GTRE: A Physiological-Semantic Foundation Model for Generalizable EEG Analysis via Preference-Guided Alignment

Mohd Azfar[1(✉)⋆] and Izhar Dad Khan[2⋆]

[1] Indian Institute of Science, Bengaluru, India
azfar45@gmail.com
[2] Basque Center on Cognition, Brain and Language, San Sebastián, Spain
i.khan@bcbl.eu

Abstract. Clinical EEG interpretation relies on expert narratives, yet existing self-supervised methods often ignore this textual context. We propose a two-stage multimodal framework using a novel **Graph Temporal Relational Encoder (GTRE)** that learns montage-agnostic features grounded in 3D brain topology. **Stage 1** establishes domain-invariant representations across datasets via multi-domain grounding and prototype-based regularization. **Stage 2 (Masked-DPO)** introduces a preference-guided objective using expert-validated summary triplets (Fleiss' $\kappa = 0.71$, 90.1% accuracy) to prioritize clinically salient features. By employing a similarity-scaled margin and pathology-aware masking, our model adaptively focuses on diagnostic intent. Results show superior cross-institutional transfer, achieving 96.0% accuracy with only 5% supervision and a +12.7% gain for under-represented seizure types. This highlights the power of preference-based alignment for developing clinically-grounded EEG foundation models.

Keywords: Epilepsy · EEG · Seizure Classification · Transformer · Contrastive Learning · Multimodal Pretraining · Preference Based Pretraining · Clinical Text

1 Introduction

Electroencephalogram (EEG) recordings sourced from diverse clinical environments exhibit significant heterogeneity due to varying electrode configurations, hardware specifications, annotation protocols, and patient demographics. These inconsistencies induce substantial domain shifts across benchmark datasets such as TUSZ [1] and TUAB [2], often rendering models trained on a single source unreliable for real-world deployment. While self-supervised EEG methods have shown promise [5], cross-modal alignment of EEG with clinical text remains underexplored. To address these challenges, we leverage weak supervision from clinical text including neurologist reports and annotations to learn domain-invariant, semantically grounded EEG representations.

⋆ Equal contribution. ✉ Corresponding author.

