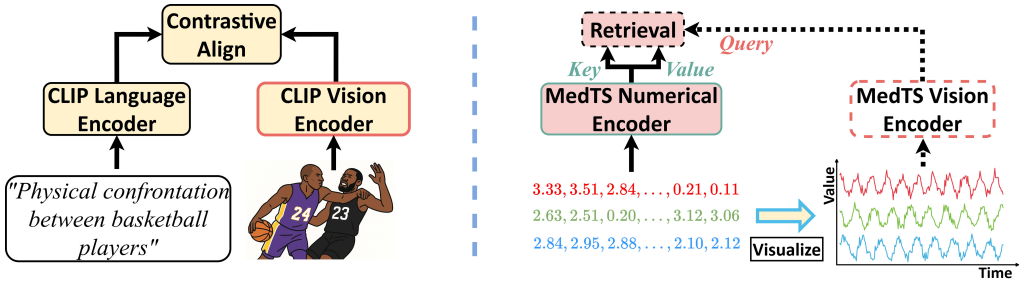
Revitalizing Medical Time Series with Vision-Informed Retrieval: A Vision-Language Perspective

Guoqi Yu **Juncheng Wang** **Shujun Wang**✉
Department of Biomedical Engineering and Sports Technology
The Hong Kong Polytechnic University
✉ Correspondence to: Shujun Wang (e-mail: shu-jun.wang@polyu.edu.hk)



(a) Vision-language aligment in VLMs (e.g., CLIP) *(b) Injecting vision priors into MedTS modeling*

Figure 1: Existing Medical Time Series (MedTS) models operate primarily on numerical sequences, overlooking the waveform morphology that clinicians rely on. We propose to use a frozen CLIP Vision Encoder to turn a deterministic waveform rendering into a morphology-aware *Vision Query*, which retrieves relevant temporal and channel evidence from numerical MedTS features.

Abstract

Medical time series (MedTS) underpin many clinical classification tasks, yet existing methods usually represent them only as numerical sequences and underuse the morphology that is explicit in waveform inspection. To bridge this gap, we introduce Vision-Informed Retrieval (ViRe), which uses a frozen VLM-derived waveform representation as a morphology-aware Query to guide retrieval from raw numerical MedTS features. Specifically, a Vision Query is extracted using pre-trained vision-language models (VLMs) to obtain morphology-aware priors from waveform plots. A tailored attention-based cross-modal retrieval mechanism then uses the Vision Query to select morphology-relevant temporal and channel evidence from the numerical representation. ViRe demonstrates strong effectiveness against ten established baselines, yielding an overall **6.42%** relative improvement over the previous state of the art across six public benchmarks. Code, training scripts, and reproducibility materials are publicly available in the **GitHub Repo**.

1 Introduction

Medical time series (MedTS), such as electrocardiography (ECG) [1] and electroencephalography (EEG) [2], provide continuous records of heart and brain activity, forming the cornerstone of modern diagnostics. Many routine clinical tasks, including epilepsy detection [3], sleep staging [4], arrhythmia screening [5], and cardiovascular risk stratification [6], are generally cast as MedTS classification. These tasks depend on both localized temporal events and relationships among channels.

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

The advent of artificial intelligence enables fast, automated classification, greatly improving diagnostic efficiency. Early machine learning relied on extracting handcrafted features (*e.g.*, band power, Hjorth parameters) from numerical input [7, 8]. Deep learning shifted this traditional paradigm, enabling convolutional and recurrent networks to predict directly from raw numerical data [9, 10]. More recently, Transformer models further improved performance by capturing longer temporal dependencies [11]. Across these architectures, however, each record remains a numerical sequence or matrix. Yet the same recordings are also inspected as waveforms by clinicians, where local shape and cross-channel structure are directly visible to the trained eye.

Despite the architectural advances, prevailing models adopt a reductive premise: *they treat MedTS purely as numerical sequences and underuse the morphology-centric structure that is explicit in waveform inspection*. In real clinical workflows, however, raw EEG and ECG recordings are visually reviewed by neurologists or cardiologists [12]. They summarize waveform morphology (*e.g.*, P–QRS–T complexes, ST-segment deviations, epileptiform discharges, and cross-channel co-activations) into structured reports and quantitative indices for referring clinicians [13, 14]. Although clinicians primarily consume these reports rather than raw waveforms, visual waveform patterns remain crucial for establishing the diagnostic criteria of MedTS [15, 16]. By focusing solely on raw numerical matrices, current MedTS models largely overlook this morphology-centric perspective and underutilize the strong inductive structure tied to waveform shape, making it harder to reach diagnoses that align with expert reasoning [11]. *Bridging this mismatch between numerics-oriented representations and waveform-based clinical reasoning is therefore critical for building MedTS models that are accurate, reliable, and clinically well-grounded.*

This mismatch points to an opportunity: since vision representations are essential for diagnosis, MedTS models should also reason over waveform visualizations, rather than rely entirely on numerical signals. However, training models that can extract such waveform representations aligned with human concepts from scratch is data-hungry and brittle under subject/device shifts [17, 18]. A natural remedy is to leverage an existing visual representation extractor that has already been trained to align image structure with human knowledge, and use it to provide compact, morphology-aware priors. This makes **Vision Language Models** (VLMs) (*e.g.*, CLIP [19]) particularly attractive to our problem. Trained with natural-language supervision, their vision encoders are optimized to organize images in a space that is aligned with **human-describable concepts** [20, 21]. By feeding waveforms into these encoders, the resulting embeddings provide a semantically structured summary of image morphology, *e.g.,* trends, spikes, and fluctuations, serving as high-level *vision priors* [22]. We therefore ask, *can such vision priors reshape and augment MedTS representations learned from raw numerical signals?*

In this work, we answer this question by proposing **ViRe** (**Vi**sion-**I**nformed **Re**trieval), a clinically grounded framework that bridges how models and clinicians "see" MedTS. ViRe adopts a dual view: *(i) a numerical view* modeled by time series Transformers, and *(ii) a waveform view* encoded by a pre-trained ***frozen*** CLIP vision encoder. Inspired by vision-to-text alignment in image–text models, we treat the waveform embedding as a *Vision Query* that retrieves *evidence* from the numerical sequence [23]. Via a tailored attention-based retrieval, this Vision Query *selects* and aggregates morphology-relevant numerical tokens, while the temporal and channel features remain the retrieved evidence. Across six MedTS benchmarks, ViRe consistently surpasses strong baselines such as Medformer [11] by **6.42%** on average. Our contributions are threefold.

- We identify the mismatch between numerics-oriented representation learning of existing MedTS models and waveform-based clinical reasoning, which motivates a morphology-aware prior.
- We deterministically visualize MedTS as waveforms and extract high-level morphology-aware priors using a pre-trained *frozen* CLIP Vision Encoder without any fine-tuning.
- ViRe introduces asymmetric dual-path cross-attention, where a *VLM-derived global Vision Query* retrieves temporal and channel evidence from numerical MedTS tokens.

2 Related Work

Medical Time Series Analysis. Recent MedTS classification has progressed from handcrafted feature pipelines to deep representation learning [24, 25]. Convolutional and recurrent models learn directly from raw EEG/ECG, while recent Transformers further capture long-range temporal dependencies and inter-channel interactions with attention [9, 11].

Yet prevailing methods remain mainly numerical, treating MedTS as sequences or matrices and learning from signal values alone, without any explicit notion of waveform shape. Clinical interpretation is instead morphology-centric: EEG and ECG are reviewed as waveforms, where diagnostic cues are recognized visually and summarized into structured reports and quantitative indices [26, 27, 14, 12]. This mismatch limits current classifiers; ViRe instead uses visual morphology as a retrieval prior that selects clinically relevant evidence from the numerical representation.

Vision-Language Models for Time Series. Vision-Language Models (VLMs) (*e.g.*, CLIP [19], ViLT [28]) learn concept-aligned visual embeddings through large-scale image-text pretraining. This makes VLMs attractive for extracting compact morphology-aware priors from waveform plots, whose visual structures encode clinically meaningful patterns. Recent progress on VLMs for time series largely follows three directions. First, some works visualize time series and use vision/VLM representations for forecasting. ViTime [29] makes the image representation the primary modeling space, while Time-VLM [30] fuses visual/textual augmentations with temporal features. Another line explicitly aligns time series with their textual information in a CLIP-style manner, in which the textual source is processed using CLIP Language Encoder, for zero-shot recognition (*e.g.*, TS-CLIP) [31]. Finally, studies also investigate whether VLMs can classify time series competitively when fed with visualizations [32]. These directions assign distinct roles to visual and textual information: ViTime predicts in an image-centered space, Time-VLM fuses visual/textual augmentations with temporal features, TS-CLIP aligns time series with text, and visualization-based classifiers operate directly on rendered signals. ViRe uses a frozen CLIP waveform representation as a global Vision Query that retrieves temporal and channel tokens, linking visual morphology to numerical evidence.

3 Methodology

This section first introduces the preliminaries of the medical time series (MedTS) (*e.g.*, EEG and ECG) classification-related foundations. Then, we present the proposed **ViRe** (**V**ision-**I**nformed **Re**trieval) framework, which aligns the numerical MedTS signals with their vision priors.

3.1 Preliminaries

Problem Formulation. *Considering a MedTS sample $X \in \mathbb{R}^{T \times C}$, where T denotes the number of timestamps and C is the number of channels, our objective is to learn a function (model) that predicts the corresponding label $\widehat{Y} \in \mathbb{R}^{K}$. K denotes the number of classes, e.g., various disease types or different stages of the same disease, depending on the specific diagnostic task.*

MedTS naturally forms a hierarchical structure: each dataset contains multiple subjects (*patients*), whose records are organized into sessions (*clinical visits*), further divided into trials (*repeated measurements*), and finally into short segments (*samples*) that are fed to diagnostic models [33, 7, 5, 34, 35]. Clinicians, however, make decisions at the ***subject*** level. To align with the real-world practice, we follow the **Subject-Independent** setting [11, 36, 37]: *the splitting is performed by subject, and samples from the same patient are assigned exclusively to the training, validation, or test set*. This closely simulates real-world practice, where models must generalize to unseen patients, and therefore provides a clinically meaningful comparison across all methods.

3.2 The ViRe Framework

The proposed **ViRe** is illustrated in Figure 2. Given a MedTS sample $X \in \mathbb{R}^{T \times C}$, we first tokenize it into the **Temporal embedding** [11] and the **Channel embedding** [38] along two complementary axes, *i.e.*, the temporal and channel dimension. Two Transformer Encoders are applied to extract temporal and channel features from these embeddings, respectively. In parallel, the same MedTS is deterministically visualized into a waveform image and fed into a *frozen* **CLIP Vision Encoder** [19]. The frozen encoder summarizes the rendered waveform into a compact morphology-aware prior for numerical evidence retrieval. ViRe then performs ***Vision-Informed Retrieval*** to align the vision and numerical modalities: the vision feature is used as the shared *Query* (Q), while the Temporal and Channel features provide *Keys* and *Values* (K, V) to the corresponding cross-attention Encoder. This yields two vision-guided summaries of temporal and channel evidence. Their sum is projected to the classification logits, so the visual representation guides retrieval from both numerical token streams.

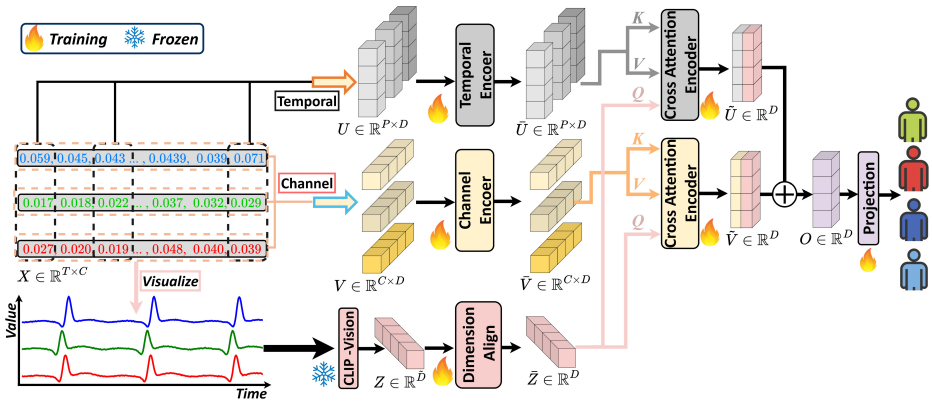


Figure 2: **Overview of ViRe.** Raw MedTS is embedded into *Temporal* and *Channel* tokens, while its waveform visualization is processed by a ***frozen*** CLIP Vision Encoder to produce a global *Vision Query*. Cross-attention Encoders retrieve vision-guided temporal and channel summaries, which are fused and projected to produce the final diagnostic prediction.

