TOPOJEPA: TOPOLOGY-GROUNDED JOINT-EMBEDDING PREDICTION FOR EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised EEG foundation models based on masked autoencoding have shown strong potential for learning transferable representations from large-scale unlabeled data. However, reconstruction objectives inherently emphasize electrode-level signal details, which may not constitute the most transferable structure for downstream EEG decoding. This raises a fundamental question: what should an EEG model predict to capture the intrinsic organization of neural dynamics? We introduce TopoJEPA, a Topology-Grounded Joint-Embedding Predictive Architecture built on the principle of *Grounding Before Prediction*. Instead of directly predicting raw signals or unconstrained latent representations, TopoJEPA first constructs a target space aligned with EEG's intrinsic spatiotemporal organization and then learns predictive dynamics within this space. Specifically, Topological Target Grounding (TTG) maps heterogeneous electrode configurations onto a unified hierarchical topology and learns hemisphere- and region-level representations through reconstruction. To prevent reconstruction-specific signal details from dominating the learned target space, a Reconstruction-Specific Branch isolates such residual information in an auxiliary pathway while retaining reconstruction fidelity. Building on the grounded representations established by TTG, Topology-Grounded Autoregressive Prediction (TGAP) then models their spatiotemporal evolution in joint-embedding space. Across ten datasets covering six representative EEG decoding tasks, TopoJEPA achieves competitive performance against existing self-supervised and EEG foundation-model baselines, with particularly strong linear-probing performance across multiple tasks. These results demonstrate the effectiveness of constructing a stable, topology-grounded target space before predictive learning, enabling transferable EEG representations that capture meaningful spatiotemporal structure while reducing reliance on reconstruction-specific signal details. Code is available at `https://anonymous.4open.science/r/TopoJEPA-F8B3`.

1 INTRODUCTION

Decoding brain activity from electroencephalography (EEG) underpins a wide range of applications, from emotion recognition and clinical assessment to brain-computer interfaces (Liu et al., 2022; Obeid & Picone, 2016; Lawhern et al., 2018). Yet EEG decoding remains difficult to scale: signals vary substantially across individuals and recording configurations, while task-specific annotations are costly to obtain. EEG foundation models offer a promising alternative by learning reusable representations from large-scale unlabeled recordings through self-supervised pretraining (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025; Jiang et al., 2025b).

Masked reconstruction has become a common choice because it provides a simple and scalable learning signal by recovering missing waveform segments or signal-derived tokens (Jiang et al., 2024; Wang et al., 2025). However, reconstruction-based learning naturally favors local, low-level signal details. Accurate recovery of waveform content does not necessarily require the model to organize its representation around higher-level structure, which can limit downstream performance when supervised adaptation is unavailable (Mohammadi Foumani et al., 2024; Panchavati et al.,

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.

2026). It raises a fundamental question: *what should an EEG foundation model predict to capture higher-level latent representations?*

Joint-embedding predictive architectures (JEPAs) provide a natural direction for EEG pretraining by shifting prediction from raw signals to latent representations (Assran et al., 2023). Recent EEG JEPAs have mainly focused on improving the construction of predictive targets, through exponential moving average (EMA) target encoders (Mohammadi Foumani et al., 2024; Guetschel et al., 2024; Hojjati et al., 2026), explicit representation regularization (Panchavati et al., 2026), or shallower latent targets (Segal et al., 2026). Further discussion of JEPA-based approaches is provided in Appendix A. However, they encounter a more fundamental challenge: as the target representation and predictive objective evolve jointly during training, learning a stable EEG latent space that reliably captures task-relevant intrinsic structure and dynamics becomes particularly challenging. This calls for a fundamental rethinking of how the predictive space itself should be established. Inspired by SALT's static target encoder in video representation learning (Li et al., 2026), we propose a novel concept called *Grounding Before Prediction*, which first learns and fixes a topology-grounded latent space before modeling its spatiotemporal dynamics.

We instantiate this principle in TopoJEPA, a Topology-Grounded Joint-Embedding Predictive Architecture. TopoJEPA first maps heterogeneous electrode configurations into a unified hierarchical topology and grounds the resulting EEG latent space, before learning temporal dependencies and hemisphere-to-region dynamics over the fixed targets. A dedicated reconstruction-specific branch further prevents reconstruction-specific information from dominating the predictive space. We evaluate TopoJEPA on ten datasets spanning six representative EEG decoding tasks under both linear probing and full fine-tuning. TopoJEPA shows particularly strong gains under linear probing, suggesting that grounding the latent space before prediction yields representations that are more directly transferable to downstream tasks with minimal adaptation. The key contributions of this work are:

- To obtain a stable EEG latent space that preserves hierarchical topology while supporting temporal dynamics modeling, we introduce Grounding Before Prediction, a pretraining principle that first establishes a stable, topology-grounded EEG latent space and then learns predictive dynamics within this fixed space.
- To instantiate this principle, we propose TopoJEPA, a two-stage framework that first establishes a fixed topology-grounded target space through Topological Target Grounding (TTG), with reconstruction-specific information handled by a separate pathway, and then learns temporal and hemisphere-to-region dynamics within this space through Topology-Grounded Autoregressive Prediction (TGAP).
- To broadly evaluate the learned representations, we validate TopoJEPA on ten datasets spanning six EEG decoding tasks under both linear probing and full fine-tuning. Its strong linear probing performance indicates that TopoJEPA learns high-quality representations that are directly useful for downstream decoding.

2 METHOD

2.1 OVERVIEW

TopoJEPA follows the principle of Grounding Before Prediction with two decoupled pretraining stages. TTG first maps heterogeneous EEG montages into a fixed set of topology tokens, comprising two hemisphere tokens and seven region tokens, and grounds their representations through waveform and spectral reconstruction. Auxiliary reconstruction tokens bypass the Target Encoder and carry reconstruction-specific information directly to the decoder. TGAP freezes the target pathway and trains a separate context pathway to learn the temporal and cross-level dynamics of the grounded topology representations. A Spatiotemporal Context Encoder regulates information flow through spatiotemporal attention constraints, while a Latent Predictor maps the resulting context representations to predicted topology representations. After pretraining, the context pathway is retained for downstream EEG decoding.

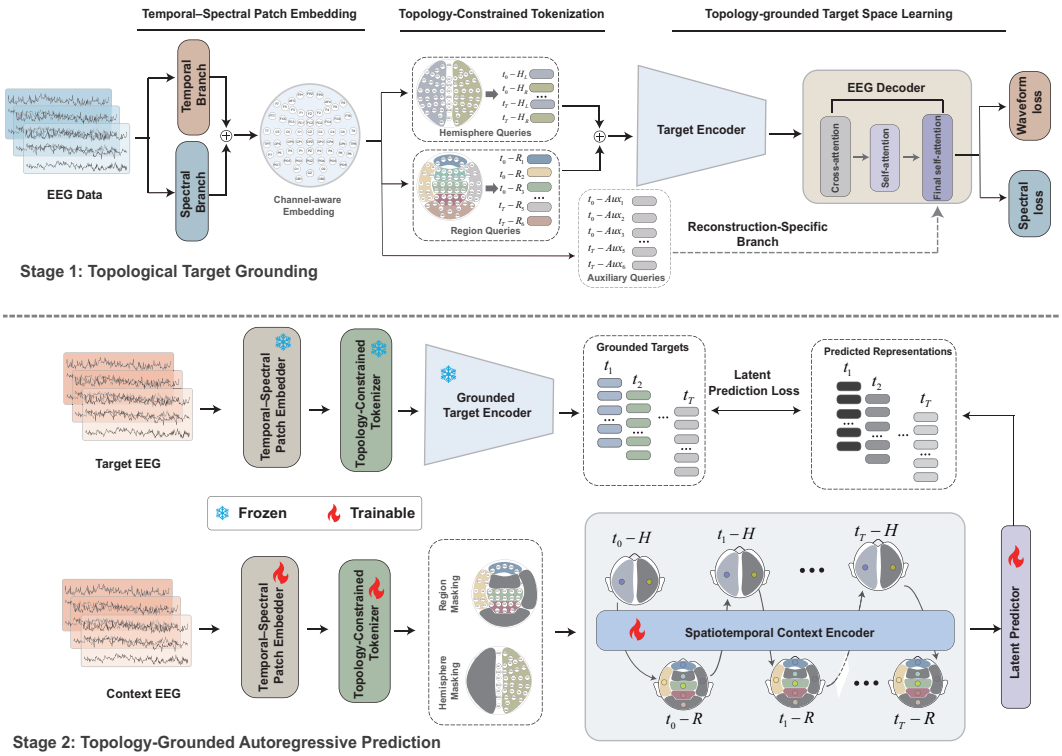


Figure 1: Overview of the TopoJEPA framework. TTG first learns grounded topology representations through waveform and spectral reconstruction, while auxiliary reconstruction tokens are routed through a separate decoder pathway. TGAP then freezes the grounded target pathway and predicts future topology representations from context-masked EEG using spatiotemporal attention constraints.

2.2 Topological Target Grounding

Patch Embedding. Let $\mathbf{X} \in \mathbb{R}^{C\times T\times P}$ denote an EEG window with C electrodes and T temporal patches, each containing P samples. For the patch $\mathbf{x}_{c,t}$ from electrode c at time step t, we compute

$$\mathbf{e}_{c,t} = \phi(\mathbf{x}_{c,t}) + \gamma(\mathbf{r}_c),$$

where ϕ extracts temporal and spectral features, $\mathbf{r}_c$ is the coordinate of electrode c, and γ encodes its spatial position. Stacking the embeddings across electrodes gives $\mathbf{E}_t \in \mathbb{R}^{C\times d_{\mathrm{T}}}$, where d_{T} is the topology representation dimension. Details of the temporal and spectral branches are provided in Appendix B.2.

Topology-Constrained Tokenization. To obtain consistent spatial representations across electrode montages, we map the channel embeddings to a fixed set of topology roles. We use $K = 9$ topology queries, comprising two hemisphere roles and seven region roles. Let $\mathbf{q}_k$ denote the learned query for topology role k and $\mathcal{G}_k$ its associated electrode group. The regional partition follows the scalp-sensor organization used in prior work (Ye et al., 2025). We further introduce D auxiliary reconstruction queries that access the full montage and are used only for reconstruction. At time step t, each query aggregates its available channel embeddings:

$$\mathbf{u}_{t,k} = \mathrm{QueryAttn}\left(\mathbf{q}_k, \mathbf{E}_t; \mathcal{V}_{k,t}\right) + \boldsymbol{\eta}_k + \mathbf{p}_{t+o}, \qquad k = 1, \ldots, K+D, \tag{1}$$

where $\mathcal{V}_{k,t} \subseteq \mathcal{G}_k$ is the set of electrodes available to query k, $\boldsymbol{\eta}_k$ is its role embedding, and $\mathbf{p}_{t+o}$ is the temporal positional embedding with a shared random offset o. For $k \le K$, $\mathbf{u}_{t,k} \in \mathbb{R}^{d_{\mathrm{T}}}$ is the topology token for role k. The remaining D tokens serve as auxiliary reconstruction tokens. Query definitions and montage-specific electrode assignments are provided in Appendix B.2 and Table 6.

Target Grounding and Reconstruction-Specific Branch. The topology tokens are stacked across time as $U \in \mathbb{R}^{KT \times d_{\mathrm{T}}}$ and encoded by the Target Encoder F_θ:

$$Z = F_\theta(U), \qquad V_t = [\mathbf{u}_{t,K+1}, \ldots, \mathbf{u}_{t,K+D}], \tag{2}$$

where θ denotes the Target Encoder parameters, $Z \in \mathbb{R}^{KT \times d_{\mathrm{T}}}$ contains the topology representations, and $V_t \in \mathbb{R}^{D \times d_{\mathrm{T}}}$ contains the auxiliary reconstruction tokens at time step t. We use $Z_t \in \mathbb{R}^{K \times d_{\mathrm{T}}}$ to denote the topology representations at time step t. The auxiliary tokens bypass the Target Encoder, allowing reconstruction-specific information to be handled separately from the topology representations.

At each time step, the EEG Decoder uses C fixed channel queries, each given by a non-learnable spatial encoding of the corresponding electrode coordinate. Through cross-attention to Z_t, these queries map the K topology representations back to C channel-wise representations. After channel-wise self-attention, this gives $Y_t^{\mathrm{dec}} \in \mathbb{R}^{C \times d_{\mathrm{D}}}$, where d_{D} is the decoder dimension. The auxiliary tokens are introduced only in the final decoder block:

$$\widetilde{Y}_t^{\mathrm{dec}} = \mathrm{Rows}_{1:C}\left[\mathcal{D}_{\mathrm{aux}}\left(\begin{bmatrix} Y_t^{\mathrm{dec}} \\ \pi_v(V_t) \end{bmatrix}\right)\right], \tag{3}$$

where $\mathcal{D}_{\mathrm{aux}}$ denotes the final decoder block, $\pi_v : \mathbb{R}^{d_{\mathrm{T}}} \to \mathbb{R}^{d_{\mathrm{D}}}$ projects the auxiliary tokens to the decoder dimension, and $\mathrm{Rows}_{1:C}$ retains the refined channel representations.

Two reconstruction heads predict waveform patches and standardized Fourier amplitudes. The Stage 1 objective is

$$\mathcal{L}_{\mathrm{rec}} = \lambda_{\mathrm{raw}}\mathcal{L}_{\mathrm{raw}} + \lambda_{\mathrm{freq}}\mathcal{L}_{\mathrm{freq}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{raw}}$ and $\mathcal{L}_{\mathrm{freq}}$ are the waveform and spectral reconstruction losses, respectively, and λ_{raw} and λ_{freq} control their contributions. Through this objective, TTG learns the topology representations that define the target space for Stage 2.

2.3 Topology-Grounded Autoregressive Prediction

After TTG, we freeze the grounded target pathway and train a separate context pathway to predict future topology representations within the fixed target space. The context pathway consists of its own Patch Embedding, Topology-Constrained Tokenizer, and Spatiotemporal Context Encoder. Before topology tokenization, we apply a topology-aware context mask M to the channel embeddings, alternating between hemisphere-level and region-level masks during training. The selected mask is kept fixed across all time steps to prevent masked spatial information from reappearing later in the temporal context. Ablations of the masking strategy are provided in Appendix E.4.