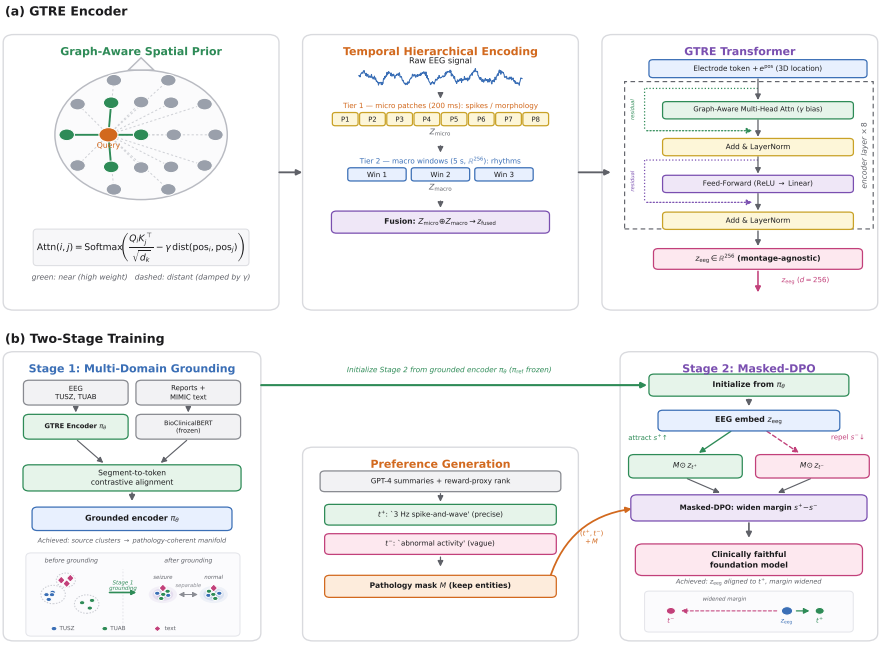

Fig. 1. GTRE architecture overview

Our contributions are threefold. **(1) Encoder:** we propose the Graph-Temporal Relational Encoder (GTRE), whose representations are biologically consistent and montage-agnostic. **(2) Multi-Domain Grounding (Stage 1):** a contrastive framework aligns EEG segments with clinical reports in a shared semantic space while mitigating global embedding drift across disparate datasets. **(3) Masked-DPO alignment (Stage 2):** a preference-based objective focuses the encoder on clinically critical features rather than superficial lexical patterns, built on preference pairs (t^+, t^-) that share semantics but differ in diagnostic precision, a pathology-aware mask (M) that restricts gradients to clinical entities, and a similarity-scaled margin that modulates supervision strength by the informativeness gap. We evaluate on seizure classification and abnormality detection, demonstrating strong in-domain, cross-domain, and few-shot performance.

2 Related Work

EEG foundation models. Recent large-scale encoders learn transferable EEG representations via masked or tokenized self-supervision: LaBraM [15] uses discrete neural tokens, FEMBA [9] scales masked pretraining, EEGFormer [7] targets transferable and interpretable features, and BioSerenity-E1 [16] studies

clinical-scale pretraining, building on MAE-style and self-supervised work [10,5]. These are largely signal-only and montage-specific. **EEG–text and multimodal alignment.** CLIP-style cross-modal objectives [11] with clinical text encoders [4] enable weak supervision, yet prior EEG work rarely grounds segments in clinical entities or refines alignment by diagnostic preference. GTRE differs through montage-agnostic spatial grounding and preference-guided refinement; we compare against these baselines at similar parameter budgets (∼32M) and comparable pretraining volume, so the reported gains are architectural rather than from scale.

3 Datasets

We construct EEG–text pairs from two clinical EEG corpora, TUSZ and TUAB, and additionally draw on ICU/neurology clinical *text* from MIMIC-IV (which contains no EEG) to enable multimodal representation learning under domain shift in both signal and language.

TUSZ (Temple University Seizure Corpus). We selected 828 single-seizure-type recordings, resampled to 250 Hz and mapped to 20 channels. Each recording includes neurologist reports, allowing precise EEG–text alignment for seizure morphology and localization, ideal for fine-grained supervision.

TUAB (Temple University Abnormal EEG Corpus). TUAB includes 3,295 abnormal recordings with associated reports describing a wide range of pathologies. Though seizures are rare, TUAB introduces linguistic and physiological diversity, supporting learning of generalizable EEG-text semantics.

MIMIC-IV. We leverage clinical narratives from the MIMIC-IV dataset [3] to enrich our clinical-text resource. These ICU and neurology notes improve the quality of GPT-generated summaries by introducing domain-specific linguistic nuances, independent of EEG data.

Text Pair Construction. We extracted relevant sections (e.g., *Findings*, *Impression*) and used GPT-4-based summarization to create concise (100–150 word) descriptions. We also curated a fixed **clinical entity vocabulary** (waveform patterns, seizure types, behavioral cues) used only for the pathology mask M and entity alignment, distinct from the *learnable* prototypes P of Sec. 3. Three certified experts validated 1000 random samples under a two-task rubric: whether (1) the source meaning from *Findings*/*Impression* was preserved, and (2) t^+ was more diagnostically informative than t^-; uncertainty terms ("possible", "unclear") were kept, not stripped, to avoid overconfident supervision. Agreement was Fleiss' $\kappa = 0.71$ [8], with 90.1% of summaries and 87.9% of preference pairs validated by $\geq$2 experts ($p < 0.01$). This pipeline yields semantically aligned EEG–text pairs across datasets, enabling contrastive pretraining over clinically heterogeneous sources despite variation in annotation granularity.

4 Methodology

To address the inherent heterogeneity across clinical corpora, where electrode configurations, sampling rates, and noise profiles vary significantly, we depart from standard channel-wise tokenized transformers [20]. Instead, we propose a **Graph-Temporal Relational Encoder (GTRE)** designed to learn montage-agnostic representations by grounding the transformer in the physical topology of the brain.

