BiCSP: Learning Unified EEG Representations via Bidirectional Event–State Consistency

Anonymous authors
Paper under double-blind review

Abstract

Electroencephalography (EEG) exhibits neural dynamics at multiple levels of organization. Transient regional activity encodes fine-grained temporal events, whereas coordinated interactions across brain regions give rise to higher-level contextual brain states. Existing self-supervised EEG representation learning methods typically optimize either local temporal modeling or global contextual modeling, leaving the semantic relationship between these complementary scales largely unconstrained.

We propose BiCSP, a hierarchical self-supervised framework for learning unified EEG representations through Bidirectional Cross-Scale Prediction. A learnable frequency-aware front-end first enhances neural signals in the spectral domain. Regional temporal patches are encoded into Event tokens, which are subsequently aggregated by a global attention encoder into a contextual State representation. BiCSP explicitly aligns these two representation levels through complementary latent prediction objectives: Event-to-State prediction reconstructs momentum-teacher State representations from Events, while State-to-Event prediction recovers masked Event targets from a context-isolated State together with visible Event tokens. Bidirectional local Event prediction and global-context masked prediction further regularize within-scale representation learning. After pretraining, Event and State embeddings are fused into a unified downstream representation and transferred using both linear probing and full fine-tuning. By explicitly enforcing semantic consistency across representation scales, BiCSP learns transferable EEG representations beyond conventional single-scale objectives.

1 Introduction

Electroencephalography (EEG) is widely used in cognitive neuroscience, brain–computer interfaces, affective computing, and neurological assessment because it provides millisecond-scale temporal resolution through non-invasive acquisition (Daube & Rubin, 2013). The growing availability of large-scale EEG datasets has shifted the focus from task-specific supervised learning toward self-supervised representation learning, where models are pretrained on unlabeled neural signals and subsequently transferred to diverse downstream tasks (Kostas et al., 2021; Baevski et al., 2022; Assran et al., 2023). The central objective is therefore to learn reusable neural representations rather than optimizing a single task-dependent classifier. Despite substantial progress, most existing EEG self-supervised methods remain effectively single-scale. Temporal prediction and masked reconstruction primarily exploit local signal continuity, whereas graph neural networks, channel attention, and contextual transformers emphasize interactions among channels or regions (Kostas et al., 2021; Yang et al., 2023; Zhang et al., 2022a; Zerveas et al., 2021; Jiang et al., 2024a). Although these approaches successfully model either local dynamics or global context, they rarely constrain the semantic correspondence between the two levels. This limitation is particularly important because neural activity is inherently hierarchical: localized oscillatory patterns emerge within regional circuits, while their coordinated interactions form higher-level functional brain configurations (Daube & Rubin, 2013). Figure 1 illustrates this hierarchical view of EEG representations. We represent regional temporal dynamics as *Event tokens* and summarize the complete EEG segment as a *Global Contextual State*. Importantly, these variables are learned latent representations rather than physiological annotations: Events correspond to regional temporal patches, while the State is a contextual embedding of the entire segment. A desirable representation should therefore satisfy two complementary properties:

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.

