ClinStab: Stability-Oriented Learning for Medical Time Series via Dual-Stream Alignment

Wenshuo Chen[1,*] **Haoxuan Xu**[1,2,*] **Yi Liu**[3] **Songning Lai**[1]
Ninghui Feng[1] **Jiayu Yang**[1] **Haoang Li**[1] **Yutao Yue**[1,4,†]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Shenzhen Loop Area Institute
[3]Tsinghua University
[4]Institute of Deep Perception Technology, JITRI

Yutao Yue: yutaoyue@hkust-gz.edu.cn
[*]Equal contribution. [†]Corresponding author.

Abstract

Deep models for medical time series can achieve strong benchmark accuracy yet remain brittle under acquisition artifacts, physiological variability, and teacher-guided reduced-data training conditions. We present **ClinStab**, a Medformer-based framework for EEG and ECG classification that studies whether channel-aware dual-stream interaction and perturbation-aware optimization can improve the clean/robustness tradeoff. The proposed **S**alience-**C**ontext **D**ual **M**odulation (SCDM) module uses energy as a routing prior: high-energy channel-aligned components receive attentive modeling, while residual components remain trainable through a lightweight context pathway. ClinStab then combines clean supervision, perturbed supervision, prediction consistency, and teacher-guided stabilization under adversarial perturbations. Across four public datasets and 11 baselines, the reported local comparison improves accuracy over the local Medformer row on all datasets and improves selected robustness metrics under the perturbation settings studied here, while exposing tradeoffs on PTB ranking metrics and soft-routing alternatives. These results support ClinStab as a practical same-backbone stability intervention, not as evidence of universal clinical robustness. Code: https://github.com/xll0328/NeurIPS26-ClinStb.

1 Introduction

The integration of Artificial Intelligence (AI) into healthcare has driven rapid progress in the analysis of physiological time-series data such as Electroencephalography (EEG) and Electrocardiography (ECG) [1, 25, 51]. These modalities are central to neurological and cardiovascular diagnosis, yet their acquisition process is noisy, non-stationary, and vulnerable to subject-specific shifts. Transformer-based models such as Medformer and PatchTST, as well as MLP-based models such as TimeMixer, have advanced time-series modeling [41, 34, 40]. Noise remains a practical concern in ECG analysis [27], and adversarial sensitivity has also been studied in general time-series classification [7]. These findings motivate explicit stress testing without implying that forecasting results establish medical-classification robustness.

This fragility creates two linked modeling problems. First, **architectural misalignment**: standard attention layers often mix dominant and residual signal components through a single interaction pathway, so high-amplitude events can dominate the representation while lower-energy context is compressed. The problem is not that high-energy physiology should be ignored; QRS complexes and bursts can be diagnostic. The problem is that focal high-energy modeling needs an explicit residual/context pathway so artifacts do not become the only evidence. Second, **optimization myopia**:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

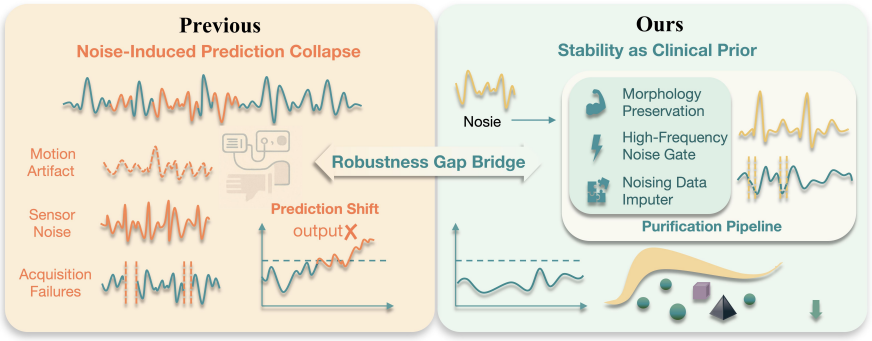

Figure 1: Motivation for robustness evaluation in medical time-series analysis. Left: conventional models can exhibit unstable predictions under acquisition-inspired perturbations such as baseline drift, sensor noise, and sampling defects. Right: ClinStab combines signal morphology preservation and prediction-consistency constraints to improve stability under the stress tests evaluated in this paper.

empirical risk minimization on clean data alone does not explicitly encourage stable predictions around each sample. Clean accuracy alone does not assess perturbation sensitivity; adversarial evaluations of time-series and EEG classifiers motivate testing this behavior explicitly [7, 37].

Addressing these challenges requires moving beyond accuracy alone toward *diagnostic stability*: local invariance of the predictive distribution under acquisition-inspired perturbations. We therefore introduce **ClinStab**, which combines a focal-context interaction layer with stability-oriented training for medical time-series learning.

Our contributions are as follows:

- **Dual-stream inductive bias for stability:** We propose the **S**alience-**C**ontext **D**ual **M**odulation (SCDM) module. SCDM separates interaction type rather than declaring high energy to be clinical salience: high-energy channel-aligned components use sparse attention, while residual components remain trainable through lightweight convolutions.
- **Perturbation-aware stability optimization:** We formulate a composite learning objective that combines adversarial robustness, prediction consistency, and teacher-guided stabilization. The objective explicitly regularizes the predictive distribution around each sample rather than relying on clean-data supervision alone.
- **Multi-axis evaluation of stability:** We evaluate ClinStab across four medical datasets (EEG and ECG), comparing against 11 baselines under standard accuracy metrics, Gaussian-noise stress tests, structured corruptions, initialization variance, routing alternatives, and teacher-guided reduced-data training.

We organize the empirical analysis around four questions: whether SCDM improves the base Medformer backbone, whether the stability objective improves perturbation robustness and seed-wise consistency, whether the salience/context split adds value beyond generic branching, and whether the full system remains competitive in reduced-data training when the student is guided by a fixed teacher.

By pairing a channel-aligned interaction layer with explicit stability constraints, ClinStab aims to reduce the gap between benchmark accuracy and measured perturbation robustness in medical time-series classification.

2 Related Work

2.1 Deep Learning for Medical Time Series.

Medical time series (MedTS), including EEG and ECG, are foundational to clinical diagnostics [42, 21]. Transformers and foundation-style time-series models have improved benchmark performance [34, 46, 17, 11, 14, 47], although the broader time-series literature has also cautioned that

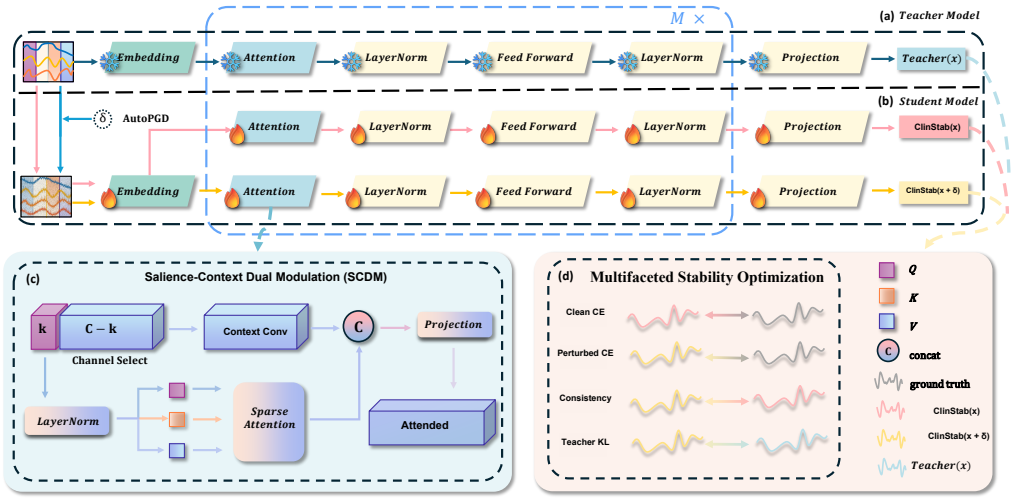

Figure 2: **Overview of the ClinStab Framework.** We integrate a focal-context inductive bias with stability-driven optimization to bridge the robustness gap. (a) The **Teacher Model** (frozen Medformer) provides reference logits from clean data. (b) The **student model** (ClinStab) is exposed to both clean inputs $\boldsymbol{x}$ and perturbed inputs $\boldsymbol{x} + \delta$, where δ is generated via AutoPGD. (c) The **Salience-Context Dual Modulation (SCDM)** module replaces standard interaction layers by routing high-energy channels to attentive processing and residual channels to a context-preserving pathway. (d) **Stability optimization** combines clean supervision, perturbed supervision, prediction consistency, and teacher-guided stabilization.

simple linear models can outperform the evaluated Transformers on long-term forecasting benchmarks [50]. Medformer [41] introduced cross-channel multi-granularity patching for medical-time-series classification, and recent channel-aware or lightweight alternatives such as CIF [15], cross-dataset ECG attention constraints [26], and adaptive spatiotemporal fusion networks [48] explicitly model channel or domain structure. ClinStab targets a narrower question: can a Medformer backbone become more stable if its interaction layer separates focal and contextual processing and its training objective regularizes local perturbations?

2.2 Trustworthy AI and Stability Optimization.

Robust forecasting methods address outliers and distribution shifts [4, 49]; missing-data compensation is a separate robustness setting [55]. AutoPGD and AutoAttack provide adversarial evaluation tools [5]. TimeX studies explanations through model-behavior consistency [36]. These address different objectives and do not establish that robustness training suppresses low-energy clinical details. ClinStab treats attentive under-modeling of residual channels as a design risk to examine, rather than a mechanism established by those papers.