Topographical Coordinate Embedding. Standard transformers treat EEG channels as an unordered list, ignoring spatial relationships critical for localized seizure detection. We map each channel c_i from dataset $D_m \in \{\texttt{TUSZ}, \texttt{TUAB}\}$ to 3D coordinates (x_i, y_i, z_i) based on the International 10–20 system. These are projected through a spatial MLP to produce a **location embedding** $e_i^{\text{pos}} = \text{MLP}_{\text{geo}}(x_i, y_i, z_i) \in \mathbb{R}^d$, which is added to temporal features $h_{m,i}^{\text{eeg}}$. This treats each electrode as a spatially anchored token, enabling robustness to varying montages and improving cross-domain transfer.

Graph-Aware Self-Attention. To capture the dynamic functional connectivity and seizure localization described in neurologist reports, we modify the self-attention mechanism to include a **Geometric Bias**. The attention weight between channel i and channel j is augmented by their physical proximity on the scalp:

$$\text{Attn}(i,j) = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}} - \gamma \cdot \text{dist}(\text{pos}_i, \text{pos}_j)\right) \tag{1}$$

where $\text{dist}(\cdot)$ is the Euclidean distance between electrodes and γ a learnable scalar. This is a **learnable soft bias**, not a hard mask: the full $Q_i K_j^\top$ content score is preserved, so distant electrodes still attend freely. It regularizes montage noise while remaining content-agnostic, and since γ is learned it does not suppress genuine long-range coordination from volume conduction or global rhythmic synchrony (e.g., generalized seizures).

Temporal Hierarchical Processing. To preserve high-frequency morphology (e.g., spikes) over long recordings, we use two scales. Signals are tokenized into overlapping 200 ms **micro-patches** Z_{micro}, whose intra-patch positional encoding retains local transients (a token does *not* collapse all temporal information); these are aggregated by self-attention into fixed, non-overlapping 5 s **macro-segments** $Z_{\text{macro}} \in \mathbb{R}^{256}$ encoding rhythmic context. The 5 s windows define the S segments used for segment-to-token alignment.

By unifying spatial coordinates and temporal hierarchies, the GTRE ensures that the learned representations are not only semantically grounded via text alignment but also biologically consistent across heterogeneous clinical domains.

4.1 Stage 1: Fine-Grained Segmental Alignment and Domain-Invariant Pretraining

The inherent heterogeneity of EEG across datasets stems from varying electrode montages and hardware-specific noise profiles. To ensure the encoder learns

physiologically grounded features that generalize across such shifts, we propose a two-fold alignment strategy: (1) local segment-to-token grounding and (2) global prototype-based regularization.

Fine-Grained Segment-to-Token Alignment. Instead of global session-level alignment, we decompose the EEG representation h_m^{eeg} into S temporal segments $\{s_1, \ldots, s_S\}$ and tokenize the clinical report T_m into medical entities $E = \{e_1, \ldots, e_V\}$ using a frozen BioClinicalBERT [4] encoder. Here, $m \in \{1, \ldots, M\}$ indexes domains and n denotes sample indices within each domain. A local similarity matrix $A \in \mathbb{R}^{S \times V}$ is defined as $A_{j,k} = \cos(z_{s,j}, z_{e,k})$, measuring alignment between EEG segments and clinical entities.

To emphasize localized pathological patterns (e.g., spike-and-wave morphology), we introduce a **Segmental CLIP loss** [11]:

$$\mathcal{L}^m_{\text{seg-clip}} = -\frac{1}{N_m} \sum_{n=1}^{N_m} \log \frac{\exp\left(\max_{j,k} A^{(n)}_{j,k}/\tau\right)}{\sum_{n'=1}^{N_m} \exp\left(\max_{j,k} A^{(n')}_{j,k}/\tau\right)}. \tag{2}$$

Here τ denotes temperature. This objective encourages the **Graph-Temporal Relational Encoder** to identify neural segments that justify clinical entities while suppressing loosely correlated textual content.

Domain-Invariant Prototype Regularization. To mitigate inter-domain shifts, we maintain K global prototypes $P \in \mathbb{R}^{K \times d}$ representing shared clinical concepts [6]. Each segment embedding z_s is softly assigned via $q(z_s) = \text{softmax}(z_s P^\top/\tau)$. Let $\bar{q}^m$ denote the mean assignment within domain m, and $\bar{q}$ the global mean across domains. Cross-domain consistency is enforced by minimizing $\mathcal{L}_{\text{dom}} = \sum_{m=1}^{M} D_{\text{KL}}(\bar{q}^m \| \bar{q})$. The complete Stage 1 objective is $\mathcal{L}_{\text{stage1}} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}^m_{\text{seg-clip}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}}$, where λ_{dom} controls the strength of domain invariance. This joint formulation yields representations that are both semantically precise and robust to dataset-specific bias.