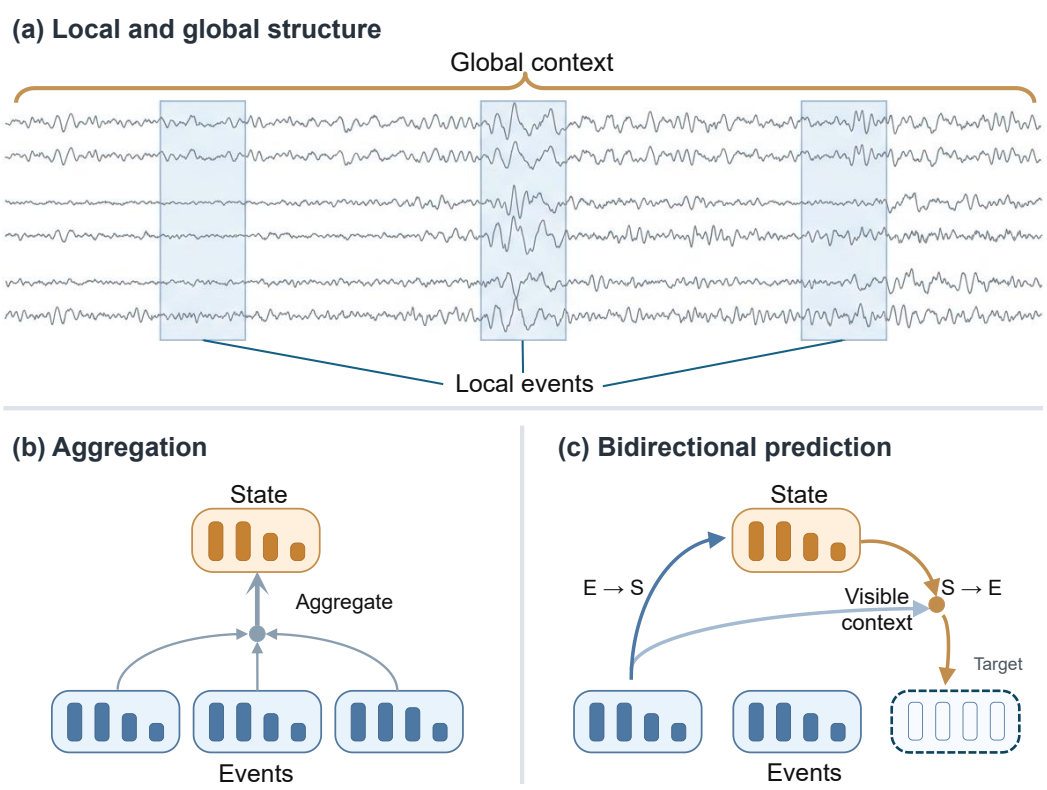

Figure 1: Conceptual motivation for hierarchical EEG representations. (a) EEG contains local regional patterns embedded in a broader global context. (b) Regional Event tokens are aggregated into a Global Contextual State Vector. (c) The two representation levels can constrain one another through bidirectional cross-scale prediction. The illustration is conceptual; Events are learned regional temporal representations rather than annotated clinical events, and the State is a learned contextual summary rather than a predefined physiological category.

the State should preserve sufficient information to explain its constituent Events, and the Events should collectively support inference of the corresponding State. We argue that this cross-scale correspondence should be treated as an explicit self-supervised learning objective instead of emerging implicitly through architectural aggregation. To this end, we propose **BiCSP** (Bidirectional Cross-Scale Prediction), a hierarchical framework for learning unified EEG representations. The overall pipeline can be summarized as

$$\text{EEG} \rightarrow \text{Frequency-aware Signal} \rightarrow \text{Event Tokens} \rightarrow \text{Global State} \rightarrow \text{Cross-Scale Consistency}.$$

A learnable frequency-aware front-end first enhances informative spectral components. The signal is then partitioned into regional temporal patches and encoded into Event tokens through a temporal Transformer. A global attention encoder aggregates all Events into a contextual State representation. Rather than treating the State as the endpoint of the encoder, BiCSP explicitly couples the two representation levels through two complementary latent prediction tasks. Event-to-State prediction learns to infer the teacher State from Event representations, whereas State-to-Event prediction reconstructs masked Event targets from a context-isolated State together with the remaining visible Events. A momentum target encoder provides stable stop-gradient supervision for both directions (Baevski et al., 2022; Assran et al., 2023). Our contributions are summarized as follows:

1. We formulate EEG self-supervised learning as a hierarchical cross-scale representation problem by explicitly modeling the correspondence between regional Event representations and contextual Global States.
2. We propose BiCSP, a unified self-supervised framework that jointly optimizes Event-to-State and State-to-Event latent prediction using momentum-teacher supervision.

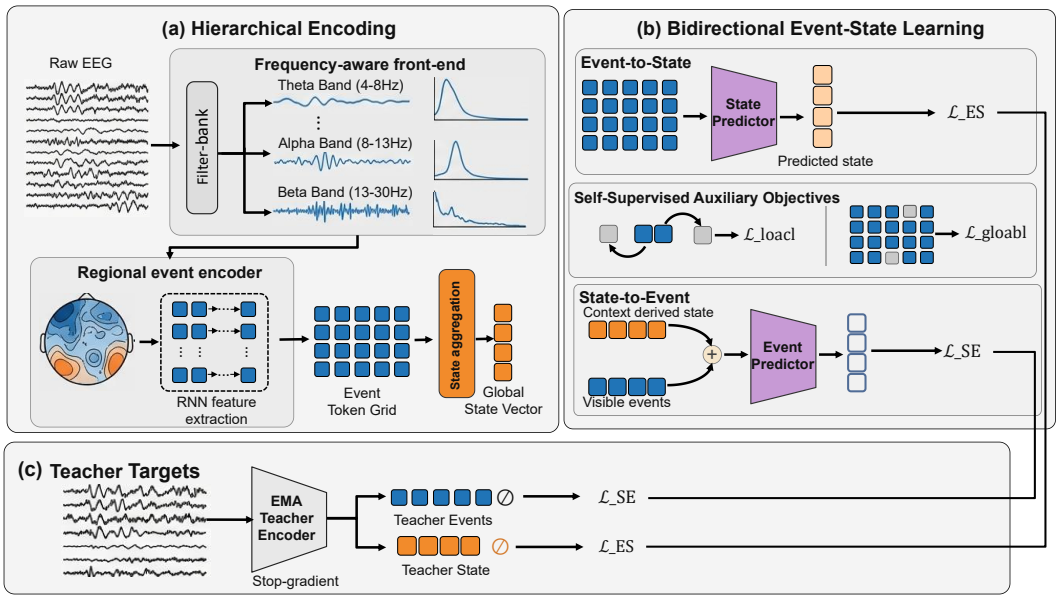

Figure 2: Overview of BiCSP. A frequency-aware front-end first transforms the EEG signal into regional temporal patches, which are encoded as Event tokens and aggregated into a contextual State representation. Four complementary self-supervised objectives jointly optimize the representation hierarchy: local bidirectional Event prediction, global-context masked prediction, Event-to-State prediction, and State-to-Event prediction. These objectives establish bidirectional consistency between local Event representations and the global contextual State, yielding a unified EEG representation for downstream tasks.