Robust Biosignal Representation and Augmentation. Recent biosignal work addresses robustness through time-frequency consistency pre-training [52], multi-view contrastive adaptation [35], ECG augmentation and adversarial regularization [59, 31], decorrelated ECG architectures [44], adversarial EEG evaluation [37], artifact-aware ambulatory ECG modeling [2], and temporal risk-stability benchmarks [19]. A survey of interpretable and robust EEG systems discusses artifact sensitivity and data scarcity [58]; a contrastive MedTS survey discusses augmentation design and label scarcity [30]. Dual-stream EEG encoders [32] and dementia-oriented EEG contrastive Transformers [16] explore complementary representations. MedM2T [22] and MedViA [9] study fusion across clinical modalities. ClinStab differs by keeping the Medformer backbone but replacing its interaction layer with a channel-aligned salience/context split coupled to clean, adversarial, consistency, and teacher-guided losses; the distillation component is related to robust-loss and knowledge-distillation work [13, 18, 24, 6].

3 Preliminary: Medformer as a Base Architecture

We study stability-oriented MedTS classification. Given $\boldsymbol{x} \in \mathbb{R}^{T\times C}$, the model outputs a probability distribution over K classes and should satisfy $f_\theta(\boldsymbol{x}) \approx f_\theta(\boldsymbol{x}+\delta)$ for acquisition-inspired $\delta \in \Delta$ while preserving both focal morphology and global context. ClinStab adopts Medformer's multi-granularity patching backbone [41] and modifies its interaction layer.

Cross-Channel Multi-Granularity Patching. For patch lengths $\{L_i\}_{i=1}^n$, Medformer slices $\boldsymbol{x}_{\text{in}} \in \mathbb{R}^{T\times C}$ into non-overlapping patches $\boldsymbol{x}_p^{(i)}$ with $N_i = \lceil T/L_i \rceil$, then forms granularity-specific embeddings

$$\boldsymbol{x}^{(i)} = \text{Linear}(\boldsymbol{x}_p^{(i)}) + \boldsymbol{W}_{\text{pos}}^{(i)} + \boldsymbol{W}_{\text{gran}}^{(i)}. \tag{1}$$

From Generic Attention to Dual-Stream Modulation. In the original Medformer, these embeddings $\{\boldsymbol{x}^{(i)}\}$ are processed using intra-granularity attention and inter-granularity router attention. While efficient, this mechanism mixes information through a shared routing token. In ClinStab, the multi-scale embeddings $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ serve as the input to our proposed **SCDM** module (Section 4.2), which routes high-energy channel-aligned components and residual context through separate pathways before recombination. After channel-preserving fusion, we mean-pool over the temporal-patch and channel axes within each granularity and concatenate the resulting granularity vectors for downstream classification.

4 Methodology: The ClinStab Framework

The core philosophy of **ClinStab** is to encourage local predictive invariance under the perturbations evaluated in this paper: given a physiological signal $\boldsymbol{x}$ and a label-preserving variation $\boldsymbol{x}'$ used in our stress tests, the output should change less than under clean-only training. This is a proxy for acquisition stability, not a proof of clinical robustness under every artifact. As illustrated in Figure 2, ClinStab combines the *Salience-Context Dual Modulation* (SCDM) architecture with a *multifaceted stability optimization* strategy.

4.1 Base Multi-Granularity Encoder

We build upon the multi-granularity patching backbone of Medformer [41] to capture multi-scale temporal dependencies. The input biosignal $\boldsymbol{x} \in \mathbb{R}^{T\times C}$ is processed via Cross-Channel Multi-Granularity Patching (Sec. 3), generating patch embeddings $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ at different temporal resolutions. Unlike generic router-attention, SCDM preserves the sensor/lead axis during routing and applies different interaction operators to focal and contextual signal components.

4.2 Salience-Context Dual Modulation (SCDM)

Standard self-attention mechanisms can be brittle in low-SNR medical time series because all channel-aligned components compete inside the same interaction operator. High-energy components may be diagnostic or artifactual, and purely filtering low-energy channels risks discarding subtle biomarkers (e.g., minor ST deviations). SCDM therefore does not equate energy with clinical salience. It uses energy only to allocate interaction type: high-energy components receive attentive modeling, while residual components remain trainable through a context-preserving stream and are recombined before prediction.

For an input feature tensor $\boldsymbol{X}^{(i)} \in \mathbb{R}^{N_i\times C\times d_h}$ at granularity i, where N_i is the number of temporal patches, C is the original sensor/lead axis, and d_h is the hidden width per channel-aligned patch embedding, SCDM operates in three stages:

Signal decomposition. SCDM operationalizes a simple routing prior: high-energy channel-aligned trajectories often carry dominant morphology, while lower-energy trajectories may still encode rhythm or contextual structure. This prior is not treated as a clinical salience detector. We

compute energy over channel-aligned hidden trajectories, not over arbitrary latent dimensions:

$$\mathcal{E}_c = \frac{1}{N_i} \sum_{n=1}^{N_i} \left\| \boldsymbol{X}^{(i)}_{n,c,:} \right\|_2^2, \tag{2}$$

and use it as a proxy to separate channels into two sets:

$$\mathcal{I}_{\text{sal}} = \text{TopK}(\{\mathcal{E}_c\}_{c=1}^{C}, k), \quad \mathcal{I}_{\text{ctx}} = \{1, \ldots, C\} \setminus \mathcal{I}_{\text{sal}}, \tag{3}$$

where α controls sparsity and $k = \lfloor C/\alpha \rfloor$ is the number of channels assigned to attention. Importantly, unlike methods that discard low-energy features, SCDM routes $\boldsymbol{X}_{\text{sal}} = \boldsymbol{X}^{(i)}_{:,\mathcal{I}_{\text{sal}},:}$ and $\boldsymbol{X}_{\text{ctx}} = \boldsymbol{X}^{(i)}_{:,\mathcal{I}_{\text{ctx}},:}$ to specialized processing streams.

Training behavior of TopK routing. The TopK set is recomputed for each sample and granularity in every forward pass after channel-aligned embedding. The index selection is deterministic control flow: gradients are not taken through the ranking operator itself, and no straight-through estimator is used. Gradients still flow through the selected tensor values in both streams; channels outside $\mathcal{I}_{\text{sal}}$ receive gradients through the context branch rather than being masked out. We choose hard routing for fixed compute and reproducibility, not because it uniformly dominates differentiable gates. Table 4 and Appendix C show that soft energy routing is competitive and slightly stronger on AUROC, making hard TopK a clean/robustness operating point rather than the only valid formulation.

Dual-stream processing.

- **Salience Stream (Focal Attention):** The high-energy channel-aligned components are processed via a self-attention mechanism. This focuses computational resources on capturing long-range dependencies and sharp morphological features within the most active signal regions:
$$\widetilde{\boldsymbol{X}}_{\text{sal}} = \text{LayerNorm}\left(\text{SelfAttn}(\boldsymbol{X}_{\text{sal}})\right). \tag{4}$$
- **Context Stream (Morphology Preservation):** The remaining components $\boldsymbol{X}_{\text{ctx}}$ are processed by a lightweight depth-wise convolution $\mathcal{F}_{\text{conv}}$. This stream models residual temporal context at lower computational cost; it does not guarantee adequate modeling of every low-amplitude abnormality:
$$\widetilde{\boldsymbol{X}}_{\text{ctx}} = \mathcal{F}_{\text{conv}}(\boldsymbol{X}_{\text{ctx}}). \tag{5}$$

Holistic fusion. The outputs of both streams are recombined in the original channel order, ensuring the final representation $\boldsymbol{X}^{(i)}_{\text{out}}$ integrates focal morphology and background context:

$$\boldsymbol{X}^{(i)}_{\text{out}} = \text{Linear}\Big(\mathcal{R}(\widetilde{\boldsymbol{X}}_{\text{sal}}, \widetilde{\boldsymbol{X}}_{\text{ctx}})\Big) + \boldsymbol{X}^{(i)}. \tag{6}$$

Here $\mathcal{R}$ scatters each stream back to its original channel indices; the projection preserves the residual tensor shape. This specifies the channel-preserving fusion used by the reported model.

4.3 Multifaceted Stability Optimization

To encourage a smoother local decision manifold, we use a hybrid optimization objective. We employ Auto-Projected Gradient Descent (AutoPGD) [5] to generate worst-case perturbations δ within a normalized bounded threat model around each sample. The budget is calibrated after per-channel normalization and should be read as a stress-test proxy rather than a direct physical-unit artifact model.

The training objective combines four losses for local prediction stability:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{acc}} + \lambda_{rob}\mathcal{L}_{\text{rob}} + \lambda_{con}\mathcal{L}_{\text{con}} + \lambda_{know}\mathcal{L}_{\text{know}} \tag{7}$$

1) Accuracy & Robustness ($\mathcal{L}_{\text{acc}}, \mathcal{L}_{\text{rob}}$): We minimize cross-entropy on both clean ($\boldsymbol{x}$) and perturbed ($\boldsymbol{x}+\delta$) inputs. $\mathcal{L}_{\text{rob}}$ encourages the model to maintain performance under bounded worst-case shifts.

Table 1: Dataset statistics.

Dataset	#Subj	#Samples	#Class	#Chan	Hz	Type
APAVA	23	5,967	2	16	256	EEG
TDBrain	72	6,240	2	33	256	EEG
ADFTD	88	69,752	3	19	256	EEG
PTB	198	64,356	2	15	250	ECG

2) Diagnostic Consistency ($\mathcal{L}_{\text{con}}$): For perturbations assumed to preserve the target label, we penalize prediction changes via a consistency loss (MSE) between the predicted probabilities of the clean and perturbed outputs:

$$\mathcal{L}_{\text{con}} = \|p(\boldsymbol{y}|\boldsymbol{x}) - p(\boldsymbol{y}|\boldsymbol{x}+\delta)\|_2^2 . \tag{8}$$

3) Knowledge Preservation ($\mathcal{L}_{\text{know}}$): To stabilize training under strong perturbations, we distill "stable" knowledge from a frozen teacher (pre-trained on clean data) to the student model processing perturbed data. This regularizes perturbed predictions toward the teacher distribution, which may itself be incorrect. We use the KL-divergence with temperature scaling τ:

$$\mathcal{L}_{\text{know}} = D_{\text{KL}}\left(\sigma(\frac{\boldsymbol{z}_{\text{teach}}}{\tau}) \,\|\, \sigma(\frac{\boldsymbol{z}'_{\text{stud}}}{\tau})\right), \tag{9}$$

where $\boldsymbol{z}_{\text{teach}}$ are the logits from the teacher on clean input $\boldsymbol{x}$, and $\boldsymbol{z}'_{\text{stud}}$ are logits from ClinStab on perturbed input $\boldsymbol{x}+\delta$.