The Spatiotemporal Context Encoder further constrains information flow across topology roles and time. For a query at (t, k) and a key at (s, k'), let $\ell_k = 0$ for hemisphere roles and $\ell_k = 1$ for region roles. The attention constraint is defined as

$$A_{(t,k),(s,k')} = \mathbf{1}[(s < t \wedge k' \neq k) \vee (s = t \wedge \ell_{k'} < \ell_k)], \tag{5}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. The first term allows information from other topology roles at preceding time steps, while the second allows region representations to access hemisphere representations within the same time step. The resulting attention pattern is illustrated in Appendix Figure 9(d). With context mask M and attention constraint A, the context pathway produces

$$Z^{\mathrm{ctx}} = \mathcal{C}_\phi(\mathbf{X}; M, A), \qquad Z^{\mathrm{ctx}} \in \mathbb{R}^{KT \times d_{\mathrm{E}}}, \tag{6}$$

where $\mathcal{C}_\phi$ denotes the trainable context mapping with parameters ϕ, and d_{E} is the context representation dimension. We denote the context representation for topology role k at time step t by $\mathbf{z}_{t,k}^{\mathrm{ctx}} \in \mathbb{R}^{d_{\mathrm{E}}}$. The Latent Predictor P_ψ maps the context representations to the grounded target space:

$$\widehat{Z}_{2:T} = P_\psi\left(Z_{1:T-1}^{\mathrm{ctx}}\right), \qquad \widehat{Z}_{2:T} \in \mathbb{R}^{K(T-1) \times d_{\mathrm{T}}}, \tag{7}$$

where ψ denotes the predictor parameters and $\widehat{\mathbf{z}}_{t+1,k} \in \mathbb{R}^{d_{\mathrm{T}}}$ is the predicted topology representation for role k at the next time step. Finally, for a minibatch of size B, we optimize

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{BK(T-1)} \sum_{b=1}^{B} \sum_{t=1}^{T-1} \sum_{k=1}^{K} \mathrm{SmoothL1}\left(\widehat{\mathbf{z}}_{b,t+1,k}, [\mathrm{sg}(\mathcal{T}_{\Theta^\star}(\mathbf{X}_b))]_{t+1,k}\right), \tag{8}$$

where $\mathcal{T}_{\Theta^\star}$ denotes the target mapping learned during TTG with frozen parameters $\Theta^\star$, and sg denotes stop-gradient. Through this objective, TGAP learns latent spatiotemporal dynamics within the fixed topology target space established by TTG.

3 EXPERIMENTS

3.1 PRETRAINING SETUP

Pretraining Datasets. We pretrain TopoJEPA on 3,179.3 hours of unlabeled EEG collected across emotion elicitation, reading, and clinical settings. The corpus combines the SEED series, including SEED, SEED-IV, and SEED-VII (Zheng & Lu, 2015; Zheng et al., 2019; Jiang et al., 2025a), ZuCo 1.0 and 2.0 (Hollenstein et al., 2018; 2020), and a 3,000-hour subset of TUH (Obeid & Picone, 2016). Both pretraining stages use the same corpus, with all recordings resampled to 200 Hz and segmented into eight-second windows. Detailed information on the pretraining datasets and preprocessing is provided in Appendix B.1.

Optimization and Implementation. Both stages are optimized with AdamW (Loshchilov & Hutter, 2019) and cosine learning-rate decay (Loshchilov & Hutter, 2017). Detailed model and pretraining configurations, together with the topology construction across electrode montages, are provided in Appendix B.2.

3.2 DOWNSTREAM EVALUATION

Downstream EEG Tasks and Baselines. We evaluate TopoJEPA on ten datasets spanning six downstream EEG decoding tasks with diverse recording settings, electrode configurations, and prediction objectives. The benchmarks include emotion recognition on SEED-V (Liu et al., 2022) and DEAP (Koelstra et al., 2012), motor imagery on BCIC-IV-2a (Tangermann et al., 2012), cognitive workload on EEGMat (Zyma et al., 2019) and STEW (Lim et al., 2018), task difficulty on BAT (Faller et al., 2019) and SFT (Jao et al., 2023), vigilance assessment on SADT (Cao et al., 2019; Cui et al., 2023) and SEED-VIG (Zheng & Lu, 2017), and clinical EEG classification on ADFTD (Miltiadous et al., 2023).

We compare against five supervised EEG models, EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), TSception (Ding et al., 2023), LGGNet (Ding et al., 2024), and EEG-Deformer (Ding et al., 2025), as well as five pretrained EEG foundation models, LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), EEGPT (Wang et al., 2024), LUNA (Döner et al., 2025), and CSBrain (Zhou et al., 2025). We use published results when the evaluation protocol is directly comparable and mark them with †. Otherwise, supervised baselines follow their official implementations, while EEG foundation models are initialized from officially released pretrained checkpoints and evaluated using the corresponding official code. Complete dataset and baseline details are provided in Appendices C.1 and C.2, respectively.

Evaluation Settings and Metrics. For pretrained models, we evaluate both linear probing and full fine-tuning to characterize representation quality and downstream adaptability. For classification, we report Balanced Accuracy and Weighted F1, together with Cohen's Kappa for multiclass tasks and the Area Under the Receiver Operating Characteristic Curve (AUROC) for binary tasks. For vigilance regression, we report Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (COR). All reproduced experiments use three training seeds. Detailed downstream task settings are provided in Appendices C.3 and D.

3.3 COMPARISON WITH EXISTING METHODS

Table 1 compares TopoJEPA with supervised and pretrained EEG models on SEED-V and DEAP valence under full fine-tuning and linear probing. Complete results across all six task families are provided in Appendix D.

On SEED-V and DEAP, TopoJEPA remains competitive under full fine-tuning and shows a substantially stronger advantage under linear probing, achieving the strongest linear-probing results across all reported metrics on both datasets. The larger gains with a frozen encoder indicate that much of the downstream-relevant information is already encoded in the pretrained representation, rather than being acquired through task-specific adaptation. This pattern extends to the broader benchmark. TopoJEPA achieves the strongest results on BAT, SFT, and SADT under both evaluation settings,

Table 1: Comparison of emotion recognition performance on SEED-V and DEAP.

Model	Params	SEED-V			DEAP valence		
		Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Weighted F1	AUROC
Supervised learning							
EEGNet	2.1K	0.2323 ± 0.0056	0.0435 ± 0.0059	0.1776 ± 0.0105	0.5146 ± 0.0039	0.5172 ± 0.0050	0.5199 ± 0.0097
EEGConformer	220.3K	0.2975 ± 0.0022	0.1177 ± 0.0032	0.2713 ± 0.0043	0.5261 ± 0.0198	0.5350 ± 0.0175	0.5480 ± 0.0287
TSception	24.4K	0.3452 ± 0.0157	0.1857 ± 0.0169	0.3546 ± 0.0153	0.5544 ± 0.0051	0.5709 ± 0.0051	0.5856 ± 0.0081
LGGNet	304.7K	0.3317 ± 0.0094	0.1649 ± 0.0107	0.3370 ± 0.0078	0.5368 ± 0.0063	0.5444 ± 0.0095	0.5490 ± 0.0113
EEG-Deformer	540.9K	0.3145 ± 0.0039	0.1374 ± 0.0031	0.2913 ± 0.0058	0.5593 ± 0.0053	0.5683 ± 0.0075	0.5923 ± 0.0128
Full fine-tuning							
LaBraM-Base	5.8M	0.3976 ± 0.0138†	0.2386 ± 0.0209†	0.3974 ± 0.0111†	0.5276 ± 0.0049	0.5279 ± 0.0030	0.5364 ± 0.0116
CBraMod	4.9M	0.4091 ± 0.0097†	0.2569 ± 0.0143†	0.4101 ± 0.0108†	0.5647 ± 0.0241	0.5706 ± 0.0229	0.5842 ± 0.0242
EEGPT-Large	25.3M	0.2449 ± 0.0095	0.0568 ± 0.0108	0.2489 ± 0.0030	0.5356 ± 0.0015	0.5437 ± 0.0065	0.5534 ± 0.0081
LUNA-Base	6.6M	0.3730 ± 0.0098†	0.1831 ± 0.0103†	0.3389 ± 0.0091†	0.5424 ± 0.0219	0.5582 ± 0.0215	0.5549 ± 0.0190
CSBrain	8.9M	**0.4197 ± 0.0033**†	**0.2785 ± 0.0034**†	**0.4280 ± 0.0023**†	0.5917 ± 0.0224	0.6001 ± 0.0177	**0.6467 ± 0.0182**
TopoJEPA	19.8M	0.4157 ± 0.0108	0.2724 ± 0.0150	0.4185 ± 0.0090	**0.5941 ± 0.0061**	**0.6057 ± 0.0128**	0.6284 ± 0.0150
Linear probing							
LaBraM-Base	5.8M	0.2521 ± 0.0267†	0.0854 ± 0.0342†	0.2543 ± 0.0265†	0.5116 ± 0.0053	0.4875 ± 0.0163	0.5159 ± 0.0107
CBraMod	4.9M	0.2536 ± 0.0257†	0.0842 ± 0.0384†	0.2568 ± 0.0275†	0.4987 ± 0.0022	0.4342 ± 0.0102	0.4682 ± 0.0365
EEGPT-Large	25.3M	0.2527 ± 0.0066	0.0688 ± 0.0065	0.2477 ± 0.0056	0.5458 ± 0.0154	0.5543 ± 0.0075	0.5660 ± 0.0058
LUNA-Base	6.6M	0.2371 ± 0.0042	0.0492 ± 0.0063	0.2291 ± 0.0114	0.4998 ± 0.0094	0.4868 ± 0.0207	0.5049 ± 0.0028
CSBrain	8.9M	0.2563 ± 0.0010	0.0704 ± 0.0020	0.2591 ± 0.0067	0.5016 ± 0.0076	0.4612 ± 0.0278	0.5523 ± 0.0131
TopoJEPA	19.8M	**0.3729 ± 0.0035**	**0.2161 ± 0.0057**	**0.3795 ± 0.0063**	**0.5767 ± 0.0028**	**0.5918 ± 0.0034**	**0.6099 ± 0.0030**

† Published results from prior work.

leads key linear-probing metrics on BCIC-IV-2a, EEGMat, and SEED-VIG, and obtains the highest full-fine-tuning Balanced Accuracy on STEW and ADFTD.

3.4 Analysis of Target-Space Grounding

We first investigate two questions: whether predictive learning benefits from grounding the target space before prediction, and what factors shape the quality of the grounded space. To answer the first, we compare TopoJEPA with an EMA-based JEPA whose target representations continue to evolve during predictive learning. To answer the second, we vary three aspects of the Stage 1 target-encoder grounding process: the training data used for grounding, target encoder size, and the number of grounding epochs, while keeping the Stage 2 configuration fixed. Representations after TTG and TGAP are evaluated on SEED-V using the same frozen-encoder linear-probing protocol. Full experimental settings and metrics are provided in Appendix E.1.

Why is grounding performed before prediction? The dashed line in Figure 2 denotes the EMA-based JEPA reference, whose target representations continue to evolve during predictive learning. Under the Medium-80 setting, the Stage 1 target encoder achieves lower linear-probing Balanced Accuracy than this reference (30.04% vs. 33.96%), while the corresponding Stage 2 context encoder reaches 37.29%. This result supports the benefit of learning prediction against a fixed, pre-established target space rather than a continuously evolving one. It also shows that the quality of a grounded target space is not reflected by the downstream separability of the target encoder alone: a weaker Stage 1 target encoder can still provide better targets for Stage 2. We therefore further examine how training data for grounding, target encoder size, and the number of training epochs for grounding affect the resulting target space.

What makes target-space grounding effective? In Figure 2(a), we vary the training data used for the target encoder. Single-source grounding leads to markedly different linear-probing performance of the resulting Stage 2 context encoder across the SEED series, ZuCo series, and TUH, while increasing the mixed-source corpus from 10% to 50% and 100% raises Balanced Accuracy from 34.12% to 35.70% and 37.29%, respectively. This suggests that effective target-space grounding benefits from exposure to diverse EEG observations rather than relying on a single dataset or montage. In Figure 2(b), we vary the target encoder size. The corresponding context encoder improves from 35.42% with the Small target encoder to 37.29% with Medium, but only marginally to 37.40%

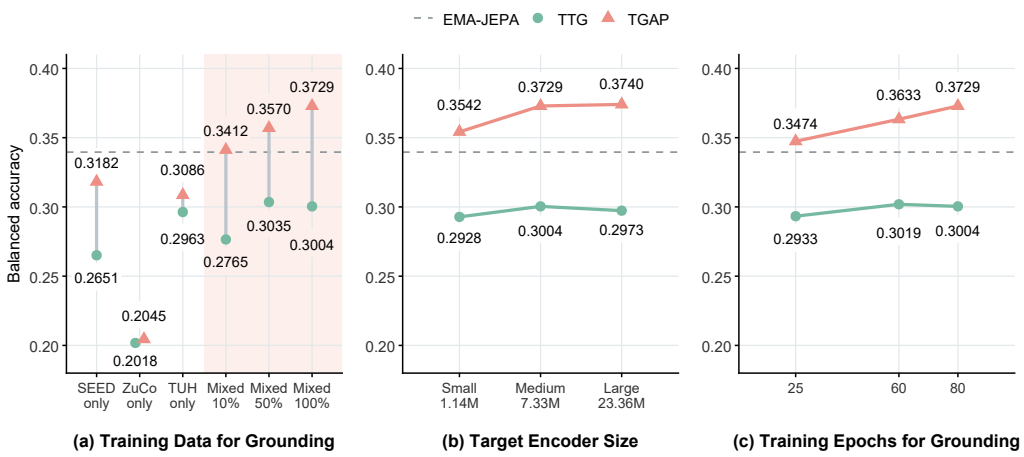


Figure 2: **Effect of target-space grounding on SEED-V.** Linear-probing Balanced Accuracy of the TTG encoder and resulting TGAP model under different grounding configurations. (**a**) Training Data for Grounding: from individual source domains (SEED, ZuCo, and TUH) to mixed-source subsets using 10%, 50%, and 100% of the full corpus. (**b**) Target Encoder Size: from Small to Medium and Large. (**c**) Training Epochs for Grounding: from 25 to 60 and 80 epochs. The Stage 2 configuration is fixed across all comparisons, and the dashed line denotes the EMA-based JEPA reference.

with Large, while the target encoder itself shows similar linear-probing performance across model sizes. This indicates that increasing the target encoder size beyond a moderate scale provides limited benefit for constructing the target space. In Figure 2(c), we vary the number of Stage 1 grounding epochs from 25 to 80. Over this range, the target encoder's Balanced Accuracy changes only from 29.33% to 30.04%, whereas that of the corresponding context encoder increases from 34.74% to 37.29%. Thus, continued grounding can improve the target space for subsequent predictive learning even after the target encoder's own linear separability has largely saturated. Overall, these results suggest that the quality of the target space is not determined by the target encoder's own downstream performance, but benefits more from exposure to diverse EEG observations.

3.5 Analysis of the Reconstruction-Specific Branch

We examine whether reconstruction-specific information should be separated from the topology target space using five variants. **(A)** Auxiliary tokens are added to the Stage 2 prediction targets, extending the prediction order from hemisphere $\rightarrow$ region to hemisphere $\rightarrow$ region $\rightarrow$ auxiliary. **(B)** Auxiliary tokens are fed into the Stage 1 target encoder, while Stage 2 continues to predict only topology representations. **(C)** Auxiliary tokens are included in both the Stage 1 target encoder and the Stage 2 prediction targets. **(D)** Auxiliary tokens and their separate Stage 1 pathway are removed, forcing reconstruction to rely solely on topology representations. **(E)** Topology-constrained tokenization is replaced with electrode-level tokens in both stages. Figure 3 summarizes the downstream and reconstruction results, with full details provided in Appendix E.2.