3. We learn unified Event–State representations that consistently transfer across heterogeneous EEG benchmarks under both frozen linear probing and full fine-tuning protocols.

Design principles. BiCSP is built upon three design principles. First, the representation hierarchy is explicit: regional Event tokens are learned before the Global State is constructed, exposing an interpretable intermediate representation instead of collapsing local dynamics into a single embedding. Second, cross-scale semantic consistency is learned bidirectionally rather than being implicitly induced by architectural aggregation, enabling Event and State representations to mutually constrain one another. Third, all supervisory signals are latent representations generated by a slowly evolving momentum teacher, avoiding reconstruction in the raw signal space while providing stable self-supervised optimization. In BiCSP, the distinction between Event and State is operational rather than physiological. Event representations preserve fine-grained regional temporal dynamics, whereas State representations capture the contextual organization of the entire EEG segment. Their complementary fusion constitutes a unified EEG representation that is transferable across diverse downstream neural decoding tasks.

2 Method

2.1 Overview and Problem Formulation

Given an EEG segment $X \in \mathbb{R}^{C\times T}$, BiCSP learns a unified representation by explicitly modeling two complementary levels of neural dynamics: regional temporal Events and a segment-level contextual State. The framework consists of a trainable online encoder and a momentum target encoder. The online branch receives stochastic views of the EEG signal and is optimized by four complementary self-supervised objectives, while the target branch processes the complete segment to provide stable latent targets.

The overall architecture is illustrated in Figure 2. The input EEG is first transformed by a learnable frequency-aware front-end and partitioned into regional temporal patches. These patches are en-

coded into Event tokens, which preserve localized temporal and regional information. A global attention encoder subsequently aggregates the Event tokens into a contextual State representation. BiCSP then imposes predictive constraints both within and across these representation levels. Specifically, local bidirectional prediction and global-context masked prediction regularize Event representations, while Event-to-State and Context-Isolated State-to-Event prediction explicitly enforce cross-scale consistency.

Let $X_i \in \mathbb{R}^{C\times T}$ denote an EEG segment. The channels are grouped into R regions and the temporal dimension is divided into N patches, defining the region–time index set

$$\mathcal{I} = \{1, \dots, R\} \times \{1, \dots, N\}. \tag{1}$$

For a stochastic training view, $V \subseteq \mathcal{I}$ denotes the visible locations and $M = \mathcal{I} \setminus V$ denotes the target locations. The online encoder produces Event tokens $E^\theta(X; V)$ and a contextual State $s^\theta(X; V)$, whereas the momentum encoder produces complete-segment targets $E^\xi(X; \mathcal{I})$ and $s^\xi(X; \mathcal{I})$. Teacher outputs are detached before being used as prediction targets.

The total pretraining objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{global}}\mathcal{L}_{\text{global}} + \lambda_{S\to E}\mathcal{L}_{S\to E} + \lambda_{E\to S}\mathcal{L}_{E\to S}. \tag{2}$$

2.2 Hierarchical EEG Representation

Frequency-aware signal transformation. EEG contains neural activity distributed across multiple frequency ranges. We therefore first apply a learnable frequency-aware front-end to transform the input signal before constructing the representation hierarchy. Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the orthonormal real Fourier transform and its inverse. For each channel,

$$\hat{X} = \mathcal{F}(X), \tag{3}$$

and the enhanced signal is obtained as

$$\tilde{X} = \mathcal{F}^{-1}\left(\hat{X} \odot W_\theta \odot G_\theta(\log(1 + |\hat{X}|)) \odot M_{\text{EEG}}\right), \tag{4}$$

where W_θ denotes a learnable complex modulation, G_θ is a learnable frequency gate, and M_{EEG} selects the implemented EEG frequency ranges. In our implementation, the retained bands are 0.5–4 Hz, 4–8 Hz, 8–12 Hz, 12–30 Hz, and 30–45 Hz. A learnable residual gate combines the enhanced signal with the original input, followed by normalization. The frequency-aware module is removed in the corresponding ablation.

Regional Event encoding. The enhanced signal is partitioned into R channel regions and N temporal patches of length P. For each region–time location (r, t), the corresponding regional patch is projected into a d-dimensional embedding:

$$h_{r,t} = \text{Proj}(\text{Norm}(x_{r,t})) + p_t + q_r, \tag{5}$$

where p_t and q_r are learnable temporal and regional embeddings. Tokens belonging to each region are refined by a temporal Transformer. The resulting Event collection is

$$E(X; V) = \{e^V_{r,t}\}_{(r,t)\in V}. \tag{6}$$

Each Event token therefore retains localized temporal dynamics while maintaining its regional identity.

Global State encoding. The Event tokens are subsequently aggregated into a segment-level contextual representation. A learnable query attends to the Event collection through multi-head attention:

$$s(X; V) = \mathcal{A}_\phi(E(X; V)) \in \mathbb{R}^d. \tag{7}$$

The State encoder consists of attention, feed-forward transformation, residual connections, and normalization. Unlike an Event token, which represents a localized regional temporal pattern, the State provides a compact contextual summary of the available EEG segment.

