LATENT-SPACE PREDICTIVE LEARNING FOR EEG FOUNDATION MODELS VIA GEOMETRIC MODELING AND MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models aim to learn general-purpose neural representations that can transfer across heterogeneous downstream tasks. However, existing models often rely heavily on end-to-end fine-tuning, while their frozen representations remain substantially less effective, indicating limited representation transferability. We identify three key limitations in current pretraining paradigms: a focus on low-level waveform reconstruction rather than transferable contextual representation learning, insufficient integration of electrode geometry into inter-channel modeling, and local information leakage under random masking. To address these limitations, we propose **EEG-GPRL**, a **G**eometry-aware **P**redictive **R**epresentation **L**earning framework that learns transferable contextual EEG representations by predicting teacher representations in latent space. EEG-GPRL incorporates Geometry-aware Attention (GeoAttn) to couple channel interactions with relative electrode geometry. We propose Spatiotemporal Gaussian Field Masking (ST-GFM) to construct correlated masked contexts to reduce local information leakage. VISReg is used to prevent representational collapse. Across nine datasets spanning six EEG decoding paradigms, EEG-GPRL achieves the highest balanced accuracy on eight datasets under both KNN and frozen probing, demonstrating strong transferability across heterogeneous tasks. Ablation studies further validate the complementary contributions of its individual components.

1 INTRODUCTION

Foundation models aim to learn general-purpose representations from large-scale data that can be reused across heterogeneous downstream tasks (Liu et al., 2026). This objective is particularly relevant to EEG, whose applications span diverse domains, including brain-computer interfaces (Chaudhary et al., 2016; Abiri et al., 2019), mental workload estimation (Ke et al., 2014), emotion recognition (Koelstra et al., 2011), sleep staging (Supratak et al., 2017), and disorder diagnosis (Acharya et al., 2018). Accordingly, EEG foundation models are expected not only to support downstream finetuning, but more importantly to learn transferable representations that remain broadly reusable across tasks (Liu et al., 2026). However, recent studies show that existing EEG foundation models often achieve competitive performance primarily after fine-tuning, while their frozen representations perform substantially worse (Yang et al., 2026; Liu et al., 2026). This discrepancy raises the concern that current EEG pretraining paradigms may function primarily as initialization for task-specific optimization, rather than learning representations with strong intrinsic transferability.

We identify three fundamental challenges in current EEG pretraining paradigms that may limit representation transferability. **(1) Pretraining paradigm**. The widely used masked signal reconstruction objective emphasizes recovery in waveform space (El Ouahidi et al., 2025; Wang et al., 2025; Zhou et al., 2025), which may bias learned representations toward signal-specific structures rather than more transferable contextual information. **(2) Model architecture**. EEG electrodes are arranged over the scalp according to a specific spatial geometry, and signals recorded at different electrodes exhibit spatial dependencies related to their relative locations. Existing models encode channel identity or electrode position through embeddings and geometric input transformations (Yang et al., 2023; Jiang et al., 2024b; Wang et al., 2024; Xiong et al., 2025; Yi et al., 2023; Hong et al., 2025), but relative electrode geometry is typically not explicitly incorporated into channel-wise attention.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(3) Masking strategy. Several existing methods adopt random masking (Wang et al., 2025; Döner et al., 2026; Ma et al., 2026; Chen et al., 2026). Random masking is susceptible to local information leakage because spatially or temporally adjacent observations may remain visible around masked tokens, providing redundant local cues and reducing the need to integrate broader spatiotemporal context.

To address these challenges, we propose **EEG-GPRL**, a **G**eometry-aware **P**redictive **R**epresentation **L**earning framework for EEG foundation model pretraining. EEG-GPRL replaces waveform reconstruction with prediction of contextual teacher features, directing pretraining toward dependencies in an evolving latent space, while VISReg regularizes this space to prevent representational collapse. To model EEG spatiotemporal structure, we design a geometry-aware model that factorizes temporal and channel attention. We propose Geometry-aware Attention (GeoAttn), whose coordinate-aware rotary interactions introduce relative electrode topology into channel-wise attention alongside signal content. We further propose a Spatiotemporal Gaussian Field Masking (ST-GFM) that generates correlated masks according to electrode locations and temporal proximity, reducing local information leakage from neighboring electrodes and adjacent temporal windows.

We evaluate EEG-GPRL on nine datasets spanning six EEG decoding paradigms using KNN probing and frozen probing. EEG-GPRL consistently achieves strong transfer performance across heterogeneous tasks, while ablation studies further support our analysis and the effectiveness of the proposed components.

Our main contributions are summarized as follows:

- We identify three key factors that may limit transferable EEG representation learning: the masked signal reconstruction pretraining objective, insufficient modeling of electrode topology, and local information leakage under random masking.
- We propose **EEG-GPRL**, a latent-space predictive pretraining framework that learns to infer contextual EEG representations rather than reconstruct signal waveforms, promoting transferable representation learning.
- We propose Geometry-aware Attention (GeoAttn) to incorporate relative electrode topology into channel-wise attention, and Spatiotemporal Gaussian Field Masking (ST-GFM) to reduce local information leakage and encourage learning from broader spatiotemporal context.
- We evaluate EEG-GPRL on nine datasets covering six EEG decoding paradigms, where KNN probing and frozen probing consistently demonstrate improved representation transferability across heterogeneous downstream tasks.

2 Related Works

Pretraining Objectives of EEG Foundation Models. Masked signal reconstruction defines prediction targets in waveform space (Wang et al., 2025; Zhou et al., 2025; Döner et al., 2026; El Ouahidi et al., 2025), which may emphasize signal-specific characteristics rather than more transferable contextual representations. Spectral and statistical properties have therefore been incorporated as auxiliary targets (Hong et al., 2025; Chen et al., 2026; Wang et al., 2026). LaBraM and CodeBrain adopt discrete token prediction (Jiang et al., 2024b; Ma et al., 2026), where the prediction targets are determined by learned tokenizers. EEGPT further introduces representation alignment while retaining signal reconstruction (Wang et al., 2024). In contrast, EEG-GPRL performs contextual prediction directly in an evolving latent space, avoiding predefined reconstruction or tokenizer-based targets and shifting the learning objective toward contextual representation prediction. EEG-JEPA also brings latent prediction and structured electrode–time masking to EEG (Li et al., 2026), while EEG-GPRL explicitly incorporates relative electrode geometry into channel attention and samples spatiotemporal masks from a Gaussian field.

Geometric Modeling in EEG Foundation Models. Existing EEG foundation models capture channel structure through channel embeddings, structured attention, or spatial organization (Jiang et al., 2024b; Wang et al., 2024; 2025; Zhou et al., 2025). Other methods further incorporate electrode coordinates or montage geometry through positional features or input alignment (El Ouahidi et al., 2025; Döner et al., 2026; Hong et al., 2025). These designs incorporate geometry through token features or input alignment, but leave relative electrode displacement implicit into inter-channel

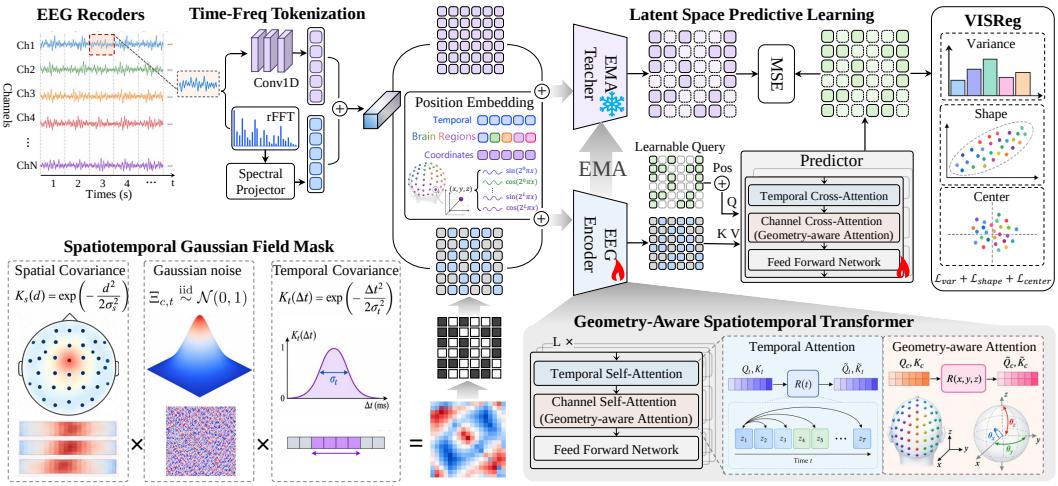


Figure 1: Overview of EEG-GPRL. EEG recordings are converted into time-frequency tokens and encoded by a geometry-aware spatiotemporal Transformer. ST-GFM generates spatially and temporally correlated masks, while an EMA teacher provides latent targets for masked prediction. GeoAttn incorporates relative electrode geometry into channel attention, and VISReg regularizes the latent space to prevent representational collapse.

attention. Our GeoAttn introduces coordinate-aware rotary interactions, allowing relative electrode topology to explicitly modulate channel-wise attention.

