

Context-Conditioned Spatio-Temporal Modeling of Multimodal Neurophysiological Signals for Interictal Epileptiform Discharge Detection

Zhan He*
Yiting Wei*

zhan.he@stud.uni-heidelberg.de
yiting.wei@uni-heidelberg.de
Heidelberg University
Heidelberg, Germany

Nima TaheriNejad
Heidelberg University
Heidelberg, Germany
TU Wien

Vienna, Austria
nima.taherinejad@ziti.uni-heidelberg.de
nima.taherinejad@tuwien.ac.at

Abstract

Interictal epileptiform discharges (IEDs) are the principal electrophysiological biomarker of epilepsy, yet their automatic detection from short physiological windows remains difficult: the discharges are rare, morphologically diverse, buried in ocular and muscular artifacts, and their statistics vary strongly across patients and acquisition sessions. We tackle this problem using multimodal recordings that pair scalp electroencephalography (EEG) with electrocardiography (ECG) and electromyography (EMG), and argue that a fixed feed-forward network is poorly matched to such heterogeneous, shifting inputs. We instead propose a context-conditioned spatio-temporal network that decouples *what* is computed from *how* it is adapted. A signal trunk first extracts per-channel temporal morphology with a shared convolutional frontend and a dilated temporal convolutional network then models cross-channel and cross-modal dependencies with a channel-wise Transformer over the 26 physiological nodes. In parallel, a lightweight context branch summarizes each recording into a latent vector by fusing a variational time-domain statistical embedding and a modality-aware spectral embedding through a learned per-dimension convex combination. This latent conditions the entire trunk through feature-wise linear modulation (FiLM) at twenty-three normalization and feature points, allowing the network to re-parameterize its computation per recording without any test-time weight update. A five-way discharge-subtype head provides auxiliary weak supervision on positive windows. On the NeuroMM 2026 Track 1 benchmark, the proposed model reaches 0.9531 area under the precision–recall curve (AUPRC) as a single model, and multi-split probability averaging further raises it to 0.9693. Ablations show that temporal modeling, cross-channel aggregation, fused context conditioning, and moderate auxiliary supervision each contribute, with context conditioning being the single largest architectural factor.

Keywords

Interictal epileptiform discharge, EEG, multimodal physiological signals, temporal convolution, Transformer, feature-wise linear modulation, context conditioning, auxiliary learning

1 Introduction

Epilepsy is among the most common serious neurological disorders. In 2021, it affected an estimated 51.7 million people worldwide and accounting for a large share of the global neurological disease burden [8]. Clinically, epilepsy is defined by an enduring predisposition to generate epileptic seizures [7], and its diagnosis and monitoring rely heavily on EEG, which records cortical electrical activity at millisecond resolution [12]. Between overt seizures, the brain of a person with epilepsy intermittently produces IEDs—transient spikes and sharp waves lasting tens to a few hundred milliseconds. Because they can be captured in routine recordings without waiting for a seizure to occur, IEDs serve as one of the primary electrophysiological biomarkers for establishing epileptogenicity, localizing epileptic foci, and guiding clinical treatment. However, identifying IEDs remains a labor-intensive visual task that requires experts to inspect hours of multichannel EEG recordings and is subject to substantial inter-rater variability, making reliable automated detection of considerable clinical value [15].

Although EEG provides the most direct representation of IEDs, automatic detection remains challenging. EEG signals are highly non-stationary and exhibit a low signal-to-noise ratio, while ocular, muscular, and movement artifacts frequently resemble genuine epileptiform discharges [3, 18, 24]. More importantly, the amplitude, spectral characteristics, and channel reliability of EEG recordings vary substantially across patients and even across recording sessions, leading to severe distribution shifts that have been widely recognized as a major obstacle to deployable EEG deep learning systems [23, 24]. In addition, IEDs occur only sparsely relative to background activity, resulting in severe class imbalance [28]. Consequently, clinically useful detectors should prioritize accurate ranking of candidate events rather than relying solely on a fixed classification threshold [13, 18, 26].

These challenges suggest that robust IED detection requires models capable of adapting to heterogeneous clinical recordings. On the one hand, relying solely on EEG makes it difficult to distinguish true epileptiform discharges from physiological artifacts, whereas complementary physiological signals such as EMG and ECG provide valuable contextual information and have been shown to improve epileptic event detection when jointly modeled with EEG [30]. On the other hand, the substantial statistical variability across recordings calls for models that can adapt their internal computation to the characteristics of individual recordings rather

*Both authors contributed equally to this research.

than applying identical transformations to every sample. Furthermore, temporal statistical information and spectral representations capture complementary properties of physiological signals, and effectively integrating these sources of contextual information is essential for improving robustness and generalization [10].

Motivated by these observations, we propose a context conditioned spatio-temporal network for multimodal IED detection. The proposed framework jointly models EEG, ECG, and EMG signals to learn temporal morphology, cross-channel interactions, and cross-modal dependencies, while leveraging recording-level contextual information to dynamically modulate feature representations throughout the network. By adapting its computation to each individual recording, the proposed framework is better suited to handle the heterogeneous and distribution-shifted nature of real-world clinical data.

Our contributions are summarized as follows.

- **A context-conditioned multimodal spatio-temporal framework for IED detection.** We propose a unified framework that jointly models EEG, ECG, and EMG signals, combining spatio-temporal feature learning with recording-level context conditioning to improve robustness against cross-patient and cross-session distribution shifts.
- **A statistical-spectral context encoding mechanism.** We develop a recording-level context representation by integrating temporal statistical information with modality-aware spectral features, enabling adaptive feature modulation throughout the network without requiring any test-time optimization.
- **An auxiliary multi-task learning strategy.** We introduce discharge-subtype prediction as an auxiliary supervision task to enhance representation learning and improve the robustness and generalization capability of the proposed detector.