2.3 WITHIN-LEVEL PREDICTIVE LEARNING

Before enforcing correspondence between Events and States, BiCSP first regularizes the Event representation at its own scale. We employ two complementary objectives that capture temporal dynamics and contextual dependencies.

Bidirectional Event prediction. For each region, a forward GRU predicts the next Event token from preceding Events, while a backward GRU predicts the preceding Event from the reversed temporal sequence. Let $\hat{e}^{\rightarrow}_{r,t}$ and $\hat{e}^{\leftarrow}_{r,t}$ denote the two predictions, and let $e^{\xi}_{r,t}$ denote the corresponding momentum-teacher Event. The local objective is

$$\mathcal{L}_{\text{local}} = \frac{1}{2}\left[d\left(\hat{e}^{\rightarrow}_{r,t}, e^{\xi}_{r,t+1}\right) + d\left(\hat{e}^{\leftarrow}_{r,t}, e^{\xi}_{r,t-1}\right)\right], \tag{8}$$

where $d(\cdot,\cdot)$ measures latent discrepancy. This objective encourages Event representations to preserve local temporal dependencies in both directions.

Global-context masked prediction. To complement local temporal prediction, we randomly mask region–time locations from the online Event collection. A predictor receives the online State together with the remaining visible Events and predicts the teacher representations at the masked positions:

$$\mathcal{L}_{\text{global}} = \frac{1}{|M|}\sum_{(r,t)\in M} d\left(\hat{e}_{r,t}, e^{\xi}_{r,t}\right). \tag{9}$$

Unlike waveform reconstruction, this objective predicts latent Event representations conditioned on global context. It therefore encourages the State to encode contextual information useful for recovering missing regional activity.

2.4 BIDIRECTIONAL CROSS-SCALE PREDICTION

The central idea of BiCSP is to explicitly constrain the correspondence between the Event and State representation levels. The forward encoder naturally aggregates Events into a State, but architectural aggregation alone does not guarantee that Event representations retain sufficient information to support global inference. Conversely, predicting a local Event from contextual information does not necessarily require the learned State to capture meaningful global structure. We therefore introduce two complementary prediction directions:

$$E^{\theta} \rightarrow s^{\xi}, \qquad (s^{\theta}, E^{\theta}_{\text{visible}}) \rightarrow e^{\xi}. \tag{10}$$

These objectives encourage mutual predictive compatibility between local and global representations without requiring the two mappings to be exact inverses.

Event-to-State prediction. Given online Event tokens, an attention-based pooling predictor produces a predicted State:

$$\hat{s} = P_{E\rightarrow S}\left(E^{\theta}(X;V)\right). \tag{11}$$

The prediction is supervised by the complete-segment State generated by the momentum teacher:

$$s^{\xi} = \text{sg}\left[s_{\xi}(X;\mathcal{I})\right], \tag{12}$$

yielding

$$\mathcal{L}_{E\rightarrow S} = d(\hat{s}, s^{\xi}). \tag{13}$$

This objective requires the collection of regional Event representations to contain sufficient information for predicting the contextual State.

Context-Isolated State-to-Event prediction. The reverse direction predicts local Event representations from contextual information. For each target location $(r,t) \in M$, Context Isolation removes a local temporal neighborhood around the target from the online context. The resulting online branch produces a context State s^{θ}_{ctx} and visible Event collection E^{θ}_{ctx}. A position query $q_{r,t}$ specifies the target location, and the predictor estimates

$$\hat{e}_{r,t} = P_{S\rightarrow E}\left(q_{r,t}, s^{\theta}_{\text{ctx}}, E^{\theta}_{\text{ctx}}\right). \tag{14}$$

The target is the corresponding Event representation generated from the complete segment by the momentum encoder:

$$e^{\xi}_{r,t} = \mathrm{sg}\left[e_{\xi,r,t}(X;\mathcal{I})\right]. \tag{15}$$

The resulting objective is

$$\mathcal{L}_{S\to E} = \frac{1}{|M|}\sum_{(r,t)\in M} d\left(\hat{e}_{r,t}, e^{\xi}_{r,t}\right). \tag{16}$$

Importantly, State-to-Event prediction is not reconstruction from the State alone. Visible Event tokens provide fine-grained contextual information, while the State summarizes the broader organization of the available signal. Context Isolation prevents the predictor from relying on immediately adjacent temporal tokens as a shortcut, encouraging prediction from broader contextual information.

Latent discrepancy. For a prediction u and stop-gradient target v, we use a combination of cosine discrepancy and normalized mean-squared error:

$$d(u,v) = 1 - \frac{u^\top v}{\|u\|_2\|v\|_2} + 0.1\left\|\frac{u}{\|u\|_2} - \frac{v}{\|v\|_2}\right\|_2^2. \tag{17}$$

All prediction objectives operate in latent space, avoiding the need to reconstruct raw EEG waveforms.