Masking Strategies. Random masking is widely used in EEG pretraining (Wang et al., 2025; Döner et al., 2026), but can leave correlated neighbors visible, making masked content partly accessible from local context. More structured schemes have therefore been explored. MMM combines random and region-wise masking (Yi et al., 2023), REVE extends block masking jointly across spatial and temporal dimensions (El Ouahidi et al., 2025), and SAMBA introduces temporally informed masking (Hong et al., 2025). These structured schemes define mask regions through explicit grouping rules, whereas ST-GFM samples a Gaussian field whose covariance depends on electrode distance and temporal separation, providing tunable spatiotemporal correlations that encourage neighboring patches to be masked together and reduce local information leakage.

3 Methods

As illustrated in Figure 1, EEG-GPRL learns EEG representations through masked prediction in latent space. Each EEG recording is divided into non-overlapping 1-s windows, and a shared Time-Frequency Tokenizer maps them to EEG tokens, to which temporal and spatial positional encodings are applied. In the online branch, the Spatiotemporal Gaussian Field Mask (ST-GFM) masks spatiotemporally coupled EEG context tokens, while an EEG encoder processes the remaining visible contexts through temporal attention and channel attention with Geometry-aware Attention (GeoAttn). Meanwhile, an exponential moving average (EMA) teacher, updated from the online branch, encodes the unmasked EEG tokens into contextual targets. The predictor uses learnable queries to infer the representations of masked contexts from the visible EEG contexts. VISReg regularizes the learned latent space to prevent representational collapse. After pretraining, the tokenizer and online encoder are retained for downstream tasks.

3.1 Model Architecture

Let $X \in \mathbb{R}^{C\times T\times S}$ denote an EEG recording segmented into T non-overlapping 1-s windows, where C is the number of electrodes and S is the number of samples per window. We use $x_{c,t} \in \mathbb{R}^S$ to denote the signal segment from electrode c in window t.

Time-Frequency Tokenization. The Time-Frequency Tokenizer maps each channel-wise signal segment $x_{c,t}$ to a token through two shared branches. A 3-layer convolutional network (CNN1D) extracts local temporal features, while a linear layer projects the magnitude spectrum obtained by the FFT for real-valued inputs:

$$e_{c,t} = \text{CNN1D}(x_{c,t}) + W_{\text{freq}} \, |\text{rFFT}(x_{c,t})| \in \mathbb{R}^D, \tag{1}$$

where $W_{\text{freq}} \in \mathbb{R}^{D \times (\lfloor S/2 \rfloor + 1)}$ is the spectral projection. The two branches provide temporal and spectral descriptions of the EEG segment, and their outputs are added to form the token $e_{c,t}$. Their parameters are shared across all electrodes and windows, producing EEG tokens $E \in \mathbb{R}^{C \times T \times D}$.

Spatiotemporal Positional Encoding. Each EEG token is assigned an absolute positional encoding that combines electrode geometry, scalp-region identity, and temporal order. Let $p_c \in \mathbb{R}^3$ and $r_c \in \{1, \ldots, 5\}$ denote the coordinate and scalp-region label of electrode c, respectively. A learnable region table $E_{\text{reg}} \in \mathbb{R}^{5 \times D}$ assigns a shared embedding to electrodes within the same anatomical region, including frontal, central, temporal, parietal, and occipital. Using $K = D/6$ frequency bands, we construct the coordinate encoding $\gamma(p_c) = \left[\left(\sin(2^k \pi p_c^{(a)}), \cos(2^k \pi p_c^{(a)}) \right)_{a \in \{x,y,z\}} \right]_{k=0}^{K-1} \in \mathbb{R}^D$, where the brackets denote concatenation over coordinate axes and frequency bands. The absolute channel encoding is then given by $e_c^{\text{pos}} = \text{MLP}(\gamma(p_c)) + E_{\text{reg}}[r_c]$. Temporal order is represented by a fixed sinusoidal encoding $e_t^{\text{pos}} \in \mathbb{R}^D$. For $m = 0, \ldots, D/2 - 1$ and temporal scale $T_{\max} = 512$, its components are $e_{t,2m}^{\text{pos}} = \sin\left(t T_{\max}^{-2m/D}\right)$ and $e_{t,2m+1}^{\text{pos}} = \cos\left(t T_{\max}^{-2m/D}\right)$. The input to the EEG encoder is then

$$h_{c,t}^0 = e_{c,t} + e_c^{\text{pos}} + e_t^{\text{pos}}. \tag{2}$$

Here, e_c^{pos} encodes the electrode position, while e_t^{pos} encodes its temporal position.

Geometry-Aware Spatiotemporal Transformer. Starting from $h_{c,t}^0$, the EEG encoder stacks L factorized Transformer blocks, each comprising temporal attention, geometry-aware channel attention, and a feed-forward network with pre-layer normalization and residual connections.

Temporal attention follows vanilla self-attention with rotary position embeddings (Su et al., 2021) to incorporate relative temporal offsets into query-key interactions. Let $d_h = D/H$ be the dimension of each of H attention heads, and let $R(\alpha)$ denote a two-dimensional rotation. The temporal rotation is defined as

$$R(\alpha) = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}, \qquad R_{\text{t}}(t) = \text{blkdiag}_{m=1}^{d_h/2} \, R(t\omega_m), \tag{3}$$

where ω_m is the rotation frequency for feature pair m. Queries and keys are rotated as $\widetilde{q}_{c,t}^{\text{time}} = R_{\text{t}}(t) q_{c,t}^{\text{time}}$ and $\widetilde{k}_{c,t}^{\text{time}} = R_{\text{t}}(t) k_{c,t}^{\text{time}}$. Their scaled dot products determine the attention weights used to aggregate value features across temporal windows within each electrode.

To introduce electrode topology into channel attention we propose geometry-aware self-attention (GeoAttn) through coordinate-aware rotary interactions. GeoAttn rotates queries and keys according to their three-dimensional electrode coordinates to incorporate relative electrode topology. Each group of six feature dimensions contains three pairs, rotated according to the x, y, and z coordinates:

$$R_{\text{ch}}(p_c) = \text{blkdiag}_{m=1}^{d_h/6} \left[R(\omega_m p_c^{(x)}), R(\omega_m p_c^{(y)}), R(\omega_m p_c^{(z)}) \right]. \tag{4}$$

Rotating the query and key at their respective electrode coordinates gives:

$$A_{t;i,j}^{\text{ch}} = \frac{(q_{i,t}^{\text{ch}})^\top R_{\text{ch}}(p_i)^\top R_{\text{ch}}(p_j) k_{j,t}^{\text{ch}}}{\sqrt{d_h}} = \frac{(q_{i,t}^{\text{ch}})^\top R_{\text{ch}}(p_j - p_i) k_{j,t}^{\text{ch}}}{\sqrt{d_h}}, \tag{5}$$

The coordinate-aware rotary interaction $R_{\text{ch}}(p_j - p_i)$ enables GeoAttn to incorporate relative electrode location into query-key compatibility. Applied to channel attention, it combines signal content with a geometric prior on inter-electrode relationships.

A feed-forward network (FFN) then applies two linear transformations with an intervening nonlinear activation to each EEG token independently, mixing information across feature dimensions. Temporal attention, geometry-aware channel attention, and the FFN together form a block, with L stacked blocks producing contextual EEG representations for latent-space prediction.

3.2 SPATIOTEMPORAL GAUSSIAN FIELD MASKING (ST-GFM)

EEG electrodes are geometrically arranged over the scalp, and volume conduction introduces spatial dependencies among their recordings. Random masking commonly leaves temporally or spatially adjacent observations visible, exposing redundant information about the masked content through neighboring electrodes or temporal windows. This may allow prediction with less reliance on broader context. We therefore propose Spatiotemporal Gaussian Field Masking (ST-GFM) to generate spatiotemporally correlated masks according to electrode locations and temporal proximity, encouraging prediction from broader context.

Given electrode coordinates $\{p_c\}_{c=1}^{C}$ and T temporal windows, we construct spatial and temporal covariance matrices $K_s \in \mathbb{R}^{C\times C}$ and $K_t \in \mathbb{R}^{T\times T}$ using Gaussian kernels:

$$(K_s)_{c,c'} = \exp\left(-\frac{\|p_c - p_{c'}\|_2^2}{2\sigma_s^2}\right), \qquad (K_t)_{t,u} = \exp\left(-\frac{(t-u)^2}{2\sigma_t^2}\right). \tag{6}$$

Here, σ_s controls the spatial correlation scale in electrode-coordinate units, and σ_t controls the temporal correlation scale in window units. Larger scales encourage masking scores to remain similar across broader electrode neighborhoods or longer temporal intervals.