4.2 Stage 2: Fine-Grained Preference Refinement via Masked-DPO

While Stage 1 establishes global cross-modal grounding, clinical reports often contain stylistic shorthand or loosely written descriptions that introduce semantic ambiguity. To distinguish clinically precise summaries (t^+) from vague ones (t^-), we introduce a preference-based refinement stage using **Masked Direct Preference Optimization (Masked-DPO)**.

Automated Preference Distillation. To scale beyond manual validation, we employ a **Reward Model Distillation** strategy. Using a seed set of doctor-validated triplets $(X_{\text{eeg}}, t^+, t^-)$, we train a lightweight clinical reward proxy. This proxy automatically ranks multiple GPT-generated summaries for unlabeled recordings (e.g., TUAB), prioritizing summaries containing verified clinical entities (e.g., "3Hz spike-and-wave") over generic descriptions (e.g., "abnormal activity").

Masked-DPO objective. Standard preference objectives treat the entire text as a single unit, which is suboptimal due to boilerplate language in medical

Table 1. Examples of fine-grained preference triplets (t^+, t^-) used for Masked-DPO. Entities highlighted in **bold** represent masked clinical tokens M that drive the preference gradient.

Perturbation Strategy	Preferred Detailed Summary (t^+)	Non-Preferred/Vague Summary (t^-)
Entity Underspecification	EEG shows **generalized 3 Hz spike-and-wave** discharges consistent with **absence seizures**.	EEG shows *generalized abnormal activity* suggestive of possible seizure events.
Topographical Deletion	Frequent **right temporal sharp waves** observed during drowsiness, suggesting **focal epilepsy**.	*Sharp waves* observed during drowsiness; interpretation pending.
Causal De-alignment	Patient exhibited **tonic stiffening** following **aura**; consistent with **focal to bilateral tonic–clonic** seizure.	*Tonic activity* observed; unclear if related to focal or generalized onset.
Artifact Disambiguation	**Burst suppression** likely due to **Propofol**; no **epileptiform discharges** noted.	Low amplitude EEG in sedated patient; seizure activity *not clearly visualized*.

reports. We define a **Pathology-Aware Mask** $M \in \{0,1\}^V$, where $M_k = 1$ only for diagnostic entities identified via medical NER [4]. The preference loss is computed exclusively over masked tokens, ensuring that optimization focuses on pathological content rather than stylistic variations. Let z_e denote the EEG embedding from the GTRE; the similarity-scaled Masked-DPO loss is defined as:

$$\mathcal{L}_{\text{m-dpo}} = -\log \sigma\big(\beta\left[(s_\theta^+ - s_{\text{ref}}^+) - (s_\theta^- - s_{\text{ref}}^-)\right] - \lambda(1-\delta)\big) \tag{3}$$

where $s_\theta^\pm = \cos\big(z_e,\, M \odot z_{t^\pm}\big)$ are *normalized dot-product* similarities (not generative likelihoods) between the EEG embedding and masked entity embeddings under the learned encoder θ and the frozen Stage-1 reference, and $\delta = \cos(z_{t^+}, z_{t^-})$. Eq. 3 is thus a Bradley–Terry preference over similarity margins; the adaptive term $\lambda(1-\delta)$ prevents overfitting when $t^+ \approx t^-$ ($\delta \approx 1$).

Consistency-Based Prototype Regularization. To preserve the domain-invariant structure learned in Stage 1, we enforce consistency between original and refined embeddings $(z_{\text{old}}, z_{\text{new}})$ using shared prototypes $P \in \mathbb{R}^{K\times d}$. Soft assignments p_{old} and p_{new} are aligned via $\mathcal{L}_{\text{proto}} = -\sum_{k=1}^{K} p_{\text{old}}^{(k)} \log p_{\text{new}}^{(k)}$. The final training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{m-dpo}} + \lambda \mathcal{L}_{\text{proto}}$, where λ controls the retention of structured representations from Stage 1. This formulation promotes alignment with clinically richer text while preserving the generalizable EEG-text semantics learned earlier.