2.5 Unified Representation and Optimization

Momentum target encoder. The online encoder parameters are denoted by θ, while the target encoder parameters are denoted by ξ. The two encoders share the same hierarchical architecture. The target encoder is updated by exponential moving average:

$$\xi \leftarrow m\xi + (1-m)\theta, \qquad m \in [0,1). \tag{18}$$

The target encoder is not updated by backpropagation, and all target representations are detached before computing prediction losses. Using a shared momentum target space across the four objectives provides stable latent supervision throughout pretraining.

Unified downstream representation. After pretraining, the online encoder produces Event tokens E and a Global State s. Event tokens are mean-pooled across regions and temporal positions and concatenated with the State:

$$z = \mathrm{LayerNorm}\left([\mathrm{mean}(E); s]\right) \in \mathbb{R}^{2d}. \tag{19}$$

This representation combines fine-grained regional temporal information from Events with global contextual information from the State.

For linear probing, the pretrained encoder is frozen and only a task-specific linear classifier is optimized. For full fine-tuning, the encoder and downstream classifier are jointly optimized. The momentum encoder and pretraining prediction heads are discarded during downstream adaptation.

3 Experiments

3.1 Experimental Setup

Benchmarks. We evaluate BiCSP on five heterogeneous EEG benchmarks: DREAMER, Crowdsourced, STEW, N-Back, and TUEV. Unless a benchmark-specific protocol is required, signals are resampled to 128 Hz and segmented into two-second windows with a one-second stride. Subject- or recording-level partitioning is performed before window extraction.

We compare BiCSP with representative EEG self-supervised methods, including BENDR Kostas et al. (2021), MAEEG, TS-TCC, TF-C, BIOT Yang et al. (2023), and EEG2Rep. We evaluate pretrained representations using both linear probing and full fine-tuning. Binary tasks are evaluated using accuracy and AUROC, while TUEV is evaluated using balanced accuracy and weighted F1. All results are reported as mean $\pm$ sample standard deviation over five random seeds. Additional implementation and reproducibility details are provided in Appendix.

Table 1: Linear probing results on five EEG benchmarks. We report mean ± sample standard deviation over five independent seeds. Acc. and AUROC are reported for binary benchmarks, while B-Acc. and W-F1 are reported for TUEV.

Model	DREAMER		Crowdsourced		STEW		N-Back		TUEV	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	B-Acc	W-F1
BENDR	51.48±1.87	51.68±1.98	70.46±4.14	70.56±4.12	63.03±1.07	63.03±1.07	62.18±3.44	60.91±5.14	37.38±3.11	61.31±3.18
MAEEG	54.24±2.04	53.98±2.08	75.21±2.11	75.01±2.01	67.99±1.86	68.58±1.88	62.30±4.27	60.56±5.08	37.23±2.99	61.38±3.08
TS-TCC	53.60±2.68	53.87±2.28	77.75±2.94	77.83±2.11	64.54±1.62	64.64±1.73	52.73±4.13	56.96±4.94	35.98±2.85	60.67±2.98
TF-C	52.52±1.88	52.26±2.08	64.82±7.23	65.32±6.12	58.84±2.36	58.69±2.43	58.27±4.23	58.13±5.93	30.12±4.06	56.23±4.05
BIOT	53.35±2.41	53.42±1.91	76.23±4.56	76.33±4.12	67.54±2.08	67.69±3.60	51.80±3.08	55.27±5.50	40.02±1.87	65.98±2.01
EEG2Rep	58.45±1.82	55.19±1.95	81.66±2.93	81.67±2.65	69.04±1.04	69.10±1.23	62.26±4.84	62.97±3.51	43.25±3.12	69.95±3.21
BiCSP (Ours)	**58.70±1.63**	**57.77±2.17**	**86.37±2.69**	**94.94±1.39**	61.94±0.38	65.90±0.53	**72.99±0.66**	**73.54±2.30**	**45.89±2.87**	**71.00±2.54**

Preprocessing and Input Representation. BiCSP uses a frequency-aware front-end followed by regional temporal tokenization. The front-end explicitly models EEG frequency information in five canonical bands: 0.5–4 Hz, 4–8 Hz, 8–12 Hz, 12–30 Hz, and 30–45 Hz. The temporal patch length is set to 8 samples. When a benchmark does not provide anatomically or functionally defined channel groups, individual EEG channels are treated as independent regions. The same representation architecture is used across all benchmarks without introducing dataset-specific architectural modifications. The representation dimension is $d = 32$, the temporal Event encoder contains two Transformer layers, and the momentum teacher uses a momentum coefficient of 0.99.

Baselines. We compare BiCSP with representative self-supervised EEG representation-learning methods, including BENDR (Kostas et al., 2021), MAEEG (Chien et al., 2022), TS-TCC (Eldele et al., 2021), TF-C (Zhang et al., 2022b), BIOT (Yang et al., 2023), and EEG2Rep (Mohammadi Foumani et al., 2024). For baselines reproduced in our experiments, we use the same downstream data partitions and evaluation metrics as BiCSP whenever the released implementation permits. Results taken from prior work or official checkpoints are identified separately and are not treated as directly controlled comparisons.