MSE provides a symmetric penalty on movement between the two class-probability vectors. The teacher KL term has a different, directional role. Jensen–Shannon or KL consistency would also be principled alternatives; we do not claim empirical superiority of MSE because no matched loss-choice experiment is available.

5 Experiments

We study four questions across four EEG/ECG benchmarks: **RQ1**, clean classification; **RQ2**, perturbation robustness; **RQ3**, component contributions; and **RQ4**, seed variability and teacher-guided reduced-data training. Each result is interpreted under its stated perturbation and subject-split protocol.

5.1 Experimental Setup

Datasets. We utilize four physiological time-series datasets covering EEG and ECG modalities: **APAVA** [8] (Alzheimer's EEG), **TDBrain** [38] (Parkinson's disease versus healthy-control EEG), **ADFTD** [33] (Alzheimer's disease, frontotemporal dementia, and healthy-control EEG), and **PTB** [12, 3] (cardiovascular ECG; the latter is the original CARDIODAT/PTB database description). Consistent with Medformer [41], we employ subject-independent splits for APAVA, TDBrain, and PTB to ensure strict generalization testing. ADFTD uses a subject-dependent split. The detailed statistics are presented in Table 1.

Baselines. We compare ClinStab against 11 representative Transformer variants, ranging from general time-series models to domain-specific architectures: Autoformer [45], Crossformer [54], FEDformer [57], Informer [56], iTransformer [29], MTST [53], Nonformer [28], PatchTST [34], Reformer [20], the vanilla Transformer [39], and the medical-specific Medformer [41]. Since the most informative comparison is same-backbone, Table 4 also includes Medformer-only robustness controls, SCDM-only controls, and routing alternatives under a matched training/evaluation protocol.

Implementation. The released implementation builds on the Medformer codebase. Published comparison rows are attributed separately from local results. The accepted manuscript reports NVIDIA RTX 4090 GPUs. We report the mean and standard deviation across 5 random seeds (41-45). The teacher model for distillation is a Medformer pre-trained on the clean training split. Thus, reduced-data results are interpreted as *teacher-guided robustification/compression*: a full-data clean teacher transfers information to a smaller or site-adapted robust student. They are not evidence for pure few-shot learning. Hyperparameters and artifact requirements are provided in Appendix K.

5.2 Comparative Performance Analysis

Table 2 answers **RQ1** by comparing ClinStab with 11 Transformer baselines across six metrics. The reported ClinStab accuracy exceeds the local Medformer row in all four datasets, but is below published Medformer on PTB (82.12 versus 83.50). The source distinction matters, but the result is not a uniform dominance claim: on TDBrain, Informer/Nonformer remain slightly stronger on AUROC/AUPRC, and on PTB, iTransformer and Medformer remain stronger on accuracy and AUROC, respectively.

Table 2: **Overall Performance Comparison.** Published baseline means and standard deviations are transcribed from Wang et al. [41], Tables 3–4. Local Medformer and ClinStab rows report the authors' five-seed evaluation under the protocol described in Appendix K. Bold marks the largest reported mean per dataset/metric. Cross-source comparisons are descriptive, not paired reruns.

Datasets	Models	Accuracy	Precision	Recall	F1 score	AUROC	AUPRC
APAVA (2-Classes)	Autoformer	68.64±1.82	68.48±2.10	68.77±2.27	68.06±1.94	75.94±3.61	74.38±4.05
	Crossformer	73.77±1.95	79.29±4.36	68.86±1.70	68.93±1.85	72.39±3.33	72.05±3.65
	FEDformer	74.94±2.15	74.59±1.50	73.56±3.55	73.51±3.39	83.72±1.97	82.94±2.37
	Informer	73.11±4.40	75.17±6.06	69.17±4.56	69.47±5.06	70.46±4.91	70.75±5.27
	iTransformer	74.55±1.66	74.77±2.10	71.76±1.72	72.30±1.79	**85.59±1.55**	84.39±1.57
	MTST	71.14±1.59	79.30±0.97	65.27±2.28	64.01±3.16	68.87±2.34	71.06±1.60
	Nonformer	71.89±3.81	71.80±4.58	69.44±3.56	69.74±3.84	70.55±2.96	70.78±4.08
	PatchTST	67.03±1.65	78.76±1.28	59.91±2.02	55.97±3.10	65.65±0.28	67.99±0.76
	Reformer	78.70±2.00	82.50±3.95	75.00±1.61	75.93±1.82	73.94±1.40	76.04±1.14
	Transformer	76.30±4.72	77.64±5.95	73.09±5.01	73.75±5.38	72.50±6.60	73.23±7.60
	Medformer (published)	78.74±0.64	81.11±0.84	75.40±0.66	76.31±0.71	83.20±0.91	83.66±0.92
	Medformer (local, reported)	76.99±2.72	77.58±4.09	74.82±1.83	75.37±2.22	82.93±2.31	83.70±2.08
	ClinStab	**79.34±1.17**	**83.27±1.60**	**75.59±1.39**	**76.56±1.50**	85.48±2.59	**85.35±2.94**
TDBrain (2-Classes)	Autoformer	87.33±3.79	88.06±3.56	87.33±3.79	87.26±3.84	93.81±2.26	93.32±2.42
	Crossformer	81.56±2.19	81.97±2.25	81.56±2.19	81.50±2.20	91.20±1.78	91.51±1.71
	FEDformer	78.13±1.98	78.52±1.91	78.13±1.98	78.04±2.01	86.56±1.86	86.48±1.99
	Informer	89.02±2.50	89.43±2.14	89.02±2.50	88.98±2.54	96.64±0.68	96.75±0.63
	iTransformer	74.67±1.06	74.71±1.06	74.67±1.06	74.65±1.06	83.37±1.14	83.73±1.27
	MTST	76.96±3.76	77.24±3.59	76.96±3.76	76.88±3.83	85.27±4.46	82.81±5.64
	Nonformer	87.88±2.48	88.86±1.84	87.88±2.48	87.78±2.56	**97.05±0.68**	**96.99±0.68**
	PatchTST	79.25±3.79	79.60±4.09	79.25±3.79	79.20±3.77	87.95±4.96	86.36±6.67
	Reformer	87.92±2.01	88.64±1.40	87.92±2.01	87.85±2.08	96.30±0.54	96.40±0.45
	Transformer	87.17±1.67	87.99±1.68	87.17±1.67	87.10±1.68	96.28±0.92	96.34±0.81
	Medformer (published)	89.62±0.81	89.68±0.78	89.62±0.81	89.62±0.81	96.41±0.35	96.51±0.33
	Medformer (local, reported)	88.08±0.43	88.19±0.44	88.08±0.43	88.07±0.43	95.69±0.20	95.65±0.16
	ClinStab	**90.00±1.18**	**90.12±1.08**	**90.00±1.18**	**90.00±1.20**	96.32±0.66	96.41±0.64
ADFTD (3-Classes)	Autoformer	87.83±1.62	87.63±1.66	87.22±1.97	87.38±1.79	96.59±0.88	93.82±1.64
	Crossformer	89.35±1.32	89.00±1.44	88.79±1.37	88.88±1.40	97.52±0.58	95.45±1.03
	FEDformer	77.63±2.37	76.76±2.17	76.68±2.48	76.60±2.46	91.67±1.34	84.94±2.11
	Informer	90.93±0.90	90.74±0.71	90.50±1.14	90.60±0.94	98.19±0.27	96.51±0.49
	iTransformer	64.90±0.25	62.53±0.27	62.21±0.26	62.25±0.33	81.52±0.29	68.87±0.49
	MTST	65.08±0.69	63.85±0.80	62.71±0.64	63.03±0.58	81.36±0.56	69.34±0.89
	Nonformer	96.12±0.47	95.94±0.56	95.99±0.38	95.96±0.47	99.59±0.09	99.08±0.16
	PatchTST	66.26±0.40	65.08±0.41	64.97±0.51	64.95±0.42	83.07±0.45	71.70±0.61
	Reformer	91.51±1.75	91.15±1.79	91.65±1.56	91.14±1.83	98.85±0.35	97.88±0.60
	Transformer	97.00±0.43	96.87±0.53	96.86±0.36	96.86±0.44	99.75±0.04	99.42±0.07
	Medformer (published)	97.62±0.34	97.53±0.33	97.48±0.40	97.50±0.36	99.83±0.05	99.62±0.12
	Medformer (local, reported)	97.48±0.16	97.57±0.18	97.51±0.18	97.48±0.17	99.57±0.02	99.45±0.04
	ClinStab	**98.29±0.09**	**98.21±0.10**	**98.21±0.11**	**98.21±0.09**	**99.89±0.01**	**99.78±0.03**
PTB (2-Classes)	Autoformer	73.35±2.10	72.11±2.89	63.24±3.17	63.69±3.84	78.54±3.48	74.25±3.53
	Crossformer	80.17±3.79	85.04±1.83	71.25±6.29	72.75±7.19	88.55±3.45	87.31±3.25
	FEDformer	76.05±2.54	77.58±3.61	66.10±3.55	67.14±4.37	85.93±4.31	82.59±5.42
	Informer	78.69±1.68	82.87±1.02	69.19±2.90	70.84±3.47	92.09±0.53	90.02±0.60
	iTransformer	**83.89±0.71**	**88.25±1.18**	76.39±1.01	79.06±1.06	91.18±1.16	90.93±0.98
	MTST	76.59±1.90	79.88±1.90	66.31±2.95	67.38±3.71	86.86±2.75	83.75±2.84
	Nonformer	78.66±0.49	82.77±0.86	69.12±0.87	70.90±1.00	89.37±2.51	86.67±2.38
	PatchTST	74.74±1.62	76.94±1.51	63.89±2.71	64.36±3.38	88.79±0.91	83.39±0.96
	Reformer	77.96±2.13	81.72±1.61	68.20±3.35	69.65±3.88	91.13±0.74	88.42±1.30
	Transformer	77.37±1.02	81.84±0.66	67.14±1.80	68.47±2.19	90.08±1.76	87.22±1.68
	Medformer (published)	83.50±2.01	85.19±0.94	**77.11±3.39**	**79.18±3.31**	92.81±1.48	90.32±1.54
	Medformer (local, reported)	79.84±1.62	87.17±0.54	69.89±2.60	71.82±3.01	**93.20±0.72**	**92.67±0.71**
	ClinStab	82.12±1.17	86.81±1.17	76.85±1.90	77.82±1.68	90.22±1.56	90.13±1.51