Implementation Details GTRE uses a shared Transformer backbone (L=8 layers, H=8 heads, FFN dimension 800, latent dimension d=256 so $Z_{\text{macro}} \in \mathbb{R}^d$, sequence length 100). Each recording yields 200 ms patches Z_{micro} forming S macro-segments aligned against V clinical entities, and spatial grounding maps 3D electrode coordinates into $\mathbb{R}^d$ via a 3-layer MLP. We train for 20 epochs (Adam, batch 1024, learning rate 5×10^{-4} decayed $\times 0.96$ after 80% of training). Stage 1 grounding uses λ_{dom}=0.1 with K=4 global prototypes; Stage 2 refinement uses λ_{reg}=0.3, K=8 semantic prototypes, and Masked-DPO with

Table 2. Comprehensive downstream classification accuracy across diverse clinical EEG datasets. Statistical significance was confirmed via paired t-tests ($p < 0.05$) with a mean performance variation of $\pm 0.4\%$.

Model	TUSZ (Seizure)	TUAB (Abnormal)	CHB-MIT (Seizure)
From Scratch	85.2	87.3	87.9
Stage 1 Only	83.0	86.2	86.5
Stage 1 + Stage 2 (Ours)	**94.3**	**88.8**	**91.5**
Few-shot Finetuned (Ours)	**96.0**	**89.1**	**94.3**
Foundation / Large-Scale Models			
LaBraM (2024) [15]	95.2	85.0	91.5
FEMBA (2025) [9]	93.0	82.5	89.7
EEGFormer (2024) [7]	93.6	83.0	91.0
BioSerenity-E1 (2025) [16]	94.0	83.5	90.3
MAEEG (2022) [10]	89.2	80.3	86.8
Specialized SOTA Baselines			
ScatterFormer (2023) [19]	95.7	87.8	92.6
Attention-TSS (2024) [12]	95.5	–	–
WaveNet+LSTM [14]	89.5	89.0	–
Deep-ConvNet [17]	90.0	86.0	88.8
TSCNN [13]	80.0	–	88.0
ChronoNet [18]	90.5	87.0	90.0

preference scale β=0.1 and adaptive margin λ=0.3. Prototypes are learnable, Xavier-initialized, and optimized jointly with the encoder. No explicit signal preprocessing is applied; raw EEG is tokenized into fixed-length sequences. Stage 1 uses subject-disjoint EEG–text pairs across datasets for domain invariance, while Stage 2 uses a labeled subset whose dispreferred summaries (t^-) are controlled LLM perturbations of preferred ones (t^+). All evaluations (downstream, transfer, few-shot at 5% labels) use held-out subjects and standard splits to prevent leakage.

5 Results

We evaluate on three clinical EEG corpora (TUSZ, TUAB, CHB-MIT), covering downstream performance, cross-domain generalization, rare-class learning, and robustness to sensor variation. Unless noted, a **single-layer MLP** head is used: **few-shot** fine-tunes the encoder and head on only 5% of labels, **SOTA** comparisons use **full fine-tuning** on all labels with subject-disjoint splits, and **zero-shot** transfer assigns classes by cosine similarity to medical-entity embeddings (no task-specific training).

Every component of GTRE contributes measurably. On **downstream classification** (Table 2), two-stage pretraining lifts TUSZ accuracy to 94.3% (+9.1 over from-scratch), on par with fully-supervised specialists (ScatterFormer 95.7%, Attention-TSS 95.5%); with **few-shot** fine-tuning on 5% labels it reaches 96.0%, surpassing those baselines, and gains over matched-capacity foundation models (LaBraM, FEMBA) thus reflect architecture rather than scale. The benefit is largest for **rare seizure types** (Table 3(a)): TCSZ and TNSZ gain +12.7 and +11.1 points and Masked-DPO adds +10.4 on SPSZ, so entity-aware

Table 3. Component-wise results: **(a)** rare-class gains (TUSZ); **(b)** ablations (TUAB, Acc/F1); **(c)** cross-domain zero-shot accuracy; **(d)** entity-level alignment; **(e)** robustness to sensor/montage variation. In (c), † denotes auxiliary MIMIC-IV notes added during training