The frozen visual branch supplies only the retrieval Query while all Keys and Values remain numerical, and the same Query conditions both tokenizations. Thus, vision selects *what evidence to retrieve* without becoming an independent diagnostic pathway of its own.

Numerical modeling. ViRe represents each sample $X \in \mathbb{R}^{T\times C}$ through two complementary views: a **Temporal embedding** that captures dynamics across timestamps and a **Channel embedding** that preserves channel-wise semantics and explicitly models inter-channel dependencies.

Along the temporal axis, we split the signal into non-overlapping segments (with length L) and treat each segment across all channels as a token:

$$\begin{aligned} U_{i,:} &= \mathrm{vec}\big(X_{(i-1)L:iL,:}\big)W_t + b_t + W^{tpos}_{i,:}, \\ & i = 1,\dots,P, \quad P = \left\lceil \tfrac{T}{L} \right\rceil, \\ W_t &\in \mathbb{R}^{LC\times D}, \quad b_t \in \mathbb{R}^{D}, \quad W^{tpos} \in \mathbb{R}^{P\times D}. \end{aligned} \tag{1}$$

where $\mathrm{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ flattens a matrix into a vector. W^{tpos} is the position embedding [39]. Stacking all temporal tokens yields $U \in \mathbb{R}^{P\times D}$. Such tokenization eases the modeling of trend and seasonality patterns [40]. Orthogonally, we summarize each channel by aggregating its entire trajectory into a token:

$$\begin{aligned} V_{j,:} &= X_{:,j}^{\top} W_c + b_c + W^{cpos}_{j,:}, \quad j = 1,\dots,C, \\ W_c &\in \mathbb{R}^{T\times D}, \quad b_c \in \mathbb{R}^{D}, \quad W^{cpos} \in \mathbb{R}^{C\times D}. \end{aligned} \tag{2}$$

With another position embedding W^{cpos}, this forms the Channel embedding $V \in \mathbb{R}^{C\times D}$. Summarizing each channel's full trajectory in one token preserves channel-specific semantics, which is crucial for modeling inter-channel correlations [41–43]. Separate Transformer Encoders then produce the temporal feature $\bar{U} \in \mathbb{R}^{P\times D}$ and channel feature $\bar{V} \in \mathbb{R}^{C\times D}$ used as Keys and Values in retrieval.

Vision Query. To inject vision priors, we convert the same $X \in \mathbb{R}^{T\times C}$ into a waveform image via a deterministic visualization operator (detailed in Appendix D.3). It plots and stacks signals from all channels into a single image:

$$I = \mathcal{V}(X) \in \mathbb{R}^{H\times W\times 3}. \tag{3}$$

The image I is fed into a pre-trained *frozen* **CLIP Vision Encoder** $f_{\mathrm{CLIP}}(\cdot)$ [19]:

$$Z = f_{\mathrm{CLIP}}(I) \in \mathbb{R}^{\tilde{D}}, \tag{4}$$

yielding a high-level visual feature that embeds text–image alignment knowledge. A ***Dimension Align layer*** projects Z into the shared numerical latent space:

$$\bar{Z} = ZW_z + b_z, \qquad W_z \in \mathbb{R}^{\tilde{D}\times D}, \quad b_z \in \mathbb{R}^{D}. \tag{5}$$

The aligned feature $\bar{Z} \in \mathbb{R}^{D}$ serves as the global *Vision Query*. Sharing the latent dimension with $\bar{U}$ and $\bar{V}$ lets the query attend to both numerical views through a common cross-attention layer.

Cross-Modal Alignment. Given the Temporal embedding $\bar{U} \in \mathbb{R}^{P\times D}$, Channel embedding $\bar{V} \in \mathbb{R}^{C\times D}$, and the vision-derived Query $\bar{Z} \in \mathbb{R}^{D}$, ViRe performs Cross-Modal Alignment via two cross-attention Encoders [44–46]. We reshape the Vision Query into a shared token $Q \in \mathbb{R}^{1\times D}$; the Temporal and Channel embeddings provide the following two numerical Key–Value sets:

$$\begin{aligned} \widetilde{U} &= \mathrm{CrossAttn}(Q, \bar{U}, \bar{U}) \in \mathbb{R}^{1\times D}, \\ \widetilde{V} &= \mathrm{CrossAttn}(Q, \bar{V}, \bar{V}) \in \mathbb{R}^{1\times D}. \end{aligned} \tag{6}$$

We add the two retrieved summaries and project the result:

$$\begin{aligned} O &= \widetilde{U} + \widetilde{V} \in \mathbb{R}^{1\times D}, \\ \widehat{Y} &= OW_y + b_y, \qquad W_y \in \mathbb{R}^{D\times K},\ b_y \in \mathbb{R}^{K}. \end{aligned} \tag{7}$$

producing the final logits $\widehat{Y} \in \mathbb{R}^{K}$ for MedTS diagnosis. The same query is shared by both retrieval paths, so their outputs remain directly comparable before fusion while each path attends over a different numerical organization of the same underlying input waveform.

Since we employ a pre-trained *frozen* CLIP Vision Encoder as the vision feature extractor, training requires no auxiliary objectives. We use the standard classification objective, namely cross-entropy between the predicted labels $\widehat{Y}$ and the ground-truth Y [47, 48]. During training, one augmentation operation per input is applied consistently to the signal and its rendering (Appendix C); at validation and test time, only the deterministic rendering pipeline is used.

4 Experimental Results

4.1 Experiment Setting

4.1.1 Datasets

We evaluate six subject-disjoint public benchmarks spanning EEG and ECG. **APAVA** [37] and **ADFTD** [49] cover Alzheimer's-related EEG classification; **TDBrain** [50] targets Parkinson's disease from EEG. For ECG, **PTB** [51] evaluates myocardial infarction, **PTB-XL** [35] provides five diagnostic classes, and **MIMIC** [34] evaluates heart disease versus healthy controls.

Splits are subject-disjoint, so every test sample comes from an unseen individual. Table 1 summarizes the statistics, while Appendix B gives segmentation and split details. The six benchmarks span EEG and ECG, varied cohort sizes, and different channel configurations, testing one retrieval mechanism without dataset-specific changes. EEG recordings are segmented into 1-second windows at 256 Hz, PTB-XL into 1-second windows at 250 Hz, and PTB and MIMIC into R-peak-aligned single heartbeats, following the Medformer preprocessing protocol [11].

Table 1: **Dataset statistics after preprocessing.** Length is the input sequence length.

Modality	Dataset	Subjects	Samples	Classes	Channels	Length
EEG	APAVA	23	5,967	2	16	256
	ADFTD	88	69,752	3	19	256
	TDBrain	72	6,240	2	33	256
ECG	PTB	198	64,356	2	15	300
	PTB-XL	17,596	191,400	5	12	250
	MIMIC	20,437	204,370	2	12	250

4.1.2 Baselines

We compare ViRe with ten representative architectures spanning complementary design families: Autoformer and FEDformer for decomposition/frequency modeling [52, 53]; Informer, Reformer, and Transformer for attention variants [54, 55, 39]; MTST, Nonformer, iTransformer, and PatchTST for patch-, non-stationary-, and channel-aware modeling [56, 57, 38, 58]; and Medformer as the dedicated state-of-the-art baseline for medical time series classification [11]. All baselines share the same encoder depth, model dimension, and optimizer (Appendix D.1).

Table 2: **Subject-independent classification summary.** Per-dataset Avg over Accuracy, Precision, Recall, F1, AUROC, and AUPRC. Full mean±std results are in Appendix E; **best** is bolded and *second* is blue italics.

	EEG			ECG			
Model	**APAVA**	**ADFTD**	**TDBrain**	**PTB**	**PTB-XL**	**MIMIC**	**Mean**
Autoformer	70.71	46.43	89.52	70.86	57.53	79.69	69.12
FEDformer	77.21	48.20	80.98	75.90	56.83	86.79	70.98
Informer	71.36	50.04	91.64	80.62	67.84	86.94	74.74
iTransformer	77.23	51.71	77.63	*84.95*	64.45	87.11	73.85
MTST	69.94	47.89	79.35	76.80	68.70	87.52	71.70
Nonformer	70.70	50.94	91.07	79.58	66.78	86.35	74.24
PatchTST	65.89	45.56	81.94	75.35	**69.90**	86.79	70.91
Reformer	77.02	53.19	90.84	79.51	68.04	*87.90*	76.08
Transformer	74.42	52.09	90.34	78.69	66.73	87.01	74.88
Medformer	*79.74*	*54.63*	*91.91*	84.69	69.28	87.14	*77.90*
ViRe	**92.79**	**59.83**	**95.54**	**89.08**	*69.47*	**90.68**	**82.90**

4.1.3 Implementation

We report Accuracy, Precision, Recall, F1-Score, AUROC, and AUPRC, use F1-Score for early stopping, and average five random seeds. All experiments run on one NVIDIA RTX 4090 GPU. Baselines are reproduced through the Medformer benchmark [11] using identical subject-level splits, metrics, and early-stopping criteria. Model selection uses validation subjects only; full hyperparameters, preprocessing details, and code availability are given in Appendix D.

4.2 Classification Performance

Table 2 provides one Avg per dataset, defined as the arithmetic mean of the six classification metrics. ViRe ranks first on five of six benchmarks. Relative to Medformer, Avg improves from 79.74 to 92.79 on APAVA (**16.37%**), 54.63 to 59.83 on ADFTD (**9.52%**), 91.91 to 95.54 on TDBrain (**3.95%**), 84.69 to 89.08 on PTB (**5.18%**), and 87.14 to 90.68 on MIMIC (**4.06%**). On PTB-XL, PatchTST leads with 69.90, followed closely by ViRe at 69.47 and Medformer at 69.28.

Across all six datasets, ViRe reaches an overall mean Avg of **82.90**, versus 77.90 for Medformer and 76.08 for Reformer. Against the strongest non-ViRe entry, ViRe leads APAVA, ADFTD, TDBrain, PTB, and MIMIC by 13.05, 5.20, 3.63, 4.13, and 2.78 points, respectively; on PTB-XL it ranks second, only 0.43 points below PatchTST. Full metric-wise mean±std results are provided in Appendix E.

The gains span both EEG and ECG and persist from small cohorts to larger benchmarks, indicating that the benefit is not tied to a single modality or dataset scale. PTB-XL is the main exception: ViRe remains competitive but does not surpass PatchTST, suggesting that morphology-aware retrieval complements rather than uniformly dominates strong numerical patch modeling.

Two cross-dataset patterns are especially informative. The largest gains occur on APAVA and ADFTD, the two smaller EEG cohorts, where a stable morphology prior can be valuable when subject-level supervision is limited. At the same time, positive improvements persist on TDBrain, PTB, and MIMIC despite substantial differences in channel count, cohort size, and signal modality. Together with the near-tie on PTB-XL, this pattern supports ViRe as an inductive bias for evidence selection rather than a dataset-specific capacity increase: its benefit is strongest when waveform morphology helps isolate diagnostic evidence that a numerical encoder alone may not emphasize.

4.3 Model Analysis

We isolate where the gain comes from by varying one factor at a time while keeping the numerical pipeline fixed: **(i)** the retrieval Query, which tests whether the benefit stems from the retrieval mechanism or from the VLM-derived Query content; **(ii)** the fusion operator, comparing cross-attention retrieval with addition and concatenation; **(iii)** the vision backbone, which contrasts random, ImageNet, and CLIP initializations; and **(iv)** the training-data proportion, which probes data efficiency. Studies (i)–(iii) use two EEG (ADFTD, APAVA) and two ECG (PTB, MIMIC) datasets and report Accuracy, F1-Score, and the average relative improvement over the variant without retrieval (w/o).

4.3.1 Ablation of Retrieval Query

Table 3: ***Ablation of the retrieval Query.*** Zero replaces the Vision Query with an all-zero vector, and Gaussian replaces it with a vector sampled from a Gaussian distribution.