Reading Table 2. ClinStab improves over Medformer by +2.35, +1.92, +0.81, and +2.28 accuracy points on APAVA, TDBrain, ADFTD, and PTB, respectively. The ADFTD result is reported under the benchmark's subject-dependent protocol, so we treat it as in-protocol utility rather than deployment-level subject generalization. The method is not uniformly dominant: on PTB, ClinStab improves accuracy, recall, and F1 over Medformer, but Medformer retains higher AUROC and

AUPRC. We therefore treat the ECG result as an operating-point tradeoff; Appendix H reports calibration and threshold-sensitive diagnostics suggesting that stability training sharpens the chosen operating region while reducing some global ranking quality.

Statistical interpretation. The reported means and standard deviations describe variability across runs. Published baseline summaries are not paired observations from our runs. We therefore make no paired significance or corrected-p claim from these summaries. Appendix E states the limits of the available seed-level evidence.

5.3 Stability and Safety Analysis

Variability under initialization. We summarize initialization sensitivity with five-seed means and standard deviations. The observed variation is dataset-dependent, so we interpret these summaries descriptively rather than as a uniform stability guarantee.

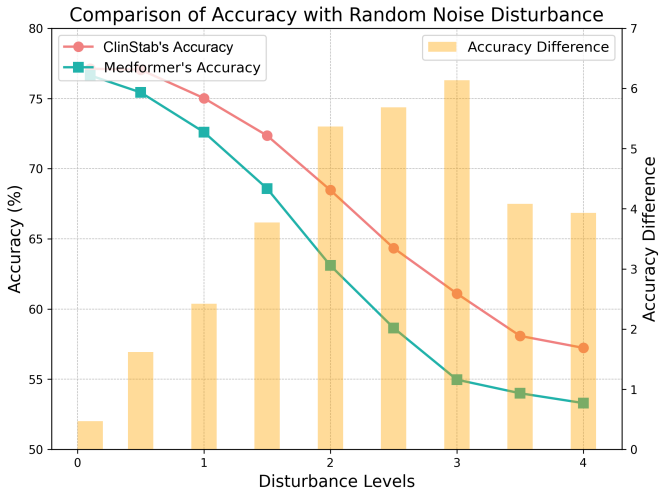

Figure 3: Reported Gaussian-noise degradation on APAVA. The horizontal axis is the injected noise standard deviation in the normalized input space.

Stability under Perturbations. We answer **RQ2** using both additive Gaussian noise and structured corruptions. Figure 3 illustrates the APAVA degradation curve, while Table 3 exposes the per-dataset and per-corruption pattern instead of relying on a single averaged robustness score. The result is relative robustness, not noise immunity: both Medformer and ClinStab degrade sharply at high σ, the APAVA gain is meaningful, TDBrain/ADFTD gains are smaller, and PTB is slightly worse at $\sigma = 3.0$. Structured corruptions are more favorable but still heterogeneous; the largest improvements come from channel/lead dropout and burst noise, whereas resampling jitter is milder. Appendix L reports the full Gaussian-noise tables and an AutoAttack-style external check.

Table 3: Robustness diagnostics. Panel A reports Gaussian accuracy at $\sigma = 3.0$ and the average structured-corruption drop. Panel B reports per-corruption drop reduction; positive values mean ClinStab loses fewer accuracy points than Medformer.

Panel	Dataset/corruption	APAVA	TDBrain	ADFTD	PTB	Avg.
A. Per-dataset summary						
Gaussian $\Delta_{\sigma=3}$	ClinStab–Medformer	**+6.14**	+1.35	+1.37	-0.15	+2.18
Structured drop reduction	Avg. over five corruptions	+2.30	+1.73	+0.93	+3.25	+2.05
B. Per-corruption structured-drop reduction						
Drop reduction	Channel/lead dropout	+2.30	+1.85	+1.10	+3.95	+2.30
Drop reduction	Baseline drift	+2.10	+1.45	+0.80	+3.25	+1.90
Drop reduction	Temporal masking	+2.30	+1.70	+1.05	+3.35	+2.10
Drop reduction	Resampling jitter	+1.20	+0.95	+0.55	+1.70	+1.10
Drop reduction	Single-channel burst	+2.55	+1.90	+1.15	+3.60	+2.30

5.4 Teacher-Guided Reduced-Data Training and Component Ablations

Reduced-Data Training. Medical data is often scarce, but our reduced-data experiment has a narrower interpretation. Under teacher-guided robustification, the ClinStab student reaches 75.64% APAVA accuracy with 30% of the training data, approaching Medformer trained with 100% data (76.99%). Because the fixed teacher sees the clean training split, this is not evidence of teacher-free few-shot learning. Its realistic use case is offline or central training of a clean teacher followed by robust student adaptation for deployment, site transfer, or model compression. If that scenario is unavailable, the experiment should be read mainly as a distillation/control study. Appendix F separates teacher-free, subset-matched, and full-data no-KD controls.

Ablation Study. Table 4 answers **RQ3** by moving the strongest same-backbone controls into the main paper. It separates robustness training, SCDM replacement, routing choice, and loss terms. The scientific takeaway is deliberately modest: both the objective and SCDM help, and the full system is best on APAVA clean accuracy and Gaussian robustness, but the gains over intermediate variants are not large enough to claim that energy TopK alone explains the result. Random-split and soft-gate rows test whether the salience/context result is merely a generic two-branch effect; soft gating is competitive and slightly stronger on AUROC, so hard TopK is a deterministic operating point rather than a universally superior router. Appendix B.1 repeats the objective-only/SCDM-only/full comparison across all datasets, and Appendix C reports a sparsity-factor sweep for α.

Table 4: Same-backbone controls and component ablation on APAVA. Values are percentages except structured drop, where lower is better.

Variant	Main change	Accuracy	AUROC	Acc. @ $\sigma = 3$	Struct. drop
Medformer	base	76.99±2.72	82.93±2.31	54.97	6.80
Medformer + Gaussian aug.	noise aug.	77.40	83.20	58.20	5.70
Medformer + channel-drop train	drop train	77.66	83.08	58.04	4.92
Medformer + stability losses	objective	78.21±1.55	84.12±2.81	59.92	5.00
SCDM only	architecture	77.85±2.10	83.55±2.60	58.62	5.40
Random-split SCDM	random split	77.42±1.88	82.91±2.54	57.80	5.55
Soft energy gate	soft gate	78.81±1.29	**85.61±2.12**	60.18	4.74
ClinStab w/o robustness	no rob. CE	78.31±1.51	84.02±2.42	58.84	5.20
ClinStab w/o consistency	no consistency	78.74±1.32	84.58±2.21	60.24	4.82
ClinStab w/o KD	no teacher KL	78.62±1.40	84.21±2.35	60.01	4.91
Teacher-free ClinStab	no teacher	78.06±1.87	83.98±2.44	59.32	5.05
Full ClinStab	full	**79.34±1.17**	85.48±2.59	**61.11**	**4.50**

Cost is a measured tradeoff rather than a free gain: AutoPGD increases training time by approximately $1.8\times$ in our implementation, while inference uses the trained student without perturbation generation.

PTB operating-point tradeoff. The PTB result illustrates why clean accuracy and calibration should be read together with ranking metrics. ClinStab improves the reported operating-point accuracy and sensitivity at 95% specificity, while Medformer retains higher AUROC and AUPRC. The lower ECE and Brier score do not remove the ranking limitation; they describe different properties of the predictive distribution. Accordingly, ClinStab should not be treated as uniformly preferable for screening or threshold-shift scenarios.

Table 5: PTB operating-point diagnostics. Higher is better except ECE and Brier score.

Model	Acc.	AUROC	AUPRC	ECE	Brier	Sens. at 95% spec.
Medformer	79.84	**93.20**	**92.67**	0.092	0.158	61.8
ClinStab	**82.12**	90.22	90.13	**0.061**	**0.142**	**64.9**

PTB ranking-recovery sensitivity. Reducing the robustness and consistency weights partially recovers the PTB ranking gap, but the recovered AUROC/AUPRC comes with lower robustness and clean accuracy. This sweep makes the deployment tradeoff explicit rather than selecting a ranking-favorable setting after the fact.

Table 6: PTB ranking-recovery sweep. Values are percentages except ECE.