Following standard Gaussian sampling via Cholesky decomposition (Rasmussen & Williams, 2006), we factorize $K_s = L_s L_s^\top$ and $K_t = L_t L_t^\top$ and sample a spatiotemporal score field $G \in \mathbb{R}^{C\times T}$:

$$\Xi_{c,t} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad G = L_s \Xi L_t^\top. \tag{7}$$

Left multiplication introduces correlations across electrodes, while right multiplication introduces correlations across temporal windows. This gives $\mathrm{Cov}(G_{c,t}, G_{c',u}) = (K_s)_{c,c'}(K_t)_{t,u}$, coupling spatial and temporal proximity within the same field without forming a full $CT \times CT$ covariance matrix. Random noise is sampled for each recording to vary the masked contexts during training.

For a masking ratio ρ, we rank the valid tokens Ω by their field scores and mask the $\mathrm{round}(\rho|\Omega|)$ highest-scoring tokens. These tokens form the masked set $\mathcal{M}$, while the remaining tokens form the visible set $\mathcal{V}$. Spatial and temporal correlations in the field encourage neighboring electrodes and adjacent windows to be masked together, reducing information leakage within local spatiotemporal neighborhoods and encouraging the model to learn neural contextual representations from more distant observations.

3.3 LATENT-SPACE PREDICTIVE PRETRAINING

Given the visible and masked token sets produced by ST-GFM, we learn EEG representations by predicting masked teacher features from visible contexts, following the joint-embedding predictive principle (Assran et al., 2023).

Exponential Moving Average Teacher. Let F_θ comprise the tokenizer, spatiotemporal positional encoding, and EEG encoder. The online branch encodes EEG contexts with an mask $\mathcal{M}$, while a teacher $F_{\bar\theta}$ with the same architecture encodes the unmasked recording:

$$Z = F_\theta(X, \mathcal{M}), \qquad Y = \mathrm{sg}[F_{\bar\theta}(X, \varnothing)], \tag{8}$$

where $Z, Y \in \mathbb{R}^{C\times T\times D}$, the second argument specifies the masked locations, and sg stops gradients. The teacher is initialized from the online branch and updated after each optimization step as $\bar\theta \leftarrow \mu\bar\theta + (1-\mu)\theta$, where $\mu \in [0,1)$ is the EMA momentum. Its representations at masked locations provide contextual targets that evolve gradually during training.

Spatiotemporal Cross-Attention Predictor. For each masked location $(c,t) \in \mathcal{M}$, we construct a query $u_{c,t} = e_{\text{pred}} + e_c^{\text{pos}} + e_t^{\text{pos}}$, combining a learnable embedding with spatial and temporal positional encodings. The predictor P_ϕ uses these queries to aggregate $Z_\mathcal{V}$, the representations at valid visible positions, through cross-attention across electrodes and temporal windows. Feed-forward transformations and an output projection produce the predicted representations. We minimize their discrepancy from the teacher targets only at masked locations:

$$\widehat{Y}_\mathcal{M} = P_\phi(U, Z, \mathcal{M}), \qquad \mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{M}|D} \sum_{(c,t)\in\mathcal{M}} \|\widehat{y}_{c,t} - y_{c,t}\|_2^2, \tag{9}$$

where U collects the queries at masked locations. This objective encourages the model to capture latent neural features by predicting the teacher's contextual representations from visible EEG contexts.

Variance-Invariance-Sketching Regularization (VISReg). To prevent representational collapse, we adopt VISReg (Wu et al., 2026) to the latent EEG contexts produced by the online encoder across a minibatch. The scale term $\mathcal{L}_{\text{scale}}$ encourages unit feature standard deviations, and the center term $\mathcal{L}_{\text{center}}$ encourages zero feature means. The sketching term $\mathcal{L}_{\text{shape}}$ matches the sorted random projections of standardized features to standard Gaussian quantiles, constraining their distributional shape. The detailed definitions are given in Appendix B. The total pretraining objective combines these three terms of VISReg with the latent prediction loss:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{center}}, \tag{10}$$

The online branch and predictor are optimized by backpropagation, while the teacher is updated by EMA. Together, latent prediction and distributional regularization encourage contextual representations while maintaining feature variation.

4 Experimental Evaluation

4.1 Experimental Setup

Pretraining Datasets. We pretrain on heterogeneous EEG recordings from TUEG Obeid & Picone (2016), PhysioNet2018 Ghassemi et al. (2018), Raw EEG Data Trujillo (2020), Resting State EEG Data Trujillo et al. (2017), Siena Scalp EEG Database Detti et al. (2020), and five OpenNeuro datasets Ataseven (2025a;b); Carrillosulub (2025); Kinley & Becker (2025); Ghaffari et al. (2025). The total collection contains 42,755.68 hours of recordings, with 18,546.35 valid hours retained after preprocessing. Dataset statistics are provided in Tab. 4.

Downstream Datasets. We evaluate on nine datasets spanning six EEG tasks: emotion recognition (SEED-V Liu et al. (2021), SEED-VII Jiang et al. (2024a)), motor imagery (PhysioNet-MI Schalk et al. (2004), BCICIV-2a Brunner et al. (2008)), sleep staging (ISRUC Khalighi et al. (2016), HMC Alvarez-Estevez & Rijsman (2021)), seizure detection (CHB-MIT Shoeb (2009)), Major Depressive Disorder (MDD) detection (Mumtaz2016 Mumtaz (2016)), and workload classification (MentalArithmetic Zyma et al. (2019)). These binary and multi-class classification datasets span different electrode configurations, recording durations, label spaces, and subject splits. Details are reported in Tab. 5.

Preprocessing. EEG recordings are band-pass filtered at 0.3–75 Hz, notch filtered at 50/60 Hz according to acquisition site, resampled to 200 Hz, and scaled by 100 μV to approximately $[-1, 1]$.

Implementation Details. EEG-GPRL is implemented in PyTorch 2.5.1 with CUDA 12.4. We pretrain for 10 epochs on eight NVIDIA A6000 GPUs using BF16 mixed precision, with a per-GPU batch size of 48 and a global batch size of 384. Each epoch takes approximately 90 minutes, giving a total pretraining time of about 15 hours. We use AdamW with weight decay 10^{-2} and a peak learning rate of 5×10^{-4}. A one-epoch linear warm-up is followed by cosine decay to 10^{-6}.

The encoder contains 12 geometry-aware spatiotemporal attention blocks with hidden dimension 240 and eight attention heads. The predictor with two-layer cross attention learns to infer latent representations at masked locations. ST-GFM masks 50% of tokens using spatial and temporal correlation scales of 0.09 and 1.0, selecting based on the validation sets. Additional pretraining settings are provided in Table 6.

Evaluation Protocols. We evaluate pretrained representations using KNN probing and frozen probing, with the backbone fixed in both settings. This directly assesses representation transferability by isolating the frozen backbone from end-to-end task-specific adaptation. EEG representations are flattened before being passed to either KNN or the task-specific head. For KNN probing, we evaluate $K \in 3, 5, 7, 9, 11$ and report the mean and standard deviation across these five settings, predicting test labels from neighboring labeled training examples. As a parameter-free readout, KNN

Table 1: KNN-probing performance comparison with existing models (%). The **best** and second-best available results are highlighted.

Methods	BCICIV-2a, 4-class			SEED-V, 5-class			MentalArithmetic, 2-class		
	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC
LaBraM	28.82 ±0.93	5.09 ±1.24	28.58 ±0.88	20.81 ±0.09	1.16 ±0.15	21.06 ±0.23	52.50 ±1.14	28.51 ±1.18	53.49 ±1.01
CBraMod	32.57 ±1.26	10.09 ±1.68	32.51 ±1.28	20.75 ±0.13	1.11 ±0.19	20.45 ±0.52	51.18 ±2.87	26.54 ±1.74	52.38 ±1.76
CSBrain	28.80 ±0.74	5.07 ±0.99	28.63 ±0.81	20.53 ±0.04	0.74 ±0.07	20.36 ±0.47	49.44 ±2.18	28.58 ±1.99	54.56 ±2.23
REVE	26.27 ±1.05	1.69 ±1.40	24.83 ±1.30	20.00 ±0.05	0.07 ±0.08	19.08 ±0.52	53.12 ±3.44	27.75 ±1.82	52.15 ±2.34
CodeBrain	30.50 ±1.68	7.34 ±2.24	29.92 ±1.67	20.88 ±0.16	1.29 ±0.21	20.82 ±0.38	53.82 ±1.49	31.89 ±2.28	58.92 ±2.37
Ours	**35.07** ±1.68	**13.43** ±2.24	**34.60** ±1.68	**29.42** ±0.75	**11.77** ±0.99	**29.85** ±0.76	**60.07** ±2.02	**49.39** ±2.77	**64.12** ±2.05