	ADFTD		APAVA		PTB		MIMIC		
Query Type	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Avg. Gain**
w/o	54.79±1.81	51.73±2.22	83.86±2.87	82.45±3.69	81.45±3.44	74.58±6.89	84.92±1.11	84.81±1.12	–
Zero	54.72±2.82	51.23±2.72	84.34±1.72	83.73±1.96	84.62±2.55	80.21±3.96	86.13±0.14	86.06±0.13	1.92%
Gaussian	54.46±2.21	51.54±1.93	82.89±2.66	81.65±3.12	82.83±2.15	77.34±3.64	86.50±0.11	86.44±0.11	0.76%
Vision	**57.83**±1.82	**54.02**±2.22	**91.43**±1.03	**91.19**±1.03	**88.26**±1.12	**85.81**±1.52	**88.61**±0.12	**88.54**±0.12	**7.72%**

Table 3 evaluates the impact of different retrieval Query. Removing retrieval (w/o) gives the worst results. Content-free retrieval queries provide modest gains: an all-zero Query (Zero) improves the average by 1.92%, and a Gaussian-initialized Query (Gaussian) by 0.76%. The Vision-Informed Query, derived from the VLM vision-language space, achieves the strongest improvement (7.72%). For example, on APAVA, Accuracy increases from 83.86 to 91.43 (9.03%) and F1 from 82.45 to 91.19 (10.60%). ***The results identify the Vision Query as the main source of the retrieval gain, providing a structured VLM-derived prior for evidence selection.***

4.3.2 Ablation of Modality Fusion

Table 4: ***Ablation of modality fusion.*** Cross-attention retrieval is compared with simple addition (*Add*) and concatenation (*Concat*) of the vision prior and the numerical features.

	ADFTD		APAVA		PTB		MIMIC		
Fusion Strategy	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Avg. Gain**
w/o	54.79±1.81	51.73±2.22	83.86±2.87	82.45±3.69	81.45±3.44	74.58±6.89	84.92±1.11	84.81±1.12	–
Add	56.67±1.46	53.53±2.12	85.95±1.77	85.45±1.71	84.81±2.17	81.26±3.47	87.57±0.06	87.49±0.07	4.05%
Concat	56.86±1.54	53.55±1.21	84.23±2.55	83.02±3.11	83.60±2.25	79.13±3.47	87.88±0.18	87.81±0.18	3.02%
Retrieval	**57.83**±1.82	**54.02**±2.22	**91.43**±1.03	**91.19**±1.03	**88.26**±1.12	**85.81**±1.52	**88.61**±0.12	**88.54**±0.12	**7.72%**

Table 4 studies how different fusion strategies impact ViRe. The w/o variant simply adds the Temporal and Channel branches without using external priors, and gives the poorest performance. Replacing it with naive multimodal fusion already helps: both simple Add (4.05%) and Concat (3.02%) consistently outperform w/o on all datasets. Add and Concat show that vision priors already improve numerical modeling under simple fusion. Cross-attention retrieval further raises the overall gain to 7.72%. Its token-wise, content-dependent interaction provides the strongest integration of vision priors with numerical temporal and channel evidence across all four datasets.

4.3.3 Ablation of Vision Backbone

Table 5: ***Ablation of the vision backbone:*** Random ViT, ImageNet ViT, and CLIP-Vision.

	ADFTD		APAVA		PTB		MIMIC		
Vision Backbone	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Avg. Gain**
w/o	54.79±1.81	51.73±2.22	83.86±2.87	82.45±3.69	81.45±3.44	74.58±6.89	84.92±1.11	84.81±1.12	–
ViT (Random Init)	37.39±3.11	27.34±1.92	80.62±0.60	77.18±0.54	80.65±1.68	73.50±2.79	83.67±0.14	83.58±0.13	-11.81%
ViT (ImageNet)	52.21±0.73	50.55±1.13	84.48±2.80	84.86±3.67	82.06±3.54	75.72±5.83	86.52±0.21	86.42±0.21	0.34%
CLIP-Vision	**57.83**±1.82	**54.02**±2.22	**91.43**±1.03	**91.19**±1.03	**88.26**±1.12	**85.81**±1.52	**88.61**±0.12	**88.54**±0.12	**7.72%**

Table 5 ablates how different vision backbones affect ViRe. ***Simply plugging in a generic vision backbone (e.g., ViT) is insufficient for our task.*** Using a randomly initialized ViT even *hurts* performance, leading to an average degradation of 11.81%. Pretraining ViT on ImageNet recovers some performance but yields only a marginal gain (0.34%), indicating that pure-image pretraining brings limited benefit for our sequence-waveform alignment objective. In contrast, employing the ***CLIP Vision Encoder***, which is pretrained on text-image pairing tasks, delivers a clear advantage: it achieves the best performance on all datasets, improves the average performance by 7.72%, and outperforms the ImageNet-pretrained ViT by 7.11%. These results demonstrate that the gains of ViRe stem neither from the ViT architecture alone nor from generic vision pretraining, but from the ***cross-modal text-image alignment*** learned by VLMs, which forms a well-aligned vision-language latent space and produces human-describable vision priors for morphology-aware retrieval.

4.3.4 Generalizability Analysis

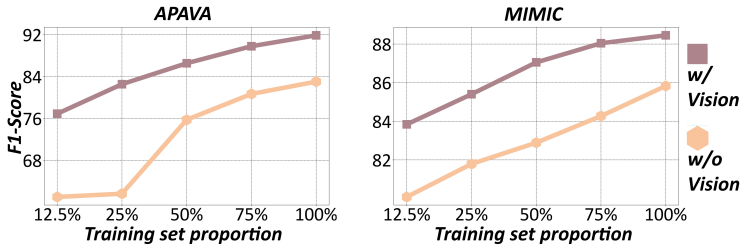


Figure 3: F1-Score with and without vision priors across training-set proportions.

We vary the training-set proportion without changing the model to probe the generalizability of the vision priors. Figure 3 shows that vision priors consistently improve F1-Score under all data regimes, with the largest separation under limited supervision. The frozen morphology prior therefore improves data efficiency and reduces reliance on abundant labeled subjects.

4.4 Further Analysis of Vision Feature

4.4.1 Analysis of Retrieved Features

Table 6: **Clustering quality of retrieved features.** Raw denotes no retrieval. Lower DBI and higher NMI, Homogeneity, and Completeness indicate better clustering.

	ADFTD				PTB				MIMIC			
Metrics/Features	Vision	Zero	Gaussian	Raw	Vision	Zero	Gaussian	Raw	Vision	Zero	Gaussian	Raw
DBI ↓	**4.034**	7.962	18.098	9.607	**1.156**	1.466	1.722	1.385	**0.893**	1.216	32.475	1.562
NMI ↑	**0.101**	0.054	0.004	0.042	**0.425**	0.345	0.117	0.194	**0.398**	0.327	0.002	0.276
Homogeneity ↑	**0.112**	0.055	0.004	0.040	**0.441**	0.371	0.122	0.202	**0.403**	0.332	0.002	0.274
Completeness ↑	**0.102**	0.053	0.003	0.039	**0.411**	0.332	0.113	0.186	**0.401**	0.329	0.003	0.281

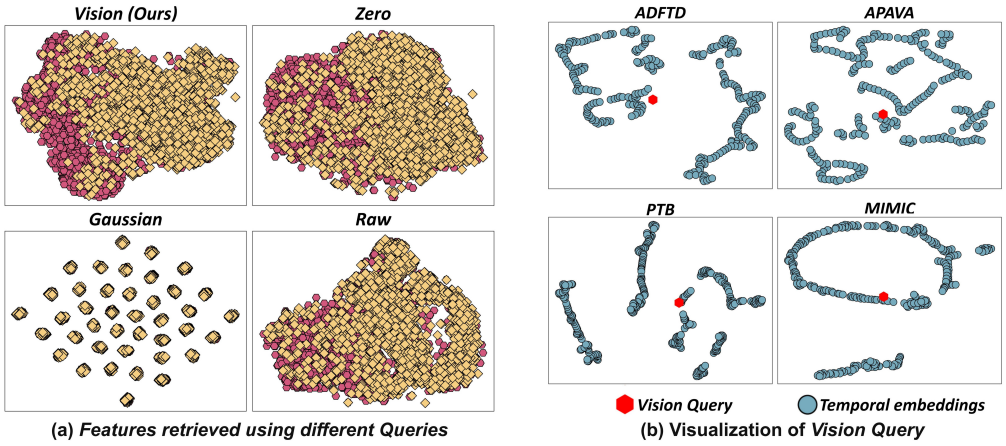


Figure 4: **(a)** t-SNE visualization of retrieved features under different queries. **(b)** t-SNE geometry of the Vision Query and the temporal embeddings across four datasets.

Table 6 and Figure 4**(a)** analyze the structure of the retrieved feature space under different Query types. Quantitatively, the vision-derived features consistently yield the lowest DBI and the highest NMI, Homogeneity, and Completeness, indicating well-formed and label-consistent clusters. In contrast, the Zero and Gaussian Queries substantially degrade clustering quality, with DBI increasing sharply and the supervised scores dropping close to zero in several cases, especially on ADFTD and MIMIC. The Raw features (no retrieval) preserve some structure but are clearly inferior to the vision-derived features. The t-SNE plots in Figure 4**(a)** echo these trends: with the ***Vision Query***, samples from different classes form compact, well-separated manifolds, whereas Raw features exhibit large overlapping regions. The Zero Query only partially improves separability, and the Gaussian Query destroys the class structure, producing a scattered and highly overlapped representation space.

4.4.2 Geometric Insight into the Vision Query

Figure 4**(b)** provides a geometric view of how the *Vision Query* interacts with the temporal embeddings. In all cases, the *Vision Query* is not an outlier; instead, it is embedded near the densest region of the representation manifold across datasets. This indicates that the vision-derived Query lies in a semantically meaningful region of the MedTS representation space, making it well-positioned to attend to and aggregate temporal information. Combined with the clustering results in Table 6, these findings suggest that the *Vision Query* provides a structured and semantically aligned entry point for cross-attention over numerical embeddings, and that retrieval emphasizes regions that are already informative in the numerical manifold rather than arbitrary directions of the space.

4.4.3 Attention Alignment with Clinical Morphology

Table 7: **Quantitative Attention Alignment on ECG Datasets.** MAR and CAR measure attention-density enrichment on high-curvature and QRS regions, respectively; higher is better.

	MAR ↑		CAR ↑	
Dataset	**ViRe**	**Gaussian**	**ViRe**	**Gaussian**
PTB	**1.81**	1.46	**1.65**	1.33
PTB-XL	**1.67**	1.38	**1.51**	1.27
MIMIC	**1.72**	1.41	**1.63**	1.35

We further quantify whether the retrieval attention concentrates on meaningful temporal regions of ECG. Following Appendix G.2, the Morphology-Aware Attention Ratio (MAR) measures attention-density enrichment on high-curvature timestamps, and the Clinical Alignment Ratio (CAR) measures enrichment on the clinically recognized QRS complex; both divide the attention density inside the target region by that in its complement, so values above one indicate concentration. As reported in Table 7, the CLIP-derived Vision Query consistently exceeds a non-semantic Gaussian Query on both metrics across PTB, PTB-XL, and MIMIC. The qualitative map in Appendix G.1 shows the same behavior on individual samples, where attention peaks at synchronous cross-lead variations.

4.5 Additional Clinical Evidence

PTB-XL diagnosis results and patient-matched controls. On the official PTB-XL test set, ViRe improves 11 of 13 diagnoses ($p = 0.0225$), with the largest gains on HYP and LVH. Zero masking of the visual input reduces macro-F1 by 6.19 points, while cross-patient visual shuffling produces a 15.61-point reduction (Table 8). Positive directions also span infarction, ST–T change, ischemia, and rhythm-related categories, extending the pattern beyond HYP and LVH. Repeated subject-disjoint splits confirm this pattern: ViRe wins 7/7 splits on PTB-XL, 8/10 on APAVA, and 9/10 on PTB, and the paired statistical evidence and confidence intervals are reported in Appendix I.