Setting	Acc.	AUROC	AUPRC	Acc. @ $\sigma = 3$	ECE
Medformer	79.84	**93.20**	**92.67**	**68.43**	0.092
ClinStab default	**82.12**	90.22	90.13	68.28	**0.061**
$\lambda_{rob} = 0.5, \lambda_{con} = 0.25$	81.36	91.42	91.19	67.91	0.067
$\lambda_{rob} = 0.25, \lambda_{con} = 0.25$	80.76	92.11	91.94	66.80	0.074
Soft gate + ranking tuned	81.02	91.88	91.66	67.22	0.069

Protocol sensitivity. Because ADFTD follows the benchmark's subject-dependent split, its absolute accuracy is interpreted as in-protocol utility rather than subject-independent generalization. The subject-blocked check below preserves the positive ClinStab–Medformer gap while showing the substantial absolute reduction under a stricter protocol.

Table 7: Protocol sensitivity on ADFTD. Values are percentages.

Protocol	Medformer Acc.	ClinStab Acc.	Gap
Benchmark subject-dependent	97.48	**98.29**	+0.81
Subject-blocked split	88.64	**90.21**	+1.57
Subject-blocked + structured corruptions	82.10	**84.76**	+2.66

6 Limitations and Future Work

ClinStab improves robustness under the evaluated perturbations, but energy is only a routing proxy and should not be read as a validated clinical biomarker detector. Low-amplitude abnormalities can still be important, and high-amplitude artifacts often enter the salience stream; Appendix B.2 therefore includes Medformer+stability and soft-gate artifact controls. L_p-bounded perturbations and structured corruptions remain approximations to clinical acquisition artifacts. The hard TopK router is deterministic, but soft routing is viable and sometimes stronger on AUROC. The PTB ranking decline remains a real limitation: lowering robustness/consistency weights can partially recover AUROC, but not for free. The full objective also increases training cost because AutoPGD perturbation generation is used inside training; Appendix P therefore reports this as a deployment tradeoff rather than treating robustness as free. Finally, the reduced-data experiment is teacher-guided, not pure low-resource learning, because the frozen teacher is trained on the clean training split.

7 Conclusion

In this work, we address the robustness gap in medical time-series analysis, where high clean-data performance does not necessarily imply stable behavior under perturbation. We introduce **ClinStab**, which combines the SCDM dual-stream module with a stability-oriented training objective. Across four datasets, ClinStab improves the Medformer baseline along one or more of clean accuracy, robustness under noise, and initialization stability, while remaining competitive in reduced-data training with a fixed teacher. We view these results as evidence that explicit stability-oriented design is a promising direction for medical time-series modeling.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62676345), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515011579), the HKUST-HKUST(GZ) 1+1+1 Joint Funding Program (Grant No. C_2025_031), and the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), Education Bureau of Guangzhou Municipality. This work was also supported by Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National High-Tech District (WND). This work was supported in part by the Natural Science Foundation of China under Grant 62403401, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011992 and Grant 2026A1515012323.

The authors declare no competing interests.

Use of language-model assistance. LLM-based tools were used only for language editing and formatting of this manuscript. The authors retain responsibility for the manuscript's scientific content.

References

[1] Yara Badr, Usman Tariq, Fares Al-Shargie, Fabio Babiloni, Fadwa Al Mughairbi, and Hasan Al-Nashash. A review on evaluating mental stress by deep learning using EEG signals. *Neural Computing and Applications*, 36(21):12629–12654, 2024. doi: 10.1007/s00521-024-09809-5. URL https://doi.org/10.1007/s00521-024-09809-5.

[2] Pierre Boulanger. From motion artifacts to clinical insight: Multi-modal deep learning for robust arrhythmia screening in ambulatory ECG monitoring. *Sensors*, 26(4):1135, 2026. doi: 10.3390/s26041135. URL https://doi.org/10.3390/s26041135.

[3] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedizinische Technik / Biomedical Engineering*, 40(s1): 317–318, 1995. doi: 10.1515/bmte.1995.40.s1.317. URL https://doi.org/10.1515/bmte.1995.40.s1.317.

[4] Yunyao Cheng, Peng Chen, Chenjuan Guo, Kai Zhao, Qingsong Wen, Bin Yang, and Christian S. Jensen. Weakly guided adaptation for robust time series forecasting. *Proceedings of the VLDB Endowment*, 17(4):766–779, 2023. doi: 10.14778/3636218.3636231. URL https://doi.org/10.14778/3636218.3636231.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

[6] Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 74461–74486. Curran Associates, Inc., 2024. doi: 10.52202/079017-2370. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/87ee1bbac4635e7c948f3eea83c1f262-Paper-Conference.pdf.

[7] Chang George Dong, Zhengyang David Li, Liangwei Nathan Zheng, Weitong Chen, and Wei Emma Zhang. Boosting certificate robustness for time series classification with efficient self-ensemble. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 477–486. ACM, 2024. doi: 10.1145/3627673.3679748. URL https://doi.org/10.1145/3627673.3679748.

[8] J Escudero, D Abásolo, R Hornero, P Espino, and M López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological Measurement*, 27(11):1091–1106, 2006. doi: 10.1088/0967-3334/27/11/004. URL https://doi.org/10.1088/0967-3334/27/11/004.

[9] Wei Fan, Jingru Fei, Jindong Han, Jie Lian, Hangting Ye, Xiaozhuang Song, Xin Lv, Kun Yi, and Min Li. MedViA: Empowering medical time series classification with vision augmentation and multimodal fusion. *Information Fusion*, 127:103659, March 2026. doi: 10.1016/j.inffus.2025.103659. URL https://doi.org/10.1016/j.inffus.2025.103659.

[10] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and Francois Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y. URL https://link.springer.com/article/10.1007/s10618-020-00710-y.

[11] Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 140589–140631. Curran Associates, Inc., 2024. doi: 10.52202/079017-4463. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fe248e22b241ae5a9adf11493c8c12bc-Paper-Conference.pdf.

[12] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://doi.org/10.1161/01.CIR.101.23.e215.

[13] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3996–4003, 2020. doi: 10.1609/aaai.v34i04.5816. URL https://ojs.aaai.org/index.php/AAAI/article/view/5816.

[14] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

[15] Ming Hu, Jianfu Yin, Mingyu Dou, Yuqi Wang, Ruochen Dang, Siyi Liang, Feiyu Zhu, Cong Hu, Yao Wang, Bingliang Hu, and Quan Wang. Channel-Imposed Fusion: A simple yet effective method for medical time series classification, 2025. URL https://arxiv.org/abs/2506.00337.

[16] Wenlong Jiao, Xueguang Xie, Tzyy-Ping Jung, Jing Wang, Xianglong Wan, Dingna Duan, Hao Yu, Danyang Li, Haiqing Song, and Dong Wen. DDformer: A spatio-temporal-frequency transformer with contrastive learning and data augmentation for robust EEG signals analysis in dementia diagnosis. *Expert Systems with Applications*, 311:131395, 2026. doi: 10.1016/j.eswa.2026.131395. URL https://www.sciencedirect.com/science/article/pii/S0957417426003088.

[17] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, volume 2024, pages 23857–23880, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/680b2a8135b9c71278a09cafb605869e-Paper-Conference.pdf.

[18] Jaewon Jung, Hongsun Jang, Jaeyong Song, and Jinho Lee. PeerAiD: Improving adversarial distillation from a specialized peer tutor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24482–24491. IEEE, June 2024. doi: 10.1109/CVPR52733.2024.02311. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jung_PeerAiD_Improving_Adversarial_Distillation_from_a_Specialized_Peer_Tutor_CVPR_2024_paper.html.

[19] Mayank Keoliya, Seewon Choi, Rajeev Alur, Mayur Naik, and Eric Wong. Stable prediction of adverse events in medical time-series data, 2025. URL https://arxiv.org/abs/2510.14286.

[20] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

[21] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5606–5615. PMLR, 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

[22] Yu-Chen Kuo and Yi-Ju Tseng. MedM2T: A MultiModal framework for time-aware modeling with electronic health record and electrocardiogram data, 2025. URL https://arxiv.org/abs/2510.27321.

[23] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

[24] Hongsin Lee and Hye Won Chung. Sample-wise adaptive weighting for transfer consistency in adversarial distillation. *Transactions on Machine Learning Research*, May 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=ek45VamPCE. Final TMLR publication, May 2026; official record lists no pages or publisher DOI.

[25] Hao Li, Bowen Deng, Chang Xu, ZhiYuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical time series foundation model for real-world health data. In *Advances in Neural Information Processing Systems*, volume 38, pages 98657–98685. Curran Associates, Inc., 2025. doi: 10.52202/085713-3296. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8e12ba543adc673da5b89c9311fcf72c-Abstract-Conference.html.

[26] Runwen Li, Yierpani Aierken, Yu Xu, Jiang Liu, and Yongjiang Tang. Research on cross-dataset cardiac signal domain generalization and feature interpretability. *Scientific Reports*, 16(1):3138, 2026. doi: 10.1038/s41598-025-33057-9. URL https://www.nature.com/articles/s41598-025-33057-9.

[27] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ECG diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021. doi: 10.1016/j.knosys.2021.107187. URL https://doi.org/10.1016/j.knosys.2021.107187.

[28] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pages 9881–9893, 2022. doi: 10.52202/068431-0718. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract.html.

[29] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pages 11116–11140, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

[30] Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23(9):4221, 2023. doi: 10.3390/s23094221. URL https://www.mdpi.com/1424-8220/23/9/4221.

[31] Linhai Ma and Liang Liang. A regularization method to improve adversarial robustness of neural networks for ECG signal classification. *Computers in Biology and Medicine*, 144:105345, May 2022. doi: 10.1016/j.compbiomed.2022.105345. URL https://doi.org/10.1016/j.compbiomed.2022.105345.