2 Related Work

One active line of recent work makes the spatial structure of the electrode array explicit through graph models. Tang et al. modeled dynamic brain connectivity with a graph neural network and self-supervised pre-training, improving both seizure detection and the classification of rare seizure types, at the cost of requiring a large unlabeled corpus for pre-training and an EEG-only montage [25]. Hajisafi et al. extended this direction with NeuroGNN, a dynamic graph network that jointly captures spatial, temporal, semantic, and taxonomic correlations among electrodes and brain regions and reports state-of-the-art detection and classification, but its graph is constructed from EEG electrode geometry alone and does not accommodate non-cortical modalities [9]. These models confirm that cross-channel relations are informative—an intuition our channel-wise Transformer also exploits—yet neither conditions the network on the statistics of the recording in front of it.

A second line targets interictal discharges directly and, in particular, generalization across patients. Zheng et al. proposed ScatterFormer, a locally-invariant scattering Transformer for patient-independent multispectral detection of epileptiform discharges, explicitly optimizing for cross-patient invariance, but it relies on fixed scattering features and a single EEG modality [31]. Chung

et al. trained deep models for automated detection and multiclass classification of focal IEDs by cortical region, a setting close to our joint binary-plus-subtype objective, yet used region-specific models on a single-center cohort, which limits transfer to new sites and montages [6]. Albaqami et al. proposed MP-SeizNet, a multi-path CNN-BiLSTM that fuses wavelet features with raw EEG for seizure-type classification, but the large gap it reports between seizure-wise and patient-wise F1 again exposes the cross-patient generalization bottleneck [1].

A third line broadens the input beyond cortical channels or develops richer physiological-signal representations, including approaches designed for cross-dataset transfer. Zhang et al. showed that fusing wearable EEG with EMG and accelerometry improves tonic-clonic seizure detection over any single modality, confirming that non-cortical signals carry complementary epileptic and artifact information, though their pipeline targets convulsive events with hand-designed modality fusion rather than end-to-end conditioning [30]. Yang et al. proposed BIOT, a biosignal Transformer that tokenizes each channel to learn across datasets with mismatched channels, lengths, and modalities including EEG and ECG, directly confronting distribution shift, but it addresses heterogeneity through tokenization rather than by adapting a task-specific backbone to each recording [29]. Jiang et al. pushed this further with LaBraM, a large EEG foundation model pre-trained on thousands of hours of data for cross-dataset generalization, at the cost of substantial pre-training data and compute [14]. Liang et al. introduced EEG-VL, which maps multichannel EEG into visual representations and couples a pretrained vision encoder with a large language model for seizure detection [17]. More recently, Ma et al. proposed E²-LLM, which aligns pretrained EEG representations with large language models for interpretable neural-signal analysis [20]. These approaches, however, remain centered on EEG and do not explicitly exploit complementary physiological modalities such as ECG and EMG.

In summary, recent work has advanced graph-based spatial modeling, patient-independent discharge detection, and cross-dataset foundation models, and has confirmed the value of both multimodal signals and discharge-type labels—but most systems remain EEG-only, apply a fixed processing pipeline to every recording, and treat type labels as a separate downstream task.

3 Methodology

3.1 Overview

Figure 1 summarizes the model. It comprises a signal trunk and a context branch. The trunk maps a 26-channel physiological window into node-level temporal representations, aggregates them across channels, and produces a detection logit. The context branch reduces an (unlabeled) context batch, namely the current forward batch of windows itself, into a single latent vector $\mathbf{z}$, which conditions the trunk at every normalization and feature-modulation point. The two branches interact only through $\mathbf{z}$: the trunk never sees the raw context statistics, and the context branch never sees trunk gradients through the detached context input.