Table 8: **PTB-XL diagnosis results and patient-matched visual intervention controls.**

A. Diagnosis F1				
Diagnosis	**ViRe**	**Medformer**	**Gain**	**Patient-bootstrap 95% CI**
MI	72.21	71.35	+0.85	[-1.42,+3.08]
STTC	74.91	73.44	+1.47	[-0.99,+3.97]
HYP	**59.49**	44.17	**+15.32**	**[+10.39,+20.25]**
AMI	72.93	71.01	+1.92	[-1.23,+5.14]
ISCA	37.07	30.15	+6.92	[-2.03,+15.94]
LVH	**66.51**	46.73	**+19.78**	**[+13.98,+25.48]**
AFIB	67.08	62.61	+4.47	[-2.20,+10.98]
B. Visual intervention controls			Holm-adjusted $p = 0.0008$	
Intervention	**Macro-F1 drop**		**95% CI**	
Zero-mask visual input	**-6.19**		**[-8.06,-4.44]**	
Cross-patient visual shuffling	**-15.61**		**[-17.85,-13.28]**	

Concept decoding from the frozen vision prior. On **MEETI** [59], we probe the frozen visual representation with lightweight concept heads and obtain positive confidence intervals for all 22 supported attributes. For continuous attributes, ρ denotes Spearman's rank correlation coefficient; prolonged QT is evaluated by AUROC. Table 9 reports representative amplitude, interval, synchrony, and QT results, and Table 10 summarizes the remaining representation analyses.

Table 9: **Representative MEETI concept-decoding results** [59].

Attribute	Metric	CLIP	Random	Gain
Amplitude range	ρ	**0.903**	0.501	**+0.402**
PR interval	ρ	**0.633**	0.248	**+0.384**
Cross-lead synchrony	ρ	**0.684**	0.320	**+0.364**
Prolonged QT	AUROC	**0.904**	0.656	**+0.248**

Morphology-aware representation analysis. Because **MEETI** pairs ECG waveforms with images, measurements, and clinical reports, we further evaluate the frozen visual representation without retraining, testing alignment with clinical text, whether same-patient representation changes follow measured waveform changes, and whether nearby visual embeddings share morphology-related attributes. As summarized in Table 10, all 12 text directions improve, all 5 longitudinal measurements show positive gains in direction accuracy, and all 21 retrieval attributes have positive confidence intervals, even though the visual encoder is never trained on the benchmark labels.

Table 10: **Morphology-aware representation analyses on MEETI.**

Zero-shot text alignment	**12/12 directions**
Key results	AF AUROC **+0.423**; QTc ρ **+0.284**; amplitude **+0.239**; synchrony **+0.232**.
Same-patient natural changes	**5/5 measurements**
Key results	Direction accuracy: HR **+0.128**; PR **+0.154**; QRS **+0.127**; QT **+0.238**; QTc **+0.144**.
Raw-space retrieval	**21/21 positive CIs**
Key results	Bradycardia **+0.337**; HR **+0.331**; amplitude **+0.262**; QT/QRS **+0.256/+0.245**.

The interventions in Table 8 and the probes in Table 9 and Table 10 are complementary: replacing the visual evidence degrades diagnosis, whereas the frozen visual space already organizes recognizable waveform attributes, so the visual branch acts as a morphology-aware retrieval prior rather than a diagnostic shortcut that bypasses the numerical evidence altogether.

5 Conclusion

We introduce ViRe, a Vision-Informed Retrieval framework that complements numerical MedTS modeling with a morphology-aware prior from a frozen CLIP vision encoder, whose global *Vision Query* retrieves temporal and channel evidence by cross-attention. Across six public EEG and ECG benchmarks, ViRe achieves the best overall performance on five datasets and improves the strongest baseline, Medformer, by **6.42%** on average and by **16.37%** on APAVA.

Mechanism analyses favor vision-informed queries and CLIP-Vision over content-free queries and generic image pretraining, and PTB-XL diagnosis results and MEETI concept decoding link the visual prior to recognizable waveform structure. ViRe thus lets waveform morphology guide numerical feature selection, and we hope it encourages further use of frozen visual priors in clinical time series.

Acknowledgments and Disclosure of Funding

This work was partially supported by the Research Grants Council (RGC) of Hong Kong under the Collaborative Research Fund (CRF) (No. C5055-24G), the Start-up Fund of The Hong Kong Polytechnic University (No. P0045999), the Seed Fund of the Research Institute for Smart Ageing (No. P0050946), the Tsinghua-PolyU Joint Research Initiative Fund (No. P0056509), and the University Grants Committee (UGC) funding of The Hong Kong Polytechnic University (No. P0053716).

References

[1] Majd AlGhatrif and Joseph Lindsay. A brief review: history to understand fundamentals of electrocardiography. *Journal of community hospital internal medicine perspectives*, 2(1):14383, 2012.

[2] Michael X Cohen. Where does EEG come from and what does it mean? *Trends in neurosciences*, 40(4):208–218, 2017.

[3] Ihsan Ullah, Muhammad Hussain, Hatim Aboalsamh, et al. An automated system for epilepsy detection using EEG brain signals based on deep learning approach. *Expert Systems with Applications*, 107:61–71, 2018.

[4] Yingying Jiao, Yini Deng, Yun Luo, and Bao-Liang Lu. Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks. *Neurocomputing*, 408:100–111, 2020.

[5] Yanrui Jin, Zhiyuan Li, Mengxiao Wang, Jinlei Liu, Yuanyuan Tian, Yunqing Liu, Xiaoyang Wei, Liqun Zhao, and Chengliang Liu. Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis. *Communications Medicine*, pages 1–8, February 2024. ISSN 2730-664X. doi: 10.1038/s43856-024-00464-4.

[6] Rushuang Zhou, Lei Lu, Zijun Liu, Ting Xiang, Zhen Liang, David A Clifton, Yining Dong, and Yuan-Ting Zhang. Semi-supervised learning for multi-label cardiovascular diseases prediction: a multi-dataset study. *TPAMI*, 2023.

[7] Katerina D Tzimourta, Vasileios Christou, Alexandros T Tzallas, Nikolaos Giannakeas, Loukas G Astrakas, Pantelis Angelidis, Dimitrios Tsalikakis, and Markos G Tsipouras. Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: A systematic review. *International journal of neural systems*, 31 (05):2130002, 2021.

[8] Salah S Al-Zaiti, Christian Martin-Gill, Jessica K Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O Alrawashdeh, Richard E Gregg, Stephanie Helman, Nathan T Riek, Karina Kraevsky-Phillips, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, 2023.

[9] Aniqa Arif, Yihe Wang, Rui Yin, Xiang Zhang, and Ahmed Helmy. EF-Net: Mental state recognition by analyzing multimodal EEG-fNIRS via CNN. *Sensors*, 24(6):1889, 2024.

[10] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *ICLR*, 2021.

[11] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=jfkid2HwNr.

[12] Petr Nejedly, Vaclav Kremen, Vojtech Sladky, Jan Cimbalnik, Petr Klimes, Filip Plesinger, Ivo Viscor, Miroslav Pail, Jan Halamek, Bruce H. Brinkmann, Milan Brazdil, Pavel Jurak, and Gregory Worrell. Exploiting graphoelements and convolutional neural networks with long short term memory for classification of the human electroencephalogram. *Scientific Reports*, 9(1): 11383, 2019. doi: 10.1038/s41598-019-47854-6.

[13] F Lopes Da Silva. EEG analysis: theory and practice. *Electroencephalography: basic principles, clinical applications and related fields*, pages 1125–1159, 1999.

[14] Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the American college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *Journal of the American College of Cardiology*, 49 (10):1109–1127, 2007.

[15] Mustafa Aykut Kural, Lene Duez, Vibeke Sejer Hansen, Stefan Rampp, Reinhard Schulz, Hatice Tankisi, Richard Wennberg, Bo Bibby, Michael Scherg, and Sandor Beniczky. Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study. *Neurology*, 94 (20):e2139–e2147, 2020. doi: 10.1212/WNL.0000000000009439.

[16] William O. Tatum. EEG interpretation: Common problems. *Clinical Practice*, 9(5):527–538, 2012. doi: 10.2217/cpr.12.51.

[17] Cuong V Nguyen and Cuong D Do. Transfer learning in ECG diagnosis: Is it effective? *PloS one*, 20(5):e0316043, 2025.

[18] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in biology and medicine*, 141:105114, 2022.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[20] Shizhan Gong, Haoyu LEI, Qi Dou, and Farzan Farnia. Boosting the visual interpretability of CLIP via adversarial fine-tuning. In *ICLR*, 2025. URL https://openreview.net/forum?id=khuIvzxPRp.

[21] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (spliCE). In *NeurIPS*, 2024. URL https://openreview.net/forum?id=7UyBKTFrtd.

[22] Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of VLMs for chart understanding, 2025. URL https://arxiv.org/abs/2503.18435.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[24] Golshan Fahimi, Seyed Mahmoud Tabatabaei, Elnaz Fahimi, and Hamid Rajebi. Index of theta/alpha ratio of the quantitative electroencephalogram in Alzheimer's disease: a case-control study. *Acta Medica Iranica*, pages 502–506, 2017.

[25] Nagarajan Ganapathy, Ramakrishnan Swaminathan, and Thomas M. Deserno. Deep Learning on 1-D Biosignals: a Taxonomy-based Survey. *Yearbook of Medical Informatics*, 27(1):98–109, August 2018. ISSN 0943-4747. doi: 10.1055/s-0038-1667083.

[26] Mark H. Libenson. *Practical Approach to Electroencephalography*. Saunders/Elsevier, Philadelphia, 1 edition, 2010.

[27] Lawrence J Hirsch, Michael WK Fong, Markus Leitinger, Suzette M LaRoche, Sandor Beniczky, Nicholas S Abend, Jong Woo Lee, Courtney J Wusthoff, Cecil D Hahn, M Brandon Westover, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. *Journal of clinical neurophysiology*, 38(1):1–29, 2021.

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021.

[29] Luoxiao Yang, Yun Wang, Xinqi Fan, Israel Cohen, Jingdong Chen, and Zijun Zhang. ViTime: Foundation model for time series forecasting powered by vision intelligence. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=XInsJDBIkp.

[30] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *ICML*, 2025. URL https://openreview.net/forum?id=b5h60xQnzM.

[31] Ziwen Chen, Xiaoyuan Zhang, and Ming Zhu. TS-CLIP: Time series understanding by CLIP. In *EMNLP*, pages 4646–4664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.231.

[32] Vinay Prithyani, Mohsin Mohammed, Richa Gadgil, Ricardo Buitrago, Vinija Jain, and Aman Chadha. On the feasibility of vision-language models for time-series classification. *arXiv preprint arXiv:2412.17304*, 2024. doi: 10.48550/arXiv.2412.17304.

[33] Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *NeurIPS*, 36, 2024.

[34] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset (version 1.0). PhysioNet, 2023.

[35] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[36] Andreas Miltiadous, Emmanouil Gionanidis, Katerina D Tzimourta, Nikolaos Giannakeas, and Alexandros T Tzallas. DICE-net: a novel convolution-transformer architecture for Alzheimer detection in EEG signals. *IEEE Access*, 2023.

[37] J Escudero, Daniel Abásolo, Roberto Hornero, Pedro Espino, and Miguel López. Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091, 2006.

[38] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. *ICLR*, 2024.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[40] Xiyuan Zhang, Xiaoyong Jin, Karthick Gopalswamy, Gaurav Gupta, Youngsuk Park, Xingjian Shi, Hao Wang, Danielle C. Maddix, and Bernie Wang. First de-trend then attend: Rethinking attention for time-series forecasting. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=GLc8Rhney0e.

[41] Wenzhen Yue, Yong Liu, Xianghua Ying, Bowei Xing, Ruohao Guo, and Ji Shi. Freeformer: Frequency enhanced transformer for multivariate time series forecasting. *arXiv preprint arXiv:2501.13989*, 2025.

[42] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pages 1185–1196, 2025.

[43] Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *ICML*, 2025. URL https://openreview.net/forum?id=490VcNtjh7.

[44] Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting, 2025. URL https://openreview.net/forum?id=baSU1eVLwS.

[45] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=INAeUQO4lT.

[46] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, 2025.

[47] Arshia Afzal, Grigorios Chrysos, Volkan Cevher, and Mahsa Shoaran. REST: Efficient and accelerated EEG seizure analysis through residual state updates. In *ICML*, 2024. URL https://openreview.net/forum?id=9GbAea74O6.