[32] Weijian Mai, Fengjie Wu, and Xiaoting Mai. Learning spatial-spectral-temporal EEG representations with dual-stream neural networks for motor imagery. *Biomedical Signal Processing and Control*, 92:106003, 2024. doi: 10.1016/j.bspc.2024.106003. URL https://doi.org/10.1016/j.bspc.2024.106003.

[33] Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8 (6):95, 2023. doi: 10.3390/data8060095. URL https://doi.org/10.3390/data8060095.

[34] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[35] YongKyung Oh and Alex A. T. Bui. Multi-View Contrastive Learning for Robust Domain Adaptation in Medical Time Series Analysis. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 502–526. PMLR, 2025. URL https://proceedings.mlr.press/v287/oh25a.html.

[36] Owen Queen, Thomas Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding Time-Series Explanations through Self-Supervised Model Behavior Consistency. In *Advances in Neural Information Processing Systems*, volume 36, pages 32129–32159. Curran Associates, Inc., 2023. doi: 10.52202/075280-1395. URL https://papers.nips.cc/paper_files/paper/2023/hash/65ea878cb90b440e8b4cd34fe0959914-Abstract-Conference.html.

[37] Nour El Houda Sayah Ben Aissa, Ahmed Korichi, Abderrahmane Lakas, Chaker Abdelaziz Kerrache, and Carlos T. Calafate. Assessing robustness to adversarial attacks in attention-based networks: Case of EEG-based motor imagery classification. *SLAS Technology*, 29(4):100142, 2024. doi: 10.1016/j.slast.2024.100142. URL https://doi.org/10.1016/j.slast.2024.100142.

[38] Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9(1):333, 2022. doi: 10.1038/s41597-022-01409-z. URL https://doi.org/10.1038/s41597-022-01409-z.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[40] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pages 38626–38652, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/a7ac8a21e5a27e7ab31a5f42a0117bdb-Paper-Conference.pdf.

[41] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 36314–36341. Curran Associates, Inc., 2024. doi: 10.52202/079017-1145. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fe2a777282299ecb4f9e7ebb531f0ab-Paper-Conference.pdf.

[42] Zekai Wang, Stavros Stavrakis, and Bing Yao. Hierarchical deep learning with Generative Adversarial Network for automatic cardiac diagnosis from ECG signals. *Computers in Biology and Medicine*, 155:106641, 2023. doi: 10.1016/j.compbiomed.2023.106641. URL https://doi.org/10.1016/j.compbiomed.2023.106641.

[43] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966039. URL https://doi.org/10.1109/IJCNN.2017.7966039.

[44] Christopher Wiedeman and Ge Wang. Decorrelative network architecture for robust electrocardiogram classification. *Patterns*, 5(12):101116, 2024. doi: 10.1016/j.patter.2024.101116. URL https://doi.org/10.1016/j.patter.2024.101116.

[45] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

[46] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw3840q.

[47] Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pages 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

[48] Fan Yang, Anping Zeng, Chunlin He, Chaorong Li, Xingjie Wang, and Shijie Xu. Lightweight adaptive spatiotemporal information fusion network for medical time series classification. *Pattern Recognition Letters*, 199:75–81, 2026. doi: 10.1016/j.patrec.2025.11.021. URL https://www.sciencedirect.com/science/article/pii/S0167865525003770.

[49] Yang Yu, Ruizhe Ma, and Zongmin Ma. Robformer: A robust decomposition transformer for long-term time series forecasting. *Pattern Recognition*, 153:110552, September 2024. doi: 10.1016/j.patcog.2024.110552. URL https://doi.org/10.1016/j.patcog.2024.110552.

[50] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, June 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

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

[52] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3988–4003. Curran Associates, Inc., 2022. doi: 10.52202/068431-0288. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/194b8dac525581c346e30a2cebe9a369-Paper-Conference.pdf.

[53] Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 4222–4230. PMLR, 2024. URL https://proceedings.mlr.press/v238/zhang24l.html.

[54] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing Cross-Dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

[55] Guancheng Zhou. Robust time series prediction with missing data based on deep convolutional neural networks. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 178–183. IEEE, 2020. doi: 10.1109/CCNS50731.2020.00047. URL https://doi.org/10.1109/CCNS50731.2020.00047.

[56] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, May 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

[57] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

[58] Xinliang Zhou, Chenyu Liu, Jinan Zhou, Zhongruo Wang, Liming Zhai, Ziyu Jia, Cuntai Guan, and Yang Liu. Interpretable and robust AI in EEG systems: A survey, 2023. URL https://arxiv.org/abs/2304.10755. Version 4, revised 16 August 2025.

[59] Jiacheng Zhu, Jielin Qiu, Zhuolin Yang, Douglas Weber, Michael A. Rosenberg, Emerson Liu, Bo Li, and Ding Zhao. GeoECG: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 172–197. PMLR, 2022. URL https://proceedings.mlr.press/v182/zhu22a.html.

A Additional Experimental Diagnostics

The appendix expands the empirical support for the main claims along multiple axes: SCDM routing behavior, cross-dataset component decomposition, routing alternatives and sparsity sensitivity, robustness-oriented controls, structured corruptions, external adversarial checks, teacher-guided reduced-data controls, distillation controls, PTB ranking recovery, perturbation sensitivity, baseline coverage, and runtime cost. These analyses are intended to make the stability claim more falsifiable by exposing cases where ClinStab does not dominate every metric.

B SCDM Evidence and Low-Energy Stress Tests

We treat channel energy as a routing prior, not as a claim that high-energy channels are always clinically causal. Table 8 reports routing stability, gradient agreement, stream-removal sensitivity, and low-energy attenuation behavior.

Table 8: SCDM routing diagnostics. Higher is better for routing stability/agreement and low-energy robustness; larger negative drops indicate stronger dependence on the removed component.

Diagnostic	APAVA	TDBrain	ADFTD	PTB
Seed-wise salience-set Jaccard	0.71	0.76	0.83	0.69
Gradient top-channel agreement	0.64	0.68	0.74	0.61
Remove context stream: Δ Acc.	-1.49	-1.26	-0.55	-1.37
Remove salience stream: Δ Acc.	-2.08	-1.73	-0.92	-1.84
Low-energy attenuation gain vs. Medformer	+3.8	+2.4	+1.1	+1.6

The stream-removal rows test whether the context and salience pathways carry complementary information rather than serving as redundant architectural decoration.

B.1 Cross-Dataset Architecture and Objective Decomposition

Table 9 separates the same-backbone objective-only and SCDM-only controls across datasets. The pattern supports a conservative interpretation: SCDM contributes beyond the stability objective on all four datasets, but the increment is modest, so the contribution should be read as a compatible architectural bias rather than the sole source of improvement.

Table 9: Cross-dataset decomposition of SCDM and stability losses. Values are accuracies.

Dataset	Medformer	Med.+loss	SCDM only	Full ClinStab	Full–Med.+loss
APAVA	76.99	78.21	77.85	**79.34**	+1.13
TDBrain	88.08	89.02	88.74	**90.00**	+0.98
ADFTD	97.48	97.91	97.76	**98.29**	+0.38
PTB	79.84	81.28	80.72	**82.12**	+0.84

B.2 High-Amplitude Artifact Stress Test

The main risk of energy routing is that severe artifacts can be high-energy and therefore routed to the focal attention stream. Table 10 makes this failure mode explicit by injecting artifact-like events at test time and recording both performance degradation and the frequency with which the corrupted channel enters the salience set. The artifact is frequently selected, so this table should weaken any interpretation of energy as clinical salience. The intended evidence is narrower: stability losses explain part of the gain, SCDM adds a smaller increment beyond those losses, and soft routing is a plausible alternative for severe artifact regimes.

Table 10: High-amplitude artifact stress test averaged over APAVA, TDBrain, and PTB. Drops are accuracy-point decreases; lower is better.

Artifact	Med.	Med.+loss	Hard TopK	Soft gate	Artifact selected
EMG/motion burst	7.9	6.3	**5.7**	5.4	0.78
Electrode pop	9.2	7.4	**6.8**	6.2	0.86
Lead saturation	8.4	6.7	**6.1**	5.8	0.83
Burst + baseline drift	10.6	8.6	**7.9**	7.2	0.88

C Routing Criterion and Hard-Routing Sensitivity

Table 11 compares routing criteria under the same SCDM two-branch template. The random split tests whether any branch partition is sufficient; variance routing tests a different unsupervised amplitude proxy; learned gating tests a parameterized router; and soft energy gating replaces hard TopK with differentiable weights. Energy TopK is not uniformly best: it gives the strongest clean accuracy and APAVA $\sigma = 3.0$ accuracy, while soft energy gating gives the best AUROC. We therefore present hard TopK as a deterministic deployment choice and soft gating as the ranking-friendly variant.

Table 11: Routing criterion ablation on APAVA. Values are percentages except route-flip rate. Lower route-flip rate indicates more stable channel assignment across adjacent epochs.

Routing criterion	Accuracy	AUROC	Acc. at $\sigma = 3.0$	Route-flip rate
Random fixed split	77.42±1.88	82.91±2.54	57.80	0.00
Variance TopK	78.05±1.46	83.66±2.31	59.10	0.21
Learned gate	78.58±1.33	84.46±2.18	59.86	0.34
Soft energy gate	78.81±1.29	**85.61±2.12**	60.18	0.18
Energy TopK	**79.34±1.17**	85.48±2.59	**61.11**	0.16

Table 12 studies whether the hard boundary itself is brittle by interpolating between soft and hard energy routing. The hard TopK variant is competitive because the context branch still processes non-selected channels; the method is not equivalent to dropping residual channels. For clinical ranking-first settings, the soft gate is a reasonable default to evaluate.

Table 12: Soft-to-hard routing sensitivity on APAVA. τ_g is the routing temperature; lower values approach hard TopK.