Figure 3(a–b) shows that variants A and C, both of which include auxiliary tokens as Stage 2 prediction targets, consistently reduce downstream performance. This suggests that reconstruction-specific information is not necessarily useful as a target for predictive learning. Variants B and D further clarify why this information should be handled separately. In B, auxiliary tokens enter the target encoder and interact with the topology representations during encoding, which also degrades Stage 2 performance. This suggests that reconstruction-specific information can interfere with the learned target space through such interactions. In D, removing the auxiliary branch likewise reduces downstream performance, which is consistent with the topology representations having to accommodate more reconstruction-specific information under the reconstruction objective. Together, these results support using a separate auxiliary pathway to absorb reconstruction-specific information while keeping

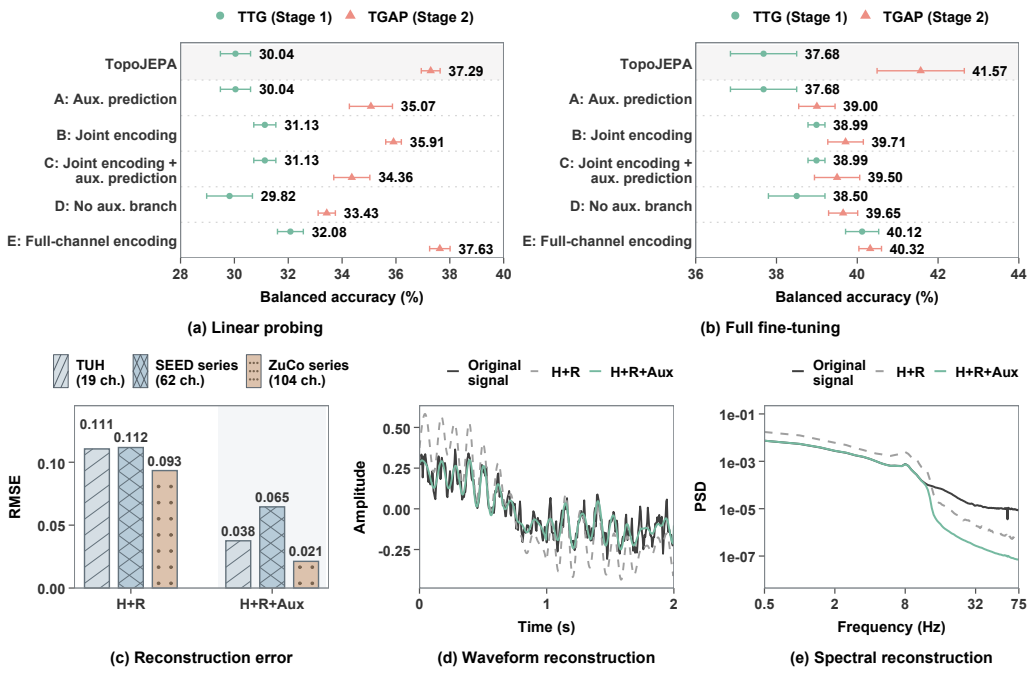


Figure 3: **Analysis of the Reconstruction-Specific Branch.** (**a–b**) SEED-V downstream performance after TTG and TGAP under linear probing and full fine-tuning. (**c**) Reconstruction error across the three pretraining montages. (**d–e**) Representative TUH waveform and power spectral density (PSD) reconstruction with and without auxiliary-token information. H and R denote hemisphere and region representations; Aux denotes auxiliary tokens. Waveform amplitudes use the model-input scale (microvolt values divided by 100).

the topology target space focused on the representations used for predictive learning. Variant E retains electrode-level tokens instead of the nine topology tokens and achieves comparable linear-probing performance but lower full-fine-tuning accuracy than TopoJEPA (40.32% vs. 41.57%). For 62-channel EEG, this increases the sequence length from 9 to 62 tokens per time step and raises the quadratic self-attention cost by approximately $47.46\times$, since $(62/9)^2 \approx 47.46$. The additional computation therefore brings no consistent downstream benefit.

The reconstruction results in Figure 3(c–e) further clarify the role of the auxiliary pathway. Retaining auxiliary tokens reduces RMSE by approximately 42–77% and improves waveform recovery, but also produces stronger high-frequency attenuation in the TUH example. We hypothesize that this reflects a reconstruction bias: while the auxiliary pathway helps recover fine-grained signal details, allowing the same bias to enter target encoding or prediction may reduce the model's ability to preserve task-relevant high-frequency information, contributing to the weaker transfer observed in variants. Taken together, these results suggest that reconstruction-specific information is useful for signal recovery, but is better handled through a separate pathway rather than incorporated into the topology representations or prediction targets.

3.6 Analysis of Spatiotemporal Attention Constraints

We examine whether TGAP benefits from explicitly constraining information flow across time and topology roles. Table 2 compares TopoJEPA with variants that remove all spatiotemporal attention constraints or individually remove the temporal constraint, hemisphere-to-region constraint, and same-role history constraint. All variants use the same input masking strategy. Variant definitions and the corresponding attention patterns are provided in Appendix E.3 and Figure 9.

As shown in Table 2, the temporal constraint has a clear impact on representation quality. Both w/o Spatiotemporal Constraints and w/o Temporal Constraint underperform TopoJEPA across the

Table 2: Ablation of spatiotemporal attention constraints.

	Full fine-tuning			*Linear probing*		
Strategy	Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Cohen's Kappa	Weighted F1
w/o Spatiotemporal Constraints	0.3925 ± 0.0008	0.2426 ± 0.0010	0.3979 ± 0.0009	0.3409 ± 0.0034	0.1752 ± 0.0075	0.3414 ± 0.0112
w/o Temporal Constraint	0.3943 ± 0.0036	0.2437 ± 0.0056	0.3992 ± 0.0051	0.3251 ± 0.0031	0.1544 ± 0.0053	0.3265 ± 0.0047
w/o Hemi.-to-Reg. Constraint	0.4009 ± 0.0047	0.2533 ± 0.0072	<u>0.4075 ± 0.0049</u>	**0.3768 ± 0.0040**	**0.2184 ± 0.0051**	**0.3822 ± 0.0030**
w/o Same-Role History Constraint	<u>0.4038 ± 0.0064</u>	<u>0.2540 ± 0.0085</u>	0.4073 ± 0.0080	0.3667 ± 0.0012	0.2063 ± 0.0024	0.3672 ± 0.0044
TopoJEPA	**0.4157 ± 0.0108**	**0.2724 ± 0.0150**	**0.4185 ± 0.0090**	<u>0.3729 ± 0.0035</u>	<u>0.2161 ± 0.0057</u>	<u>0.3795 ± 0.0063</u>

reported metrics, with the largest degradation appearing under linear probing. Allowing access to future context therefore does not improve transfer, even though it makes the latent prediction task easier. Instead, preserving the temporal-causal constraints of the prediction task is important for learning intrinsic EEG dynamics that remain useful for downstream tasks. Removing the same-role history constraint also weakens performance. Allowing a topology role to directly access its own previous representations creates an easier temporal path, whereas restricting this access encourages prediction to integrate information from other topology roles across time. Removing the hemisphere-to-region constraint slightly improves linear-probing performance, while TopoJEPA remains stronger under full fine-tuning. This suggests that the hemisphere-to-region structure contributes less to frozen representation separability but provides useful organization when the representations are adapted to downstream tasks.

Taken together, these ablations show that the Stage 2 attention constraints shape how temporal and hierarchical information is integrated, helping TGAP learn representations that are more useful for downstream transfer within the fixed topology-grounded target space.

4 Conclusion

TopoJEPA follows the principle of *Grounding Before Prediction*: it first establishes a stable, topology-grounded target space and then learns EEG spatiotemporal dynamics within this space through autoregressive latent prediction. Across diverse downstream tasks, TopoJEPA demonstrates strong linear-probing performance, suggesting that its learned representations preserve task-relevant information while requiring minimal adaptation. Overall, our results highlight the importance of grounding predictive learning in a structured target space, providing an effective foundation for learning transferable EEG representations that capture meaningful spatiotemporal organization. Further discussion and limitations are provided in Appendix F.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks with required disclosure; such tasks were either conducted without AI assistance or were not applicable. We used generative AI tools only for recommended-disclosure tasks, including improving manuscript readability and consistency and suggesting alternative wording and titles. All AI-assisted content was manually reviewed and revised by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses previously released EEG datasets and involves no new human-subject data collection. All datasets are used solely for research under their respective access conditions. As EEG may contain sensitive physiological and clinical information, particularly in clinical recordings, TopoJEPA is intended for research purposes only. The reported results should not be interpreted as supporting clinical diagnosis or deployment, and potential biases arising from the underlying datasets should be considered in future applications.

REPRODUCIBILITY

To support reproducibility, we provide the following:

- **Code:** Source code, training configurations, preprocessing scripts, and evaluation scripts are available at `https://anonymous.4open.science/r/TopoJEPA-F8B3/`.
- **Datasets:** All pretraining and downstream datasets are existing EEG benchmarks and are described in Appendices B and C. Raw data are not redistributed and should be obtained from their original sources.
- **Configurations:** Model architectures, optimization settings, preprocessing procedures, and downstream evaluation protocols are detailed in Appendices B–D.
- **Ablations:** Complete settings and results for the ablation studies are provided in Appendix E.

REFERENCES

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.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6(1):19, 2019. doi: 10.1038/s41597-019-0027-4.

Jian Cui, Zirui Lan, Olga Sourina, and Wolfgang Müller-Wittig. EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7921–7933, 2023.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023.

Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. EEG-Deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70682–70708, 2025.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 86048–86073, 2024.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, 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, pp. 22541–22577, 2025.

Josef Faller, Jennifer Cummings, Sameer Saproo, and Paul Sajda. Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. *Proceedings of the National Academy of Sciences*, 116(13):6482–6490, 2019.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. NeurIPT: Foundation model for neural interfaces. In *Advances in Neural Information Processing Systems*, volume 38, pp. 150819–150861, 2025.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. In *Proceedings of the 9th Graz Brain-Computer Interface Conference*, pp. 11–16, 2024.

Amirabbas Hojjati, Lu Li, Ibrahim Hameed, Anis Yazidi, Pedro G. Lind, and Rabindra Khadka. From video to EEG: Adapting joint embedding predictive architecture to uncover visual concepts in brain signal analysis. *arXiv preprint arXiv:2507.03633*, 2026. Version 5.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291, 2018. doi: 10.1038/sdata.2018.291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.18/.

Ping-Keng Jao, Ricardo Chavarriaga, and José del R. Millán. EEG-based online regulation of difficulty in simulated flying. *IEEE Transactions on Affective Computing*, 14(1):394–405, 2023.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

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, 2025a. doi: 10.1109/TAFFC.2024.3485057.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations*, 2025b.

Sander Koelstra, Christian Mühl, 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, 2012. doi: 10.1109/T-AFFC.2011.15.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *International Conference on Learning Representations*, 2026.

W. L. Lim, O. Sourina, and L. P. Wang. STEW: Simultaneous task EEG workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2106–2114, 2018. doi: 10.1109/tnsre.2018.2872924.

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, 2022.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ziling Lu, Zongsheng Li, Xinke Shen, Kexin Lou, Yingyue Xin, Xiaoqi Chen, Shinan Wang, Xiang Chen, Jiahao Fan, Chenyu Huang, Xin Xu, Zhoujie Hou, Chen Wei, and Quanying Liu. OmniEEG-Bench: A standardized evaluation benchmark for EEG foundation models. *arXiv preprint arXiv:2606.00815*, 2026.

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

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Saarang Panchavati, Uddhav Panchavati, Hiroki Nariai, Corey Arnold, and William Speier. Laya: A LeJEPA approach to EEG via latent prediction over reconstruction. *arXiv preprint arXiv:2603.16281*, 2026.

Roy Segal, Yoni Svechinsky, and Tomer Fekete. STST-JEPA: Shallow-target spatio-temporal joint embedding prediction architecture for EEG self-supervised learning. *arXiv preprint arXiv:2607.06629*, 2026.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Guagnyu 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*, volume 37, pp. 39249–39280, 2024.

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 *International Conference on Learning Representations*, 2025.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Hua Xie, Lifang He, and Yu Zhang. Pre-training graph contrastive masked autoencoders are strong distillers for EEG. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66358–66377. PMLR, 2025.

Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, 2024.

Wenchao Yang, Weidong Yan, Wenkang Liu, Yulan Ma, and Yang Li. THD-BAR: Topology hierarchical derived brain autoregressive modeling for EEG generic representations. In *Advances in Neural Information Processing Systems*, volume 38, pp. 150290–150314, 2025.

Mengqing Ye, C. L. Philip Chen, and Tong Zhang. Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19489–19500, 2025.

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

Tongtian Yue, Xuange Gao, Shuning Xue, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. BrainGPT: Unleashing the potential of EEG generalist foundation model by autoregressive pretraining. *arXiv preprint arXiv:2410.19779*, 2025. Version 2.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. *Journal of Neural Engineering*, 14(2):026017, 2017.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019.

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*, volume 38, pp. 87150–87195, 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. doi: 10.3390/data4010014.

A Related Work

A.1 General EEG Foundation Models

Foundation models are increasingly explored for EEG representation learning, with the goal of learning transferable representations from large-scale unlabeled recordings. Existing approaches broadly follow two pretraining paradigms: reconstruction-based pretraining and autoregressive pretraining.

Reconstruction-Based Pretraining. A major line of work learns EEG representations by recovering information hidden from the input. Within this paradigm, recent methods mainly differ in how they model spatial structure and accommodate heterogeneous recording configurations. CBraMod (Wang et al., 2025) and CSBrain (Zhou et al., 2025) reconstruct masked EEG while modeling spatiotemporal dependencies, whereas LUNA (Döner et al., 2025) and REVE (El Ouahidi et al., 2025) emphasize adaptation to varying electrode layouts through latent channel aggregation and coordinate-aware representations, respectively. NeurIPT (Fang et al., 2025) further revisits the masking strategy itself, introducing amplitude-aware masking to reduce reliance on local signal continuity. Other methods move the recovery target beyond the raw waveform. LaBraM (Jiang et al., 2024) predicts discrete neural codes produced by a learned tokenizer, while EEGPT (Wang et al., 2024) combines masked signal reconstruction with alignment to momentum-updated spatiotemporal representations. Despite these differences, these methods recover information within the same EEG segment and therefore do not explicitly model how neural representations evolve over time.