[48] Jintai Chen, Kuanlun Liao, Kun Wei, Haochao Ying, Danny Z Chen, and Jian Wu. ME-GAN: Learning panoptic electrocardio representations for multi-view ECG synthesis conditioned on heart diseases. In *ICML*, pages 3360–3370. PMLR, 2022.

[49] Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023.

[50] 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.

[51] Ary L Goldberger, Luis AN 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.

[52] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *NeurIPS*, 34:22419–22430, 2021.

[53] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, pages 27268–27286. PMLR, 2022.

[54] 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 *AAAI*, volume 35, 2021.

[55] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2019.

[56] Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *International Conference on Artificial Intelligence and Statistics*, pages 4222–4230. PMLR, 2024.

[57] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *NeurIPS*, 35:9881–9893, 2022.

[58] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*, 2023.

[59] Deyun Zhang, Xiang Lan, Shijia Geng, Qinghao Zhao, Sumei Fan, Mengling Feng, and Shenda Hong. MEETI: A multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. *Scientific Data*, 13:527, 2026. doi: 10.1038/s41597-026-06796-1.

[60] Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-Patch: Hierarchical patch GNN for irregular multivariate time series. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 41494–41519, 2025.

[61] Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6):105:1–105:17, 2022. doi: 10.1145/3516367.

[62] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

[63] Shuhan Zhong, Weipeng Zhuo, Sizhe Song, Guanyao Li, Zhongyi Yu, and S.-H. Gary Chan. MTM: A multi-scale token mixing transformer for irregular multivariate time series classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 4074–4085, 2025. doi: 10.1145/3711896.3737058.

A Relationship to Numerical-Only Modeling

Setup. For a training set $\{(X_i, Y_i)\}_{i=1}^n$, ViRe renders $X \in \mathbb{R}^{T\times C}$ as $I = \mathcal{V}(X)$, extracts the frozen CLIP feature $Z = f_{\mathrm{CLIP}}(I)$, and projects it to the global Vision Query $Q = ZW_z + b_z \in \mathbb{R}^{1\times D}$. The numerical backbone produces temporal and channel tokens $\bar{U} \in \mathbb{R}^{P\times D}$ and $\bar{V} \in \mathbb{R}^{C\times D}$. ViRe computes $\widetilde{U} = \mathrm{CrossAttn}(Q, \bar{U}, \bar{U})$ and $\widetilde{V} = \mathrm{CrossAttn}(Q, \bar{V}, \bar{V})$, followed by $\widehat{Y} = (\widetilde{U} + \widetilde{V})W_y + b_y$. The numerical counterpart uses mean pooling, $\widehat{Y}_{\mathrm{Num}} = (\mu(\bar{U}) + \mu(\bar{V}))W_y + b_y$.

Lemma 1 (Cross-attention contains mean pooling). For standard scaled dot-product attention, there exists a parameter setting such that, for any $Q \in \mathbb{R}^{1\times D}$ and $K \in \mathbb{R}^{S\times D}$,

$$\mathrm{CrossAttn}(Q, K, K) = \mu(K). \tag{A.1}$$

Proof. For one head, set $W_Q = 0$ (equivalently $W_K = 0$). Every attention logit is zero, the softmax becomes uniform, and the head returns $S^{-1}\sum_{s=1}^{S} k_s W_V$. After concatenating heads and applying the output projection, the remaining linear map of $\mu(K)$ is absorbed into (W_y, b_y). □

Theorem 1 (ViRe contains mean-pooled numerical modeling). Let $\mathcal{F}_{\mathrm{ViRe}}$ and $\mathcal{F}_{\mathrm{Num}}$ denote the ViRe and numerical hypothesis classes. For empirical cross-entropy risk $\widehat{\mathcal{L}}$, the inclusion $\mathcal{F}_{\mathrm{Num}} \subseteq \mathcal{F}_{\mathrm{ViRe}}$ implies

$$\mathcal{L}^\star_{\mathrm{ViRe}} := \inf_{f\in\mathcal{F}_{\mathrm{ViRe}}} \widehat{\mathcal{L}}(f) \le \inf_{f\in\mathcal{F}_{\mathrm{Num}}} \widehat{\mathcal{L}}(f) =: \mathcal{L}^\star_{\mathrm{Num}}. \tag{A.2}$$

Proof. Take any $f \in \mathcal{F}_{\mathrm{Num}}$. By Lemma 1, choose the two retrieval blocks so that $\widetilde{U} = \mu(\bar{U})$ and $\widetilde{V} = \mu(\bar{V})$ for every input, while retaining the same numerical backbone and classifier. ViRe exactly reproduces f, establishing both the hypothesis-class inclusion and empirical-risk inequality. □

Proposition 1 (Complementary structure after numerical compression). Let $S = (\bar{U}, \bar{V})$ and let $\bar{Z}$ denote the projected visual representation. Both are deterministic representations of X and can preserve different aspects of its structure. Positive conditional information, $I(Y; \bar{Z} \mid S) > 0$, yields

$$R^\star(S, \bar{Z}) = H(Y \mid S, \bar{Z}) < H(Y \mid S) = R^\star(S). \tag{A.3}$$

Proof. Under log-loss, Bayes risk equals conditional entropy, and $H(Y \mid S) - H(Y \mid S, \bar{Z}) = I(Y; \bar{Z} \mid S) > 0$ gives the result. □

Proposition 2 (Vision-conditioned retrieval realizes adaptive pooling). For a token set $K = [k_1, \ldots, k_S]^\top$, one retrieval head produces

$$\mathrm{Attn}(Q, K) = \sum_{s=1}^{S} \alpha_s(Q, K) k_s W_V, \qquad \alpha(Q, K) = \mathrm{softmax}\left(\frac{QW_Q(KW_K)^\top}{\sqrt{d}}\right). \tag{A.4}$$

The weights therefore form a sample-dependent pooling rule indexed jointly by the Vision Query and the numerical tokens. Uniform pooling is recovered by Lemma 1, whereas non-uniform logits yield selective aggregation over temporal or channel evidence. Applying this operator to $\bar{U}$ and $\bar{V}$ gives ViRe two complementary adaptive retrieval paths over the original numerical representation.

Proof. Lemma 1 gives the uniform-weight member of the family. Whenever two projected key scores differ, the softmax assigns distinct weights to their tokens. Because the scores depend jointly on Q and K, changing either the visual query or the numerical tokens changes the pooling coefficients continuously. The operator therefore spans both uniform and sample-adaptive aggregation. □

Dual-path aggregation. The temporal path applies this adaptive weighting over P temporal tokens, emphasizing time-localized morphology such as sharp transitions, recurrent complexes, and long-range waveform structure. The channel path applies the same principle over C channel tokens, emphasizing cross-channel patterns and channel-specific morphology. Their summed representation combines these two views before classification, linking the global vision prior to the original numerical evidence at both temporal and channel resolutions of the input signal.

B Data Preprocessing and Train-validation-test Split

We utilize **three** EEG (APAVA, ADFTD, and TDBrain) and **three** ECG (PTB, PTB-XL, and MIMIC) datasets under the Subject-Independent setting [33], where samples from the same subject are exclusively divided into training, validation, or test sets. The data preprocessing and train-validation-test split protocol follows Medformer [11]. Since MIMIC is not included in the Medformer benchmark, we download and preprocess it with the identical protocol to keep the comparison fair.

APAVA. The Alzheimer's Patients' Relatives Association of Valladolid (APAVA) dataset [37] is a public two-class EEG benchmark containing recordings from 23 subjects, i.e., 12 Alzheimer's disease (AD) patients and 11 healthy controls (HC). For each trial, we extract 9 half-overlapping windows, where each window corresponds to a 1-second sequence with 256 timestamps. In total, this yields 5,967 samples. We reserve subjects 15,16,19,20 for validation and 1,2,17,18 for testing, and use the remaining subjects, together with all of their samples, for training.

ADFTD. The Alzheimer's Disease and FronTotemporal Dementia (ADFTD) dataset [49] is a public EEG dataset with three classes recorded over 19 channels, including 36 AD patients, 23 Frontotemporal Dementia (FTD) patients, and 29 HCs. We first resample each trial from 500 Hz to 256 Hz, then slice it into non-overlapping 1-second windows with 256 timestamps, discarding any trailing segments shorter than 1 second. This yields 69,752 samples. We perform a subject-wise split, allocating 60%, 20%, and 20% of subjects (and all corresponding samples) to the training, validation, and test sets, respectively, so that no subject appears in more than one partition.

TDBrain. TDBrain [50] is a large permissioned EEG dataset with 33 channels collected from 1,274 individuals. Each subject provides two recordings (eyes-open and eyes-closed). We consider a balanced subset consisting of 25 Parkinson's disease (PD) subjects and 25 HCs, using only the eyes-closed recordings. Each trial is partitioned into non-overlapping 1-second segments (256 timestamps), and segments shorter than 1 second are removed. This results in 6,240 samples. We assign subjects 18,19,20,21,46,47,48,49 to the validation set and 22,23,24,25,50,51,52,53 to the test set; all remaining subjects, together with all of their segments, are used exclusively for training.

PTB. The PTB dataset [51] is a public ECG dataset collected from 290 subjects with 15 leads and 8 labels (7 cardiac conditions plus healthy control). In this work, we select 198 subjects from the myocardial infarction and HC categories. We downsample the original 500 Hz recordings to 250 Hz and standardize the signals. We then convert each recording into single-heartbeat samples: R-peaks are detected across all leads, outlier intervals are filtered out, and each beat is extracted around its R-peak. To enforce a fixed length, we zero-pad shorter beats using the maximum beat duration across all channels as the reference. This procedure produces 64,356 heartbeat samples, which are split subject-wise into training, validation, and test sets at 60% : 20% : 20%.

PTB-XL. PTB-XL [35] is a large-scale public ECG dataset with 12 leads from 18,869 subjects and 5 diagnostic categories (4 diseases plus healthy control). To avoid label inconsistency, we remove subjects whose diagnoses differ across trials, leaving 17,596 subjects. Each record spans 10 seconds and is available at 100 Hz and 500 Hz; we use the 500 Hz version, resample it to 250 Hz, and apply standard scaling. Next, we segment each trial into non-overlapping 1-second windows (250 timestamps) and discard any remainder shorter than 1 second. This yields 191,400 samples. We use a subject-level 60% : 20% : 20% split, keeping all windows from each subject together.

MIMIC. MIMIC [34] is an ECG dataset with binary labels indicating heart disease versus healthy control. Each recording is 10 seconds long and sampled at 500 Hz. We curate a subset of patients with consistent diagnostic labels across records, then downsample signals to 250 Hz and standardize them. R-peaks are detected on all leads to segment each record into individual heartbeats, and noisy or outlier beats are removed. Each heartbeat is aligned to its R-peak and zero-padded to a uniform length determined by the maximum beat duration observed in the dataset. This preprocessing yields 204,370 samples. Finally, we split subjects (and all corresponding heartbeats) into training, validation, and test sets using a 60% : 20% : 20% subject-level split, keeping all heartbeats of a subject together.

Summary. All six datasets follow the same pipeline: subject-level partitioning is fixed before preprocessing, each subject contributes to exactly one split, and the numerical input and its rendered waveform image are derived from the same preprocessed segment. The visual branch thus never observes information unavailable to the numerical branch, and Table 1 describes both modalities.

C Data Augmentation

During training, each input is augmented by exactly one augmentation operation, which is sampled uniformly at random from the six transformations summarized in Table 11.

Table 11: Training-time augmentations and default settings.

Operation	**Transformation**	**Default**
Temporal flipping	Reverse the sequence along the time axis.	$prob = 0.5$
Channel shuffling	Randomly permute the channel order.	$prob = 0.5$
Temporal masking	Mask timestamps shared across all channels.	$ratio = 0.1$
Frequency masking	Suppress randomly selected frequency bands and transform the signal back to the time domain.	$ratio = 0.1$
Jittering	Add random noise sampled from $[0, 1]$ and scale its magnitude.	$scale = 0.1$
Dropout	Randomly set a fraction of signal values to zero.	$ratio = 0.1$

For each training instance, we draw one operation a uniformly from the six-operation pool $\mathcal{A}$. The numerical and visual inputs are

$$X' = a(X), \qquad I' = \mathcal{V}(X'). \tag{C.1}$$

Both temporal and channel branches therefore receive the same transformed signal, so masked intervals, channel permutations, and local perturbations stay synchronized.