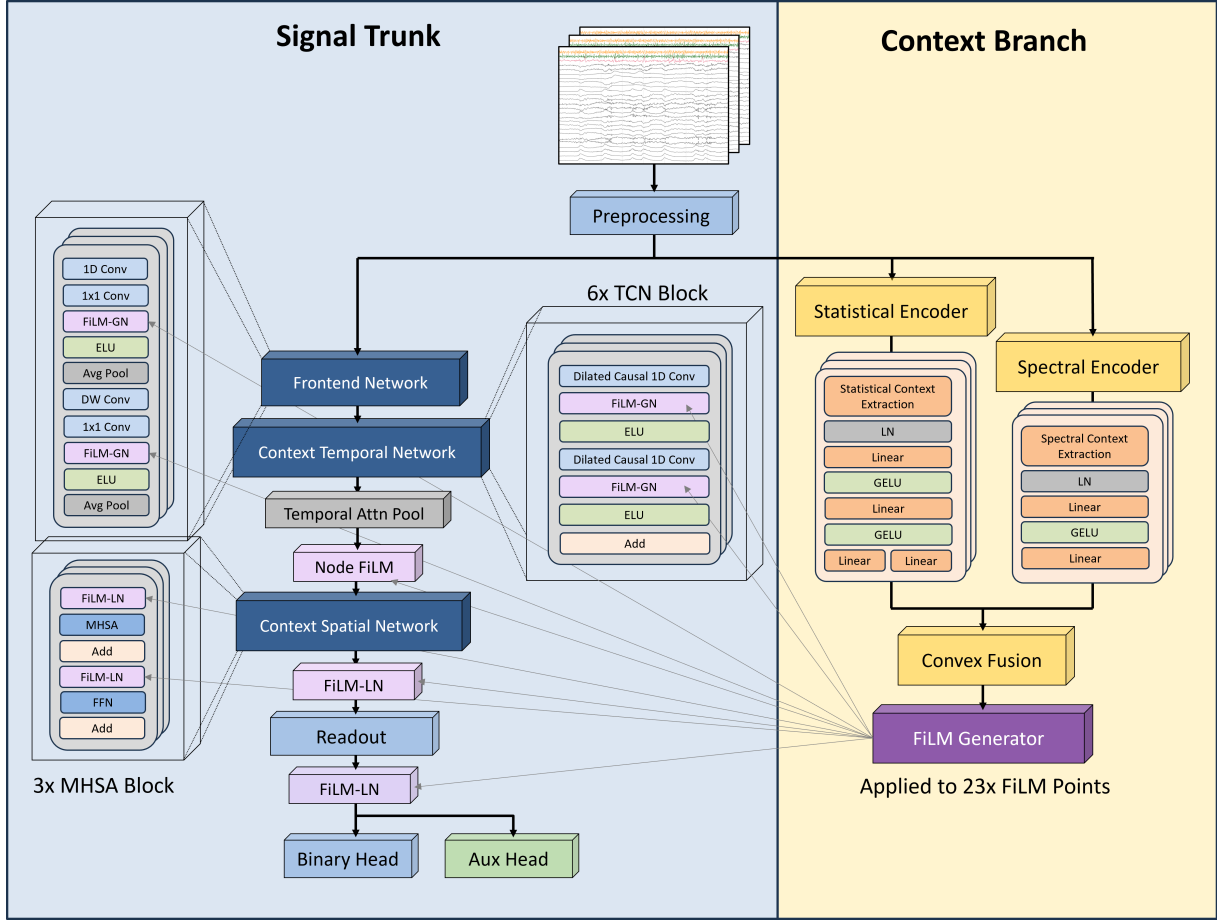


Figure 1: Overall architecture. The signal trunk (blue shaded) processes the multimodal window through robust normalization, a shared per-node frontend, a dilated temporal convolutional network, temporal attention pooling with node-level modulation, a channel-wise Transformer, and two prediction heads. The context branch (yellow shaded) extracts prior statistical and spectral descriptors, encodes them into time and frequency encoder, and fuses them by a per-dimension convex gate into the context latent z , which drives all twenty-three FiLM points (pink blocks).

3.2 Input Representation and Preprocessing

Each sample is a 26-channel window comprising 23 EEG-side channels, one ECG channel, and two EMG channels, of length 2000 samples at 500 Hz (4 seconds). We apply a single zero-phase, forward-backward 0.3 Hz high-pass biquad to remove slow baseline drift while preserving the timing and shape of fast transients; no down-sampling or band-limiting is performed, so discharge morphology is left intact. Each channel is then normalized independently per sample with a robust median/MAD z-score,

$$\tilde{x}_{c,t} = \text{clip}\left(\frac{x_{c,t} - m_c}{1.4826 \text{MAD}_c + \epsilon}, -8, 8\right), \quad (1)$$

where $m_c = \text{median}_t(x_{c,t})$, $\text{MAD}_c = \text{median}_t|x_{c,t} - m_c|$, and $\epsilon = 10^{-3}$. The median/MAD estimator and the ± 8 clip make the normalization insensitive to the very spikes and artifacts we wish to detect, which a mean/standard-deviation scaling would let dominate. During training we additionally apply EEG channel dropout, which zeroes a small random subset (up to three) of the EEG channels per

sample, regularizing the model against channel-level acquisition failures while leaving the ECG and EMG channels intact.

3.3 Signal Trunk

The trunk treats each of the 26 channels as a node and proceeds in two stages: within-node temporal modeling, then cross-node spatial modeling.

Per-node frontend. A single EEGNet-style stem is shared across all nodes and maps each raw one-dimensional sequence to an $F_2=64$ -dimensional temporal feature map. It applies a temporal convolution, a 1×1 convolution that lifts the channel dimension, a FiLM-conditioned group normalization, an ELU non-linearity, average pooling, and dropout; then a depthwise convolution, a pointwise 1×1 convolution, a second FiLM-conditioned group normalization, ELU, average pooling, and dropout. The two pooling stages reduce the 2000-sample input to a 125-step feature map per node. This is the same lightweight, physiologically motivated inductive bias

popularized for compact EEG decoding [16], here made context-adaptive through its two FiLM normalizations.

Context temporal network. Temporal context is modeled by six residual blocks of a dilated causal temporal convolutional network shared across nodes [5, 11]. Block l contains two weight-normalized dilated causal convolutions of kernel size 5 with dilation 2^l for $l = 0, \dots, 5$ (i.e. 1, 2, 4, 8, 16, 32), each followed by a FiLM-conditioned group normalization, an ELU, and dropout, with a residual connection:

$$\mathbf{h}^{(l+1)} = \text{ELU}\left(\mathbf{h}^{(l)} + G_l\left(\mathbf{h}^{(l)}, \mathbf{z}\right)\right), \quad (2)$$

where G_l is the two-convolution, FiLM-modulated transformation and $\mathbf{z}$ is the context latent. The exponentially growing dilation gives the six-block stack a receptive field that spans the whole window while keeping the block count small. A temporal attention-pooling layer then collapses each node from 125 steps to a single 64-dimensional vector, yielding a node representation of shape $[B, 26, 64]$.

Node modulation and channel-wise Transformer. Before spatial mixing, a node-level FiLM transform modulates the pooled node features and a learned node positional embedding is added. This is the only point where FiLM acts directly on a feature map rather than on a normalized activation. Cross-channel and cross-modal structure is then modeled by three Transformer blocks operating over the 26 nodes [27]. Each block applies a FiLM-conditioned layer normalization, multi-head self-attention with four heads, a residual addition, a second FiLM-conditioned layer normalization, a feed-forward network with ELU activation, and a second residual addition. Separating the stages in this way lets the temporal network specialize in within-channel discharge morphology while the Transformer learns which channels and modalities agree—so that, for example, a sharp transient consistent across neighboring EEG nodes but absent from EMG is treated differently from one that coincides with a muscle burst.