(a) Rare-class gains (TUSZ)

Type	Scratch	S1	S1+2	Δ
GNSZ	86.2	84.8	**93.5**	+7.3
FNSZ	87.4	85.3	**94.7**	+7.3
CPSZ	85.1	83.6	**92.8**	+7.7
SPSZ	78.5	74.2	**88.9**	+10.4
TCSZ	75.0	70.9	**87.7**	+12.7
TNSZ	68.3	65.1	**79.4**	+11.1
ABSZ	79.1	76.3	**87.5**	+8.4

(b) Ablation (TUAB): Acc / F1

Full	**88.8**	**85.2**
w/o Geo. Bias	86.3	81.6
w/o Loc. Embed.	84.9	80.1
w/o Mask	83.7	79.0
w/o Proto.	82.1	78.3
w/o Margin	83.0	79.4
Raw Reports	80.5	76.8

(c) Cross-Domain Zero-Shot Accuracy

Train	Test	Scr.	S1	S1+2
TUAB	TUSZ	66.7	71.0	**75.5**
TUAB+MIMIC-IV†	TUSZ	70.2	74.5	**78.3**
TUSZ+MIMIC-IV†	TUAB	72.1	76.7	**80.4**
Unseen Hospital Transfer				
(Pretrained)	CHB-MIT	87.9	86.5	**91.5**

(d) Entity-Level Alignment

Clinical Entity	Prec.	Rec.	F1
3Hz Spike-Wave	0.92	0.89	0.90
R. Temporal Sharp	0.88	0.85	0.86
Burst Suppr.	0.94	0.91	0.92
Triphasic Waves	0.85	0.82	0.83
Avg (Ours)	**0.90**	**0.87**	**0.88**
Avg (Stage 1)	0.76	0.72	0.74

(e) Robustness (TUSZ Acc)

Scenario	Model	Full	Part.	Extr.
Ch. Dropout	Std	85.2	81.4	68.2
	GTRE	**94.3**	**93.8**	**91.2**
Rand. Montage	Std	85.2	62.4	45.1
	GTRE	**94.3**	**92.5**	**90.4**
Add. Noise	Std	85.2	78.9	71.3
	GTRE	**94.3**	**93.5**	**92.8**

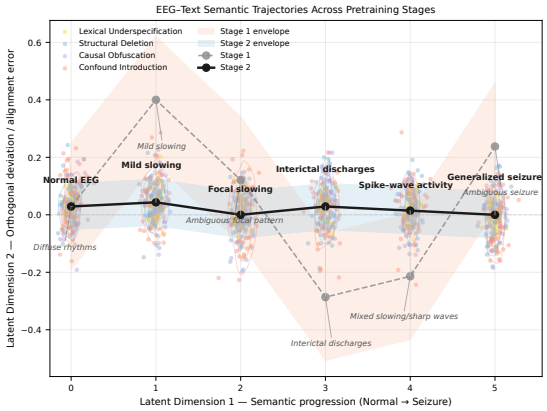

Fig. 2. Latent EEG–text trajectories.

alignment offsets class imbalance. Our **ablations** (Table 3(b)) isolate each design choice: removing the *pathology-aware mask* (−5.1 Acc), *location embedding* (−3.9 Acc), or *geometric bias* (−2.5 Acc) each degrades performance, confirming the geometric prior *outperforms* unconstrained attention rather than merely