Downstream Evaluation. We consider two complementary evaluation protocols. In *linear probing*, the pretrained encoder is frozen and only a linear classification head is optimized. This setting evaluates the amount of task-relevant information directly accessible from the learned representation. In *fine-tuning*, the complete encoder and downstream classification head are jointly optimized, measuring the adaptability of the pretrained representation to each downstream task. During downstream evaluation, the momentum teacher and all self-supervised prediction heads are discarded. Only the learned EEG representation is retained. For binary classification benchmarks, we report Accuracy (Acc.) and Area Under the Receiver Operating Characteristic curve (AUROC). For TUEV, we report Balanced Accuracy (B-Acc.) and Weighted F1 (W-F1) to account for its class imbalance and multi-class setting. Each experiment is repeated using five independent random seeds, and we report the mean and sample standard deviation.

3.2 Main Results

Linear Probing. We first evaluate the representations under frozen linear probing. This setting isolates the information encoded by the pretrained representation without allowing downstream optimization to modify the encoder. The frozen evaluation shows that the BiCSP representation contains directly accessible task-relevant information across several heterogeneous EEG benchmarks. On Crowdsourced, BiCSP achieves 86.37% accuracy and 94.94% AUROC. On N-Back, it reaches 72.99% accuracy and 73.54% AUROC, while on TUEV it obtains 45.89% balanced accuracy and 71.00% weighted F1. On STEW, the representation achieves 61.94% accuracy and 65.90% AUROC. These results also illustrate that representation quality can depend on the downstream metric and benchmark characteristics. For example, the relative performance of different pretrained representations is not identical across Accuracy and AUROC. We therefore report all metrics separately rather than collapsing the results into a single aggregate score.

Fine-Tuning. We next allow the pretrained encoder to adapt to each downstream task. Fine-tuning evaluates whether the representation provides a useful initialization while permitting task-specific optimization. Fine-tuning consistently improves the BiCSP representation on the currently com-

Table 2: Fine-tuning results on five EEG benchmarks. We report mean ± sample standard deviation over five independent seeds.

Model	DREAMER		Crowdsourced		STEW		N-Back		TUEV	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	B-Acc	W-F1
BENDR	54.45±2.11	53.02±3.11	83.78±2.35	83.80±2.63	69.74±2.11	69.77±2.03	59.91±9.35	62.01±5.49	41.17±2.89	67.31±2.96
MAEEG	53.63±2.61	52.08±2.36	86.75±3.50	86.21±3.41	72.46±3.67	72.50±3.22	57.85±5.02	60.40±8.97	41.23±3.65	67.38±3.69
TS-TCC	58.16±2.11	55.05±1.79	89.22±1.22	89.22±1.22	71.00±2.98	71.03±3.02	59.72±5.52	58.14±8.33	40.98±2.55	68.67±2.89
TF-C	52.82±1.66	52.86±1.85	82.93±4.02	82.90±4.32	68.65±1.12	68.70±1.75	63.33±5.88	65.32±6.11	40.12±3.66	66.23±3.85
BIOT	53.45±2.01	53.53±2.13	87.95±3.52	87.78±3.09	69.88±2.15	70.11±2.57	56.29±12.10	59.06±13.15	46.02±1.68	69.98±1.99
EEG2Rep	60.37±1.52	59.42±1.45	94.13±2.11	94.13±2.17	73.60±1.47	74.40±1.50	59.58±2.00	62.76±2.34	52.95±1.58	75.08±1.21
BiCSP (Ours)	**60.74±2.12**	**61.66±2.34**	90.52±0.71	**96.68±0.52**	**75.34±1.12**	**84.72±1.28**	**77.31±2.23**	**82.82±3.73**	**55.64±1.70**	72.46±2.19

pleted binary benchmarks. On Crowdsourced, fine-tuning reaches 90.52% accuracy and 96.68% AUROC. On STEW, the corresponding results are 75.34% and 84.72%, while N-Back reaches 77.31% accuracy and 82.82% AUROC. On TUEV, BiCSP achieves 55.64% balanced accuracy and 72.46% weighted F1. The comparison also indicates that the behavior of BiCSP depends on the evaluation metric. For example, on Crowdsourced, EEG2Rep obtains higher accuracy (94.13%) than BiCSP (90.52%), whereas BiCSP obtains higher AUROC (96.68% versus 94.13%). This difference motivates reporting multiple evaluation metrics rather than interpreting a single metric as a complete characterization of representation quality.

3.3 Ablation Studies

We conduct an ablation study on four EEG tasks, namely N-Back, Crowdsourced, STEW, and DREAMER. We compare the full BiCSP model with four variants that remove the frequency-aware front-end, Event self-supervision, State self-supervision, or the State-to-Event objective, respectively. All variants use the same data splits, optimization settings, and downstream fine-tuning protocol. Figure 3 show that the complete model achieves the best or near-best fine-tuning accuracy across all four tasks, demonstrating the complementary contributions of the proposed components. Removing the frequency-aware front-end causes a substantial degradation on Crowdsourced, STEW, and DREAMER, suggesting that frequency-aware modeling may be particularly beneficial for affective and heterogeneous EEG signals. Removing Event self-supervision leads to a pronounced performance drop on Crowdsourced, while its effect is smaller on N-Back and STEW, suggesting that local temporal prediction is particularly useful for capturing fine-grained EEG dynamics under complex cross-subject variability. Removing State self-supervision consistently reduces performance, supporting the contribution of global contextual State representations. Finally, removing the State-to-Event objective also degrades performance, particularly on STEW, supporting the role of bidirectional cross-scale alignment between global State representations and local Event tokens. Overall, the ablation results show that the frequency-aware encoder, local Event objective, global State objective, and State-to-Event prediction provide complementary contributions to the learned BiCSP representation.