Readout and heads. A final FiLM-conditioned layer normalization is followed by a readout that concatenates node-wise mean pooling and node-wise attention pooling into a 128-dimensional vector, which passes through a last FiLM-conditioned layer normalization. A two-layer MLP produces the binary IED logit. During training only, a second two-layer MLP predicts one of five discharge subtypes from the same pooled vector (Section 3.5).

3.4 Fused Statistical–Spectral Context

The context branch (Figure 2) summarizes a context batch $\mathbf{c}$; when no external context is supplied, the current batch is reused, and in all cases $\mathbf{c}$ is detached so no trunk gradient flows into it. We distinguish a prior feature extractor Φ from an encoding network E : Φ produces interpretable physical descriptors, and E maps them to a latent.

Statistical encoder. Φ_{time} computes six statistics per channel—mean, log standard deviation, median, log scaled MAD, log root-mean-square, and log line length—and concatenates them over the 26 channels into a 156-dimensional descriptor. The encoder E_{time} is a layer normalization followed by two linear–GELU stages and

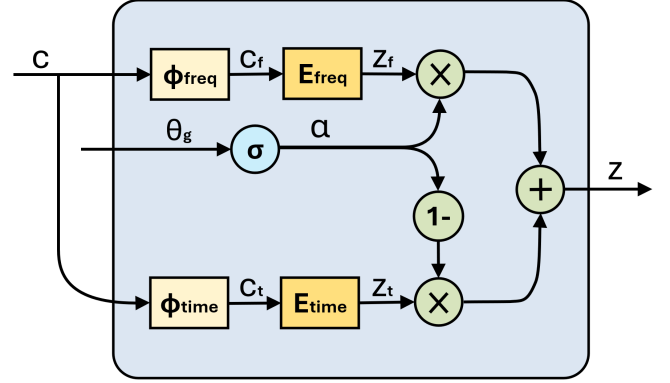


Figure 2: Statistical–spectral convex fusion. Prior descriptors Φ_{time} and Φ_{freq} are extracted from the detached context batch and encoded into $\mathbf{z}_{\text{time}}$ (variational) and $\mathbf{z}_{\text{freq}}$ (zero-initialized). A learned per-dimension gate $\alpha = \sigma(\theta_g)$ forms the context latent $\mathbf{z}$ that drives FiLM generation.

two parallel linear heads producing a mean μ and a log-variance $\log \sigma^2$ over a 32-dimensional latent. We treat this as a variational encoding and, during training, sample

$$\mathbf{z}_{\text{time}} = \mu + \tau \epsilon \odot \exp\left(\frac{1}{2} \log \sigma^2\right), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

with a small noise scale τ ; at inference $\mathbf{z}_{\text{time}} = \mu$. The log-variance head is initialized to a small value, so the injected noise is negligible early in training and acts as a mild stochastic regularizer that discourages the context latent from overfitting recording-specific idiosyncrasies.

Spectral encoder. Φ_{freq} computes eighteen modality-aware descriptors, six per modality, from a Hann-windowed periodogram. For EEG it takes the log band powers in the delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), low-beta (13–16 Hz), high-beta (16–30 Hz), and gamma (30–80 Hz) bands, *mean-pooled across the 23 EEG channels* so that the EEG contributes a single representative six-vector rather than a per-channel one. For ECG it takes log RMS, log mean absolute slope, log power in the heart-rate band (0.7–3.5 Hz), the normalized dominant heart-rate frequency, log power in a low-frequency modulation band (0.04–0.5 Hz), and log power in the QRS band (5–25 Hz). For EMG it takes log RMS, the zero-crossing rate, the normalized mean and median frequencies, and log power in the 20–50 Hz and 50–150 Hz bands. The encoder E_{freq} is a layer normalization, a linear–GELU stage, and a final linear projection to the same 32-dimensional latent; its final layer is zero-initialized, so $\mathbf{z}_{\text{freq}} = \mathbf{0}$ at the start of training.

Convex fusion. The context latent is a per-dimension convex combination of the two encodings,

$$\begin{aligned} \mathbf{c}_t &= \Phi_{\text{time}}(\mathbf{c}), \\ \mathbf{c}_f &= \Phi_{\text{freq}}(\mathbf{c}), \\ \mathbf{z}_t &= E_{\text{time}}(\mathbf{c}_t), \\ \mathbf{z}_f &= E_{\text{freq}}(\mathbf{c}_f), \\ \alpha &= \sigma(\theta_g), \\ \mathbf{z} &= (1 - \alpha) \odot \mathbf{z}_t + \alpha \odot \mathbf{z}_f, \end{aligned} \quad (4)$$

where $\theta_g \in \mathbb{R}^{32}$ is a learned per-dimension gate and $\odot$ is element-wise product. Unlike a single scalar mixing weight or an unconstrained additive fusion [4], the per-dimension convex gate lets each latent coordinate independently decide how much to rely on the spectral description, while the convex constraint keeps the fused latent on the same scale as its constituents. Because $\mathbf{z}_{\text{freq}}$ starts at zero, $\mathbf{z} = \mathbf{z}_{\text{time}}$ at initialization: the model begins as a purely statistics-conditioned network and admits spectral information only as it earns gradient, a warm start that we found important for stable training. After training, α is a one-tensor readout of how much each dimension came to depend on frequency content.