Only one operation is sampled in each forward pass. This avoids compounding several perturbations into an unrealistic waveform, while keeping the selected transformation and its default strength explicit. The sampling rule and settings in Table 11 are shared across datasets.

Augmentation is disabled for validation and testing. Every held-out sample is processed by the deterministic preprocessing and rendering pipeline used for model selection and final evaluation.

Rationale. The pool targets nuisance factors that are common in clinical recordings rather than label-relevant morphology. Temporal flipping and channel shuffling discourage the encoders from memorizing absolute positions or a fixed electrode order; temporal masking and dropout imitate transient electrode dropout and missing samples; frequency masking removes narrow bands in the way that filtering or line-noise suppression does; and jittering models low-amplitude sensor noise. Because the rendering is regenerated from the augmented signal, the visual branch is exposed to the same nuisance variation and cannot rely on rendering-specific artifacts.

D Implementation Details

D.1 Implementation Details of All Baselines

All baseline methods are implemented on top of Medformer [11], which unifies competing approaches within a shared training pipeline, enabling a consistent comparison. We benchmark ten Transformer baselines in this common pipeline: Autoformer [52], FEDformer [53], Informer [54], iTransformer [38], MTST [56], Nonformer [57], PatchTST [58], Reformer [55], Medformer [11], and the vanilla Transformer [39], all trained and evaluated under identical settings.

For Medformer, we reproduce the reported results using the authors' official implementation. For the remaining baselines, we standardize the architecture by using a 6-layer encoder, setting the attention embedding dimension D to 128, and the hidden size of the feed-forward network to 256. We train all models with the Adam optimizer using a learning rate of $1\mathrm{e}{-4}$. The batch size is fixed to $\{32, 32, 128, 128, 128, 128\}$ for APAVA, TDBrain, ADFTD, PTB, PTB-XL, and MIMIC, respectively. Each model is trained for 100 epochs with early stopping (patience $= 10$) based on validation macro F1-Score. We checkpoint the model achieving the best validation F1-Score and report test performance accordingly. We evaluate Accuracy, macro-Precision, macro-Recall, macro-F1, macro-AUROC, and macro-AUPRC. All experiments are repeated with five random seeds under fixed train, validation, and test splits, and results are reported as mean$\pm$std.

Autoformer. Autoformer [52] replaces standard self-attention with an auto-correlation operator tailored for time series forecasting It further incorporates a decomposition module that separates the input into trend/cyclical and seasonal components to facilitate representation learning.

FEDformer. FEDformer [53] exploits Fourier-domain representations through frequency-enhanced blocks and frequency-domain attention, and it introduces a decomposition mechanism that replaces layer normalization in the Transformer to improve its modeling capacity.

Informer. Informer [54] introduces ProbSparse attention and a one-shot generative forecasting paradigm to reduce both the computational and memory costs of long-sequence modeling.

iTransformer. iTransformer [38] forms tokens by embedding entire channels and correspondingly swaps dimensions in normalization and feed-forward modules.

MTST. MTST [56] uses multi-scale tokens with heterogeneous patch lengths.

Nonformer. Nonformer [57] models non-stationary dynamics with de-stationary attention and applies paired normalization and denormalization to mitigate over-stationarization.

PatchTST. PatchTST [58] constructs patch tokens from single-channel temporal segments, enlarging the receptive field of each token and improving long-horizon temporal modeling.

Reformer. Reformer [55] approximates dot-product attention with locality-sensitive hashing and uses reversible residual layers to reduce attention complexity and training memory.

Transformer. The vanilla Transformer [39], introduced in "Attention Is All You Need," can be adapted to time series by treating each multivariate timestamp as a token.

Medformer. Medformer [11] is a multi-granularity patching Transformer for medical time series classification, capturing local and long-range dependencies.

Common evaluation protocol. All methods use the same subject-disjoint partitions, preprocessing, label definitions, validation-based early stopping, six evaluation metrics, and five random seeds. The test set is evaluated only after the best validation checkpoint has been selected.

Input controls. Each baseline receives the same numerical input under the partitions used by ViRe. Architecture-specific modules follow the corresponding released implementations, while dataset statistics and evaluation settings remain fixed across methods.

D.2 Implementation Details of the Proposed ViRe

ViRe uses shared settings across datasets, summarized in Table 12. Code and training scripts are publicly available in the GitHub Repo, together with the executable visualization notebook and the precomputed CLIP features of all six benchmark datasets used in this paper.

Table 12: **Implementation configuration of ViRe.**

Item	Setting
Batch size	$B = 128$
Learning rate	$1e{-4}$
Model dimension	$D = 128$
Temporal Encoder depth	$M = 6$ for all datasets
Channel Encoder depth	$N = 6$ by default; $N = 0$ for TDBrain
Temporal granularity L	Dataset order: APAVA, TDBrain, ADFTD, PTB, PTB-XL, MIMIC; $L = \{1, 3, 8, 1, 6, 6\}$

Training protocol. ViRe is optimized with Adam at a learning rate of $1e{-4}$ for up to 100 epochs, with early stopping (patience $= 10$) on validation macro-F1. We use five random seeds under fixed subject-disjoint splits and report mean$\pm$std. The CLIP encoder remains frozen throughout optimization, so the vision branch introduces no additional trainable backbone parameters.

Dataset-specific configuration. The model dimension is fixed at $D = 128$ and the temporal encoder depth at $M = 6$. The channel depth and temporal granularity L follow Table 12; all remaining optimization and evaluation settings are shared across datasets.

Visual feature caching. Waveform rendering is deterministic. During training, the frozen CLIP features are cached once per sample, so repeated optimization does not rerun the visual encoder. This is also the execution mode used for the training-cost measurements in Appendix F.

Trainable components. Optimization updates the numerical encoders, dimension-alignment layer, two retrieval blocks, and classifier. The CLIP parameters remain fixed in every experiment.

Checkpoint evaluation. The best validation macro-F1 checkpoint is evaluated on the test split for each seed, using the same macro-averaged metrics as the baselines in Appendix D.1.

Vision branch. The frozen CLIP Vision Encoder is used in inference mode: images are produced by the operator in Appendix D.3, encoded once, and the resulting $\tilde{D}$-dimensional features are cached. Only the Dimension Align layer that maps them to $D = 128$ is trained, so the visual branch adds only one linear projection to the total trainable parameter count.

Retrieval cost. Because a single Vision Query attends over P temporal and C channel tokens, each retrieval block costs $O((P + C)D)$ per sample, which is negligible compared with the quadratic self-attention cost of the two Transformer Encoders over P and C tokens.

D.3 Visualization Operator

ViRe deterministically transforms each preprocessed multichannel MedTS sample into a CLIP-readable stacked waveform image. Each channel is rendered on a separate panel with fixed axes, after which the panels are vertically concatenated. This standardized layout preserves channel morphology and cross-channel timing while removing decorative plot elements. Algorithm 1 specifies the operator, and an executable version is available as a notebook in the GitHub Repo. The same operator is applied without modification to training, validation, and test samples, and Figure 5 shows the rendering of a twelve-lead PTB sample produced by this operator with all decorations removed.

Algorithm 1 Visualization Operator $\mathcal{V}(\cdot)$

Require: Preprocessed MedTS sample $X \in \mathbb{R}^{C \times T}$, sampling rate f_s
Ensure: Stacked RGB waveform image I_{stack}
Construct $t \leftarrow [0, 1, \ldots, T-1]/f_s$ and initialize $\mathcal{P} \leftarrow [\,]$.
for $c = 1$ to C **do**
    Render $X_{c,:}$ against t on an independent canvas.
    Remove ticks, spines, legends, grids, and axis decorations.
    Export RGB panel P_c and append it to $\mathcal{P}$.
end for
$I_{\text{stack}} \leftarrow \text{VConcat}(P_1, \ldots, P_C)$.
return I_{stack}.

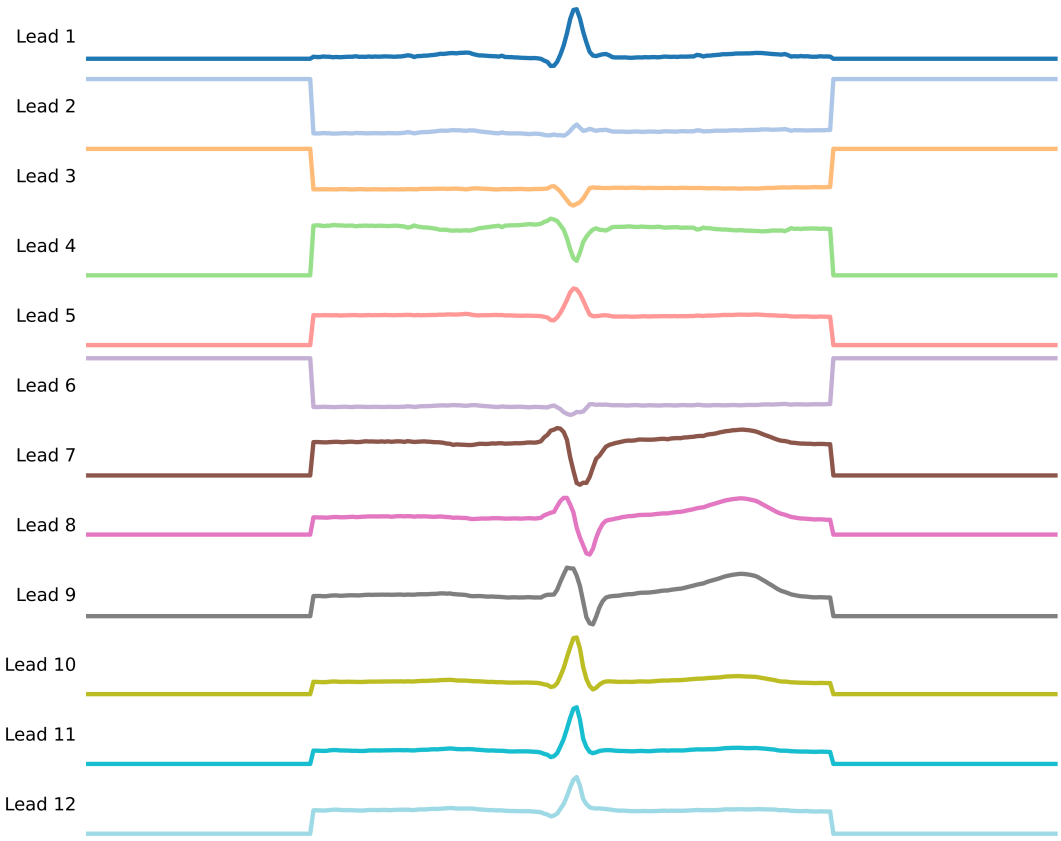


Figure 5: Visualization example. Stacked twelve-lead PTB waveform rendered by $\mathcal{V}(\cdot)$.

E Full Classification Results

Table 13: **Full subject-independent classification results.** Mean±std over five seeds; best is bolded and second-best is shown in blue italics.