Methods	PhysioNet-MI, 4-class			Mumtaz2016, 2-class			SEED-VII, 7-class		
	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC	Bal. Acc.	Kappa	Weighted F1
LaBraM	32.84 ±2.00	10.46 ±2.67	32.79 ±1.99	71.01 ±9.54	72.84 ±4.51	79.21 ±5.39	15.62 ±0.10	1.65 ±0.13	15.37 ±0.18
CBraMod	28.25 ±0.85	4.33 ±1.13	27.89 ±0.80	69.57 ±4.28	72.64 ±2.52	76.37 ±3.88	16.35 ±0.17	2.54 ±0.20	16.18 ±0.22
CSBrain	29.66 ±0.53	6.21 ±0.71	29.60 ±0.51	68.26 ±2.55	71.18 ±1.98	74.94 ±2.93	15.74 ±0.32	1.77 ±0.38	15.54 ±0.31
REVE	**33.49** ±0.65	**11.99** ±0.87	**33.38** ±0.65	58.38 ±2.78	62.97 ±0.69	64.65 ±1.24	15.13 ±0.08	1.04 ±0.10	14.17 ±0.27
CodeBrain	30.99 ±1.05	7.98 ±1.40	31.01 ±1.03	66.39 ±2.83	67.72 ±1.28	71.50 ±1.35	16.64 ±0.14	2.87 ±0.17	16.40 ±0.12
Ours	30.67 ±0.74	6.63 ±0.99	30.80 ±0.72	**84.03** ±0.50	**89.43** ±0.43	**90.43** ±1.09	**25.16** ±0.25	**12.95** ±0.33	**25.53** ±0.28

Methods	ISRUC, 5-class			CHB-MIT, 2-class			HMC, 5-class		
	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC	Bal. Acc.	Kappa	Weighted F1
LaBraM	22.12 ±0.38	2.00 ±0.55	15.15 ±2.49	49.64 ±0.47	1.41 ±0.19	48.45 ±1.56	22.09 ±2.39	2.68 ±2.49	21.63 ±6.17
CBraMod	26.80 ±0.65	8.32 ±0.62	24.91 ±0.27	62.66 ±5.21	12.11 ±7.38	75.88 ±5.10	32.61 ±1.28	14.98 ±1.71	34.29 ±1.50
CSBrain	19.27 ±0.18	0.01 ±0.32	22.91 ±0.14	61.87 ±2.74	6.75 ±2.06	72.74 ±4.31	26.40 ±0.30	7.09 ±0.35	27.39 ±0.62
REVE	35.28 ±1.31	20.65 ±1.74	37.26 ±0.98	50.00 ±0.00	4.79 ±0.93	54.06 ±1.43	22.65 ±1.04	4.22 ±0.30	23.55 ±5.47
CodeBrain	26.41 ±0.80	9.87 ±0.98	30.98 ±0.94	62.36 ±1.65	12.27 ±2.26	64.56 ±2.64	32.18 ±0.72	14.60 ±1.06	34.13 ±1.33
Ours	**56.08** ±1.38	**46.42** ±2.58	**59.55** ±1.87	**73.28** ±1.28	**31.96** ±2.23	**83.29** ±2.40	**45.69** ±0.88	**32.73** ±2.57	**49.27** ±2.60

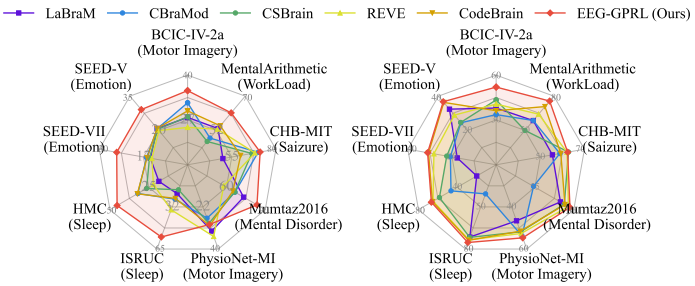
probing provides a stringent assessment of the learned representation space, as strong performance requires task-relevant structure to be directly reflected in feature similarity rather than recovered by a learned downstream decision function. For frozen probing, we train an MLP head on the fixed representations, as in prior works (Wang et al., 2025; Zhou et al., 2025; Ma et al., 2026). In contrast, frozen probing allows the MLP to transform the fixed representations using downstream supervision, evaluating whether task-discriminative information can be extracted by a lightweight supervised readout. Together, the two protocols characterize the transferability of pretrained representations without adapting the backbone.

Baselines. We compare EEG-GPRL with EEG foundation models LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), REVE (El Ouahidi et al., 2025), and CodeBrain (Ma et al., 2026). We reproduce these baselines using the pretrained model publicly released and evaluate all models on the same training, validation, and test splits.

Evaluation Metrics. We report balanced accuracy (Bal. Acc.), AUC-PR, and AUROC for binary classification. For multi-class classification, we report balanced accuracy, Cohen's kappa, and weighted F1. For frozen probing, we report the mean and standard deviation over five downstream training runs with different random seeds.

4.2 Results

Figure 2 visualizes balanced accuracy under the two probing protocols, while Tables 1 and 2 report the full metrics. Since the pretrained backbone remains fixed in both settings, these evaluations largely isolate the quality and transferability of the pretrained representations from improvements introduced by end-to-end task-specific adaptation.



(a) KNN Probing Performance (b) Frozen Probing Performance

Figure 2: Radar plot of balanced accuracy across downstream datasets under (a) KNN probing and (b) frozen probing.

Table 2: Frozen-probing performance comparison with existing models (mean ± std, %). The **best** and second-best available results are highlighted.

	BCICIV-2a, 4-class			SEED-V, 5-class			MentalArithmetic, 2-class		
Methods	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC
LaBraM	37.96 ±1.15	17.29 ±1.49	33.54 ±1.62	32.40 ±0.55	15.42 ±0.72	31.98 ±0.88	58.62 ±9.31	67.55 ±2.12	79.23 ±2.06
CBraMod	33.59 ±0.55	11.46 ±0.74	22.13 ±0.32	24.47 ±0.37	5.34 ±0.49	16.77 ±1.15	58.68 ±2.46	42.94 ±2.63	63.18 ±1.87
CSBrain	43.40 ±0.06	24.54 ±0.08	42.41 ±0.15	24.64 ±0.47	5.99 ±0.52	24.57 ±0.96	50.00 ±2.21	45.37 ±2.68	65.78 ±0.36
REVE	40.62 ±1.23	20.83 ±1.54	34.92 ±1.37	28.95 ±0.79	10.91 ±0.95	29.30 ±1.28	64.24 ±7.63	61.61 ±2.24	78.72 ±1.66
CodeBrain	36.20 ±1.04	14.93 ±1.39	27.49 ±1.38	36.32 ±0.31	20.58 ±0.54	36.34 ±0.76	70.83 ±1.96	60.67 ±2.16	75.69 ±0.61
Ours	**52.00** ±1.27	**36.00** ±1.69	**50.49** ±1.78	**36.73** ±0.35	**20.65** ±0.58	**36.43** ±0.63	**75.69** ±7.81	**78.40** ±7.01	**88.74** ±3.26
	PhysioNet-MI, 4-class			Mumtaz2016, 2-class			SEED-VII, 7-class		
Methods	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC	Bal. Acc.	Kappa	Weighted F1
LaBraM	46.60 ±0.61	28.74 ±0.83	46.81 ±0.50	84.54 ±1.22	90.91 ±1.10	92.82 ±0.69	17.63 ±2.00	4.15 ±2.44	16.24 ±6.36
CBraMod	53.75 ±0.71	38.32 ±0.94	52.87 ±0.53	53.34 ±1.06	97.01 ±0.46	97.37 ±0.08	20.72 ±0.21	7.82 ±0.30	20.19 ±1.39
CSBrain	51.85 ±0.08	35.81 ±0.10	52.00 ±0.06	89.04 ±0.22	94.64 ±0.17	95.70 ±0.11	22.38 ±0.11	9.67 ±0.15	22.13 ±0.93
REVE	52.63 ±0.08	36.75 ±0.11	51.96 ±0.07	93.81 ±1.59	98.73 ±0.06	98.45 ±0.06	28.19 ±0.37	16.74 ±0.44	28.31 ±1.15
CodeBrain	51.52 ±1.25	35.36 ±1.67	51.90 ±0.88	89.78 ±0.39	97.77 ±0.25	97.71 ±0.27	30.63 ±0.73	19.35 ±0.74	31.11 ±0.81
Ours	**54.69** ±0.75	**39.58** ±1.00	**54.57** ±0.71	**95.55** ±2.66	**99.28** ±0.07	**99.32** ±0.08	**31.05** ±0.64	**19.71** ±0.64	**31.19** ±0.73
	ISRUC, 5-class			CHB-MIT, 2-class			HMC, 5-class		
Methods	Bal. Acc.	Kappa	Weighted F1	Bal. Acc.	AUC-PR	AUROC	Bal. Acc.	Kappa	Weighted F1
LaBraM	71.46 ±1.05	67.38 ±1.18	74.21 ±1.04	55.49 ±4.11	29.05 ±4.42	89.61 ±3.00	20.00 ±0.06	0.00 ±0.12	20.03 ±0.14
CBraMod	40.73 ±2.21	24.87 ±2.33	29.86 ±1.99	58.87 ±3.75	36.52 ±9.28	85.89 ±4.74	46.52 ±1.75	26.62 ±3.41	31.37 ±5.62
CSBrain	71.70 ±1.18	66.28 ±2.16	72.83 ±1.97	**62.78** ±3.76	10.06 ±2.40	76.44 ±3.88	58.41 ±1.02	50.28 ±0.65	62.55 ±0.61
REVE	73.46 ±1.57	67.78 ±0.98	74.59 ±1.02	62.11 ±3.65	**44.75** ±4.30	89.79 ±1.26	65.02 ±0.52	57.53 ±1.41	65.61 ±1.51
CodeBrain	73.76 ±0.43	65.35 ±0.79	72.45 ±0.52	59.29 ±5.08	37.06 ±4.79	88.43 ±0.67	65.65 ±0.49	56.41 ±1.03	64.90 ±0.88
Ours	**75.35** ±0.15	**68.99** ±0.70	**75.27** ±0.49	62.40 ±2.47	41.49 ±6.95	**90.54** ±0.81	**66.75** ±0.77	**58.81** ±0.21	**66.37** ±0.12