3.5 FiLM Conditioning and Auxiliary Supervision

Every conditioning point uses the same feature-wise linear modulation operator [22]. For a point with C channels, a small generator maps the shared latent to a scale and a shift,

$$\begin{aligned} (\gamma, \beta) &= s \cdot \tanh(W_2 \text{GELU}(W_1 \mathbf{z})), \\ \hat{\mathbf{h}} &= \text{Norm}(\mathbf{h}) \odot (1 + \gamma) + \beta, \end{aligned} \quad (5)$$

where $W_1 \in \mathbb{R}^{64 \times 32}$, $W_2 \in \mathbb{R}^{2C \times 64}$, Norm is the point's affine-free group or layer normalization (identity at the single node-level point), and $s = 0.5$ bounds the modulation. Each generator's output layer is zero-initialized, so at the start of training $\gamma = \beta = \mathbf{0}$ and every FiLM point is an identity map, giving the whole conditioning mechanism a warm start consistent with the zero-initialized spectral path. The trunk contains twenty-three such points: two in the frontend, twelve in the six temporal blocks, one at the node representation, six in the three Transformer blocks, and two in the final and head normalizations. Distributing conditioning this densely lets the context latent reshape the computation at every scale—filter responses, temporal residuals, attention statistics, and the pooled readout—rather than only at the input.

The primary objective is binary IED detection. During training, the same pooled 128-dimensional readout also feeds a five-way head predicting the discharge subtype of positive windows; this label is undefined for negatives and is masked out of the auxiliary loss. The total loss is

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left[\mathcal{L}_{\text{BCE}}(y_i, \hat{y}_i) + \lambda_{\text{aux}} y_i \mathcal{L}_{\text{CE}}(u_i, \hat{u}_i) \right], \quad (6)$$

where i indexes the N windows in a batch and $u_i \in \{1, \dots, 5\}$ is the discharge subtype. Since $y_i \in \{0, 1\}$, the factor y_i zeroes the auxiliary term on negative windows, so the subtype loss is applied to positive windows only. The auxiliary head is a purely training-time regularizer that encourages subtype-discriminative structure in the positive representation, in the spirit of using type labels to sharpen detection [1, 25]; it is discarded at inference, so the deployed model is exactly the binary trunk. We use $\lambda_{\text{aux}} = 0.2$ in the final single model.

4 Experiments

4.1 Dataset and Protocol

We evaluate on Track 1 of the NeuroMM 2026 Grand Challenge [21], built on the clinically grounded vEpiSet recordings [19]. The benchmark comprises 84 subjects (52 with epilepsy, 32 controls) and roughly 25,400 four-second epochs, each a 26-channel window (23 EEG, one ECG, two EMG) at 500 Hz, split patient-disjointly into 20,298 training and 5,128 validation epochs. Positive (IED-containing) epochs are markedly rare—about 2,500 against nearly 23,000 negatives in the source recordings [19]—so the problem is severely class-imbalanced. Each epoch carries a binary label indicating whether it contains an interictal epileptiform discharge and, for positive epochs, one of five discharge categories defined by spatial distribution (generalized, frontal, temporal, occipital, and centroparietal). Because discharges are rare and the clinically meaningful goal is to rank suspicious windows ahead of abundant background, we take the area under the precision–recall curve (AUPRC) as the primary metric and report accuracy (ACC), precision, recall, binary F1, and weighted F1 (WF1) as secondary threshold-dependent diagnostics.

The released recordings provide a 29-electrode sequence, Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz, PG1, PG2, A1, A2, ECG1, ECG2, EMG1, EMG2, EMG3, EMG4 [19]. Following the official baseline preprocessing [21], we map this sequence to the 26-channel model input. The first 23 channels are kept in order and comprise the 19 standard 10–20 scalp electrodes together with four auxiliary/reference electrodes (PG1, PG2, A1, A2); the three remaining physiological channels are formed as bipolar derivations $\text{ECG} = \text{ECG1} - \text{ECG2}$, $\text{EMG1} = \text{EMG1} - \text{EMG2}$, and $\text{EMG2} = \text{EMG3} - \text{EMG4}$. The resulting input order is the 23 EEG-side channels Fp1, $\dots$, A2, followed by ECG, EMG1, and EMG2. This alignment matches the channel layout visualized in Figure ?? and consumed by every model in this paper, and Figure ?? illustrates why the task benefits from all three modalities: genuine cortical discharges must be distinguished from the muscular and cardiac transients that dominate the EMG and ECG traces.

4.2 Training Details

All models are trained from scratch under the patient-disjoint split with mini-batch optimization and a cosine schedule with warmup. The final single model uses the temporal-plus-Transformer backbone with fused statistical and spectral context and $\lambda_{\text{aux}} = 0.2$. Robust median/MAD normalization, the 0.3 Hz high-pass, EEG channel dropout, and positive-only subtype supervision are applied as described above; the auxiliary head is active only during training. For the ensemble results we independently train models on different patient-wise splits and average their predicted probabilities at inference. We compare against compact EEG baselines re-trained under the same protocol—EEGNet in a small and a medium configuration [16], EEG-TCNet [11], ATCNet in a small and a medium configuration [2], and a one-dimensional ResNet-18—and against architectural and context ablations of our own model.