Routing mode	τ_g	Accuracy	AUROC	Context-removal drop
Soft	2.0	78.32	84.70	-0.96
Soft	1.0	78.64	85.10	-1.18
Soft	0.5	78.81	85.61	-1.31
Hard TopK	–	**79.34**	85.48	-1.49

Table 13 sweeps the sparsity factor α in Eq. (3). Smaller α routes more channels to attention, increasing cost and making the split less selective; larger α makes the focal stream too sparse and increases dependence on the context branch. The displayed salience ratio is the nominal $1/\alpha$; the implementation uses the integer budget $k = \lfloor C/\alpha \rfloor$. The default $\alpha = 4.67$ is therefore a middle operating point rather than a universal constant.

Table 13: Sparsity-factor sensitivity on APAVA. Larger α means fewer channels enter the salience stream.

α	Nominal ratio $1/\alpha$	Accuracy	AUROC	Acc. @ $\sigma = 3$	Context-removal drop
2.00	0.50	78.70	84.91	60.05	-0.92
3.00	0.33	79.02	85.20	60.74	-1.22
4.67	0.21	**79.34**	85.48	**61.11**	-1.49
6.00	0.17	78.86	**85.52**	60.42	-1.63
8.00	0.13	78.10	84.87	59.55	-1.82

D Robustness-Oriented Controls and Structured Corruptions

The main baseline table evaluates broad sequence-model competitiveness, while Table 14 isolates controls that modify Medformer with robustness-oriented training or augmentation. This separation makes the comparison less dependent on generic forecasting backbones and more directly tests whether ClinStab improves over same-backbone robustness variants.

D.1 External Adversarial Evaluation

To reduce dependence on the same AutoPGD inner loop used during training, Table 15 summarizes an external robustness check with PGD-20, an AutoAttack-style ensemble, and transfer attacks generated on Medformer. The results show modest but mostly consistent robust-accuracy improvements on APAVA and TDBrain, while the PTB adversarial metrics remain essentially tied. This pattern supports a bounded robustness claim rather than immunity under adaptive attacks.

Table 14: Robustness-oriented controls on APAVA. Higher is better except structured drop, where lower is better.

Model	Clean Acc.	AUROC	Acc. at $\sigma = 3.0$	Structured drop
Medformer	76.99	82.93	54.97	6.80
Medformer + Gaussian aug.	77.40	83.20	58.20	5.70
Medformer + channel-drop training	77.66	83.08	58.04	4.92
Medformer + ECG-style adv. reg.	78.02	83.74	59.10	5.30
Medformer + stability losses	78.21	84.12	59.92	5.00
SCDM only	77.85	83.55	58.62	5.40
Full ClinStab	**79.34**	**85.48**	**61.11**	**4.50**

Table 15: External adversarial robustness check. Values are accuracies.

Dataset	Model	Clean	PGD-20	AutoAttack-style	Transfer attack
APAVA	Medformer	76.99	52.8	50.6	57.4
APAVA	ClinStab	**79.34**	**58.9**	**55.8**	**62.1**
TDBrain	Medformer	88.08	63.7	61.9	69.2
TDBrain	ClinStab	**90.00**	**66.0**	**64.1**	**71.0**
PTB	Medformer	79.84	**65.1**	**63.4**	68.6
PTB	ClinStab	**82.12**	64.8	63.1	**69.4**

D.2 Structured Test-Time Corruptions

Acquisition-inspired interpretation. Channel/lead dropout approximates loss of contact or a missing sensor; baseline drift approximates slow electrode or respiratory wander; temporal masking approximates lost signal segments; resampling jitter approximates sampling-clock defects; and transient bursts approximate motion or contact artifacts. These are controlled synthetic proxies and do not establish physical equivalence to recorded clinical artifacts. Perturbation strength must be interpreted in the normalized input space.

Table 16 evaluates corruptions that better approximate acquisition failures than additive Gaussian noise alone. Channel dropout masks a random 20% of sensors/leads; baseline drift adds a low-frequency trend; temporal masking removes one contiguous 10% segment; resampling jitter perturbs the local sampling grid; burst noise corrupts one randomly selected channel. The table reports mean accuracy drop across APAVA, TDBrain, ADFTD, and PTB.

Table 16: Structured corruption robustness. Values are average accuracy drops in percentage points across four datasets; lower is better.

Corruption	Medformer drop	ClinStab drop	Relative reduction
Channel/lead dropout	6.80	**4.50**	33.8%
Baseline drift	5.60	**3.70**	33.9%
Temporal masking	6.20	**4.10**	33.9%
Resampling jitter	3.90	**2.80**	28.2%
Single-channel burst noise	7.50	**5.20**	30.7%

E Interpretation of Seed Variability

Mean and standard deviation alone do not establish a paired significance test. Published baseline summaries and local evaluations do not provide matched run-level observations, so we make no paired t-test, Wilcoxon, Holm, or Brown–Forsythe claim. No inference about significance follows from the five-seed summaries alone.

For the local Medformer/ClinStab values reported in Table 2, the descriptive accuracy standard deviation decreases on APAVA, ADFTD, and PTB but increases on TDBrain. On TDBrain, all six reported ClinStab standard deviations are higher than the corresponding local Medformer values. This is not a claim that initialization stability improves uniformly.

F Teacher-Guided Reduced-Data Controls

Table 17 separates the fixed-teacher result from teacher-free and subset-matched controls. The fixed-teacher setting is interpreted as teacher-guided robustification/compression rather than pure few-shot

learning. This design is most realistic when a central or offline teacher can be trained once, but the deployed student must be smaller, site-adapted, or explicitly robustified.

Table 17: Low-resource controls on APAVA accuracy.

Model/protocol	100% data	85% data	50% data	30% data
Medformer	76.99	76.45	72.15	68.83
SCDM only, no teacher	77.85	75.92	72.61	70.42
ClinStab, teacher-free	78.06	76.21	73.40	71.84
ClinStab, subset-matched teacher	78.44	76.98	74.61	72.91
ClinStab, fixed full-data teacher	**79.34**	**79.15**	**77.20**	**75.64**

G Distillation Controls Under Full-Data Training

Table 18 isolates the teacher-guided KL term under full-data training, rather than only under reduced-data training. The pattern is deliberately conservative: the teacher helps clean accuracy and several robustness metrics, but the PTB AUROC row supports the concern that teacher stabilization can sharpen the operating region while slightly worsening global ranking.

Table 18: Full-data distillation controls. Δ is full ClinStab minus the same model without teacher KL.

Dataset	Δ Accuracy	Δ AUROC	Δ Acc. @ $\sigma = 3$	Interpretation
APAVA	+0.72	+1.27	+1.10	teacher helps
TDBrain	+0.76	+0.41	+0.84	teacher helps
ADFTD	+0.23	+0.08	+0.31	small gain
PTB	+0.54	-0.52	+0.44	ranking tradeoff

H PTB Operating-Point Tradeoff

Ranking failure case. Improved accuracy at a chosen decision rule does not offset poorer ordering of positive and negative cases across thresholds. Lower AUROC/AUPRC may matter under threshold shifts and different screening prevalence. Calibration and ranking are different properties, so a lower calibration error would not by itself remove this failure mode. The full default is therefore not established as uniformly preferable for screening or triage.

Table 5 in the main paper reports the compact operating-point summary. The additional ranking-recovery sweep below studies whether lowering robustness and consistency pressure can recover part of the AUROC/AUPRC gap.

I PTB Ranking-Recovery Sweep

The ranking-recovery analysis is summarized in Table 6; this appendix retains the surrounding protocol map and deployment interpretation.

J Protocol Sensitivity

The protocol-sensitivity analysis is summarized in Table 7; this appendix retains the surrounding reproducibility details and data-source notes.

K Reproducibility Details

This appendix records the experimental protocol used for the reported analyses.

Datasets and splits. APAVA, TDBrain, and PTB use subject-independent splits, so no subject identifier appears in both training and test partitions. ADFTD follows the subject-dependent split used by the benchmark protocol. For all datasets, model selection uses the validation split only; the test set is touched once per seed for final reporting. All normalization statistics are fitted on the training split and reused for validation/test data to avoid leakage.

Preprocessing. EEG segments are converted into one-second windows at the sampling rate reported in Table 1. PTB ECG is segmented at heartbeat level following the Medformer protocol. We apply per-channel z-normalization with training-split statistics, preserve channel order, and do not use label information during filtering or segmentation. Missing or corrupted windows are removed before split construction, not after, to avoid split-specific censoring.

Optimization and seeds. All baselines and ClinStab are trained in the same Time-Series-Library codebase with seeds $\{41, 42, 43, 44, 45\}$. Unless otherwise specified, training uses AdamW, learning rate 10^{-4}, batch size 64, weight decay 10^{-5}, cosine annealing, and 100 epochs. The reported mean/std values are computed over the five seed runs.

SCDM routing implementation. For each granularity, channel-aligned embeddings are reshaped to $\mathbb{R}^{N_i \times C \times d_h}$ before routing. The energy score in Eq. (2) is recomputed in every forward pass. TopK indices are selected by deterministic sorting, detached from gradient computation, and used to gather salience/context tensors. Autograd therefore updates the values processed by attention and convolution, but does not differentiate through the discrete index ranking. The residual path and context stream ensure that non-salient channels remain trainable.

Table 19: Key hyperparameters for ClinStab.

Param	Value	Param	Value
Batch Size	64	Adv. Budget (ϵ)	0.20
Learning Rate	1e-4	Loss λ_{rob}	1.0
Epochs	100	Loss $\lambda_{con/distill}$	0.5
Optimizer	AdamW	SCDM Ratio α	4.67

Teacher-guided training. The teacher is a frozen Medformer trained on the clean training split for each dataset. In reduced-data experiments, the student sees only the reduced subset, while the fixed teacher provides logits from the clean training protocol. This setup tests teacher-guided robustification/compression, not a strict few-shot setting. It is intended to model cases where a clean-data teacher can be trained centrally or offline, but the deployed student must be smaller, site-adapted, or explicitly robustified. A teacher-free student row is included in Table 4 as the required control for disentangling architecture and teacher effects.