restricting it; dropping *prototype regularization* (−6.7 Acc) or using *raw reports* (−8.4 F1) is most harmful. **Cross-domain** transfer (Table 3(c)) stays strong (75.5%→78.3% on TUAB→TUSZ with auxiliary MIMIC text; 91.5% on unseen-hospital CHB-MIT), and **robustness** (Table 3(e)) holds under extreme channel dropout (91.2% vs. 68.2%) and shuffled montages (90.4% vs. 45.1%), validating the spatial prior. **Entity-level alignment** (Table 3(d)) rises from 0.74 to 0.88 F1 with Stage 2, and **latent-trajectory analysis** (Fig. 2) decomposes EEG–text embeddings along a normalized semantic axis $\mathbf{v}_{\text{sem}} = (\mathbf{z}_{\text{seiz}} - \mathbf{z}_{\text{norm}})/\|\mathbf{z}_{\text{seiz}} - \mathbf{z}_{\text{norm}}\|$; the orthogonal residual $\mathbf{z}_{\perp} = \mathbf{z}_{\text{eeg}} - (\mathbf{z}_{\text{eeg}}^{\top}\mathbf{v}_{\text{sem}})\mathbf{v}_{\text{sem}}$ captures non-diagnostic variance from institutional bias and boilerplate. Stage 1 shows large $\mathbf{z}_{\perp}$ fluctuations (drifting toward vague t^-), whereas Stage 2 (Masked-DPO) minimizes this residual, yielding consistent paths toward informative t^+.

6 Conclusion

Our two-stage GTRE with preference-guided alignment achieves robust embeddings that generalize across tasks and datasets, filtering non-diagnostic noise, separating rare seizures, and improving downstream clinical performance. While we acknowledge potential residual biases from LLM-based supervision, we address these through future efforts in multimodal expansion to foster more robust, scalable diagnostic AI.

Disclosure of Interests. The authors have no competing interests.

References

1. Shah, V., von Weltin, E., Lopez, S., McHugh, J.R., Veloso, L., Golmohammadi, M., Obeid, I., Picone, J.: The Temple University Hospital seizure detection corpus. Frontiers in Neuroinformatics **12**, 83 (2018)
2. López, S., Suarez, G., Jungreis, D., Obeid, I., Picone, J.: Automated identification of abnormal adult EEGs. In: IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–5. IEEE (2015)
3. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.W.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**(1), 1 (2023)
4. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78 (2019)
5. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.A., Gramfort, A.: Uncovering the structure of clinical EEG signals with self-supervised learning. J. Neural Eng. **18**(4), 046020 (2021)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9650–9660 (2021)

7. Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., Qiu, L.: EEGFormer: towards transferable and interpretable large-scale EEG foundation model. arXiv:2401.10278 (2024)
8. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
9. Tegon, A., Ingolfsson, T.M., Wang, X., Benini, L., Li, Y.: FEMBA: efficient and scalable EEG analysis with a bidirectional Mamba foundation model. arXiv:2502.06438 (2025)
10. Chien, H.Y.S., Goh, H., Sandino, C.M., Cheng, J.Y.: MAEEG: masked auto-encoder for EEG representation learning. In: NeurIPS 2022 Workshop on Learning from Time Series for Health. arXiv:2211.02625 (2022)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)
12. Fan, X., Xu, P., Sun, W., Yang, S., Zhao, Q., Hao, C., He, Z., Zhao, Z., Wang, Z.: EEG-based seizure type classification with temporal-spatial-spectral attention. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 4150–4157. IEEE (2024)
13. Raghu, S., Sriraam, N., Temel, Y., Rao, S.V., Kubben, P.L.: EEG based multi-class seizure type classification using convolutional neural network and transfer learning. Neural Netw. **124**, 202–212 (2020)
14. Albaqami, H., Hassan, G.M., Datta, A.: Automatic detection of abnormal EEG signals using WaveNet and LSTM. Sensors **23**(13), 5960 (2023)
15. Jiang, W.B., Zhao, L.M., Lu, B.L.: Large brain model for learning generic representations with tremendous EEG data in BCI. In: International Conference on Learning Representations (ICLR). arXiv:2405.18765 (2024)
16. Bettinardi, R.G., Rahmouni, M., Gimenez, U.: BioSerenity-E1: a self-supervised EEG model for medical applications. arXiv:2503.10362 (2025)
17. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. **38**(11), 5391–5420 (2017)
18. Roy, S., Kiral-Kornek, I., Harrer, S.: ChronoNet: a deep recurrent neural network for abnormal EEG identification. In: Artificial Intelligence in Medicine (AIME 2019). LNCS, vol. 11526, pp. 47–56. Springer (2019)
19. Zheng, R., Li, J., Wang, Y., Luo, T., Yu, Y.: ScatterFormer: locally-invariant scattering transformer for patient-independent multispectral detection of epileptiform discharges. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37(1), pp. 148–158 (2023)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)