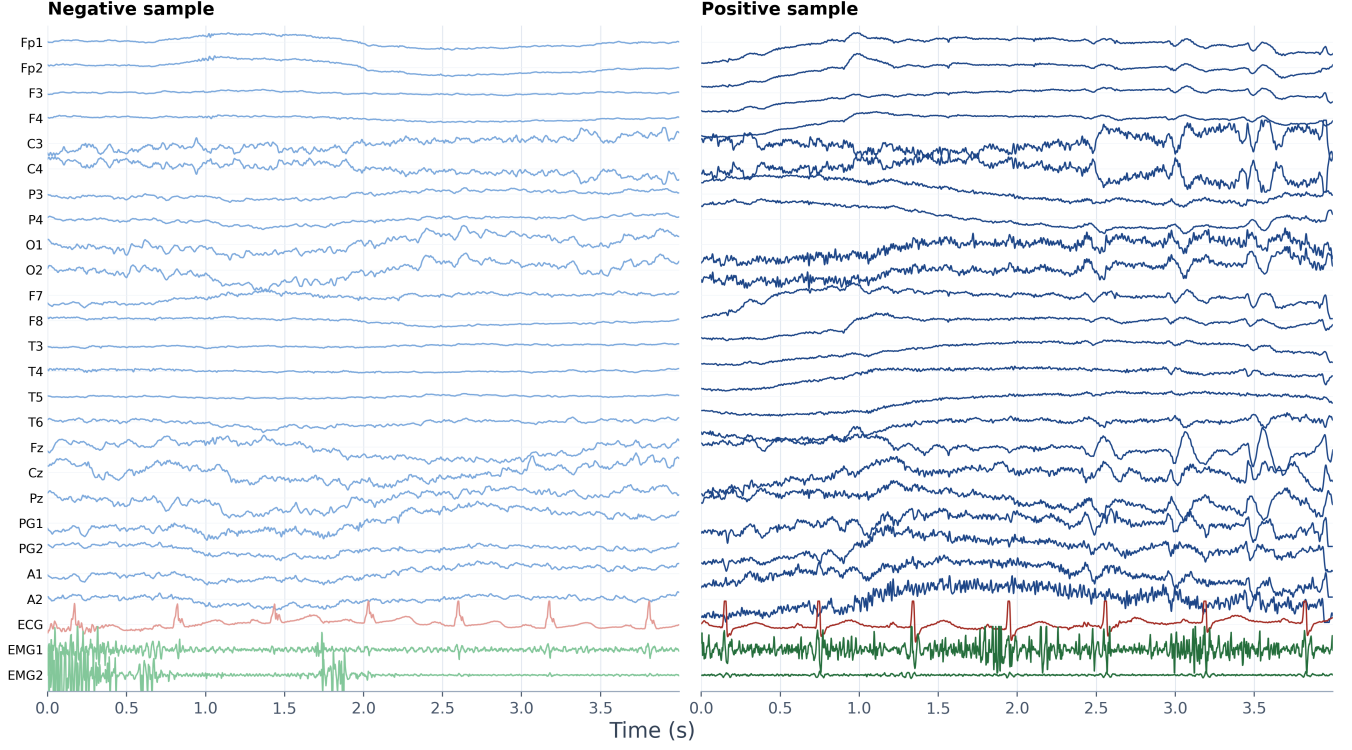


Figure 3: Representative 26-channel physiological windows from the NeuroMM 2026 Track 1 dataset [19]. The left panel shows a negative window without an IED, and the right panel shows a positive window containing an interictal epileptiform discharge. The traces are displayed after the project-defined channel alignment and median centering and scaling applied only for visualization. The model input contains 23 EEG-related channels, one bipolar ECG channel, and two bipolar EMG channels.

5 Results

5.1 Comparison with Baselines

Table 1 compares the proposed model with compact EEG baselines. The strongest baseline by AUPRC, EEGNet-m, reaches 0.8225, whereas the proposed model reaches 0.9531, an absolute gain of 0.1306. Among the baselines, ATCNet-m attains the highest precision of 0.9132 but a recall of only 0.4902, while EEGNet-m provides a more balanced operating point. The proposed model achieves both substantially better ranking and a stronger balance between precision and recall. Under the same patient-disjoint protocol, these results support the value of combining multimodal spatio-temporal modeling with recording-level context conditioning.

5.2 Architectural Component Ablation

Table 2 isolates the temporal TCN, spatial Transformer, and context-conditioning branch. The TCN-only model reaches an AUPRC of 0.9145, compared with 0.8550 for the Transformer-only model, showing that within-channel temporal morphology provides the stronger individual representation for IED detection. Combining the TCN and Transformer increases AUPRC to 0.9336, and adding context conditioning further raises it to 0.9531, an improvement of 0.0195 over the same spatio-temporal architecture without context. Context also improves the TCN-only model from 0.9145 to 0.9204,

but does not benefit the Transformer-only model, indicating that the conditioning branch is most effective when applied to representations with explicit temporal structure. The complete model achieves the best accuracy, precision, F1, and WF1, while the TCN with context produces the highest recall.

5.3 Context-Branch Ablation

Table 3 examines the two encoders within the context-conditioning branch. Statistical-only and spectral-only conditioning reach similar AUPRC values of 0.9330 and 0.9347, whereas their fusion raises AUPRC to 0.9531, corresponding to gains of 0.0201 and 0.0184, respectively. This improvement shows that time-domain statistics and modality-aware spectral descriptors provide complementary context rather than interchangeable summaries. Spectral conditioning achieves the highest recall but has substantially lower precision, while the fused representation provides the best accuracy, precision, F1, and WF1 and a more balanced operating point.

5.4 Multi-Split Ensembling

Table 4 reports probability averaging over models trained on different patient-wise splits. The single model attains an AUPRC of 0.9531. Averaging two, four, and six models raises AUPRC to 0.9598, 0.9670, and 0.9693, respectively, while precision increases to approximately 0.94 and recall remains near 0.88. The larger and steadier

Table 1: Comparison with compact EEG baselines on NeuroMM 2026 Track 1 using patient-disjoint validation.