Autoregressive Pretraining. Another line of work models EEG through causal sequence prediction. BrainGPT (Yue et al., 2025) predicts future continuous EEG tokens from preceding observations, while NeuroLM (Jiang et al., 2025b) first maps EEG into discrete neural tokens and then models their multichannel autoregressive dependencies within a language-model framework. THD-BAR (Yang et al., 2025) further incorporates anatomical hierarchy by constructing multiscale discrete representations and predicting them in a next-scale-time order. Compared with masked recovery, autoregressive learning places temporal prediction at the center of pretraining. However, existing approaches either predict relatively low-level signal tokens directly or rely on discrete tokenizers to define the states being modeled. TopoJEPA instead first learns continuous, topology-aware hierarchical representations and then models how these representations evolve across successive EEG segments within a grounded latent space.

A.2 Topology-Aware EEG Foundation Models

Topology-aware EEG models exploit the spatial organization of electrodes and brain regions to improve representation learning across heterogeneous recording settings. One line of work focuses on preserving spatial correspondence across different montages. MMM (Yi et al., 2023) maps heterogeneous electrode layouts onto a unified topology and introduces hierarchical channel organization, while LUNA (Döner et al., 2025) and REVE (El Ouahidi et al., 2025) address montage variability through latent channel aggregation and coordinate-aware representations, respectively. NeurIPT (Fang et al., 2025) similarly incorporates physical electrode locations during pretraining and performs lobe-level aggregation during downstream adaptation. EEG-DisGCMAE (Wei et al., 2025) further transfers graph structure across electrode densities through topology distillation. Together, these approaches use topology to preserve spatial correspondence across recording configurations and electrode densities, but largely focus on structure within individual segments rather than how it evolves over time. THD-BAR (Yang et al., 2025) extends topology-aware EEG modeling to temporal dynamics through autoregressive prediction over discrete multiscale tokens. TopoJEPA similarly models the temporal evolution of hierarchical EEG representations, but shifts the prediction target from discrete tokens to continuous topology-grounded latent representations.

A.3 JEPA-Based EEG Foundation Models

Latent predictive learning replaces signal reconstruction with prediction in representation space. I-JEPA (Assran et al., 2023) predicts latent features of masked image regions from visible context, while V-JEPA (Bardes et al., 2024) extends the same principle to spatiotemporally masked video. This shifts the learning objective away from reconstructing low-level input details, but also makes the quality and stability of the target representation itself critical. Latent MIM (Wei et al., 2024) further shows that target–context coupling, patch correlations, and decoder design can strongly influence representation quality and collapse behavior. SALT (Li et al., 2026) takes a different approach by separating target-space construction from predictive learning, first learning a reconstruction-based target representation and then using it for latent prediction.

Recent EEG methods adapt latent predictive learning through different choices of context and masking. EEG2Rep (Mohammadi Foumani et al., 2024) predicts masked representations from informative temporal subsequences, S-JEPA (Guetschel et al., 2024) emphasizes spatial block prediction,

and EEG-VJEPA (Hojjati et al., 2026) extends spatiotemporal latent prediction to video-like EEG representations. These methods rely on momentum-updated target encoders, whereas Laya (Panchavati et al., 2026) replaces the momentum teacher with explicit geometric regularization in a shared-encoder framework. Brain-JEPA (Dong et al., 2024) explores a related direction in fMRI, combining functional-gradient positioning with spatiotemporal prediction over predefined brain regions.

Most of these approaches learn the predictive target space together with the predictor. TopoJEPA instead follows *Grounding Before Prediction*: it first establishes a topology-grounded hierarchical target space through reconstruction, and then learns how representations evolve within this space. By grounding continuous hemisphere- and region-level representations before predictive learning, TopoJEPA provides a structured latent space for modeling temporal dynamics across hierarchical EEG representations.

B Pretraining Details

B.1 Pretraining Datasets

Table 3 summarizes the retained pretraining data.

- **SEED** (Zheng & Lu, 2015): Fifteen participants completed three sessions of film-based emotion elicitation covering positive, neutral, and negative states. EEG was recorded from 62 channels at 1,000 Hz using an ESI NeuroScan system.
- **SEED-IV** (Zheng et al., 2019): Fifteen participants completed three sessions comprising 72 film clips that elicited happiness, sadness, fear, and neutrality. EEG was recorded from 62 channels at 1,000 Hz using the same NeuroScan system.
- **SEED-VII** (Jiang et al., 2025a): Twenty participants viewed 80 film clips across four sessions, covering six basic emotions and neutrality with continuous emotion-intensity annotations. EEG was recorded from 62 channels at 1,000 Hz using an ESI NeuroScan system.
- **ZuCo 1.0** (Hollenstein et al., 2018): Twelve native English speakers performed three naturalistic reading tasks involving movie reviews and Wikipedia sentences. EEG was recorded with a 128-channel Geodesic HydroCel system at 500 Hz.
- **ZuCo 2.0** (Hollenstein et al., 2020): Eighteen participants read 739 English sentences under normal-reading and task-specific semantic-relation conditions. EEG was likewise recorded with a 128-channel Geodesic HydroCel system at 500 Hz.
- **TUH** (Obeid & Picone, 2016): The TUH EEG Corpus contains clinical recordings collected during routine care with heterogeneous electrode configurations and sampling rates. We use an unlabeled subset retaining 3,000.4 hours after preprocessing, represented by 19 standard 10–20 scalp electrodes.

Table 3: Summary of the pretraining datasets.

Type	Dataset	Channels	Hours
Emotion	SEED	62	26.4
	SEED-IV	62	38.2
	SEED-VII	62	71.9
Reading	ZuCo 1.0	104	21.7
	ZuCo 2.0	104	20.7
Clinical	TUH subset	19	3,000.4

Preprocessing. We adapt the preprocessing protocol of CBraMod (Wang et al., 2025) to our multi-corpus setting. All EEG recordings are re-referenced using common-average referencing, bandpass filtered between 0.3 and 75 Hz, notch filtered at 50 Hz (60 Hz for TUH), and resampled to 200 Hz. For TUH, recordings shorter than five minutes or missing the required channels are excluded, and the first and last 60 s are removed. To limit the dominance of clinical EEG in the mixed corpus, we cap the retained TUH data at approximately 3,000 hours. For ZuCo, we use the released preprocessed

recordings, which retain 105 EEG channels after the exclusion of peripheral and EOG channels. Our montage-matching procedure further excludes Cz because its label is absent from the GSN-HydroCel-128 coordinate template, leaving 104 channels for pretraining.

Signals are segmented into non-overlapping 8-s windows. We discard windows containing non-finite values or amplitudes above 5,000 μV, as well as windows in which amplitudes above 100 μV account for more than 0.5% of all channel–time points or 5% of any individual channel. Channel identities and electrode coordinates are preserved for topology-aware modeling. Before model input, signals are scaled by 1/100 and divided into eight 1-s patches per channel.

B.2 Model and Training Configuration

Table 4: Model and training configuration for Stage 1 (TTG).

Parameter	**Setting**
Channel feature extraction	
Patch length (samples)	200
Temporal convolution channels	$1 \to 25 \to 25 \to 25$
Kernel sizes	49, 3, 3
Strides	25, 1, 1
Padding	24, 1, 1
Activation	GELU
Temporal projection	$200 \to 256$
Spectral projection	$101 \to 256$
Spectral dropout	0.1
Fusion projection	$512 \to 256$
Topology and Transformer modules	
Hemisphere / region queries	2 / 7
Auxiliary reconstruction queries	6
Target encoder width / depth	256 / 8
EEG Decoder width / depth	128 / 2
Reconstruction-specific branch width / depth	128 / 1
Attention heads	8
Feed-forward expansion ratio	4
Training	
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	10^{-4}
Global batch size	1,024
Training epochs	80
Learning-rate schedule	Cosine
Gradient clipping	1.0
Waveform loss weight	1
Spectral loss weight	0.01

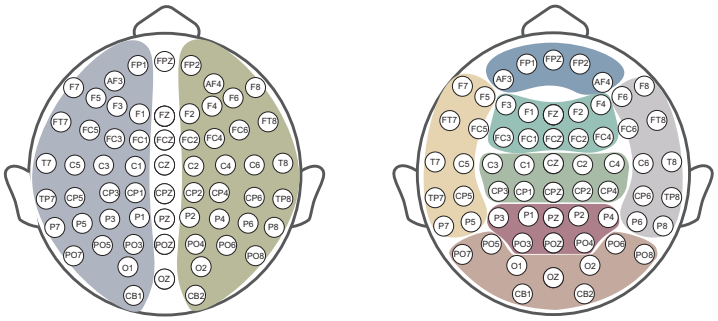
This section provides the model and training configurations for both pretraining stages, together with the topology construction used to align different electrode montages. Tables 4 and 5 summarize the configurations of TTG and TGAP, respectively.

To establish consistent spatial semantics across different electrode layouts, we use the 62-channel montage of the SEED series as the canonical topology, as illustrated in Figure 4. At the hemisphere level, electrodes are assigned to the left or right hemisphere according to their lateral coordinates, with a narrow midline band shared by both. At the regional level, electrodes are organized into seven anatomical regions following Ye et al. (2025): prefrontal, left lateral, frontocentral, right lateral, central, parietal, and occipital. Table 6 lists the resulting regional assignments for the three pretraining montages.

For other electrode layouts, such as the 128-channel ZuCo recordings and the 19-channel TUH recordings, each electrode is assigned to the nearest regional centroid in three-dimensional coordi-

Table 5: Model and training configuration for Stage 2 (TGAP).

Parameter	Setting
Channel feature extraction	
Patch length (samples)	200
Temporal convolution channels	$1 \to 25 \to 25 \to 25$
Kernel sizes	49, 3, 3
Strides	25, 1, 1
Padding	24, 1, 1
Activation	GELU
Temporal projection	$200 \to 256$
Spectral projection	$101 \to 256$
Spectral dropout	0.1
Fusion projection	$512 \to 384$
Topology and Transformer modules	
Hemisphere / region queries	2 / 7
Context encoder width / depth	384 / 10
Latent Predictor width / depth	192 / 2
Attention heads	8
Feed-forward expansion ratio	4
Training	
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	10^{-4}
Global batch size	1,024
Training epochs	80
Learning-rate schedule	Cosine
Gradient clipping	1.0
Prediction loss weight	1
SmoothL1 threshold	1



(a) Hemisphere-level electrode grouping (b) Region-level electrode grouping

Figure 4: Topology grouping of the canonical 62-channel SEED montage. (**a**) Hemisphere-level grouping and (**b**) seven-region grouping.

nate space. The centroids are computed from the corresponding regions in the canonical 62-channel layout. If no electrode is assigned to a region, the electrode closest to that region's centroid is used instead. The same topology assignment procedure is applied during both pretraining and downstream evaluation.

Table 6: Regional electrode assignments across pretraining montages.

Region	TUH (19)	SEED series (62)	ZuCo series (104)
R1: Prefrontal	FP1, FP2	FP1, FPZ, FP2, AF3, AF4	E2, E3, E9, E10, E15, E16, E18, E22, E23, E26, E27, E123
R2: Left lateral	F7, T7, P7	F7, F5, FT7, FC5, T7, C5, TP7, CP5, P7, P5	E33–E36, E38–E47, E50–E52, E57–E59, E64
R3: Frontocentral	F3, F4, FZ	F3, F1, FZ, F2, F4, FC3, FC1, FCZ, FC2, FC4	E4–E7, E11–E13, E19, E20, E24, E28–E30, E105, E106, E111, E112, E117, E118, E124
R4: Right lateral	F8, T8, P8	F6, F8, FC6, FT8, C6, T8, CP6, TP8, P6, P8	E92, E93, E95–E98, E100–E104, E108–E110, E114–E116, E120–E122
R5: Central	C3, C4, CZ	C3, C1, CZ, C2, C4, CP3, CP1, CPZ, CP2, CP4	E31, E37, E53–E55, E61, E78–E80, E86, E87
R6: Parietal	P3, P4, PZ	P3, P1, PZ, P2, P4, PO3, POZ, PO4	E60, E62, E66, E67, E71, E72, E76, E77, E84, E85, E91
R7: Occipital	O1, O2	PO7, PO5, PO6, PO8, CB1, O1, OZ, O2, CB2	E65, E69, E70, E74, E75, E82, E83, E89, E90

C Downstream Evaluation Details

C.1 Downstream Datasets

We evaluate transfer across ten EEG benchmarks spanning emotion recognition, motor imagery, cognitive workload, task difficulty, vigilance assessment, and clinical EEG classification. Table 7 summarizes their statistics, input durations, and prediction targets.

- **SEED-V** (Liu et al., 2022): SEED-V is reserved for downstream evaluation, while the other SEED variants are used during pretraining. It contains film-elicited EEG from five emotion categories: happiness, sadness, fear, disgust, and neutrality.
- **DEAP** (Koelstra et al., 2012): DEAP records EEG during music-video viewing together with self-reported affect ratings. We use the valence dimension for binary classification, assigning ratings of at least five to the high-valence class.
- **BCI Competition IV-2a** (Tangermann et al., 2012): This dataset contains cue-based motor imagery of the left hand, right hand, both feet, and tongue, forming a standard four-class motor-imagery benchmark.
- **EEGMat** (Zyma et al., 2019): EEGMat contains recordings from resting state and serial-subtraction tasks. We use the two conditions for binary cognitive-workload classification.
- **STEW** (Lim et al., 2018): STEW records EEG during multitasking together with subjective workload ratings. The ratings are grouped into three workload levels for classification.
- **Boundary-Avoidance Task (BAT)** (Faller et al., 2019): EEG is recorded while participants navigate a virtual aircraft through boundaries of varying difficulty. We use the recordings for binary task-difficulty classification.
- **Simulated Flying Task (SFT)** (Jao et al., 2023): Participants perform simulated flying under controlled changes in task difficulty. The resulting EEG is used for binary difficulty classification.
- **SADT** (Cao et al., 2019): The Sustained-Attention Driving Task records EEG during prolonged simulated driving with randomly induced lane departures. We use the recordings to distinguish alert and drowsy states.
- **SEED-VIG** (Zheng & Lu, 2017): SEED-VIG contains EEG from prolonged simulated driving with continuous vigilance annotations derived from eye-tracking-based PERCLOS. We use PERCLOS as the regression target.

- **ADFTD** (Miltiadous et al., 2023): ADFTD contains eyes-closed resting-state EEG from participants with Alzheimer's disease, frontotemporal dementia, and healthy controls. We use the three diagnostic groups for classification.

Table 7: Summary of downstream EEG datasets and task settings.

Task	Dataset	Subjects	Channels	Duration	Label
Emotion recognition	SEED-V	16	62	1 s	5-class
	DEAP valence	32	32	10 s	2-class
Motor imagery	BCIC-IV-2a	9	22	4 s	4-class
Cognitive workload	EEGMat	36	19	5 s	2-class
	STEW	45	14	4 s	3-class
Task difficulty recognition	BAT	18	64	1 s	2-class
	SFT	13	64	8 s	2-class
Vigilance assessment	SADT	11	30	3 s	2-class
	SEED-VIG	21	17	8 s	regression
Clinical EEG classification	ADFTD	88	19	10 s	3-class

C.2 Comparison Models

We compare TopoJEPA with five supervised EEG architectures and five pretrained EEG foundation models. Official implementations and pretrained checkpoints are used whenever available. Parameter counts exclude task-specific heads and modules used only during pretraining.