KNN Probing. EEG-GPRL achieves strong performance across most downstream tasks. This result is particularly informative under KNN probing, since prediction depends directly on feature similarity rather than a learned task-specific classifier. The advantage therefore suggests that task-relevant structure is already well organized in the pretrained representation space. Its consistency across heterogeneous tasks further indicates that it is broadly transferable rather than task-specific.

Frozen Probing. The advantage remains when a lightweight MLP is trained on the frozen representations, showing that discriminative information can be effectively extracted without adapting the backbone. Together, the agreement between KNN and frozen probing therefore reflects both a well-structured representation space and strong downstream accessibility. The different behavior on PhysioNet-MI further suggests that useful task information may remain encoded even when it is not fully captured by local feature similarity.

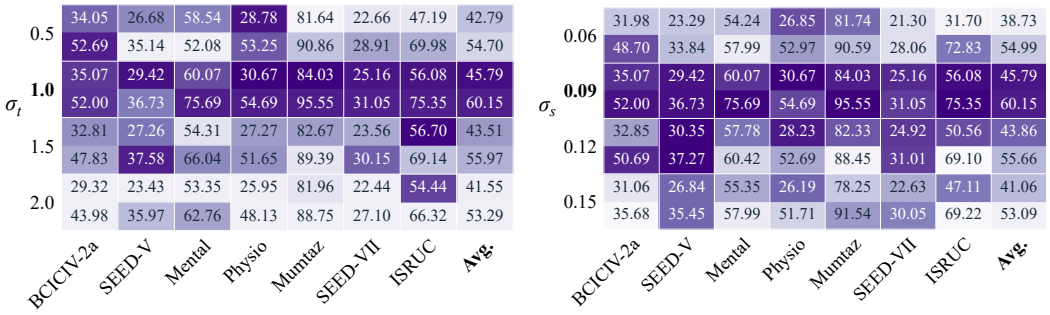
4.3 Ablation Study

Table 3 demonstrates the complementary roles of the proposed components. Replacing latent-space prediction with the raw-signal reconstruction variant yields the largest drop in transfer performance among the tested ablations. This result supports the contribution of latent predictive targets within EEG-GPRL. Building on this latent predictive objective, GeoAttn further improves transfer by explicitly incorporating electrode geometry into channel attention, even when absolute electrode positions are already encoded in the EEG token grid. VISReg is also critical to this framework: removing it causes severe degradation, with several datasets approaching near-random performance, highlighting its role in preventing representational collapse through variance and shape regularization of the latent space.

The masking ablations further suggest that transferability is associated with how effectively local information leakage is mitigated during pretraining. Random masking is particularly susceptible to this issue, as spatially or temporally adjacent observations may remain visible around masked tokens. More structured strategies partially alleviate this problem, but channel-wise and temporal masking constrain only one dimension, while block masking does not explicitly account for actual electrode geometry. ST-GFM instead couples masking with both electrode distance and temporal proximity, producing spatially and temporally coherent missing regions. This design reduces the availability of nearby observations and encourages the model to rely more on broader spatiotemporal context, consistent with its more stable transfer performance across probing protocols.

Table 3: Ablation studies under KNN Probing and Frozen Probing, measured by balanced accuracy (%, ↑). Each variant removes one component. √ and × indicate presence and removal. Avg. is the unweighted mean over all five datasets within each protocol. The best available value in each result column is bold.

Components				KNN Probing						Frozen Probing					
Lat.-Pred.	Geo-Attn	VIS-Reg	ST-GFM	BCIC-IV2a	SEED-V	Ment. Arith.	Mumtaz 2016	ISRUC	**AVG.**	BCIC-IV2a	SEED-V	Ment. Arith.	Mumtaz 2016	ISRUC	**AVG.**
√	√	√	√	**35.1** ±1.68	**29.4** ±0.75	**60.1** ±2.02	**84.0** ±0.50	**56.1** ±1.38	**52.9**	**52.00** ±1.27	**36.73** ±0.35	**75.69** ±7.81	**95.55** ±2.66	**75.35** ±0.15	**67.06**
×	√	√	√	28.1 ±1.01	20.7 ±0.08	48.3 ±2.39	59.4 ±1.17	21.7 ±0.16	35.6	34.90 ±1.33	21.01 ±0.28	53.12 ±6.43	79.61 ±1.27	41.34 ±0.22	46.00
√	×	√	√	31.6 ±1.78	26.7 ±0.51	55.8 ±1.88	81.9 ±0.65	52.7 ±1.45	49.7	47.05 ±2.24	33.72 ±0.56	65.62 ±8.89	80.34 ±4.21	73.12 ±0.46	59.97
√	√	×	√	26.3 ±0.78	20.8 ±0.09	50.7 ±1.49	75.7 ±0.93	38.2 ±1.27	42.3	24.57 ±1.06	21.11 ±0.42	59.72 ±4.52	85.76 ±1.49	55.05 ±0.31	49.24
√	√	√	Random	27.0 ±0.99	22.0 ±0.20	51.3 ±0.99	81.6 ±1.65	32.4 ±0.36	42.9	46.79 ±1.27	34.05 ±0.40	55.83 ±6.01	87.63 ±0.82	68.54 ±0.43	58.56
√	√	√	Block	27.2 ±0.55	24.5 ±0.45	51.7 ±0.93	77.0 ±0.81	51.9 ±1.20	46.5	45.66 ±1.58	35.31 ±0.49	60.42 ±8.14	92.37 ±1.85	70.13 ±0.40	60.78
√	√	√	Channel	30.0 ±1.39	21.3 ±0.13	55.4 ±2.97	81.0 ±0.91	21.1 ±0.11	41.8	44.97 ±1.10	25.82 ±0.21	63.89 ±3.62	87.30 ±0.43	70.78 ±0.19	58.55
√	√	√	Temporal	30.4 ±1.40	23.6 ±0.11	51.7 ±2.02	79.8 ±0.49	36.7 ±1.34	44.4	28.99 ±1.98	28.37 ±0.47	51.60 ±7.62	86.16 ±3.38	65.57 ±0.35	52.14



(a) Temporal correlation scale (σ_t)

(b) Spatial correlation scale (σ_s)

Figure 3: Sensitivity of ST-GFM to the temporal (σ_t) and spatial (σ_s) correlation scales. Each cell reports balanced accuracy (%) for KNN probing (top) and frozen probing (bottom). Avg. is the unweighted mean over the seven datasets. Colors are normalized separately for each dataset and probing protocol.

4.4 Hyperparameter Sensitivity

Figure 3 examines the spatial and temporal correlation scales of ST-GFM by varying one scale at a time. The selected $\sigma_s = 0.09$ and $\sigma_t = 1.0$ achieve the highest mean balanced accuracy under both KNN and frozen probing, despite some variation across individual datasets. Mean performance declines away from the selected scales, consistent with a trade-off in the spatial and temporal extent of masking. At a fixed mask ratio, smaller scales may leave nearby observations visible, whereas larger scales concentrate masked tokens into broader regions and may reduce the local context available for prediction. The consistent optimum across both probing protocols indicates that moderate spatiotemporal correlation provides a better balance for learning transferable EEG representations.