Model	AUPRC	ACC	Precision	Recall	F1	WF1
EEGNet-s	0.7060	0.8446	0.7766	0.5196	0.6226	0.8332
EEGNet-m	0.8225	0.8809	0.8371	0.6422	0.7268	0.8752
EEG-TCNet	0.8118	0.8634	0.8889	0.5098	0.6480	0.8493
ATCNet-s	0.7592	0.8628	0.7608	0.6471	0.6993	0.8589
ATCNet-m	0.8058	0.8628	0.9132	0.4902	0.6380	0.8469
ResNet-18	0.7203	0.8525	0.8306	0.5049	0.6280	0.8389
Proposed	0.9531	0.9250	0.8622	0.8284	0.8450	0.9245

Table 2: Ablation results of the architectural components

Model Components	AUPRC	ACC	Precision	Recall	F1	WF1
TCN only	0.9145	0.9141	0.7782	0.9118	0.8397	0.9163
TCN + Context	0.9204	0.9365	0.8404	0.9167	0.8769	0.9374
Transformer only	0.8550	0.8918	0.7855	0.7721	0.7787	0.8915
Transformer + Context	0.8139	0.8706	0.8540	0.5735	0.6862	0.8612
TCN + Transformer	0.9336	0.9359	0.8719	0.8676	0.8698	0.9359
TCN + Transformer + Context (Proposed)	0.9531	0.9498	0.8973	0.8995	0.8984	0.9498

Table 3: Ablation results of statistical and spectral encoders in the context-conditioning branch

Context encoder	AUPRC	ACC	Precision	Recall	F1	WF1
Statistical only	0.9330	0.9341	0.8620	0.8725	0.8672	0.9342
Spectral only	0.9347	0.8996	0.7309	0.9387	0.8219	0.9034
Statistical + Spectral (Proposed)	0.9531	0.9498	0.8973	0.8995	0.8984	0.9498

Table 4: Multi-split probability-averaging ensemble of the final model

Ensemble	AUPRC	ACC	Precision	Recall	F1	WF1
Single	0.9531	0.9468	0.8774	0.9118	0.8942	0.9471
2 models	0.9598	0.9571	0.9517	0.8701	0.9091	0.9564
4 models	0.9670	0.9553	0.9349	0.8799	0.9066	0.9548
6 models	0.9693	0.9571	0.9399	0.8824	0.9102	0.9566

improvement in AUPRC than in accuracy suggests that averaging mainly reduces split-specific ranking uncertainty caused by patient heterogeneity. Models trained on different splits observe different patient cohorts, so probability averaging can suppress errors specific to any one split. The absolute AUPRC gain of 0.0162 from six-model ensembling indicates that cross-patient generalization remains a central challenge [23, 24]. Evaluation is nevertheless limited to one benchmark and its patient-disjoint validation protocol. Cross-dataset and cross-site transfer remain untested, and relating the learned gate α to clinical recording properties is an important direction for further analysis.

6 Conclusion

We presented a context-conditioned spatio-temporal network for interictal epileptiform discharge detection from multimodal EEG, ECG, and EMG signals. The model separates within-node temporal

morphology, captured by a shared frontend and a dilated temporal convolutional network, from cross-node spatial and cross-modal structure, captured by a channel-wise Transformer, and conditions the entire trunk on a recording-level latent through feature-wise modulation at twenty-three points. The latent is a per-dimension convex combination of a variational statistical embedding and a modality-aware spectral embedding, warm-started to recover the statistical model at initialization. Discharge-subtype labels supply auxiliary weak supervision. The single model reaches 0.9531 AUPRC on NeuroMM 2026 Track 1 and multi-split ensembling reaches 0.9693; ablations attribute the gains, in order, to temporal modeling, fused context conditioning, cross-channel aggregation, and moderate auxiliary supervision. We hope the design—adapting a fixed backbone to each recording rather than enlarging it—offers a useful template for the distribution-shift-limited regime that characterizes clinical physiological signal analysis.

7 Acknowledgments

This work has been partially funded by Hector Stiftung.

