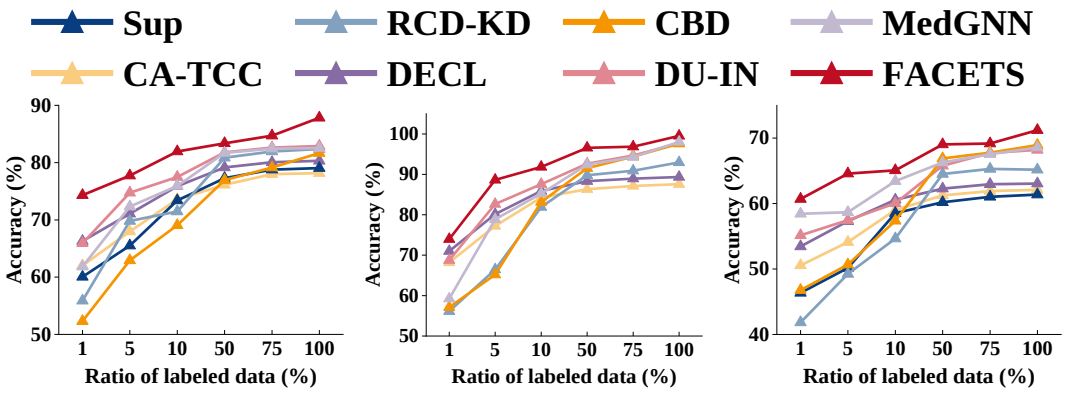


Figure 5: Extended in-domain fine-tuning comparison with 1%, 5%, 10%, 50%, 75%, and 100% of the training labels.

C.3 Corruption Robustness

The main text uses Gaussian corruption as the primary stress test. In every robustness experiment, the encoder and linear probe are trained on clean data and then held fixed; Gaussian noise, baseline wander, or impulse noise is injected only into the raw held-out test inputs. Figure 6 shows representative waveform corruption on Georgia, while Table 6 evaluates baseline wander and impulse noise under the same test-only protocol. Georgia accuracy is stable over the tested baseline-wander and impulse levels, although weighted F_1 decreases under stronger impulses. On FaultDiagnosis, performance remains high under baseline wander and mild impulse noise but drops sharply at impulse level 0.05. This failure is consistent with short discriminative fault events being overwritten across several reconstructed bands at once.

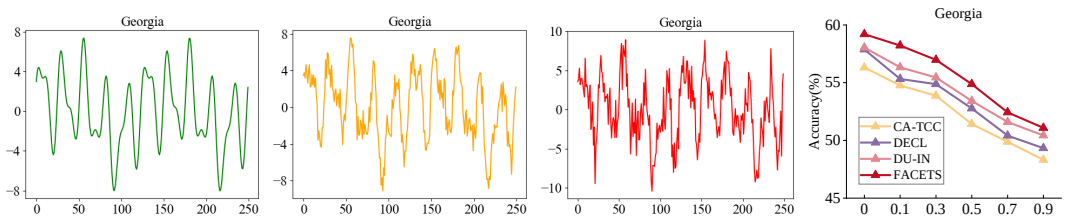


Figure 6: Gaussian-noise case study on Georgia. The encoder and linear probe are trained on clean data; zero-mean Gaussian noise is applied only to the raw held-out test inputs. Increasing corruption obscures local morphology, but FACETS retains the highest accuracy over the evaluated range.

C.4 Hyperparameter Sensitivity

We vary the auxiliary prediction weight α, contrastive temperature τ, and DWT depth on SleepEDF and PTB-XL. At decomposition depth q, the candidate set contains the q reconstructed detail

Table 6: FACETS under baseline wander and impulse noise. The encoder and linear probe are trained on clean data; corruption is applied only to the raw held-out test inputs. Entries are mean ± standard deviation (%).