3.4 Robustness to Input Perturbations.

We further evaluate the robustness of the learned representations to controlled perturbations of the EEG input on the Crowdsourced benchmark under the frozen linear-probing protocol. Following the robustness protocol, the test signals are evaluated either without corruption or under a combined perturbation that applies amplitude scaling, a non-circular temporal shift, a DC offset, and additive Gaussian noise, with perturbation severities sampled from the prescribed ranges. As shown in Fig. 4, BiCSP achieves an accuracy of 85.7% on clean inputs and 84.9% under the combined perturbation, corresponding to a drop of only 0.8 percentage points. EEG2Rep decreases from 83.9% to 82.1%, representing a 1.8-percentage-point drop, while TS-TCC decreases from 81.3% to 78.5%, representing a 2.8-percentage-point drop. TF-C exhibits a larger decrease from 74.9% to 66.4%, corresponding to an 8.5-percentage-point drop. Among the remaining baselines, BIOT, BENDR, and MAEEG experience drops of 2.8, 12.3, and 4.8 percentage points, respectively. Under this evaluation protocol, BiCSP achieves both the highest clean accuracy and the smallest performance degradation, indicating that its Event–State representation is comparatively robust to the considered signal-level distortions.

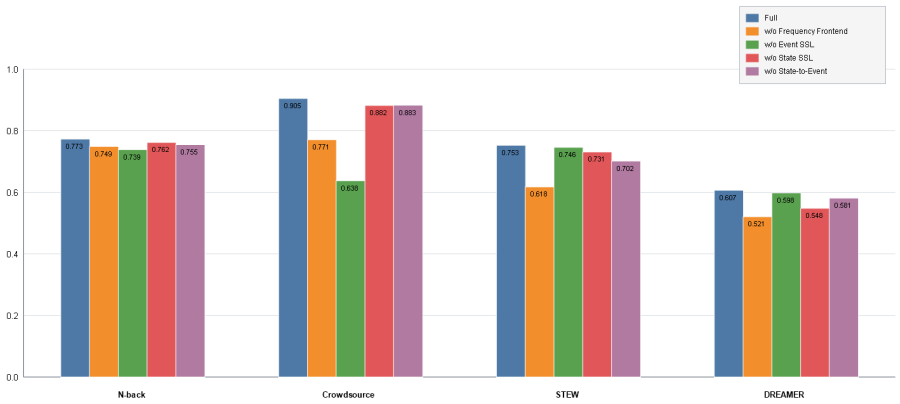

Figure 3: Fine-tuning accuracy of BiCSP and its ablated variants on four EEG datasets. "w/o" denotes removing the corresponding module or self-supervised objective while keeping all other components unchanged.

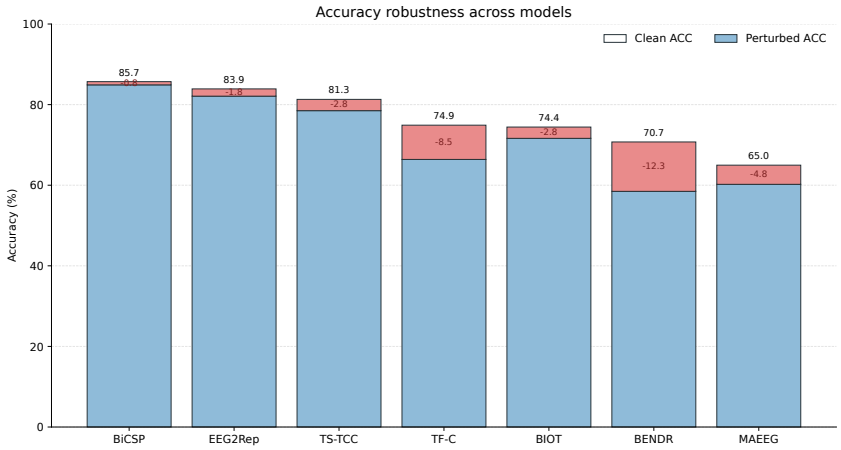

Figure 4: Robustness comparison on the Crowdsourced dataset under the combined input perturbation setting. The blue portion shows the accuracy retained under perturbation, while the red portion indicates the corresponding accuracy drop relative to the clean input. Values above the bars denote clean accuracy.