Supervised models.

- **EEGNet** (Lawhern et al., 2018): A compact CNN that combines temporal convolution, depthwise spatial filtering, and separable convolution for EEG decoding.
- **EEGConformer** (Song et al., 2023): A convolutional Transformer that first extracts local temporal and spatial features and then models their global dependencies with self-attention.
- **TSception** (Ding et al., 2023): A multiscale CNN that captures temporal and frequency patterns together with global and hemispheric spatial asymmetry.
- **LGGNet** (Ding et al., 2024): A graph neural network that combines multiscale temporal features with local and global graph modeling across functional brain regions.
- **EEG-Deformer** (Ding et al., 2025): A convolutional Transformer that models coarse-to-fine temporal patterns through a hierarchical Transformer and refines multilevel temporal information through dense information purification.

Pretrained models.

- **LaBraM** (Jiang et al., 2024): A foundation model pretrained by predicting masked discrete neural codes produced by a learned EEG tokenizer.
- **CBraMod** (Wang et al., 2025): A masked-reconstruction foundation model that separates spatial and temporal dependencies through criss-cross attention.
- **EEGPT** (Wang et al., 2024): A pretrained Transformer that combines masked EEG reconstruction with spatiotemporal representation alignment.
- **LUNA** (Döner et al., 2025): A topology-agnostic foundation model that maps variable electrode configurations into a fixed latent representation through learned queries.
- **CSBrain** (Zhou et al., 2025): A foundation model that captures cross-scale temporal and regional dependencies through multiscale tokenization and structured sparse attention.

C.3 Downstream Evaluation.

For downstream evaluation, we retain the pretrained context pathway, disable context masking, and discard the Latent Predictor. For each sample, context representations are first averaged over time

within each topology role. The resulting two hemisphere representations and seven region representations, each 384-dimensional, are then concatenated into a 3,456-dimensional feature vector. We evaluate TopoJEPA under two settings. Linear probing freezes the encoder and trains only a linear task head, while full fine-tuning jointly optimizes the encoder and task head. The corresponding training settings are summarized in Table 8.

We compare TopoJEPA with both supervised EEG models trained from scratch and pretrained models initialized from their released checkpoints. Unless otherwise noted, all reported results are obtained from our own reproductions under the same dataset partitions and task metrics. Each experiment is independently repeated with three random seeds, and test performance is reported as the mean and standard deviation across the three runs. Classification checkpoints are selected by validation Balanced Accuracy, while regression checkpoints are selected by validation MSE. Dataset-specific splits and evaluation protocols are detailed in Appendix D.

Table 8: Downstream adaptation settings for TopoJEPA.

Parameter	**Full fine-tuning**	**Linear probing**
Encoder	Trainable	Frozen
Prediction head	MLP	Linear
Hidden dimensions	(800, 200)	–
Activation	ELU	–
Head dropout	0.1	–
Optimizer	AdamW	AdamW
Encoder learning rate	10^{-4}	–
Head learning rate	5×10^{-4}	10^{-3}
Weight decay	5×10^{-4}	5×10^{-4}
Batch size	64	64
Training epochs	40	50
Learning-rate schedule	Cosine	Cosine
Gradient clipping	1.0	1.0

D Additional Downstream Results

D.1 Emotion Recognition

We evaluate emotion recognition on SEED-V and DEAP, two widely used EEG benchmarks with distinct recording settings and electrode configurations. For SEED-V, the dataset contains 62-channel EEG recordings from 16 participants, each completing three experimental sessions (Liu et al., 2022). Each session consists of 15 movie-clip trials covering five emotion categories, with three trials for each category. We segment the recordings into 1-s windows for five-class emotion recognition. For subject-dependent evaluation, following Wang et al. (2025), the 15 trials within each subject-session are divided into three consecutive groups: trials 1–5 for training, 6–10 for validation, and 11–15 for testing. For DEAP, the dataset contains recordings from 32 participants, each completing 40 trials while watching one-minute music-video clips (Koelstra et al., 2012). EEG is recorded from 32 channels, and each trial is accompanied by self-reported affective ratings on a nine-point scale. We use 10-s segments for binary valence classification and assign trials with valence ratings of at least five to the positive class. For each participant, the 40 trials are divided into 24 training, 8 validation, and 8 test trials. All segments from the same trial are kept in the same partition to avoid overlap across splits. Table 1 reports the complete results. Published SEED-V results marked with † are taken from Wang et al. (2025); Döner et al. (2025); Zhou et al. (2025).

D.2 Motor Imagery

We evaluate motor imagery decoding on BCI Competition IV-2a, a widely used four-class EEG benchmark (Tangermann et al., 2012). The dataset contains recordings from nine participants, each completing two sessions on different days. Each session consists of six runs with 48 trials per run, yielding 288 trials covering four motor imagery classes: left hand, right hand, both feet, and tongue. EEG is recorded from 22 channels at 250 Hz. We extract 4-s segments corresponding to the

motor imagery period for classification. We adopt a subject-dependent cross-session protocol while preserving the original training and evaluation sessions. For each participant, the six runs in session T are divided into runs 1–5 for training and run 6 for validation, corresponding to 240 and 48 trials, respectively. The complete session E, containing 288 trials, is reserved for testing.

Table 9 reports the complete results. Under full fine-tuning, CBraMod achieves the highest overall performance, with TopoJEPA remaining competitive across all three metrics. Under linear probing, TopoJEPA achieves the highest Balanced Accuracy and Cohen's Kappa, demonstrating the quality of the learned motor-imagery representations, while CSBrain obtains the highest Weighted F1.

Table 9: Performance comparison on motor imagery classification using BCI Competition IV-2a.

Model	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised learning			
EEGNet (Lawhern et al., 2018)	0.4824 ± 0.0111	0.3098 ± 0.0148	0.4790 ± 0.0124
EEGConformer (Song et al., 2023)	0.6115 ± 0.0106	0.4820 ± 0.0141	0.6086 ± 0.0104
TSception (Ding et al., 2023)	0.4630 ± 0.0099	0.2840 ± 0.0132	0.4225 ± 0.0283
LGGNet (Ding et al., 2024)	0.5707 ± 0.0106	0.4276 ± 0.0141	0.5702 ± 0.0107
Full fine-tuning			
LaBraM-Base (Jiang et al., 2024)	0.6164 ± 0.0154	0.4885 ± 0.0206	0.6163 ± 0.0157
CBraMod (Wang et al., 2025)	**0.6310 ± 0.0031**	**0.5081 ± 0.0041**	**0.6310 ± 0.0030**
EEGPT-Large (Wang et al., 2024)	0.3663 ± 0.0118	0.1550 ± 0.0158	0.3650 ± 0.0119
LUNA-Base (Döner et al., 2025)	0.4569 ± 0.0135	0.2759 ± 0.0180	0.4532 ± 0.0159
CSBrain (Zhou et al., 2025)	0.6098 ± 0.0189	0.4798 ± 0.0252	0.6096 ± 0.0190
TopoJEPA	<u>0.6304 ± 0.0136</u>	<u>0.5072 ± 0.0181</u>	<u>0.6303 ± 0.0138</u>
Linear probing			
LaBraM-Base (Jiang et al., 2024)	0.3282 ± 0.0002	0.1043 ± 0.0003	0.3060 ± 0.0042
CBraMod (Wang et al., 2025)	0.4579 ± 0.0062	0.2705 ± 0.0083	0.4554 ± 0.0086
EEGPT-Large (Wang et al., 2024)	0.4113 ± 0.0139	0.2150 ± 0.0185	0.4099 ± 0.0159
LUNA-Base (Döner et al., 2025)	0.4316 ± 0.0015	0.2421 ± 0.0019	0.4305 ± 0.0026
CSBrain (Zhou et al., 2025)	<u>0.4760 ± 0.0022</u>	<u>0.3013 ± 0.0029</u>	**0.4762 ± 0.0029**
TopoJEPA	**0.4806 ± 0.0172**	**0.3074 ± 0.0229**	<u>0.4736 ± 0.0206</u>

D.3 Cognitive Workload

We evaluate cognitive workload decoding on EEGMat and STEW, which capture workload induced by mental arithmetic and multitasking, respectively. For EEGMat, the dataset contains recordings from 36 participants under two conditions: an eyes-closed resting state and a mental arithmetic task based on serial subtraction (Zyma et al., 2019). EEG was recorded from 19 scalp locations at 500 Hz, with 180 s of resting-state EEG and 60 s of arithmetic-task EEG retained for each participant. We segment the recordings into 5-s windows and perform binary classification between the resting and arithmetic conditions. For subject-independent evaluation, the 36 participants are divided into 28 for training, 4 for validation, and 4 for testing. For STEW, the dataset was collected during a SIMKAP multitasking experiment, in which participants simultaneously perform visual matching, arithmetic, and information-retrieval tasks (Lim et al., 2018). EEG was recorded from 14 channels at 128 Hz, and participants rated their perceived workload on a nine-point scale after the task. We retain the 45 participants with available workload ratings and group the ratings into low (1–3), medium (4–6), and high (7–9) workload levels. The EEG recordings are segmented into 4-s windows for three-class classification. We use a subject-independent split with 32 participants for training, 4 for validation, and 9 for testing, with no participant shared across partitions.

Published results marked with † use closely related evaluation protocols and are included for reference. Table 10 reports the complete results. On EEGMat, TopoJEPA achieves the highest Balanced Accuracy and AUROC under linear probing, as well as the highest Weighted F1 under full fine-tuning. On STEW, TopoJEPA achieves the highest Balanced Accuracy and Cohen's Kappa under full fine-tuning and the highest Weighted F1 under linear probing.

Table 10: Performance comparison on cognitive workload classification using EEGMat and STEW. See Appendix C.2 for model references.

	EEGMat			STEW		
Model	Balanced Acc.	Weighted F1	AUROC	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised learning						
EEGNet	0.5023 ± 0.0040	0.6460 ± 0.0054	0.6236 ± 0.0219	0.3619 ± 0.0093	0.0501 ± 0.0132	0.3609 ± 0.0121
EEGConformer	0.6863 ± 0.0896	0.7606 ± 0.0489	0.7788 ± 0.0715	0.4962 ± 0.0232	0.2887 ± 0.0301	0.5072 ± 0.0186
TSception	0.6331 ± 0.1162	0.7188 ± 0.0264	0.8156 ± 0.0674	0.4319 ± 0.0294	0.1848 ± 0.0508	0.4405 ± 0.0324
LGGNet	0.5718 ± 0.0122	0.7165 ± 0.0112	0.7089 ± 0.0074	0.4785 ± 0.0107	0.2615 ± 0.0157	0.5109 ± 0.0102
EEG-Deformer	0.5498 ± 0.0414	0.2503 ± 0.1271	0.7718 ± 0.0438	0.3540 ± 0.0167	0.0466 ± 0.0371	0.3181 ± 0.0352
Full fine-tuning						
LaBraM-Base	0.6909 ± 0.0125†	–	0.7721 ± 0.0093†	0.4942 ± 0.0145†	0.2530 ± 0.0107†	0.4677 ± 0.0173†
CBraMod	0.7256 ± 0.0132†	–	0.7905 ± 0.0073†	0.4830 ± 0.0344†	0.2498 ± 0.0550†	0.4909 ± 0.0315†
EEGPT-Large	0.5637 ± 0.0145	0.6909 ± 0.0068	0.6068 ± 0.0336	0.5230 ± 0.0118	0.3242 ± 0.0165	**0.5563 ± 0.0012**
LUNA-Base	0.5000 ± 0.0000	0.6429 ± 0.0000	0.6834 ± 0.0115	0.4408 ± 0.0963	0.1899 ± 0.1687	0.4226 ± 0.1293
CSBrain	**0.7558 ± 0.0106**†	–	**0.8478 ± 0.0297**†	0.5286 ± 0.0229	0.3281 ± 0.0281	0.5543 ± 0.0185
TopoJEPA	0.7257 ± 0.0151	**0.8177 ± 0.0124**	0.8244 ± 0.0012	**0.5301 ± 0.0222**	**0.3338 ± 0.0280**	0.5559 ± 0.0082
Linear probing						
LaBraM-Base	0.5289 ± 0.0256	0.6745 ± 0.0278	0.7127 ± 0.0303	0.4767 ± 0.0262	0.2450 ± 0.0406	0.4783 ± 0.0286
CBraMod	0.6458 ± 0.0511	**0.7784 ± 0.0352**	0.8057 ± 0.0022	0.4596 ± 0.0285	0.2201 ± 0.0366	0.4752 ± 0.0151
EEGPT-Large	0.5683 ± 0.0612	0.6840 ± 0.0600	0.6247 ± 0.0274	**0.5154 ± 0.0429**	**0.3119 ± 0.0604**	0.5361 ± 0.0348
LUNA-Base	0.5822 ± 0.0270	0.6260 ± 0.0802	0.6341 ± 0.0335	0.4436 ± 0.0103	0.1882 ± 0.0077	0.4678 ± 0.0036
CSBrain	0.5845 ± 0.0251	0.6938 ± 0.0143	0.6486 ± 0.0210	0.4895 ± 0.0449	0.2664 ± 0.0624	0.4975 ± 0.0591
TopoJEPA	**0.7199 ± 0.0122**	0.6865 ± 0.0581	**0.8424 ± 0.0115**	0.5097 ± 0.0306	0.3047 ± 0.0415	**0.5402 ± 0.0226**

† Published results under the evaluation protocols of the original studies.

D.4 Task Difficulty Classification

We evaluate task-difficulty decoding on two EEG benchmarks based on simulated flight control, the Boundary-Avoidance Task (BAT) and the Simulated Flying Task (SFT), which induce difficulty through different visuomotor demands. For BAT, we use recordings from 18 participants performing a virtual-reality aerial navigation task with 64-channel EEG (Faller et al., 2019). Participants repeatedly navigate a simulated aircraft through rectangular boundaries under easy and hard course conditions, with each flight terminating when a boundary is missed. We segment the recordings into 1-s windows for binary difficulty classification. For subject-dependent evaluation, complete flights from each participant are divided into training, validation, and test sets using an approximately 60/20/20% split, with all windows from the same flight assigned to a single partition. For SFT, the dataset contains 64-channel EEG recordings from 13 participants performing a simulated flying task (Jao et al., 2023). Each participant completes 32 trajectories, each comprising 32 waypoints, while controlling a simulated aircraft with a joystick. Task difficulty is manipulated through waypoint size, with smaller waypoints requiring more precise control. We use 8-s segments for binary difficulty classification. For each participant, the 32 trajectories are divided into 20 for training, 6 for validation, and 6 for testing, with all windows from the same trajectory kept in the same partition.

Table 11 reports the complete results. TopoJEPA achieves the strongest performance across all three metrics under both full fine-tuning and linear probing on BAT and SFT.