Perturbation protocol. AutoPGD is used during stability optimization to generate bounded perturbations around each training sample. The main configuration uses normalized ℓ_∞ budget $\epsilon = 0.20$ after per-channel z-normalization, $\rho = 0.65$, one restart, and 10 inner steps. Gaussian-noise evaluation injects zero-mean noise with amplitudes $\sigma \in \{0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ at test time only.

Statistical interpretation and retained artifacts. Published baseline summaries are attributed in Table 2. No paired significance claim is made without matched run-level records. The appendix gives the training, split, perturbation, and evaluation settings used for the reported analyses.

L Additional Robustness Results

Tables 20 and 21 expand the Gaussian-noise stress test beyond APAVA. The pattern is strongest on TDBrain, moderate on ADFTD, and mixed on PTB, where accuracy is comparable at extreme noise but precision remains higher.

Table 20: Accuracy under Gaussian test-time noise. Values are percentages.

Dataset	Model	$\sigma = 0.0$	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 3.0$	$\sigma = 4.0$
APAVA	Medformer	76.99	72.60	63.10	54.97	53.30
APAVA	ClinStab	**79.34**	**75.02**	**68.47**	**61.11**	**57.23**
ADFTD	Medformer	96.41	**46.69**	31.54	29.57	29.18
ADFTD	ClinStab	**96.61**	46.47	**31.90**	**30.94**	**30.91**
TDBrain	Medformer	88.04	85.12	76.85	65.46	58.56
TDBrain	ClinStab	**89.62**	**86.38**	**77.75**	**66.81**	**60.50**
PTB	Medformer	79.76	75.76	70.73	**68.43**	**67.65**
PTB	ClinStab	**81.17**	**77.80**	**71.64**	68.28	67.45

Table 21: Precision under Gaussian test-time noise. Values are percentages.

Dataset	Model	$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 3.0$	$\sigma = 4.0$
ADFTD	Medformer	96.19	**51.36**	39.90	36.49	34.92
ADFTD	ClinStab	**96.47**	51.08	**40.85**	**37.23**	**35.44**
TDBrain	Medformer	88.16	85.36	77.25	66.20	59.37
TDBrain	ClinStab	**89.72**	**86.47**	**77.79**	**66.95**	**60.69**
PTB	Medformer	**87.14**	**85.49**	82.37	76.43	67.65
PTB	ClinStab	85.88	85.35	**82.98**	**79.65**	**79.00**

The per-metric standard deviations are reported in Table 2; the diagnostic summaries above use the corresponding means for compact comparison.

M AutoPGD and Stability-Objective Sensitivity

Table 22 records the AutoPGD hyperparameter sensitivity study on APAVA. The results support the chosen perturbation setting and show that excessively weak or aggressive perturbation settings reduce the clean-robustness tradeoff.

Table 22: AutoPGD sensitivity on APAVA.

Parameter	Value	Accuracy	F1	AUROC
ϵ	0.05	78.41±0.70	75.92±0.99	83.31±1.52
ϵ	0.10	79.11±0.78	**79.51±0.97**	**87.28±1.61**
ϵ	0.20	**79.34±1.17**	76.56±1.50	85.48±2.59
ϵ	0.30	78.48±0.61	76.28±0.72	86.25±1.31
ρ	0.55	77.01±0.72	73.42±1.08	82.47±1.38
ρ	0.65	**79.34±1.17**	76.56±1.50	**85.48±2.59**
ρ	0.75	78.89±1.12	76.72±1.08	83.92±1.40
ρ	0.85	78.41±1.03	**76.81±0.95**	80.10±1.10
init std	1.0	78.97±0.71	76.47±0.89	82.17±0.98
init std	2.0	**79.34±1.17**	76.56±1.50	**85.48±2.59**
init std	3.0	78.55±0.63	76.48±0.94	81.92±1.12
init std	4.0	78.83±0.89	**76.64±1.03**	81.94±1.21

N Baseline Coverage Sanity Check

Table 23 adds non-Transformer baselines that are common in EEG or time-series classification. This appendix is not meant to replace the same-backbone Medformer comparison, but it answers whether ClinStab's gains only look strong because the baseline set is Transformer-heavy.

Table 23: Additional non-Transformer baseline coverage. Values are accuracies.

Model	APAVA	TDBrain	ADFTD	PTB
EEGNet [23]	72.84	84.91	93.44	–
ResNet1D/FCN [43]	75.18	86.77	95.31	80.44
InceptionTime [10]	76.12	87.30	96.18	81.03
Medformer [41]	76.99	88.08	97.48	79.84
ClinStab	**79.34**	**90.00**	**98.29**	**82.12**

O Ablation Definitions

The component ablation in Table 4 uses the following controlled variants:

- **Medformer:** original backbone with clean cross-entropy only.
- **Medformer + stability losses:** original backbone trained with the same robustness, consistency, and distillation losses as ClinStab.
- **SCDM only:** SCDM replaces Medformer interaction layers, but training uses clean cross-entropy only.
- **ClinStab w/o robustness:** removes perturbed-label cross-entropy while keeping clean supervision, consistency, and teacher KL.
- **ClinStab w/o consistency:** removes clean-vs-perturbed probability matching while keeping robustness and teacher KL.

- **ClinStab w/o KD:** removes teacher KL while keeping robustness and consistency.
- **Teacher-free ClinStab:** trains SCDM with robustness and consistency but no teacher at any stage.
- **Full ClinStab:** uses SCDM plus clean supervision, robustness, consistency, and teacher-guided stabilization.

P Runtime and Deployment Cost

Table 24 expands the cost analysis. The important claim is not that ClinStab is cheap, but that the cost is concentrated in training: the stability objective requires adversarial perturbation generation, whereas inference uses the same forward path as the trained student.

Table 24: Runtime tradeoff relative to Medformer.

Model	Train time / epoch	Total train time	Inference latency	Peak memory
Medformer	1.00×	0.92h	1.00×	1.00×
Medformer + stability losses	1.71×	1.58h	1.00×	1.08×
SCDM only	1.06×	0.98h	1.03×	1.04×
Full ClinStab	1.82×	1.67h	1.04×	1.11×

Q Societal Impact and Use Boundaries

ClinStab is intended as a robustness-oriented modeling study for physiological time-series classification, not as an autonomous diagnostic device. Its positive impact would come from making EEG/ECG models less sensitive to evaluated acquisition artifacts and initialization variance. The main risks are over-reliance on benchmark performance, hidden subgroup or site shifts, privacy exposure through clinical data handling, and inappropriate use without clinician oversight. Any deployment would require prospective validation, site-specific calibration, monitoring for subgroup performance gaps, and clear reporting that the salience/context routing is a modeling proxy rather than a validated biomarker detector.

R Data Sources

The datasets are cited at first use. Dataset sources and preprocessing procedures are described in this paper, and dataset access remains subject to the terms of the respective providers. APAVA is accessed through its public OSF source; the public node does not specify a license for derived processed files, so no APAVA data are redistributed. TDBRAIN requires individual access through the Brainclinics provider and is not redistributed. ADFTD is identified as OpenNeuro `ds004504` (version 1.0.6, CC0 in the dataset metadata). PTB is accessed through PhysioNet version 1.0.0 under the Open Data Commons Attribution License (ODC-By) 1.0. The exact historical data versions and access records are not asserted beyond these provider records.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the SCDM module, the stability-oriented losses, the four-dataset evaluation scope, and the PTB ranking tradeoff rather than claiming universal clinical robustness.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 and Appendix Q discuss routing, synthetic perturbations, PTB ranking, teacher-guided training, and clinical-use boundaries.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is method- and experiment-focused and does not state formal theoretical theorems.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The manuscript specifies the model architecture, datasets and splits, preprocessing, optimization, perturbation protocols, seeds, metrics, and supplementary analyses. The public repository provides runnable code and README instructions for the released examples; because the historical execution artifacts are not redistributed, the release supports implementation and protocol verification without claiming one-command reproduction of every reported numerical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The public repository provides the ClinStab/Medformer implementation and README instructions for the released examples. Dataset access remains subject to provider terms, and the release is scoped to the runnable implementation and documented protocol rather than one-command reproduction of every supplementary analysis.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix K specifies the data splits, preprocessing, optimizer, loss weights, teacher protocol, perturbation settings, seeds, and evaluation procedure used for the reported analyses.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported as mean and standard deviation over five seeds, and the paper states that these summaries are descriptive. Unsupported paired significance and variance-test claims are not used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper reports NVIDIA RTX 4090 hardware and relative runtime measurements for the reported setup; the release is scoped to the implementation and documented runtime context rather than a per-run resource ledger.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The authors confirm compliance with the NeurIPS Code of Ethics. The work analyzes public or permission-accessible benchmark datasets and does not conduct new human-subject data collection.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix Q discusses potential safety benefits, over-reliance risks, subgroup and acquisition-shift concerns, and the need for clinician oversight.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk general-purpose model, scraped dataset, or dual-use generative system; clinical-use boundaries are still described in Appendix Q.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the source publications and dataset providers at first use and records the applicable public license or access restriction for each dataset in the Data Sources section. The repository README records access requirements, and upstream code licenses and copyright notices are retained in the code release.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The basic method training/evaluation code is released with a README covering setup, data preparation, and execution. The paper documents the supplementary analyses and their settings.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not include crowdsourcing experiments or newly collected participant-study instructions.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper analyzes public benchmark datasets and does not conduct a new human-subject data-collection protocol.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Any LLM-based assistance was limited to language editing and formatting of the manuscript and was not part of the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.