5 Conclusion

This work targets the limited transferability of pretrained EEG representations by identifying three key issues in current pretraining paradigms: low-level waveform-oriented objectives, insufficient integration of electrode geometry into inter-channel modeling, and local information leakage under random masking. We propose EEG-GPRL to address these issues through latent-space predictive learning, GeoAttn, ST-GFM, and VISReg. Experiments across diverse downstream tasks show strong transfer performance under both KNN and frozen probing, while ablation studies verify the complementary contributions of the proposed components.

Despite these results, our evaluation is limited to nine classification datasets and two fixed-backbone probing protocols. The scalability of EEG-GPRL with larger models, longer pretraining, and more diverse downstream settings therefore remains to be explored. Future work will extend EEG-GPRL to broader EEG tasks and investigate scaling laws across model size, data scale, and compute, together with transferability under more challenging cross-domain settings.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, discussion of experimental design, and preparation of figures and tables. All AI-assisted content was reviewed by the authors. Mathematical formulations, implementations, experiments, and reported results were independently checked and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses previously collected EEG datasets for pretraining and downstream evaluation and involves no new human-subject recruitment, intervention, or data collection. Ethical approvals and informed-consent procedures for the original studies are described in the corresponding dataset publications. The present study focuses on methodological development and representation learning; results on clinical datasets are intended for research evaluation and should not be interpreted as direct clinical validation.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the main text and appendix. The preprocessing pipeline, model architecture, pretraining objective, ST-GFM construction, and probing protocols are described in the main text. The appendix further provides dataset details, the formulation of VISReg, complete pretraining and frozen-probing hyperparameters, training dynamics, and additional analyses of the spatial structure of ST-GFM. All baseline models are evaluated using publicly released pretrained models under the same downstream data splits and evaluation protocols. Anonymous source code is available at https://anonymous.4open.science/r/EEG-GPRL-ICLR2027.

REFERENCES

Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of neural engineering*, 16(1): 011001, 2019.

U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in biology and medicine*, 100:270–278, 2018.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Nursena Ataseven. Memory reactivation levels remain unaffected by anticipated interference, 2025a.

Nursena Ataseven. Memory reactivation levels remain unaffected by anticipated interference experiment 2 dataset, 2025b.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16:1–6, 2008.

Enrique Carrillosulub. Light hypnotherapy, 2025.

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature reviews neurology*, 12(9):513–525, 2016.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-NTFM: A unified foundation model for EEG signal representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oUMiuYHW21.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7), 2020. doi: 10.3390/pr8070846.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. *Advances in Neural Information Processing Systems*, 38:70682–70708, 2026.

Yassine El Ouahidi, Jonathan Lys, Philipp Th"olke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

Sanaz Ghaffari, Arian Yavari, Sara Bonyadian, Arsalan Ghofrani, and Russell Butler. Visual eeg study (brainvision → bids), 2025.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Jiazhen Hong, Geoffrey Mackellar, and Soheila Ghane. Samba: Toward a long-context eeg foundation model via spatial embedding and differential mamba. *arXiv preprint arXiv:2511.18571*, 2025.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2024a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=QzTpTRVtrP.

Yufeng Ke, Hongzhi Qi, Feng He, Shuang Liu, Xin Zhao, Peng Zhou, Lixin Zhang, and Dong Ming. An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Frontiers in human neuroscience*, 8:703, 2014.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Isaac Kinley and Sue Becker. Cueing the future to reduce temporal discounting, 2025.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

Jinhao Li, Zhiyuan Ma, Xueqiao Han, Zhongye Xia, Xinche Zhang, Shanghong Xie, Yixuan Liu, Yongjian Li, Runmin Gan, Tianlin Huo, and Sen Song. EEG-JEPA: Structured latent prediction for EEG foundation models. *arXiv preprint arXiv:2608.00114*, 2026. URL https://arxiv.org/abs/2608.00114.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems. *arXiv preprint arXiv:2601.17883*, 2026.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=msJgEkjwh5.

Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New), 2016. URL https://doi.org/10.6084/m9.figshare.4244171.v2.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 9780262182539. URL https://gaussianprocess.org/gpml/.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE transactions on neural systems and rehabilitation engineering*, 25(11):1998–2008, 2017.

Logan Trujillo. Raw EEG Data, 2020. URL https://doi.org/10.18738/T8/SS2NHB.

Logan T Trujillo, Candice T Stanfield, and Ruben D Vela. The effect of electroencephalogram (eeg) reference choice on information-theoretic measures of the complexity and integration of eeg signals. *Frontiers in neuroscience*, 11:425, 2017.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Yiming Kang, Shijian Li, and Gang Pan. Deeperbrain: A neuro-grounded eeg foundation model towards universal bci. *arXiv preprint arXiv:2601.06134*, 2026.

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-invariance-sketching regularization for JEPA training. *arXiv preprint arXiv:2606.02572*, 2026. URL https://arxiv.org/abs/2606.02572.

Wei Xiong, Junming Lin, Jiangtong Li, Jie Li, and Changjun Jiang. Alfee: Adaptive large foundation model for eeg representation. *arXiv preprint arXiv:2505.06291*, 2025.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are eeg foundation models worth it? comparative evaluation with traditional decoders in diverse bci tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. In *Advances in Neural Information Processing Systems*, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

Appendix

A Details of Datasets

Pretraining Datasets. Table 4 summarizes the heterogeneous EEG datasets used for pretraining. These datasets differ substantially in sampling rate, channel configuration, recording duration, and acquisition setting, providing diverse neural signals for representation learning. We apply the unified preprocessing pipeline described in Sec. 4.1 to reduce acquisition-level inconsistencies. The complete collection contains 42,755.68 hours of EEG recordings, of which 18,546.35 hours are retained after preprocessing and quality control.

Downstream Datasets. Table 5 summarizes the nine downstream datasets covering emotion recognition, motor imagery, sleep staging, seizure detection, major depressive disorder detection, and workload classification. The datasets span diverse channel configurations, segment durations, label spaces, and data scales. For each dataset, we follow the train/validation/test splits used in prior EEG foundation model evaluations to ensure fair comparison with existing methods. This heterogeneous benchmark allows us to assess whether the pretrained representations remain transferable across substantially different EEG decoding paradigms.

Table 4: Information of datasets used for pre-training.

Dataset	# Rate (Hz)	# Channels	Subjects	Total Hours	# Samples	Duration (s)	Valid Hours
TUH EEG Corpus Obeid & Picone (2016)	250-1024	19	14,987	27,062	1,109,545	30	9246.21
PhysioNet2018 Ghassemi et al. (2018)	200	6	1,983	≈ 15,269	1,065,391	30	8878.26
Raw EEG Data Trujillo (2020)	256	64	48	43.08	30,978	5	43.025
Resting State EEG Data (Trujillo et al. 2017)	256	64	22	3.04	2,178	5	3.025
Siena Scalp EEG Database (Detti et al. 2020)	256	27	14	141.02	50,749	10	140.97
OpenNeuro ds006367 Ataseven (2025a)	1000	28	52	68.17	29,929	8	66.51
OpenNeuro ds006370 Ataseven (2025b)	1000	28	56	95.32	67,953	5	94.38
OpenNeuro ds006437 Carrillosulub (2025)	256	64	9	16.80	12,069	5	16.76
OpenNeuro ds006446 Kinley & Becker (2025)	2048	64	26	18.03	12,970	5	18.01
OpenNeuro ds006547 Ghaffari et al. (2025)	500	63	31	39.22	28,222	5	39.20
Total	–	–	17,228	42,755.68	2,409,984	–	18,546.35

Table 5: Overview of downstream tasks and datasets.