D.5 Vigilance Assessment

We evaluate vigilance assessment on SADT and SEED-VIG, covering both discrete drowsiness recognition and continuous vigilance estimation. For SADT, the original dataset was collected during a sustained-attention driving experiment in a virtual-reality simulator (Cao et al., 2019). We follow the processed version used by Cui et al. (2023), which contains 30-channel EEG from 11 participants. EEG is segmented into 3-s epochs preceding lane-departure events and labeled as alert or drowsy according to the corresponding reaction-time measures. For within-subject evaluation, samples from each participant are divided into training, validation, and test sets using a class-stratified 60/20/20% split. For SEED-VIG, participants perform an approximately two-hour monotonous simulated-driving task designed to induce natural fluctuations in vigilance (Zheng &

Table 11: Performance comparison on task-difficulty classification using BAT and SFT. See Appendix C.2 for model references.

	BAT			SFT		
Model	Balanced Acc.	Weighted F1	AUROC	Balanced Acc.	Weighted F1	AUROC
Supervised learning						
EEGNet	0.5512 ± 0.0141	0.5499 ± 0.0129	0.5869 ± 0.0128	0.6039 ± 0.0189	0.5953 ± 0.0283	0.6515 ± 0.0193
EEGConformer	0.5334 ± 0.0140	0.5334 ± 0.0140	0.5532 ± 0.0221	0.5657 ± 0.0336	0.5643 ± 0.0325	0.5910 ± 0.0457
TSception	0.5816 ± 0.0023	0.5853 ± 0.0018	0.5995 ± 0.0069	0.7100 ± 0.0124	0.7073 ± 0.0124	0.7885 ± 0.0075
LGGNet	0.6241 ± 0.0105	0.6184 ± 0.0164	0.6846 ± 0.0067	0.7127 ± 0.0113	0.7119 ± 0.0111	0.7745 ± 0.0090
EEG-Deformer	0.6004 ± 0.0420	0.5759 ± 0.0575	0.6848 ± 0.0606	0.5134 ± 0.0105	0.3628 ± 0.0231	0.7701 ± 0.0350
Full fine-tuning						
LaBraM-Base	0.6377 ± 0.0073	0.6325 ± 0.0062	0.6991 ± 0.0123	0.7320 ± 0.0162	0.7287 ± 0.0185	0.8171 ± 0.0187
CBraMod	0.6670 ± 0.0199	0.6571 ± 0.0269	0.7551 ± 0.0075	0.7152 ± 0.0102	0.7112 ± 0.0122	0.7826 ± 0.0041
EEGPT-Large	0.5690 ± 0.0158	0.5687 ± 0.0161	0.5993 ± 0.0168	0.7052 ± 0.0093	0.6973 ± 0.0096	0.7798 ± 0.0120
LUNA-Base	0.6515 ± 0.0087	0.6392 ± 0.0083	0.7214 ± 0.0212	0.7219 ± 0.0129	0.7201 ± 0.0149	0.8112 ± 0.0086
CSBrain	0.6233 ± 0.0025	0.6231 ± 0.0021	0.6726 ± 0.0135	0.6721 ± 0.0070	0.6701 ± 0.0064	0.7308 ± 0.0054
TopoJEPA	**0.6830 ± 0.0055**	**0.6766 ± 0.0040**	**0.7577 ± 0.0211**	**0.7709 ± 0.0059**	**0.7683 ± 0.0059**	**0.8554 ± 0.0085**
Linear probing						
LaBraM-Base	0.5710 ± 0.0042	0.5692 ± 0.0060	0.6052 ± 0.0037	0.6399 ± 0.0048	0.6297 ± 0.0059	0.6982 ± 0.0023
CBraMod	0.6098 ± 0.0062	0.6092 ± 0.0057	0.6683 ± 0.0072	0.6834 ± 0.0050	0.6820 ± 0.0047	0.7557 ± 0.0016
EEGPT-Large	0.5283 ± 0.0219	0.4828 ± 0.0754	0.5587 ± 0.0130	0.5976 ± 0.0306	0.5955 ± 0.0306	0.6411 ± 0.0531
LUNA-Base	0.5575 ± 0.0045	0.5558 ± 0.0032	0.5728 ± 0.0023	0.5942 ± 0.0120	0.5887 ± 0.0113	0.6366 ± 0.0090
CSBrain	0.5772 ± 0.0031	0.5718 ± 0.0053	0.6070 ± 0.0065	0.5900 ± 0.0081	0.5863 ± 0.0107	0.6444 ± 0.0043
TopoJEPA	**0.6632 ± 0.0045**	**0.6604 ± 0.0058**	**0.7299 ± 0.0009**	**0.7383 ± 0.0032**	**0.7341 ± 0.0040**	**0.8261 ± 0.0003**

Lu, 2017). The dataset provides 17-channel EEG together with continuous PERCLOS scores derived from eye-tracking measurements, which quantify the proportion of eye closure and serve as vigilance labels. We use 8-s EEG segments for continuous vigilance estimation. For each recording, the sequence is divided chronologically into five equal blocks, with the first three used for training, the fourth for validation, and the fifth for testing, thereby preserving the temporal order of the recording.

Table 12 reports the complete results. On SADT, TopoJEPA achieves the strongest performance across all three metrics under both full fine-tuning and linear probing. On SEED-VIG, LaBraM performs best under full fine-tuning, while TopoJEPA achieves the best RMSE, MAE, and COR under linear probing.

D.6 Clinical EEG Classification

We evaluate clinical EEG classification on ADFTD, a resting-state EEG benchmark for distinguishing Alzheimer's disease, frontotemporal dementia, and healthy controls. The dataset contains eyes-closed resting-state EEG recordings from 88 participants, including 36 patients with Alzheimer's disease, 23 with frontotemporal dementia, and 29 healthy controls (Miltiadous et al., 2023). EEG was recorded from 19 scalp channels at 500 Hz. We segment each recording into 10-s windows for three-class classification and retain at most 40 segments per participant to limit subject-level imbalance. Following the subject-level evaluation protocol of Lu et al. (2026), we construct three stratified splits, each containing 70 participants for training, 9 for validation, and 9 for testing, with no participant overlap across partitions. Each split is evaluated with three training seeds. Results are first averaged across the three runs within each split and then summarized across the three splits.

Table 13 reports the complete results. Under full fine-tuning, TopoJEPA achieves the highest Balanced Accuracy, while LaBraM performs best under linear probing.

Table 12: Performance comparison on vigilance assessment using SADT and SEED-VIG. SADT is evaluated as classification and SEED-VIG as regression. See Appendix C.2 for model references.

	SADT			SEED-VIG		
Model	Balanced Acc.	Weighted F1	AUROC	RMSE	MAE	COR
Supervised learning						
EEGNet	0.5533 ± 0.0111	0.4570 ± 0.0276	0.7705 ± 0.0193	0.2148 ± 0.0021	0.1661 ± 0.0025	0.6737 ± 0.0141
EEGConformer	0.8452 ± 0.0127	0.8451 ± 0.0127	0.9119 ± 0.0073	0.1850 ± 0.0046	0.1354 ± 0.0044	0.8071 ± 0.0266
TSception	0.7318 ± 0.0878	0.7151 ± 0.1129	0.8570 ± 0.0168	0.2435 ± 0.0681	0.1903 ± 0.0496	0.4822 ± 0.4413
LGGNet	0.7750 ± 0.0103	0.7743 ± 0.0105	0.8428 ± 0.0124	0.1934 ± 0.0050	0.1514 ± 0.0044	0.7367 ± 0.0139
EEG-Deformer	0.5406 ± 0.0159	0.4248 ± 0.0343	0.8442 ± 0.0236	0.2749 ± 0.0097	0.2206 ± 0.0105	0.3415 ± 0.3450
Full fine-tuning						
LaBraM-Base	0.8536 ± 0.0165	0.8535 ± 0.0165	0.8945 ± 0.0104	**0.1371 ± 0.0024**	**0.0984 ± 0.0041**	**0.8748 ± 0.0052**
CBraMod	0.7944 ± 0.0134	0.7940 ± 0.0132	0.8760 ± 0.0050	0.1448 ± 0.0011	0.1022 ± 0.0010	0.8605 ± 0.0020
EEGPT-Large	0.7589 ± 0.0379	0.7583 ± 0.0379	0.8345 ± 0.0415	0.1783 ± 0.0053	0.1414 ± 0.0040	0.7782 ± 0.0136
LUNA-Base	0.7589 ± 0.0102	0.7587 ± 0.0104	0.8270 ± 0.0088	0.1763 ± 0.0019	0.1386 ± 0.0008	0.8072 ± 0.0069
CSBrain	0.7733 ± 0.0230	0.7715 ± 0.0244	0.8477 ± 0.0335	0.1551 ± 0.0039	0.1123 ± 0.0051	0.8417 ± 0.0060
TopoJEPA	**0.8934 ± 0.0141**	**0.8933 ± 0.0142**	**0.9543 ± 0.0029**	0.1462 ± 0.0004	0.1061 ± 0.0002	0.8565 ± 0.0006
Linear probing						
LaBraM-Base	0.7521 ± 0.0103	0.7515 ± 0.0108	0.8106 ± 0.0036	0.2053 ± 0.0011	0.1614 ± 0.0013	0.6916 ± 0.0046
CBraMod	0.8460 ± 0.0053	0.8459 ± 0.0054	0.8983 ± 0.0020	0.2035 ± 0.0028	0.1618 ± 0.0035	0.7212 ± 0.0081
EEGPT-Large	0.7978 ± 0.0327	0.7976 ± 0.0328	0.8701 ± 0.0274	0.1771 ± 0.0039	0.1376 ± 0.0027	0.7794 ± 0.0117
LUNA-Base	0.5931 ± 0.0140	0.5857 ± 0.0181	0.6475 ± 0.0162	0.2432 ± 0.0003	0.1942 ± 0.0011	0.5315 ± 0.0052
CSBrain	0.7411 ± 0.0025	0.7409 ± 0.0026	0.8077 ± 0.0089	0.2620 ± 0.0005	0.2086 ± 0.0007	0.5718 ± 0.0035
TopoJEPA	**0.8655 ± 0.0000**	**0.8654 ± 0.0001**	**0.9374 ± 0.0007**	**0.1732 ± 0.0017**	**0.1361 ± 0.0021**	**0.7965 ± 0.0039**

Table 13: Performance comparison on clinical EEG classification using ADFTD.

Model	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised learning			
EEGNet (Lawhern et al., 2018)	0.5312 ± 0.0475	0.3542 ± 0.0970	0.5595 ± 0.0482
EEGConformer (Song et al., 2023)	0.5359 ± 0.0424	0.3640 ± 0.0639	0.5998 ± 0.0359
TSception (Ding et al., 2023)	0.5590 ± 0.0529	0.3721 ± 0.0822	0.5695 ± 0.0705
LGGNet (Ding et al., 2024)	0.5531 ± 0.0405	0.3725 ± 0.0453	0.5936 ± 0.0315
EEG-Deformer (Ding et al., 2025)	0.4713 ± 0.0857	0.2159 ± 0.1033	0.4662 ± 0.0767
Full fine-tuning			
LaBraM-Base (Jiang et al., 2024)	0.6005 ± 0.0202	**0.4592 ± 0.0428**	**0.6436 ± 0.0191**
CBraMod (Wang et al., 2025)	0.5602 ± 0.0858	0.3937 ± 0.1319	0.6057 ± 0.0833
EEGPT-Large (Wang et al., 2024)	0.5727 ± 0.0653	0.4346 ± 0.1397	0.6051 ± 0.0665
LUNA-Base (Döner et al., 2025)	0.5640 ± 0.0326	0.3658 ± 0.0603	0.5686 ± 0.0606
CSBrain (Zhou et al., 2025)	0.5408 ± 0.0596	0.3399 ± 0.0804	0.5721 ± 0.0536
TopoJEPA	**0.6038 ± 0.0340**	0.4537 ± 0.0754	0.6406 ± 0.0408
Linear probing			
LaBraM-Base (Jiang et al., 2024)	**0.5893 ± 0.0764**	**0.4311 ± 0.1166**	**0.6298 ± 0.0759**
CBraMod (Wang et al., 2025)	0.5417 ± 0.0633	0.4123 ± 0.0809	0.5648 ± 0.0497
EEGPT-Large (Wang et al., 2024)	0.5499 ± 0.0453	0.3757 ± 0.0851	0.5827 ± 0.0570
LUNA-Base (Döner et al., 2025)	0.5465 ± 0.0632	0.3781 ± 0.0922	0.5727 ± 0.0760
CSBrain (Zhou et al., 2025)	0.5387 ± 0.0678	0.3522 ± 0.0982	0.5817 ± 0.0593
TopoJEPA	0.5708 ± 0.0037	0.3825 ± 0.0026	0.6070 ± 0.0021

E Additional Ablation Studies and Analysis

E.1 Details of Target-Space Grounding Analysis

This section provides the experimental details and complete results for the target-space grounding analysis in Section 3.4. We examine three factors in TTG: grounding data, target-encoder capacity, and grounding duration. For each configuration, we first evaluate the frozen TTG target encoder by linear probing and then use it as the fixed target pathway for TGAP. Stage 2 always uses the full pretraining corpus and the default configuration, isolating the effect of the target space established during Stage 1. We additionally include EMA-JEPA as an evolving-target reference. Its target encoder is initialized from the online encoder and updated by EMA with momentum increasing from 0.996 to 1.0, while the online pathway predicts its representations from inputs masked at a ratio of 0.5. Unlike TopoJEPA, EMA-JEPA has no separate grounding stage, so its prediction targets evolve throughout pretraining. It is trained for 80 epochs with AdamW using a learning rate of 5×10^{-4} and weight decay of 0.05.

For the grounding-data study, we vary the data source and corpus size while fixing the Medium target encoder and 80 grounding epochs. The capacity study compares Small, Medium, and Large target encoders under the same 80-epoch schedule, while the duration study fixes the Medium encoder and varies grounding over 25, 60, and 80 epochs. All models are evaluated on SEED-V using the same linear-probing protocol. Table 14 reports the complete results.

Table 14: Full results for the target-space grounding analysis.