Dataset	Corruption	Level	Accuracy	Weighted F_1
Georgia	Baseline wander	0.05	55.17±0.90	43.37±2.35
Georgia	Baseline wander	0.10	55.11±0.93	43.29±2.31
Georgia	Baseline wander	0.20	55.17±0.96	43.28±2.30
Georgia	Impulse	0.01	55.17±0.72	42.21±1.43
Georgia	Impulse	0.02	55.09±0.73	41.72±1.13
Georgia	Impulse	0.05	55.19±1.00	41.50±0.97
FaultDiagnosis	Baseline wander	0.05	85.10±0.394	84.65±0.445
FaultDiagnosis	Baseline wander	0.10	85.32±0.393	84.90±0.442
FaultDiagnosis	Baseline wander	0.20	83.46±0.404	82.51±0.470
FaultDiagnosis	Impulse	0.01	86.37±0.446	86.08±0.487
FaultDiagnosis	Impulse	0.02	80.79±0.712	80.03±0.864
FaultDiagnosis	Impulse	0.05	53.21±1.363	48.65±1.581

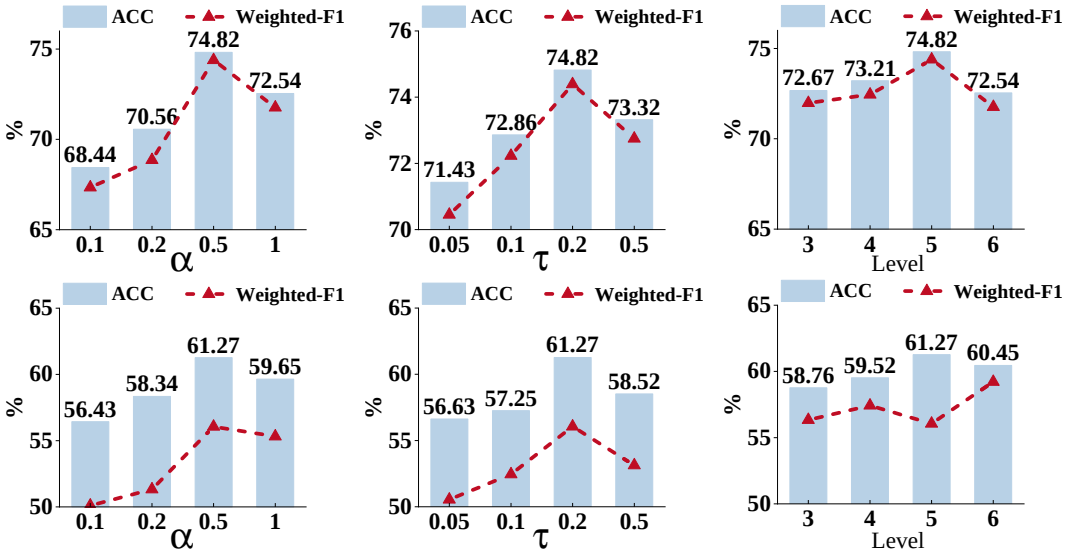


Figure 7: Sensitivity to the auxiliary prediction weight, contrastive temperature, and DWT depth. The db4 wavelet is fixed in the depth study.

components $D_1, \ldots, D_q$ and excludes the approximation component. Figure 7 shows that the reported setting ($\alpha = 0.5$, $\tau = 0.2$, depth five, $K = 5$) gives the best accuracy on both datasets. Accuracy rises from depth three to five and falls at six; PTB-XL weighted F_1 continues to improve at depth six, so the preferred decomposition can depend on the evaluation metric.

C.5 Computational Profile

Table 7 reports the computational profile of FACETS under the SleepEDF training configuration with $B = 128$, $T = 3000$, and $K = 5$ reconstructed detail components. Self-supervised pretraining uses approximately 0.6990M trainable parameters and requires 75.18 s per epoch on an NVIDIA GeForce RTX 4090. The additional pretraining cost comes primarily from processing multiple wavelet components and from the temporal prediction and probe modules. These auxiliary modules are removed after pretraining; deployment retains only the 0.3619M-parameter shared encoder, with a measured inference latency of 0.0272 ms per sample.