Datasets	Models	Accuracy	Precision	Recall	F1-Score	AUROC	AUPRC	Avg
APAVA (2-Classes) (EEG)	**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	70.71±2.63
	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	77.21±2.49
	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	71.36±5.04
	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*	77.23±1.73
	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	69.94±1.99
	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	70.70±3.81
	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	65.89±1.52
	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	77.02±1.99
	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	74.42±5.88
	Medformer	*78.74±0.64*	81.11±0.84	*75.40±0.66*	*76.31±0.71*	83.20±0.91	83.66±0.92	*79.74±0.78*
	ViRe	**91.43±1.03**	**91.07±1.16**	**91.40±0.79**	**91.19±1.01**	**95.99±0.34**	**95.67±0.44**	**92.79±0.80**
ADFTD (3-Classes) (EEG)	**Autoformer**	45.25±1.48	43.67±1.94	42.96±2.03	42.59±1.85	61.02±1.82	43.10±2.30	46.43±1.90
	FEDformer	46.30±0.59	46.05±0.76	44.22±1.38	43.91±1.37	62.62±1.75	46.11±1.44	48.20±1.22
	Informer	48.45±1.96	46.54±1.68	46.06±1.84	45.74±1.38	65.87±1.27	47.60±1.30	50.04±1.57
	iTransformer	52.60±1.59	46.79±1.27	47.28±1.29	46.79±1.13	67.26±1.16	49.53±1.21	51.71±1.28
	MTST	45.60±2.03	44.70±1.33	45.05±1.30	44.31±1.74	62.50±0.81	45.16±0.85	47.89±1.34
	Nonformer	49.95±1.05	47.71±0.97	47.46±1.50	46.96±1.35	66.23±1.37	47.33±1.78	50.94±1.34
	PatchTST	44.37±0.95	42.40±1.13	42.06±1.48	41.97±1.37	60.08±1.50	42.49±1.79	45.56±1.37
	Reformer	50.78±1.17	49.64±1.49	49.89±1.67	47.94±0.69	69.17±1.58	*51.73±1.94*	53.19±1.42
	Transformer	50.47±2.14	49.13±1.83	48.01±1.53	48.09±1.59	67.93±1.59	48.93±2.02	52.09±1.78
	Medformer	*53.27±1.54*	*51.02±1.57*	*50.71±1.55*	*50.65±1.51*	*70.93±1.19*	51.21±1.32	*54.63±1.45*
	ViRe	**57.83±1.82**	**56.44±2.49**	**53.75±3.04**	**54.02±3.19**	**76.59±1.62**	**60.33±2.78**	**59.83±2.49**
TDBrain (2-Classes) (EEG)	**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	89.52±3.28
	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	80.98±1.96
	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	91.64±1.83
	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	77.63±1.11
	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	79.35±4.17
	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*	91.07±1.79
	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	81.94±4.51
	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	90.84±1.42
	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	90.34±1.41
	Medformer	*89.62±0.81*	*89.68±0.78*	*89.62±0.81*	*89.62±0.81*	96.41±0.35	96.51±0.33	*91.91±0.65*
	ViRe	**93.96±0.75**	**94.03±0.72**	**93.96±0.75**	**93.96±0.75**	**98.65±0.34**	**98.68±0.35**	**95.54±0.61**
PTB (2-Classes) (ECG)	**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	70.86±3.17
	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	75.90±3.97
	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	80.62±1.70
	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*	*84.95±1.02*
	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	76.80±2.68
	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	79.58±1.35
	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	75.35±1.85
	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	79.51±2.17
	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	78.69±1.52
	Medformer	83.50±2.01	85.19±0.94	*77.11±3.39*	*79.18±3.31*	*92.81±1.48*	90.32±1.54	84.69±2.11
	ViRe	**88.26±1.12**	**89.54±0.78**	**83.82±1.75**	**85.81±1.52**	**93.95±0.28**	**93.08±0.72**	**89.08±1.03**
PTB-XL (5-Classes) (ECG)	**Autoformer**	61.68±2.72	51.60±1.64	49.10±1.52	48.85±2.27	82.04±1.44	51.93±1.71	57.53±1.88
	FEDformer	57.20±9.47	52.38±6.09	49.04±7.26	47.89±8.44	82.13±4.17	52.31±7.03	56.83±7.08
	Informer	71.43±0.32	62.64±0.60	59.12±0.47	60.44±0.43	88.65±0.09	64.76±0.17	67.84±0.35
	iTransformer	69.28±0.22	59.59±0.45	54.62±0.18	56.20±0.19	86.71±0.10	60.27±0.21	64.45±0.23
	MTST	72.14±0.27	63.84±0.72	60.01±0.81	61.43±0.38	88.97±0.33	65.83±0.51	68.70±0.50
	Nonformer	70.56±0.55	61.57±0.66	57.75±0.72	59.10±0.66	88.32±0.36	63.40±0.79	66.78±0.62
	PatchTST	**73.23±0.25**	**65.70±0.64**	**60.82±0.76**	**62.61±0.34**	**89.74±0.19**	**67.32±0.22**	**69.90±0.40**
	Reformer	71.72±0.43	63.12±1.02	59.20±0.75	60.69±0.18	88.80±0.24	64.72±0.47	68.04±0.52
	Transformer	70.59±0.44	61.57±0.65	57.62±0.35	59.05±0.25	88.21±0.16	63.36±0.29	66.73±0.36
	Medformer	72.87±0.23	64.14±0.42	*60.60±0.46*	*62.02±0.37*	*89.66±0.13*	66.39±0.22	69.28±0.31
	ViRe	*73.12±0.24*	*66.07±0.60*	59.62±0.60	61.59±0.47	89.65±0.17	*66.76±0.26*	*69.47±0.39*
MIMIC (2-Classes) (ECG)	**Autoformer**	77.74±6.49	77.92±6.53	77.59±6.12	77.58±6.37	84.52±6.88	82.77±7.19	79.69±6.60
	FEDformer	84.68±0.21	84.63±0.28	84.52±0.11	84.56±0.18	91.59±0.25	90.78±0.24	86.79±0.21
	Informer	84.71±0.27	84.60±0.22	84.73±0.24	84.65±0.22	91.72±0.11	91.23±0.19	86.94±0.21
	iTransformer	85.01±0.10	84.94±0.14	84.89±0.11	84.91±0.16	91.51±0.08	91.37±0.12	87.11±0.12
	MTST	85.54±0.31	85.44±0.32	85.55±0.22	85.48±0.38	91.61±0.20	91.49±0.14	87.52±0.26
	Nonformer	84.13±0.25	84.03±0.17	84.18±0.27	84.08±0.16	90.98±0.18	90.72±0.21	86.35±0.21
	PatchTST	84.79±0.29	84.71±0.31	84.81±0.36	84.73±0.32	91.10±0.28	90.61±0.32	86.79±0.31
	Reformer	*85.74±0.15*	*85.75±0.18*	*85.72±0.18*	*85.88±0.18*	*92.38±0.09*	*91.94±0.19*	*87.90±0.16*
	Transformer	84.93±0.17	84.84±0.16	84.92±0.19	84.86±0.16	91.64±0.13	90.89±0.16	87.01±0.16
	Medformer	85.10±0.33	85.02±0.35	85.12±0.38	85.04±0.38	91.44±0.23	91.12±0.22	87.14±0.32
	ViRe	**88.61±0.12**	**88.55±0.13**	**88.56±0.15**	**88.54±0.12**	**95.08±0.06**	**94.76±0.11**	**90.68±0.12**

F Computational Overhead Analysis

We measure APAVA F1-Score, per-batch latency, and peak memory at $B = 128$. Training reuses cached CLIP features; online inference includes waveform rendering and vision encoding.

Table 14: **Computational Overhead on APAVA.** F1-Score, per-batch latency ($B = 128$), and peak memory are measured. Higher is better for F1-Score; lower is better for time and memory.

Metric	ViRe	Medformer	FEDformer
F1-Score ↑	**91.19**$\pm$**1.01**	76.31$\pm$0.71	73.51$\pm$3.30
Training Time (ms) ↓	**211.9**	313.1	477.7
Training Memory (MB) ↓	**698.2**	838.4	1103.9
Inference Time (ms) ↓	267.7	**152.6**	282.6
Inference Memory (MB) ↓	740.8	**408.5**	491.6

Training. Caching removes repeated visual encoding from each optimization step. ViRe requires 211.9 ms and 698.2 MB per batch, compared with 313.1 ms and 838.4 MB for Medformer; the corresponding APAVA F1-Score is 91.19 versus 76.31 for the two models.

Inference. The complete online path requires 267.7 ms and 740.8 MB. Its latency is lower than FEDformer (282.6 ms); the higher memory cost reflects the waveform rendering and CLIP vision encoding that are executed online in this measurement rather than read from the cache.

Normalized comparison. Relative to Medformer, cached training lowers per-batch latency by 32.3% and peak memory by 16.7%, while raising the APAVA F1-Score by 14.88 points.

Deployment. Cached features suit repeated optimization, whereas online use executes the full visual front end and is represented by the inference rows in Table 14.

Trade-off. Relative to Medformer, online inference is 1.75$\times$ slower and uses 1.81$\times$ more memory (267.7 vs. 152.6 ms; 740.8 vs. 408.5 MB), while the APAVA F1-Score improves by 14.88 points and training is cheaper. Whether the online cost is acceptable depends on the deployment budget, and caching removes most of it whenever the same recordings are scored repeatedly.

Measurement protocol. All numbers are measured on the same NVIDIA RTX 4090 GPU used for the main experiments with batch size $B = 128$. Training rows use cached CLIP features, whereas inference rows execute rendering and CLIP encoding online, so the cached and online settings respectively bound the practical training and deployment cost of ViRe from below and above.

G Interpretability Analysis of ViRe

G.1 Qualitative Retrieval Attention Visualization

We inspect whether the CLIP-derived Vision Query retrieves physiologically meaningful temporal tokens. As shown in Figure 6, the temporal retrieval attention on PTB exhibits a localized peak around a morphologically salient interval, where multiple ECG leads show synchronous waveform changes. This pattern suggests that ViRe does not distribute attention uniformly over time; instead, the vision prior guides retrieval toward clinically relevant cross-lead variations.

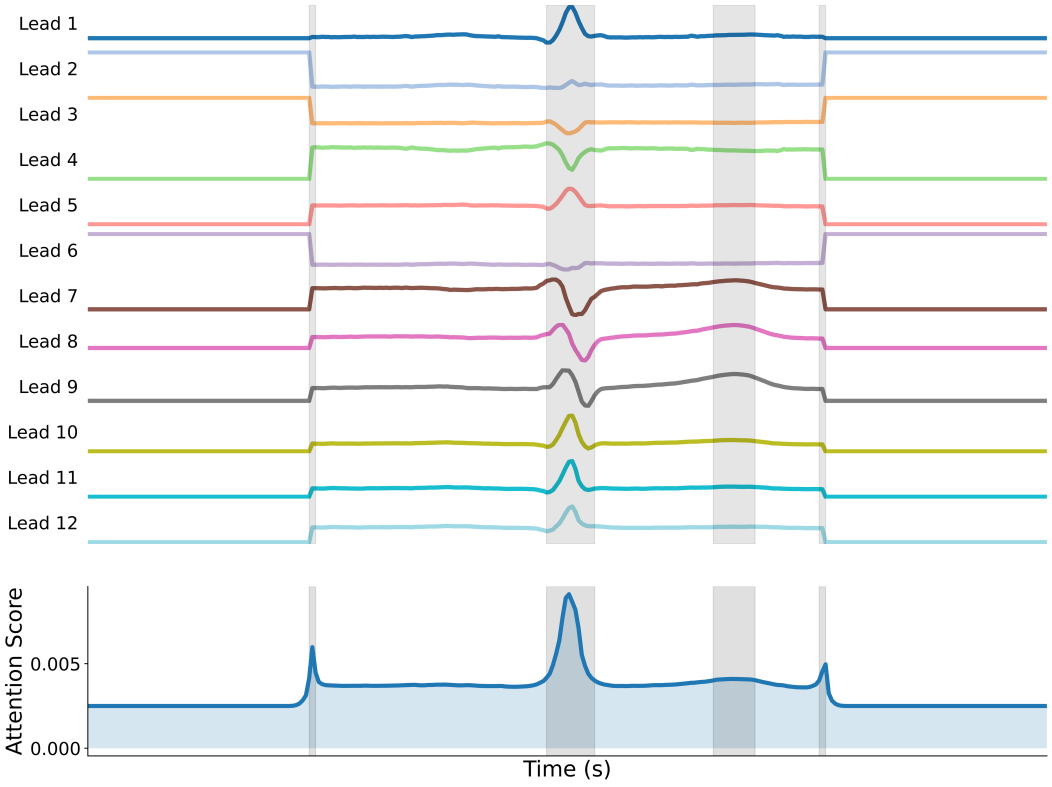


Figure 6: **Retrieval Attention Visualization on PTB.** Temporal attention from the CLIP-derived Vision Query peaks around a morphologically salient interval with synchronous cross-lead variations.

Appendix G.2 quantifies morphology and clinical alignment at the dataset level.

G.2 Quantitative Attention Analysis

We further quantify whether retrieval attention is concentrated in meaningful temporal regions. We compare the CLIP-derived Vision Query with a non-semantic Gaussian Query baseline of the same dimensionality. Both metrics are attention-density ratios; values larger than one indicate denser attention inside the target region than in its temporal complement.