	Setting	Stage	Balanced Acc.	Cohen's Kappa	Weighted F1
Grounding data	SEED only	TTG	0.2651 ± 0.0046	0.0796 ± 0.0055	0.2469 ± 0.0077
		TGAP	0.3182 ± 0.0033	0.1452 ± 0.0065	0.3187 ± 0.0083
	ZuCo only	TTG	0.2018 ± 0.0010	0.0029 ± 0.0019	0.1154 ± 0.0128
		TGAP	0.2045 ± 0.0067	0.0063 ± 0.0111	0.1445 ± 0.0393
	TUH only	TTG	0.2963 ± 0.0073	0.1191 ± 0.0132	0.2823 ± 0.0309
		TGAP	0.3086 ± 0.0084	0.1335 ± 0.0132	0.3047 ± 0.0180
	Mixed 10%	TTG	0.2765 ± 0.0045	0.0952 ± 0.0075	0.2660 ± 0.0173
		TGAP	0.3412 ± 0.0033	0.1755 ± 0.0058	0.3436 ± 0.0076
	Mixed 50%	TTG	$\mathbf{0.3035 \pm 0.0050}$	$\mathbf{0.1291 \pm 0.0071}$	$\mathbf{0.3066 \pm 0.0030}$
		TGAP	0.3570 ± 0.0037	0.1957 ± 0.0060	0.3581 ± 0.0092
	Mixed 100%	TTG	0.3004 ± 0.0056	0.1223 ± 0.0045	0.3019 ± 0.0042
		TGAP	$\mathbf{0.3729 \pm 0.0035}$	$\mathbf{0.2161 \pm 0.0057}$	$\mathbf{0.3795 \pm 0.0063}$
Encoder capacity	Small	TTG	0.2928 ± 0.0035	0.1130 ± 0.0053	0.2867 ± 0.0042
		TGAP	0.3542 ± 0.0095	0.1903 ± 0.0163	0.3533 ± 0.0178
	Medium	TTG	$\mathbf{0.3004 \pm 0.0056}$	$\mathbf{0.1223 \pm 0.0045}$	$\mathbf{0.3019 \pm 0.0042}$
		TGAP	0.3729 ± 0.0035	$\mathbf{0.2161 \pm 0.0057}$	$\mathbf{0.3795 \pm 0.0063}$
	Large	TTG	0.2973 ± 0.0020	0.1171 ± 0.0032	0.2813 ± 0.0076
		TGAP	$\mathbf{0.3740 \pm 0.0048}$	0.2134 ± 0.0071	0.3730 ± 0.0091
Grounding duration	25 epochs	TTG	0.2933 ± 0.0046	0.1169 ± 0.0059	0.2907 ± 0.0058
		TGAP	0.3474 ± 0.0027	0.1824 ± 0.0042	0.3501 ± 0.0052
	60 epochs	TTG	$\mathbf{0.3019 \pm 0.0084}$	$\mathbf{0.1252 \pm 0.0103}$	$\mathbf{0.3038 \pm 0.0090}$
		TGAP	0.3633 ± 0.0073	0.2003 ± 0.0096	0.3637 ± 0.0075
	80 epochs	TTG	0.3004 ± 0.0056	0.1223 ± 0.0045	0.3019 ± 0.0042
		TGAP	$\mathbf{0.3729 \pm 0.0035}$	$\mathbf{0.2161 \pm 0.0057}$	$\mathbf{0.3795 \pm 0.0063}$
	EMA-JEPA	—	0.3396 ± 0.0091	0.1753 ± 0.0121	0.3423 ± 0.0181

E.2 Analysis of Reconstruction-Specific Branch

The six configurations in Section 3.5 use four Stage 1 target encoders. TopoJEPA and A share the encoder with the reconstruction-specific branch, B and C share the joint-encoding encoder, and D and E each use a separate encoder. Repeated Stage 1 entries in Figure 3 (a–b) therefore represent

shared measurements. Table 15 provides balanced accuracy, Cohen's kappa, and weighted F1 for both stages under linear probing and full fine-tuning.

Table 15: Full results for the Reconstruction-Specific Branch ablation.

		Full fine-tuning			Linear probing		
Variant	Stage	Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Cohen's Kappa	Weighted F1
TopoJEPA	TTG	0.3768 ± 0.0082	0.2220 ± 0.0097	0.3834 ± 0.0078	0.3004 ± 0.0056	0.1223 ± 0.0045	0.3019 ± 0.0042
	TGAP	$\mathbf{0.4157 \pm 0.0108}$	$\mathbf{0.2724 \pm 0.0150}$	$\mathbf{0.4185 \pm 0.0090}$	0.3729 ± 0.0035	0.2161 ± 0.0057	$\mathbf{0.3795 \pm 0.0063}$
A: Auxiliary-Token Prediction	TTG	0.3768 ± 0.0082	0.2220 ± 0.0097	0.3834 ± 0.0078	0.3004 ± 0.0056	0.1223 ± 0.0045	0.3019 ± 0.0042
	TGAP	0.3900 ± 0.0045	0.2377 ± 0.0047	0.3952 ± 0.0042	0.3507 ± 0.0080	0.1859 ± 0.0099	0.3483 ± 0.0084
B: Joint Target Encoding	TTG	0.3899 ± 0.0021	0.2402 ± 0.0035	0.3966 ± 0.0031	0.3113 ± 0.0041	0.1368 ± 0.0073	0.3116 ± 0.0128
	TGAP	0.3971 ± 0.0044	0.2470 ± 0.0057	0.4036 ± 0.0052	0.3591 ± 0.0029	0.1964 ± 0.0026	0.3587 ± 0.0077
C: Joint Encoding + Prediction	TTG	0.3899 ± 0.0021	0.2402 ± 0.0035	0.3966 ± 0.0031	0.3113 ± 0.0041	0.1368 ± 0.0073	0.3116 ± 0.0128
	TGAP	0.3950 ± 0.0056	0.2460 ± 0.0079	0.4018 ± 0.0068	0.3436 ± 0.0067	0.1771 ± 0.0073	0.3404 ± 0.0071
D: No Auxiliary Branch	TTG	0.3850 ± 0.0070	0.2340 ± 0.0097	0.3914 ± 0.0081	0.2982 ± 0.0085	0.1193 ± 0.0132	0.2874 ± 0.0249
	TGAP	0.3965 ± 0.0036	0.2474 ± 0.0043	0.4025 ± 0.0032	0.3343 ± 0.0032	0.1640 ± 0.0050	0.3352 ± 0.0038
E: Full-Channel Encoding	TTG	$\mathbf{0.4012 \pm 0.0041}$	$\mathbf{0.2546 \pm 0.0037}$	$\mathbf{0.4080 \pm 0.0019}$	$\mathbf{0.3208 \pm 0.0048}$	$\mathbf{0.1827 \pm 0.0065}$	$\mathbf{0.3454 \pm 0.0107}$
	TGAP	0.4032 ± 0.0028	0.2581 ± 0.0045	0.4117 ± 0.0038	$\mathbf{0.3763 \pm 0.0038}$	$\mathbf{0.2199 \pm 0.0064}$	0.3766 ± 0.0090

Reconstruction from Retained Components. We evaluate the Stage 1 reconstruction proxy task in both the time and frequency domains using a frozen target encoder and reconstruction decoder. We consider six decoder-input conditions: hemisphere representations only (H), region representations only (R), auxiliary tokens only (Aux), both topology representation groups (H+R), all three components (H+R+Aux), and none. For each condition, components not retained are zeroed at the decoder input, while the EEG inputs and masks are kept identical. The None condition zeros H, R, and Aux simultaneously and serves as a reference. For each of the three pretraining montages, we sample 256 windows of eight seconds each from TUH (19 channels), the SEED series (62 channels), and the ZuCo series (104 channels). For waveform reconstruction, we evaluate both unmasked and 35% patch-masked inputs over the full signal. RMSE is computed over waveform samples on the model-input scale, while COR is the mean Pearson correlation across 1,600-sample channel–window pairs. Figures 5 and 6 report the aggregate reconstruction metrics and full-window waveform comparisons, respectively.

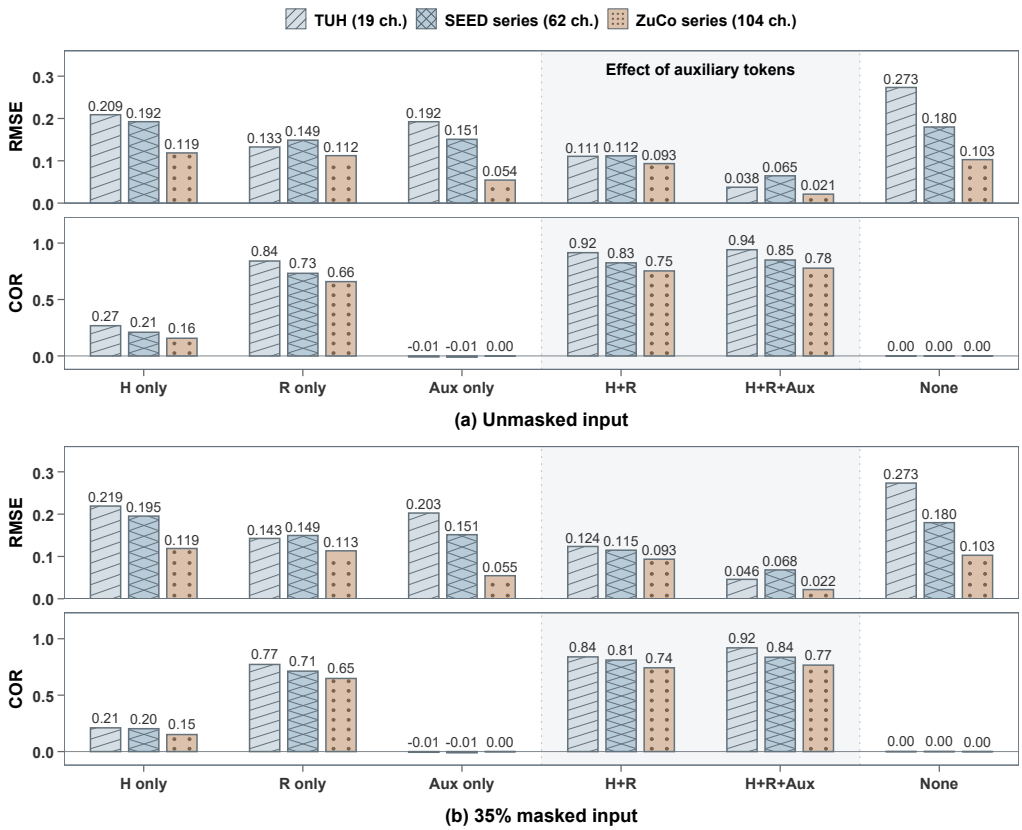


Figure 5: **Reconstruction performance with different retained components.** RMSE and COR are reported for (**a**) unmasked and (**b**) 35% patch-masked inputs across the three pretraining montages. H and R denote hemisphere and region representations; Aux denotes auxiliary tokens.

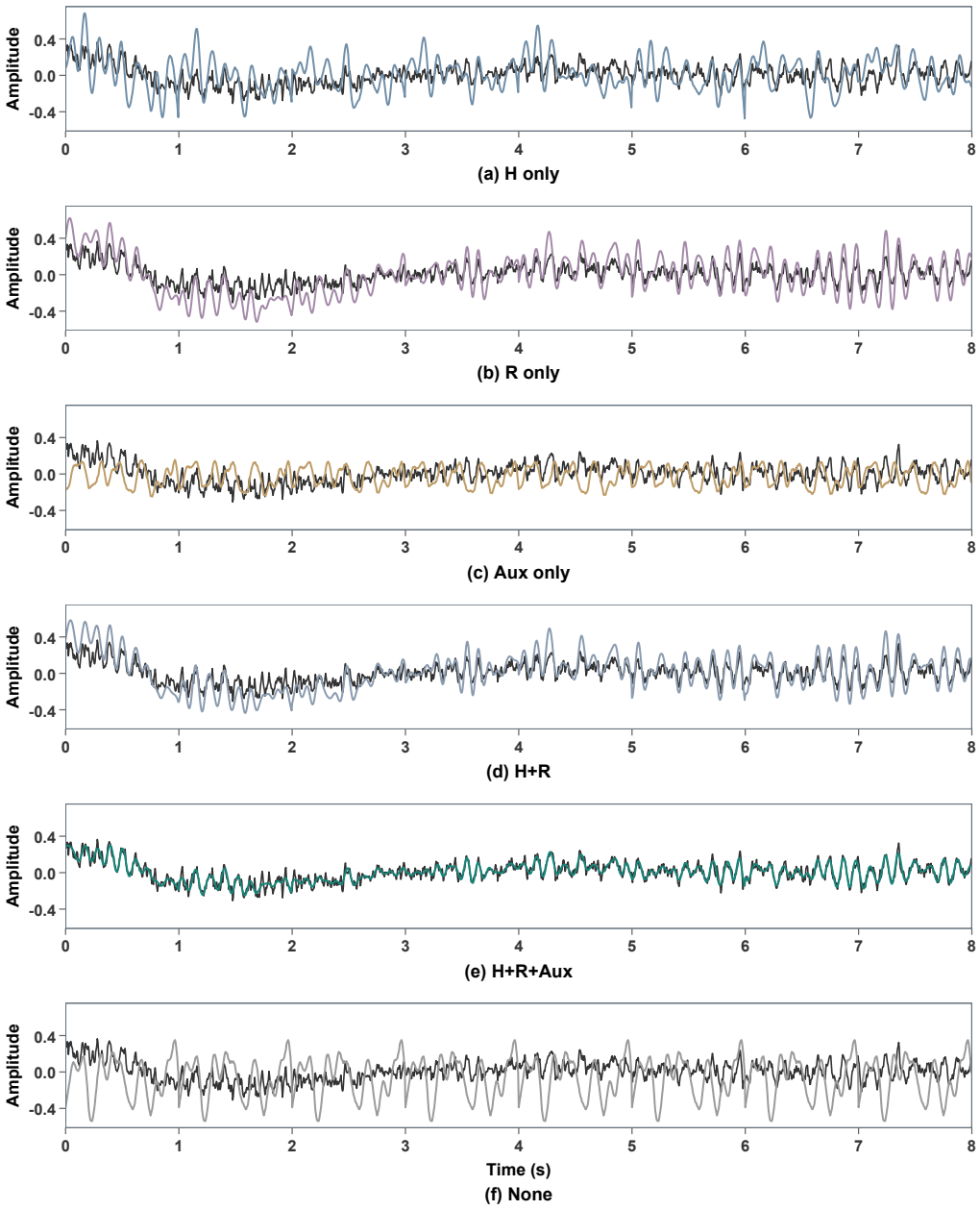


Figure 6: **Waveform reconstruction with different retained components.** Original and reconstructed signals from a representative TUH channel under different decoder-input conditions. H and R denote hemisphere and region representations; Aux denotes auxiliary tokens.

For frequency-domain reconstruction, we compute PSD from reconstructed waveforms rather than the spectral output head, using Welch's method with a 400-sample Hann window, 200-sample overlap, constant detrending, and density scaling. Spectra are averaged across channel–window pairs within each montage and converted from $\mu\mathrm{V}^2/\mathrm{Hz}$ to squared model-input units per Hz by dividing by 100^2. Figure 7 extends the spectral comparison in Figure 3 (e) to all three montages under unmasked and 35% patch-masked inputs.