BCI Tasks	Datasets	Rate (Hz)	# Channels	Duration	# Samples	Label	Train	Val	Test
Emotion	SEED-V Liu et al. (2021)	1000	62	1s	117,744	5-class	Trials per session split 5:5:5		
	SEED-VII Jiang et al. (2024a)	1000	62	5s	55,561	7-class	Trials per session split 10:5:5		
Motor Imagery	PhysioNet-MI Schalk et al. (2004)	160	64	4s	9,837	4-class	Sub 1-70	Sub 71-89	Sub 90-109
	BCICIV-2a Brunner et al. (2008)	250	22	4s	5,088	4-class	Sub 1-5	Sub 6-7	Sub 8-9
Sleep Staging	ISRUC Khalighi et al. (2016)	200	6	30s	89,240	5-class	Sub 1-80	Sub 81-90	Sub 91-100
	HMC Alvarez-Estevez & Rijsman (2021)	256	4	30s	137,243	5-class	Sub 1-100	Sub 101-125	Sub 126-154
Seizure Detection	CHB-MIT Shoeb (2009)	256	16	10s	326,993	2-class	Sub 1-19	Sub 20-21	Sub 22-23
MDD	Mumtaz2016 Mumtaz (2016)	256	19	5s	7,143	2-class	42MDD+40NC	10MDD+8NC	10MDD+10NC
Workload	MentalArithmetic Zyma et al. (2019)	500	20	5s	1,707	2-class	Sub 1-28	Sub 29-32	Sub 33-36

B Details of VISReg

We adopt Variance-Invariance-Sketching Regularization (VISReg) (Wu et al., 2026) to stabilize latent-space predictive learning and prevent representational collapse. VISReg regularizes the online encoder representations through complementary scale, center, and shape constraints. Let $B \in \mathbb{R}^{N\times D}$ collect the online encoder representations of visible EEG contexts across a minibatch, where N counts the visible tokens excluding padding. For dimension j, define

$$m_j = \frac{1}{N}\sum_{i=1}^{N} B_{ij}, \qquad s_j = \max\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(B_{ij}-m_j)^2}, \epsilon\right), \qquad \widetilde{B}_{ij} = \frac{B_{ij}-m_j}{\mathrm{sg}(s_j)}. \tag{11}$$

The scale and center terms are

$$\mathcal{L}_{\text{scale}} = \frac{1}{D}\sum_{j=1}^{D}(s_j-1)^2, \qquad \mathcal{L}_{\text{center}} = \frac{1}{D}\sum_{j=1}^{D} m_j^2. \tag{12}$$

For sketching, we draw P independent directions w_k uniformly from the unit sphere in $\mathbb{R}^D$ and sort each projected vector $\widetilde{B}w_k$. Denoting its i-th order statistic by $[\widetilde{B}w_k]_{(i)}$, the shape loss is

$$\mathcal{L}_{\text{shape}} = \frac{1}{PN}\sum_{k=1}^{P}\sum_{i=1}^{N}\left([\widetilde{B}w_k]_{(i)} - q_i\right)^2, \qquad q_i = \Phi^{-1}\left(\frac{i}{N+1}\right), \tag{13}$$

where Φ^{-1} denotes the standard Gaussian quantile function.

The three terms constrain complementary properties of the latent representation. $\mathcal{L}_{\text{scale}}$ maintains feature variation, $\mathcal{L}_{\text{center}}$ controls mean drift, and $\mathcal{L}_{\text{shape}}$ regularizes the distributional shape of the standardized representations. Together, they stabilize the latent space without introducing waveform reconstruction targets, allowing predictive learning to focus on contextual representations while reducing the risk of representational collapse.

C Implementation Details

For reproducibility, we provide the complete implementation and optimization hyperparameters omitted from the main text. Table 6 summarizes the model architecture and pretraining configuration, including the EMA and VISReg settings. Table 7 reports the optimization settings used for frozen probing.

Following the preprocessing described in Sec. 4.1, each EEG recording is divided into non-overlapping 1-s windows and mapped to time-frequency tokens. The encoder accepts up to 64 channels and 30 windows per recording. Electrode coordinates $p_c = (p_c^{(x)}, p_c^{(y)}, p_c^{(z)})$ are taken from the MNE `standard_1005` montage in meters. The same coordinates define the absolute channel encoding, the relative-coordinate rotations in GeoAttn, and the spatial covariance of ST-GFM, so these components use a consistent electrode geometry.

The encoder uses D features and H attention heads, giving a head dimension $d_h = D/H = 30$. For the temporal and spatial rotations in Sec. 3.1, the frequencies are, respectively,

$$\omega_m^{(t)} = 10000^{-2(m-1)/d_h}, \quad m = 1, \ldots, d_h/2, \quad \omega_m^{(s)} = 10000^{-6(m-1)/d_h}, \quad m = 1, \ldots, d_h/6.$$

The temporal rotation applies $\omega_m^{(t)}$ to the window index t. In channel attention, each x-, y-, and z-coordinate pair uses $\omega_m^{(s)}$ with its corresponding coordinate component, without additional coordinate rescaling.

ST-GFM uses these electrode coordinates and the window indices to construct its spatial and temporal covariance matrices. A Gaussian score field is sampled for each recording, and the highest-scoring valid tokens are masked at the fixed ratio $\rho = 0.5$. The online encoder processes the visible contexts, while the EMA teacher encodes the unmasked recording to provide targets at masked locations. A two-layer predictor estimates these targets, and VISReg regularizes the online representations. Table 6 gives the corresponding architecture and optimization settings.

For downstream evaluation, the pretrained tokenizer and encoder of EEG-GPRL and the corresponding backbones of the baselines remain frozen. In KNN probing, we flatten each model's EEG patch or token representations into one vector per sample, without additional feature-wise standardization or ℓ_2 normalization. We use the training split as the reference bank and perform exact K-nearest-neighbor searches for $K \in \{3, 5, 7, 9, 11\}$ with Euclidean distance; neither validation nor test samples enter the bank. Class votes are weighted by inverse distance, $w_i \propto (d_i + 10^{-8})^{-1}$, and normalized across the retrieved neighbors. Frozen probing instead trains a task-specific readout on fixed representations, with results reported over five downstream training seeds using the optimization settings in Table 7. The MLP flattens the channel and temporal token representations without pooling and applies three linear layers. Its two hidden layers have widths proportional to the backbone dimension D, each followed by ELU activation and dropout of 0.1.

D Training Dynamics

Figure 4 shows the optimization dynamics of latent-space predictive learning and VISReg. During the early stage of training, both the prediction and regularization objectives change rapidly before

Table 6: Hyperparameters for pretraining.

Hyperparameters	**Values**
Sampling rate f_s	200 Hz
Max Channels C	64
Maximum number of windows T	30 (1-s windows)
Mask Ratio ρ	0.5
Channel Correlation Scale σ_s	0.09
Temporal Correlation Scale σ_t	1.0
Num of Layers L	12
Predictor Layers	2
Hidden Dimension D	240
Attention Heads H	8
Head Dimension d_h	30
MLP Ratio	4
Dropout Rate	0.1
Initialization	Kaiming
# Freq Bands of Temporal Attn	$d_h/2$
# Freq Bands of Channel Attn	$d_h/3(\text{axis})/2$
Frequency Base of RoPE	10000
Epochs	10
Per-GPU Batch Size	48
Global Batch Size	384
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Max / Min LR	5e-4 / 1e-6
Weight Decay	1e-2
Gradient Clipping Norm	1.0
LR Scheduler	Cosine Decay Linear Warmup
Warmup Duration	1 epoch
Gradient Accumulation Steps	1
Mixed Precision	BF16
Initial / Final EMA Momentum	0.999 / 1.0
EMA Schedule	Linear up
VISReg Weight λ	1.0
VISReg Projection Directions	64
VISReg Target Standard Deviation	1.0

Table 7: Hyperparameters for frozen probing.

Hyperparameters	**Values**
Epochs	50
Batch Size	64 / 16 (for ISRUC)
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Maximum LR	1e-4
Minimum LR	1e-5
Weight Decay	5e-2
Gradient Clipping Norm	1.0
LR Scheduler	Cosine Decay No Warmup

entering a stable regime. Meanwhile, the feature-variance statistic approaches its target level and remains stable throughout the remaining training process. The decomposition of VISReg further shows that the scale, shape, and center terms are jointly reduced while feature variation is preserved.