References

- [1] Hezam Albaqami, Ghulam Mubashar Hassan, and Amitava Datta. 2023. MP-SeizNet: A multi-path CNN Bi-LSTM network for seizure-type classification using EEG. *Biomedical Signal Processing and Control* 84 (2023), 104780. doi:10.1016/j.bspc.2023.104780
- [2] Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. 2023. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics* 19, 2 (2023), 2249–2258. doi:10.1109/TH.2022.3197419
- [3] Andrea Apicella, Pasquale Arpaia, Giovanni D'Errico, Davide Marocco, Giovanna Mastrati, Nicola Moccaldi, and Roberto Prevete. 2024. Toward cross-subject and cross-session generalization in EEG-based emotion recognition: Systematic review, taxonomy, and methods. *Neurocomputing* 604 (2024), 128354. doi:10.1016/j.neucom.2024.128354
- [4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. In *ICLR 2017 Workshop Track*. arXiv:1702.01992 [stat.ML] doi:10.48550/arXiv.1702.01992
- [5] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018). arXiv:1803.01271 [cs.LG] doi:10.48550/arXiv.1803.01271
- [6] Yoon Gi Chung, Woo-Jin Lee, Sung Min Na, Hunmin Kim, Hee Hwang, Chang-Ho Yun, and Ki Joong Kim. 2023. Deep learning-based automated detection and multiclass classification of focal interictal epileptiform discharges in scalp electroencephalograms. *Scientific Reports* 13 (2023), 6755. doi:10.1038/s41598-023-33906-5
- [7] Robert S. Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J. Helen Cross, Christian E. Elger, Jerome Engel, Jr., et al. 2014. ILAE official report: A practical clinical definition of epilepsy. *Epilepsia* 55, 4 (2014), 475–482. doi:10.1111/epi.12550
- [8] GBD Epilepsy Collaborators. 2025. Global, regional, and national burden of epilepsy, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Public Health* 10, 3 (2025), e203–e227. doi:10.1016/S2468-2667(24)00302-5
- [9] Arash Hajisafi, Haowen Lin, Yao-Yi Chiang, and Cyrus Shahabi. 2024. NeuroGNN: Dynamic graph neural networks for precise seizure detection and classification. In *Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science, Vol. 14648)*. Springer, Singapore, 207–220. doi:10.1007/978-981-97-2238-9_16
- [10] L. Hu, K. Zhao, B. W.-K. Ling, S. Liang, and Y. Wei. 2025. Improving human activity recognition via graph attention network with linear discriminant analysis and residual learning. *Biomedical Signal Processing and Control* 100 (2025), 107053.
- [11] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. 2020. EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2958–2965. doi:10.1109/SMC42975.2020.9283028
- [12] Anice Jahanjoo, Yiting Wei, Mostafa Haghi, Philip Schörpf, and Nima TaheriNejad. 2025. Hybrid CNN-Transformer Model for Accurate Classification of Human Attention Levels Using Workplace EEG Data. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–6.
- [13] A. Jahanjoo, Y. Wei, S. Khooyooz, M. Haghi, and N. TaheriNejad. 2025. From Clean Labs to Noisy Lives: Real-World Stress Detection Using Spectrogram-Based Transformers on PPG Signals. In *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 1–7.
- [14] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. 2024. LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=QzTpTRVtrP>
- [15] R. Kemmerich, A. Wienke, U. Frischen, and B. Mathes. 2025. Evaluating machine- and deep learning approaches for artifact detection in infant EEG: Classifier performance, certainty, and training size effects. *Biomedical Physics & Engineering Express* 11, 3 (2025), 035029. doi:10.1088/2057-1976/add740
- [16] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. 2018. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering* 15, 5 (2018), 056013. doi:10.1088/1741-2552/aaec8c
- [17] Zi Liang, Peng Hu, Lian Li, Zebang Cheng, Yisu Dong, Haibo He, Qiang Lu, and Nan Lin. 2025. EEG-VL: Integrating Visual Features with Large Language Models for Automated Seizure Detection. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 4620–4627.
- [18] Nan Lin, Lian Li, Weifang Gao, Peng Hu, Gonglin Yuan, Heyang Sun, et al. 2025. Development and validation of a multimodal automatic interictal epileptiform discharge detection model: A prospective multi-center study. *BMC Medicine* 23 (2025), 479. doi:10.1186/s12916-025-04316-3
- [19] Nan Lin, Mengxuan Zheng, Lian Li, Peng Hu, Weifang Gao, Heyang Sun, Chang Xu, Gonglin Yuan, Zi Liang, Yisu Dong, et al. 2025. An EEG dataset for interictal epileptiform discharge with spatial distribution information. *Scientific Data* 12 (2025), 229. doi:10.1038/s41597-025-04572-1
- [20] Fei Ma, Han Lin, Yifan Xie, Hongwei Ren, Xiaoyu Shen, Wenbo Ding, and Qi Tian. 2026. E²-LLM: Bridging Neural Signals and Interpretable Affective Analysis. *arXiv preprint arXiv:2601.07877* (2026).
- [21] NeuroMM 2026 Organizers. 2026. NeuroMM 2026: Interictal Epileptiform Discharge Detection and Localization in Multimodal Neuro-Signals. ACM Multimedia 2026 Grand Challenge. Accessed: 2026-07-17. <https://2026.neuromm.org/>
- [22] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 3942–3951. doi:10.1609/aaai.v32i1.11671
- [23] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. 2019. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering* 16, 5 (2019), 051001. doi:10.1088/1741-2552/ab260c
- [24] Khaled Saab, Jared Dunnmon, Christopher Ré, Daniel Rubin, and Christopher Lee-Messer. 2020. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *npi Digital Medicine* 3 (2020), 59. doi:10.1038/s41746-020-0264-0
- [25] Siyi Tang, Jared A. Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L. Rubin, and Christopher Lee-Messer. 2022. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=k9bxIEfH1_-
- [26] John Thomas, Jing Jin, Prasanth Thangavel, Elham Bagheri, Rajamanickam Yuvavaraj, Justin Dauwels, Rahul Rathakrishnan, Jonathan J. Halford, Sydney S. Cash, and M. Brandon Westover. 2020. Automated Detection of Interictal Epileptiform Discharges from Scalp Electroencephalograms by Convolutional Neural Networks. *International Journal of Neural Systems* 30, 11 (2020), 2050030. doi:10.1142/S0129065720500306
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 5998–6008. <https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need>
- [28] Qitong Wang, Stephen Whitmarsh, Vincent Navarro, and Themis Palpanas. 2022. iDeaL: A deep learning framework for detecting highly imbalanced interictal epileptiform discharges. *Proceedings of the VLDB Endowment* 16, 3 (2022), 480–490. doi:10.14778/3570690.3570698
- [29] Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. 2023. BIOT: Biosignal Transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 78240–78260. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html
- [30] Jingwei Zhang, Lauren Swinnen, Christos Chatzichristos, Victoria Broux, Renee Proost, Katrien Jansen, et al. 2024. Multimodal wearable EEG, EMG and accelerometry measurements improve the accuracy of tonic-clonic seizure detection. *Physiological Measurement* 45, 6 (2024), 065005. doi:10.1088/1361-6579/ad4e94
- [31] Ruizhe Zheng, Jun Li, Yi Wang, Tian Luo, and Yuguo Yu. 2023. ScatterFormer: Locally-invariant scattering Transformer for patient-independent multispectral detection of epileptiform discharges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 148–158. doi:10.1609/aaai.v37i1.25086