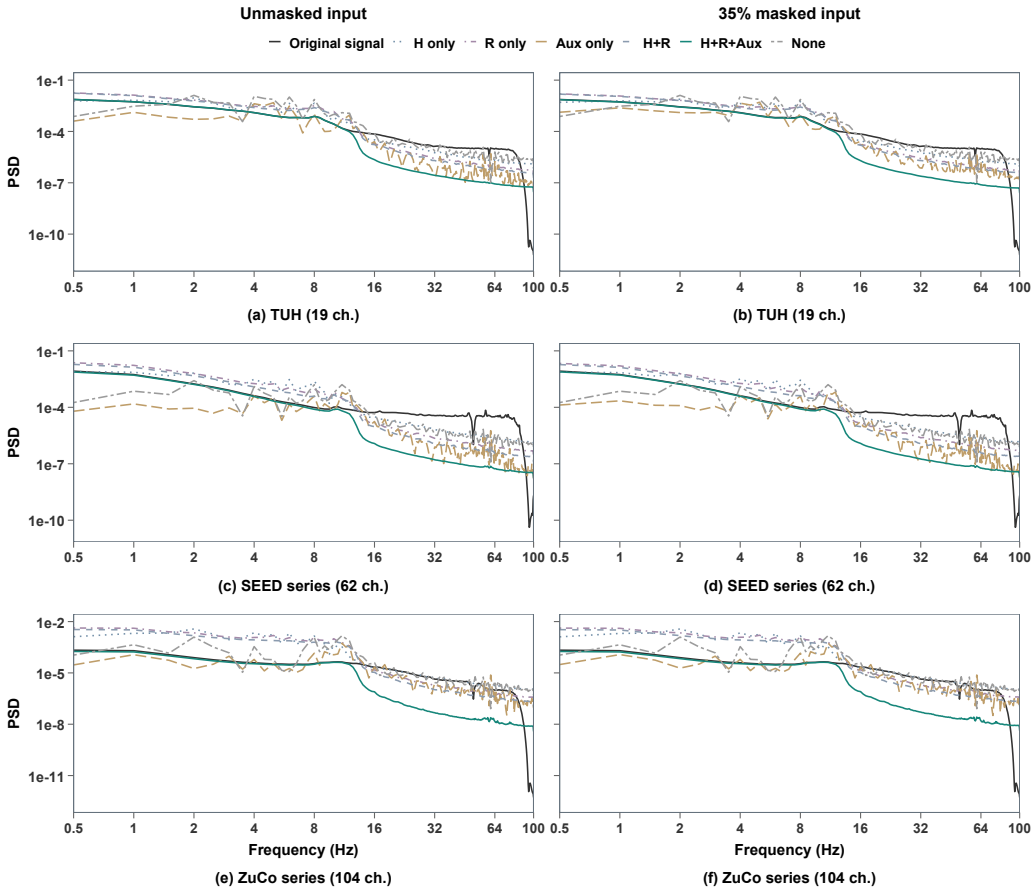


Figure 7: **Spectral reconstruction with different retained components.** Mean PSD of the original and reconstructed signals on TUH, the SEED series, and the ZuCo series under unmasked and 35% patch-masked inputs.

To further examine spatial reconstruction, Figure 8 visualizes scalp topographies from a representative unmasked SEED-series segment. Using the same decoder-input intervention, we compare reconstructions from hemisphere representations, region representations, both topology representation groups, and both groups with auxiliary tokens. Combining the two topology representation groups recovers the spatial voltage distribution better than either group alone, while retaining the auxiliary tokens produces the closest reconstruction to the original topography.

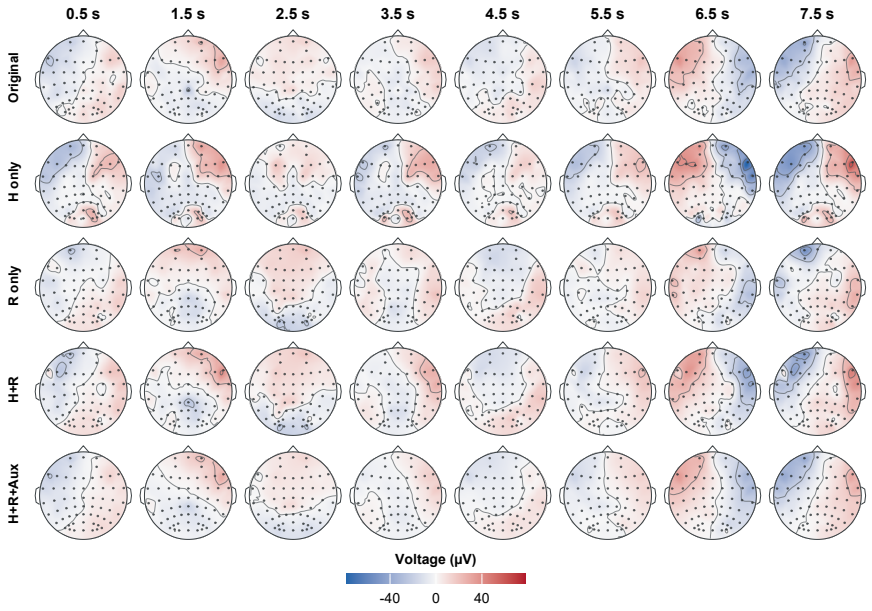


Figure 8: **Scalp topography reconstruction with different retained components.** H and R denote hemisphere and region representations; Aux denotes auxiliary tokens.

E.3 Details of Spatiotemporal Attention Constraints

This section details the attention patterns used in the spatiotemporal constraint analysis of Section 3.6. All variants use the same topology-aware context masking strategy and differ only in how the attention constraints in Eq. (5) are relaxed, with the corresponding downstream results reported in Table 2. Specifically, **w/o Spatiotemporal Constraints** removes all restrictions on information flow across time and topology roles. **w/o Temporal Constraint** removes the temporal restriction while preserving the within-step spatial hierarchy, allowing attention across different time steps, as shown in Figure 9(a). **w/o Hemisphere-to-Region Constraint** instead removes the hemisphere-to-region ordering within each time step while retaining the temporal constraints, as shown in Figure 9(b). **w/o Same-Role History Constraint** preserves temporal causality and the spatial hierarchy but allows each topology role to attend to its own previous representations, as shown in Figure 9(c), whereas **TopoJEPA** retains the full set of spatiotemporal attention constraints, as illustrated in Figure 9(d).

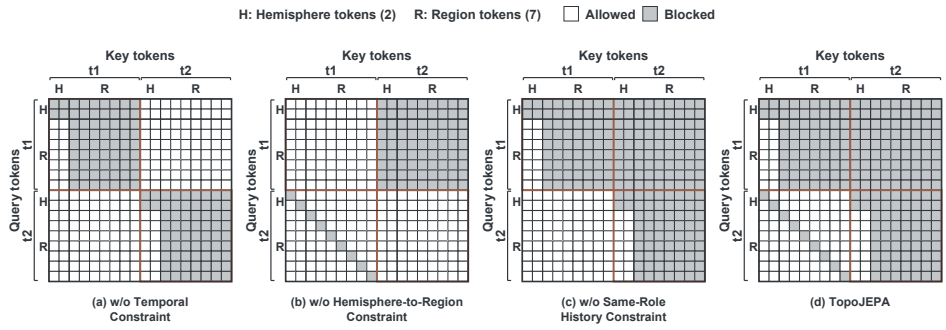


Figure 9: **Spatiotemporal attention patterns.** Context self-attention for (**a**) w/o Temporal Constraint, (**b**) w/o Hemisphere-to-Region Constraint, (**c**) w/o Same-Role History Constraint, and (**d**) TopoJEPA. H and R denote hemisphere and region tokens.

E.4 Analysis of Context Masking Strategies

While the spatiotemporal attention constraints regulate information flow among topology representations, context masking determines which spatial EEG observations are available to the context pathway. We examine how the organization and temporal consistency of this visible context affect Stage 2 representation learning, while keeping the grounded target pathway, attention constraints, and prediction objective fixed. Specifically, **w/o Context Masking** retains all EEG channels, while **Random Channel Masking** masks individual electrodes without considering the predefined topology. **Hemisphere-Only Masking** and **Region-Only Masking** restrict masking to a single topology level. **Time-Varying Mixed Masking** combines hemisphere- and region-level masking but resamples the spatial mask at each temporal patch, whereas **Time-Consistent Mixed Masking (TopoJEPA)** uses the same mixed strategy while keeping the sampled spatial mask fixed throughout the EEG window.

Table 16: Ablation of context masking strategies.

	Full fine-tuning			*Linear probing*		
Masking strategy	Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Cohen's Kappa	Weighted F1
w/o Context Masking	0.3946 ± 0.0035	0.2444 ± 0.0048	0.3996 ± 0.0051	0.3443 ± 0.0034	0.1786 ± 0.0038	0.3449 ± 0.0117
Random Channel	0.4005 ± 0.0071	0.2518 ± 0.0096	0.4057 ± 0.0075	<u>0.3597 ± 0.0056</u>	<u>0.1983 ± 0.0084</u>	0.3591 ± 0.0105
Hemisphere-Only	0.3991 ± 0.0032	0.2510 ± 0.0039	0.4062 ± 0.0034	0.3492 ± 0.0023	0.1847 ± 0.0027	0.3525 ± 0.0032
Region-Only	<u>0.4155 ± 0.0006</u>	<u>0.2701 ± 0.0024</u>	**0.4218 ± 0.0021**	0.3539 ± 0.0026	0.1903 ± 0.0033	0.3545 ± 0.0089
Time-Varying Mixed	0.3864 ± 0.0084	0.2336 ± 0.0109	0.3921 ± 0.0100	0.3595 ± 0.0041	0.1979 ± 0.0058	<u>0.3609 ± 0.0069</u>
TopoJEPA	**0.4157 ± 0.0108**	**0.2724 ± 0.0150**	<u>0.4185 ± 0.0090</u>	**0.3729 ± 0.0035**	**0.2161 ± 0.0057**	**0.3795 ± 0.0063**

As shown in Table 16, w/o Context Masking gives the weakest linear-probing performance. Random Channel Masking and the two single-level variants also perform below Time-Consistent Mixed Masking under linear probing, while Region-Only Masking remains competitive under full fine-tuning. Time-Varying Mixed Masking consistently underperforms the time-consistent strategy under both evaluation settings. These results highlight two factors in constructing the context for Stage 2 prediction. First, alternating between hemisphere- and region-level masking exposes the model to spatial contexts at different topology scales, which benefits representation learning. Second, keeping the mask fixed across time prevents masked information from re-entering the temporal context, making prediction rely on the available spatiotemporal information rather than later observations of the same spatial content.

E.5 Analysis of Differentiation among Topology Representations

We examine whether representations associated with different topology roles remain semantically differentiated after Stage 2. Figure 10(a) measures pairwise cosine similarity between context representations across topology roles. Figure 10(b) perturbs one region at a time by suppressing electrodes near its regional centroid and quantifies the resulting representation change as $1 - \cos(h, h^{\text{pert}})$, where h and h^{pert} denote the context representations before and after perturbation. Figure 10(c) further visualizes the predicted topology representations produced by the Latent Predictor using t-SNE.

The results consistently reveal topology-specific structure in the learned representations. In Figure 10(a), representations associated with different topology roles show some similarity while remaining distinguishable, with a clearer separation between hemisphere- and region-level roles. In Figure 10(b), perturbing a region produces the strongest response in the representation associated with that region, showing that the learned representations remain aligned with their assigned topology roles. Consistently, Figure 10(c) shows that predicted representations from different topology roles form distinct clusters in the t-SNE space. Together, these results indicate that Stage 2 preserves distinct role-specific semantics while retaining structured relationships among topology representations.

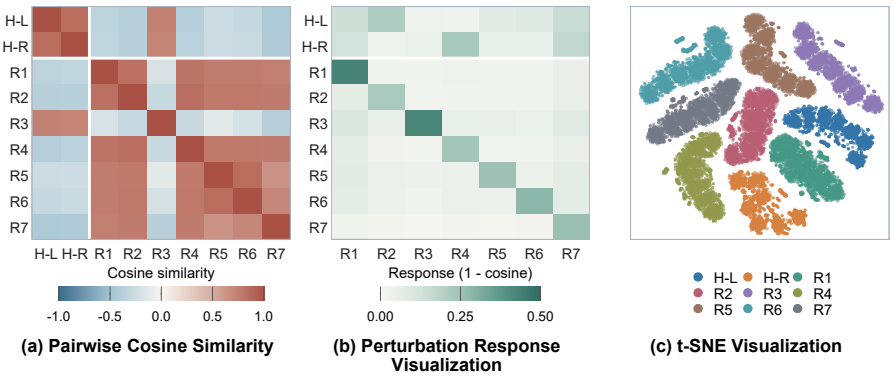


Figure 10: **Differentiation Analysis of Topology Representations.** (**a**) Pairwise cosine similarity between context representations. (**b**) Perturbation-response visualization under region-specific input perturbations. (**c**) t-SNE visualization of predicted topology representations. H-L and H-R denote the left- and right-hemisphere roles, respectively, and R1–R7 denote the seven region roles.

F Discussion

F.1 Implications.

Implications for EEG Foundation Model Design. TopoJEPA highlights the importance of separating representation-space construction from dynamics modeling in EEG pretraining. Rather than allowing the prediction target to evolve together with the predictor, Grounding Before Prediction first establishes a topology-grounded space with consistent spatial semantics and then learns richer intrinsic structure within it. Importantly, the gains of Stage 2 do not arise solely from its attention constraints or context masking. On SEED-V, removing all spatiotemporal attention constraints or removing context masking still yields linear-probing Balanced Accuracy of 34.09% and 34.43%, respectively, both above the 30.04% of the Stage 1 target encoder and competitive with the comparison methods. This suggests that the value of grounding lies in providing a stable and meaningful basis for subsequent representation learning. The Reconstruction-Specific Branch reinforces this view by keeping information needed primarily for signal recovery from dominating the representations used as prediction targets. Together, these findings suggest that the design of the target space itself should be treated as a first-class component of EEG foundation models.

Implications for EEG Foundation Representations. The second implication is that EEG spatiotemporal dynamics can be learned more effectively once a stable spatial reference space has been established. TGAP builds on the grounded topology by structuring temporal information flow and controlling the spatial context available for prediction, encouraging the model to capture intrinsic EEG dynamics rather than relying on easier shortcuts. The strong linear-probing performance across downstream tasks reflects the high quality of the learned representations. More broadly, grounding provides a stable foundation for Stage 2 to capture richer and more intrinsic information in EEG.

F.2 Limitations.

TopoJEPA makes several simplifying assumptions that leave room for further extension. First, its spatial hierarchy is fixed to two hemispheres and seven predefined scalp regions. This provides a consistent representation across electrode configurations, but may miss finer-grained spatial structure and subject-specific organization. Second, TGAP models temporal dynamics over one-second patches within eight-second windows, leaving both sub-second activity and longer-range dependencies unexplored. Third, although the pretraining corpus spans multiple recording domains, it is heavily weighted toward clinical EEG, and a more balanced corpus may lead to a more broadly transferable target space. Finally, TTG grounds the target space through waveform and spectral reconstruction. While effective in our experiments, reconstruction is only one possible proxy for EEG structure, and alternative grounding objectives may further improve the representations learned for prediction.