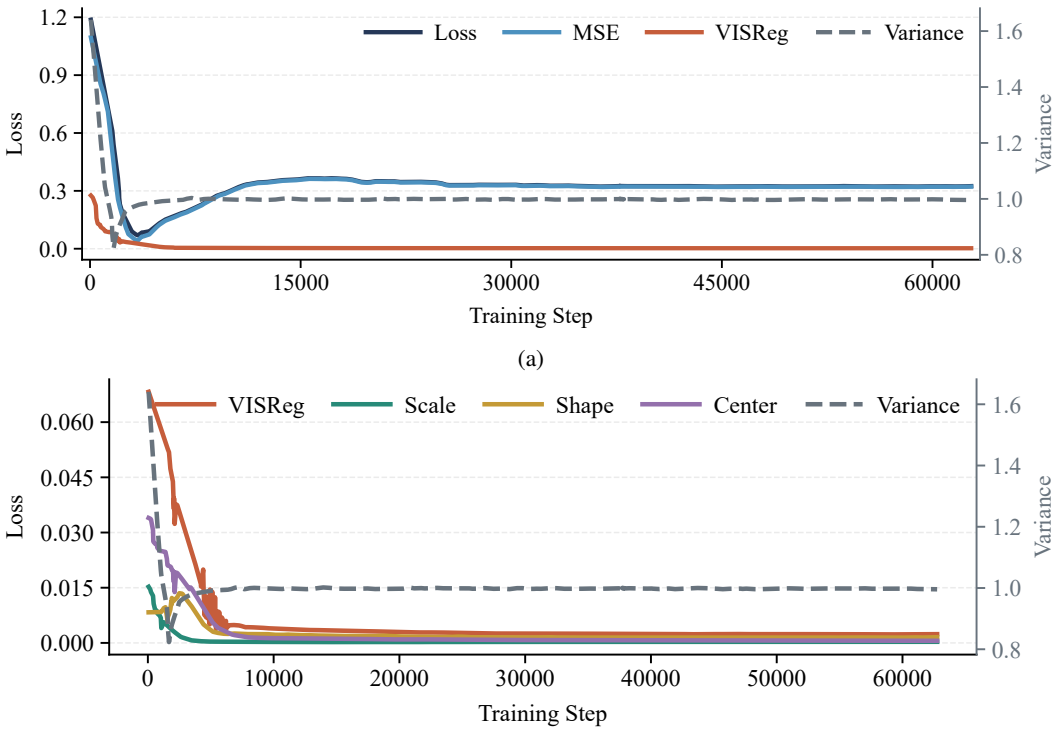


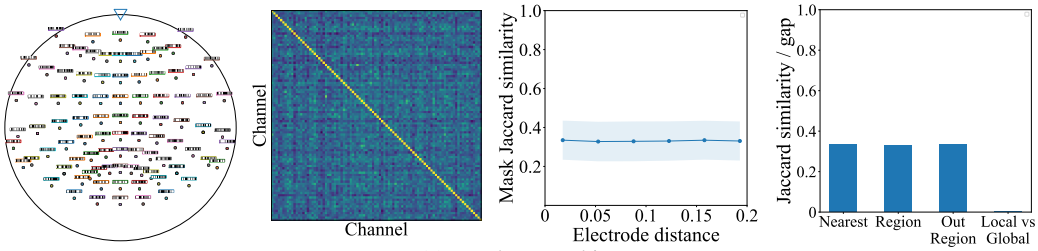
Figure 4: Training dynamics of EEG-GPRL. (a) Evolution of the total loss, latent prediction MSE, VISReg loss, and feature-variance statistic during pretraining. (b) Decomposition of VISReg into the scale, shape, and center terms. The representation variance remains stable after the initial optimization stage, indicating that feature diversity is maintained throughout latent predictive learning.

These dynamics indicate that latent prediction can be optimized without collapsing the representation space, complementing the ablation results that show a substantial degradation when VISReg is removed.

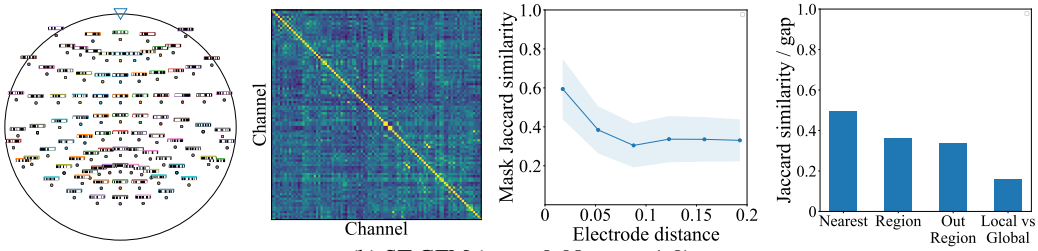
E Spatial Structure of ST-GFM

To verify that ST-GFM produces the intended geometry-dependent masking structure, Figure 5 compares random masking with ST-GFM under different spatial correlation scales σ_s, while fixing $\sigma_t = 1.0$. Random masking shows nearly distance-independent mask similarity, indicating that neighboring and distant electrodes are treated similarly. In contrast, ST-GFM produces substantially higher mask overlap between nearby electrodes, with similarity gradually decreasing as electrode distance increases. Increasing σ_s progressively extends this spatial coherence over broader electrode neighborhoods. These results verify that the Gaussian-field formulation provides explicit control over the spatial correlation of masked contexts rather than imposing a fixed geometric grouping rule.

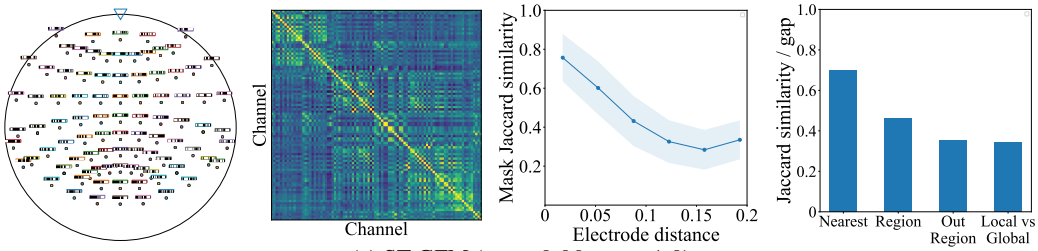
This behavior also provides a possible interpretation of the sensitivity results in Figure 3. A small σ_s induces only weak spatial coupling, which may leave more nearby observations visible around masked tokens and thus provide limited mitigation of local information leakage. In contrast, an excessively large σ_s extends mask correlation over broader regions, potentially reducing the contextual information available for prediction. The intermediate setting therefore appears to provide a better balance between mitigating local information leakage and preserving informative context.



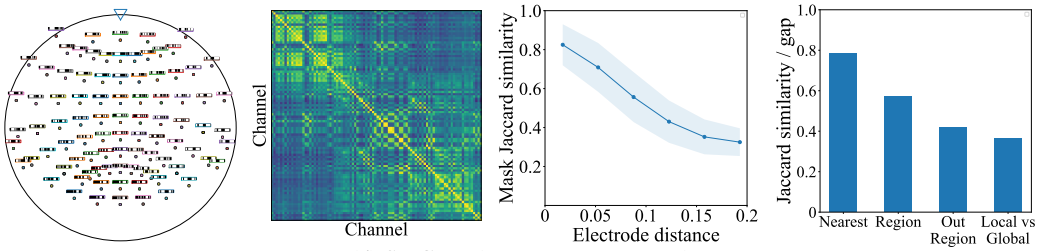
(a) Random Masking



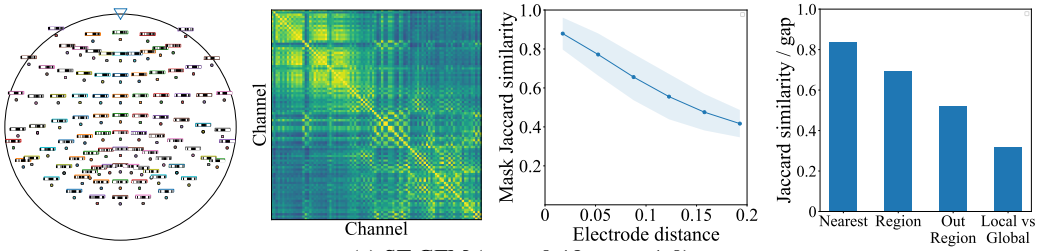
(b) ST-GFM ($\sigma_s = 0.03$, $\sigma_t = 1.0$)



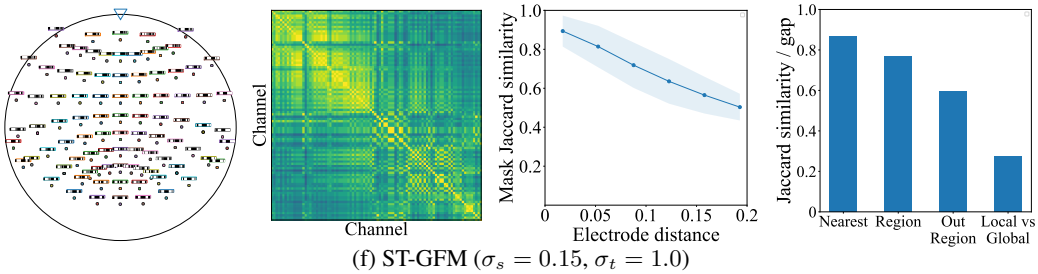
(c) ST-GFM ($\sigma_s = 0.06$, $\sigma_t = 1.0$)



(d) ST-GFM ($\sigma_s = 0.09$, $\sigma_t = 1.0$)



(e) ST-GFM ($\sigma_s = 0.12$, $\sigma_t = 1.0$)



(f) ST-GFM ($\sigma_s = 0.15$, $\sigma_t = 1.0$)

Figure 5: Spatial structure of random masking and ST-GFM under different spatial correlation scales σ_s with $\sigma_t = 1.0$. Columns show, from left to right, sampled mask patterns over electrode locations, pairwise mask Jaccard similarity, Jaccard similarity as a function of electrode distance, and summary statistics of local and global similarity. Random masking exhibits nearly distance-independent similarity, whereas ST-GFM produces stronger overlap for nearby electrodes and progressively extends spatial coherence as σ_s increases.