Let $X \in \mathbb{R}^{C\times T}$ be a channel-normalized ECG sample and let $A \in \mathbb{R}^{T}$ denote its temporal retrieval attention, with $A_t > 0$ and $\sum_{t=1}^{T} A_t = 1$. For a temporal region $\mathcal{S}$, define

$$\rho_A(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{t\in\mathcal{S}} A_t. \tag{G.1}$$

This length-normalized density makes short diagnostic intervals, such as the QRS region in ECG, comparable with their much longer complements in the same recording.

For morphology-aware alignment, we compute a channel-averaged second-order temporal variation score at every interior timestamp:

$$c_t = \frac{1}{C} \sum_{i=1}^{C} \left| X_{i,t+1} - 2X_{i,t} + X_{i,t-1} \right|, \quad t = 2, \ldots, T-1, \tag{G.2}$$

and define $\mathcal{S}_{\mathrm{curv}}$ as the top-20% timestamps ranked by c_t. For clinical alignment, $\mathcal{S}_{\mathrm{qrs}}$ denotes the timestamps covered by QRS segment(s), and $L_{\mathrm{qrs}} = |\mathcal{S}_{\mathrm{qrs}}|$. With complements denoted by bars, the Morphology-Aware Attention Ratio (MAR) and Clinical Alignment Ratio (CAR) are

$$\begin{aligned} \mathrm{MAR} &= \frac{\rho_A(\mathcal{S}_{\mathrm{curv}})}{\rho_A(\bar{\mathcal{S}}_{\mathrm{curv}})}, \\ \mathrm{CAR} &= \frac{\rho_A(\mathcal{S}_{\mathrm{qrs}})}{\rho_A(\bar{\mathcal{S}}_{\mathrm{qrs}})}. \end{aligned} \tag{G.3}$$

Thus, MAR evaluates enrichment on signal-intrinsic high-curvature morphology, whereas CAR evaluates enrichment on the clinically recognized QRS complex. Table 7 in the main text shows consistent ViRe gains over the Gaussian Query on both metrics and all three ECG datasets.

H Rendering Sensitivity Analysis

We further study how the visualization operator depends on several rendering hyperparameters. Using the F1-Score from controlled ablations on APAVA and PTB, Figure 7 summarizes the sensitivity of ViRe to ***(i)*** rendering resolution (DPI), ***(ii)*** waveform scaling (line width), and ***(iii)*** channel layout (colored *vs.* uncolored channels), with all other rendering settings fixed.

The figure reveals three consistent patterns. First, ViRe is clearly sensitive to extremely low rendering resolution: using only 5 DPI causes a pronounced performance drop on both datasets, whereas moderate-to-high resolutions (50–200 DPI) are much more stable. Second, waveform scaling also matters: overly thin or overly thick lines degrade performance, suggesting that preserving an appropriate morphological thickness is important for CLIP-based visual encoding. Third, channel coloring has only a negligible influence, indicating that ViRe mainly benefits from the global waveform morphology rather than from color-specific channel cues.

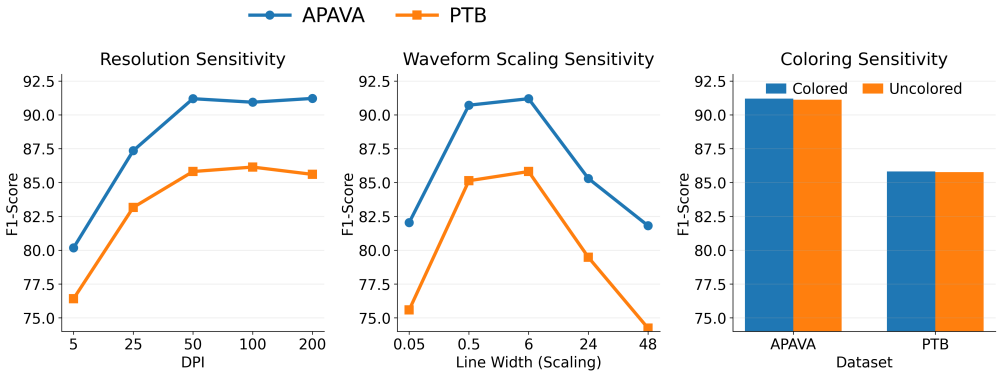


Figure 7: ***Rendering sensitivity of ViRe.*** F1-Score on APAVA and PTB under controlled variations of rendering resolution (left), waveform scaling/line width (middle), and channel layout coloring (right). The results show that ViRe is sensitive to excessively low resolution and inappropriate line width, while being largely insensitive to whether channels are colorized.

Practical recommendation. The results suggest rendering at moderate resolution (50–200 DPI) with a moderate line width, since both extremes degrade the CLIP representation, whereas channel coloring can be chosen freely. Rendering hyperparameters therefore deserve the same care as signal preprocessing when the operator is transferred to a new recording setup.

I Robustness and Transfer Analysis

Beyond the primary benchmark, we evaluate subject-partition stability on repeated subject-disjoint splits and transfer to irregularly sampled medical time series.

Subject-Partition Stability

Across held-out PTB-XL evaluations and repeated subject-disjoint splits, ViRe consistently improves over Medformer. Table 15 reports the paired statistical evidence for each evaluation.

Table 15: **Subject-partition stability and paired statistical evidence.**

Evaluation	Gain	Wins	Statistical evidence
PTB-XL, 13 diagnoses (64.93 vs. 61.57)	**+3.36**	n/a	Cluster 95% CI **[+1.46,+5.10]**; paired $p = 0.0103$.
PTB-XL, five superclasses (72.26 vs. 68.52)	**+3.74**	n/a	Cluster 95% CI **[+2.36,+5.14]**; paired $p = 0.0043$.
PTB-XL, 13 diagnoses; repeated splits	**+2.46**	**7/7**	Wilcoxon $p = 0.0156$; corrected paired t-test $p = 0.0038$; 95% CI **[+1.14,+3.77]**.
PTB-XL, five superclasses; repeated splits	**+2.42**	**7/7**	Wilcoxon $p = 0.0156$; corrected paired t-test $p = 0.0060$; 95% CI **[+0.99,+3.84]**.
APAVA; repeated splits	**+6.48**	**8/10**	Exact Wilcoxon/Holm $p = 0.0371$.
PTB; repeated splits	**+3.98**	**9/10**	Exact Wilcoxon/Holm $p = 0.0078$.

Irregular Forecasting

We follow the Hi-Patch protocol [60] while retaining ViRe's visual-query retrieval mechanism. ViRe performs best on all six metrics and reduces MSE relative to Hi-Patch by 4.7%, 6.2%, and 10.9% on Human Activity, PhysioNet, and MIMIC-III, respectively (Table 16).

Table 16: **Irregular forecasting under the Hi-Patch protocol** [60]. Lower is better.

Dataset	Metric	ViRe	Hi-Patch	t-PatchGNN	GRU-D
Human Activity	MSE $\times 10^{-3}$	**2.45±0.04**	2.57±0.02	2.66±0.03	3.94±0.29
	MAE $\times 10^{-2}$	**3.04±0.02**	3.11±0.03	3.15±0.02	4.37±0.21
PhysioNet	MSE $\times 10^{-3}$	**4.56±0.04**	4.86±0.03	4.98±0.08	5.76±0.34
	MAE $\times 10^{-2}$	**3.44±0.03**	3.62±0.07	3.72±0.03	4.53±0.15
MIMIC-III	MSE $\times 10^{-2}$	**1.56±0.09**	1.75±0.26	1.69±0.03	2.35±0.06
	MAE $\times 10^{-2}$	**6.77±0.11**	7.24±0.18	7.22±0.09	8.34±0.22

Irregular Classification

Following the MTM protocol, we compare ViRe with MTM, STraTS [61], t-PatchGNN, and GRU-D [62] on P12, P19, and PAM. ViRe ranks first on all six reported metrics (Table 17).

Table 17: **Irregular classification under the MTM protocol** [63]. Higher is better.

Dataset	Metric	ViRe	MTM	STraTS	t-PatchGNN	GRU-D
P12	AUROC	**88.5±1.2**	88.0±1.0	86.4±1.1	84.5±0.9	81.9±2.1
	AUPRC	**60.3±2.4**	58.6±4.1	53.9±3.1	50.8±2.6	46.1±4.7
P19	AUROC	**91.3±2.0**	90.3±2.0	89.7±1.8	87.0±1.4	83.9±1.7
	AUPRC	**62.3±4.3**	58.3±5.3	57.9±3.3	51.5±5.2	46.9±2.1
PAM	Accuracy	**98.3±0.6**	97.5±0.2	96.4±0.8	93.9±1.2	83.3±1.6
	F1	**98.4±0.6**	97.6±0.2	95.3±0.7	94.8±1.2	84.8±1.2

J Limitations and Societal Considerations

Limitations. ViRe is evaluated on six public EEG/ECG benchmarks with subject-independent splits, but retrospective results do not replace prospective multi-center clinical validation. Its visual prior is derived from a frozen CLIP vision encoder trained on general image-text data, which may miss clinically subtle waveform patterns that a domain-specific encoder could capture. The benchmark study covers EEG and ECG classification and, in Appendix I, irregularly sampled forecasting and classification; other physiological signals and multi-label settings remain untested. Rendering hyperparameters influence the visual prior (Appendix H), and online inference incurs the additional memory cost of on-the-fly rendering and CLIP encoding (Appendix F).

Societal considerations. ViRe is intended for decision support rather than autonomous diagnosis. It may improve data efficiency and morphology-aware EEG/ECG modeling, but risks remain, including over-reliance on automated outputs and privacy concerns in downstream use. All experiments use existing, de-identified datasets under their respective access terms, and no new patient data were collected. Clinical deployment, therefore, requires clinician oversight, privacy protection, and validation under the target distribution before any use in routine clinical practice.

Future directions. Promising extensions include vision-language encoders trained on medical waveform images, learned or adaptive rendering operators, and joint use of the frozen visual prior with textual reports, which resources such as MEETI make possible.

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 waveform-morphology motivation, the ViRe framework, and the empirical scope; the quantitative claims are supported by Section 4, Table 2, and the conclusion.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations and potential deployment considerations are discussed in Appendix J, including retrospective validation, reliance on a frozen CLIP vision encoder, and inference-time overhead.

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: [Yes]

   Justification: The paper includes theoretical analysis in Appendix A, where the setup, assumptions, Lemma 1, Theorem 1, Proposition 1, and their proofs are provided.

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: [Yes]

   Justification: The method, datasets, subject-independent splits, metrics, optimization protocol, and implementation details are described in Sections 3–4 and Appendices B–D; code and training scripts are provided in the public repository at `https://github.com/Levi-Ackman/ViRe`.

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: [Yes]

   Justification: The code and training scripts are released at `https://github.com/Levi-Ackman/ViRe`, and public dataset sources are cited with access links.

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: Section 4 specifies the evaluation metrics, random seeds, hardware, and baseline protocol, while Appendices B and D provide dataset preprocessing, subject-level splits, optimizer, batch sizes, early stopping, and hyperparameter settings.

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: All main benchmark and ablation tables report mean and standard deviation over five random seeds; Section 4 states the reporting protocol and Table 13 reports metric-wise mean$\pm$std values.

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: [Yes]

   Justification: Section 4 reports the GPU type used for experiments, Appendix F reports per-batch latency and peak memory, and Appendix D gives batch sizes and implementation settings.

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 work uses cited EEG/ECG datasets under their official access protocols and is presented as a decision-support research method rather than an autonomous clinical diagnostic system, consistent with the NeurIPS Code of Ethics.

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 J discusses positive impacts, such as morphology-aware medical time-series modeling, and negative risks, including erroneous predictions, over-reliance, and privacy concerns.

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 high-risk generative models, scraped datasets, or patient-level data; it releases code/training scripts for reproducing the proposed classification framework.

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: Existing datasets, baselines, and the CLIP vision encoder are credited through citations and URLs in Sections 3–4 and Appendices B–D; the paper does not redistribute the original datasets.

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 new asset is the implementation and training scripts for ViRe, documented by the method description, the implementation appendix, and the public repository at https://github.com/Levi-Ackman/ViRe.

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 conduct crowdsourcing experiments or collect new human-subject data; it uses existing EEG/ECG benchmark datasets.

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 does not collect new human-subject data. It uses existing datasets under their original access and preprocessing protocols, so new IRB approval by the authors is not applicable.

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: LLMs are not used as an important or non-standard component of the proposed method. The method uses a frozen CLIP vision encoder as a visual feature extractor, which is described in Section 3.