Table 7: Computational profile of FACETS on SleepEDF ($B = 128$, $T = 3000$, $K = 5$) using an NVIDIA GeForce RTX 4090. Parameter counts include only trainable modules used during self-supervised pretraining. At deployment, only the shared encoder is retained.

Component / measurement	Value
Shared encoder	0.3619M parameters
Temporal predictor	0.2774M parameters
Five subband probes	0.0597M parameters
FreqDF	<0.0001M parameters
DWT	0 trainable parameters
Total trainable pretraining parameters	$\approx$0.6990M
Pretraining time	75.18 s/epoch
Deployed parameters	0.3619M
Inference latency	0.0272 ms/sample

D Ablation Details

D.1 Exact Ablation Definitions

Table 8 summarizes the exact objectives and operators used by the ablation variants reported in Table 2.

Table 8: Objectives and operators used by the reported variants. "Probe" denotes detached relative-regularity weights.

Variant	Filtered pred.	Original pred.	CBCL	Weights	Similarity
`only FG`	yes	yes	no	–	–
`only FG`$^-$	yes	no	no	–	–
`only CBCL`	no	no	yes	uniform	cosine
`FS`	no	no	yes	probe	cosine
`FACETS`*	yes	yes	yes	probe	cosine
FACETS	yes	yes	yes	probe	directional

D.2 Supervisory-Contamination Diagnostic

We construct a controlled nuisance-decoding diagnostic on FaultDiagnosis. Four synthetic interference groups are assigned independently of the semantic class and injected during self-supervised pretraining, giving nuisance identity a four-way chance accuracy of 25%. Neither semantic nor nuisance labels are used during pretraining. After 100 epochs, the encoder is frozen and separate linear probes are trained on the full labeled training split to predict semantic class and nuisance identity. We report one seed-1 run and interpret it as mechanistic evidence rather than a population-level performance estimate. Table 9 reports the resulting semantic and nuisance probe accuracies.

Table 9: Frozen-probe diagnostic on synthetically corrupted FaultDiagnosis. Results are from a single seed-1 mechanism run; nuisance chance accuracy is 25%.

Pretraining objective	Semantic accuracy (%)	Nuisance accuracy (%)
`only FG` (prediction only)	**94.2**	96.6
FACETS	86.4	71.8
`FS` (cross-band only)	62.4	**27.6**

Prediction-only pretraining makes both semantic and nuisance information highly decodable. Removing prediction sharply reduces nuisance decodability, but also weakens semantic information. FACETS lies between these extremes, showing that auxiliary prediction trades semantic recovery against nuisance retention. The diagnostic therefore supports supervisory contamination as a mechanism, not complete nuisance disentanglement.

E Interpretation and Failure Modes

Why the deployment gap matters. The DWT components are training-only, whereas downstream tasks consume $h_i = f_\theta(x_i)$. Shared parameters couple the two paths in optimization, but they do not make Z_i and h_i equivalent. The residual $\epsilon_{\mathrm{dep},i}$ records task-relevant information available in the component set but missing from the deployed representation. An explicit component-to-original distillation loss could target this gap, but it is not part of the reported method.

When direction correction can fail. The batch mean is a reasonable estimator for coherent drift, periodic interference, or systematic sensor bias. It can be unreliable with small or compositionally biased batches, multimodal corruption, or when class structure itself contributes strongly to the mean. Robust location estimates or low-rank subspace removal are natural extensions.

Interpretation of the robustness results. The stable ranking under Gaussian noise does not mean that FACETS is invariant to corruption; all methods deteriorate. Likewise, resilience to the tested baseline-wander levels does not transfer automatically to impulses. The FaultDiagnosis collapse under strong impulse noise illustrates the boundary of cross-band redundancy: a transient can overwrite the same short discriminative event across several reconstructed components.