4 Conclusion

We presented BiCSP, a unified self-supervised framework for EEG representation learning that explicitly models the correspondence between regional Event representations and a contextual Global State. The framework combines frequency-aware signal transformation, within-level predictive learning, and bidirectional cross-scale latent prediction. Event-to-State and State-to-Event objectives provide complementary constraints between the two representation levels, while a momentum target encoder supplies stable latent supervision. The resulting Event–State representation can be transferred to downstream EEG decoding through both linear probing and full fine-tuning. More broadly, BiCSP treats cross-scale semantic correspondence as an explicit learning objective, providing a unified perspective for self-supervised EEG representation learning.

AI USE STATEMENT

We used generative AI tools to assist with language editing, manuscript organization, partial drafting, and experimental code development and debugging. The authors fully reviewed all AI-assisted content and take responsibility for the final text, citations, claims, experimental results, and code.

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 (CVPR)*, pp. 15619–15629, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M Sandino, and Joseph Y Cheng. Maeeg: Masked auto-encoder for eeg representation learning, 2022.

Jerome R Daube and David I Rubin. *Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants*. Oxford University Press, 2013.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2352–2359, 2021. doi: 10.24963/ijcai.2021/324.

Wei Jiang, Long Zhao, Jingyu Liu, Siqi Chen, Yixuan Wang, and Dong Wu. Labram: Large brain model for universal eeg representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14298–14306, 2024a.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024b.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

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. doi: 10.1145/3637528.3671600.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Advances in Neural Information Processing Systems*, 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*, volume 2025, pp. 75310–75346, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2114–2124, 2021. doi: 10.1145/3447548.3467401.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022a.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3988–4003, 2022b.

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, 2026.

A Related Work

A.1 Self-Supervised EEG Representation Learning

Self-supervised learning has become an effective paradigm for learning transferable EEG representations without requiring large-scale manual annotation. BENDR introduced masked contrastive representation learning for EEG and demonstrated transfer across recording conditions (Kostas et al., 2021). BIOT proposed a unified tokenization framework for heterogeneous biosignals with varying channel layouts and recording lengths (Yang et al., 2023). LaBraM further advanced large-scale EEG pretraining through vector-quantized neural spectrum tokenization and masked neural-code prediction (Jiang et al., 2024b). EEGPT combined masked modeling with multiple self-supervised objectives to improve downstream representation transfer (Wang et al., 2024). More recently, EEG2Rep explored latent self-prediction with informative masked inputs, showing that predicting masked representations rather than raw EEG can improve robustness to noise and signal variability (Mohammadi Foumani et al., 2024). Collectively, these approaches demonstrate the effectiveness of masked prediction, contrastive learning, and latent representation learning for EEG. However, most objectives are formulated within a single representation level, leaving the relationship between local temporal patterns and higher-level contextual representations less explicitly constrained.

A.2 Hierarchical and Multi-scale EEG Representations

EEG exhibits hierarchical and multi-scale structure across both temporal and spatial dimensions. Existing approaches capture this structure through structured tokenization, spatial–temporal modeling, and multi-scale architectures. BIOT and LaBraM preserve signal organization through structured tokenization (Yang et al., 2023; Jiang et al., 2024b), while CBraMod explicitly models spatial–temporal dependencies using criss-cross attention (Wang et al., 2025). CSBrain further introduces cross-scale spatiotemporal tokenization and structured sparse attention to explicitly model dependencies across temporal windows and brain regions (Zhou et al., 2026). These studies establish cross-scale structure as an important inductive bias for EEG representation learning. Nevertheless, existing approaches primarily realize cross-scale interaction through tokenization, aggregation, or attention. The predictive semantic correspondence between representations at different scales remains less explicitly supervised.

A.3 Latent Predictive Learning

Predicting latent representations rather than reconstructing raw input signals has emerged as an effective strategy for self-supervised representation learning. Contrastive Predictive Coding (CPC) predicts future latent variables from contextual representations (van den Oord et al., 2018). Data2vec replaces contrastive objectives with momentum-teacher latent targets and learns contextualized representations directly in embedding space (Baevski et al., 2022). I-JEPA further demonstrates that predicting masked target representations in latent space can learn semantic features without pixel-level reconstruction (Assran et al., 2023). Recent EEG studies have also explored latent predictive learning to reduce the reliance on direct raw-signal reconstruction (Mohammadi Foumani et al.,

2024). These methods motivate learning semantic targets in representation space, but their predictive relationships are generally defined within a single representation hierarchy. BiCSP extends this paradigm to an explicit Event–State hierarchy, where local Event representations and the global contextual State are jointly constrained through bidirectional Event-to-State and State-to-Event prediction.

B Dataset and Implementation Details

B.1 Data Splits and Model Selection.

We preserve official train/validation/test splits whenever they are provided. For datasets without an official split, subject- or recording-level splits are constructed before window segmentation. All hyperparameters and classification thresholds are selected using the validation set only. The final test set is evaluated once using the selected configuration.

B.2 Preprocessing and Data Splits

The preprocessing and splitting procedures described in the main paper are applied consistently across all benchmarks. EEG signals are resampled to 128 Hz and segmented into 2-second windows with a 1-second stride unless otherwise specified by the original benchmark protocol. Normalization statistics are estimated using the training set only. Subject- or recording-level splitting is performed before temporal windowing whenever such grouping information is available.

B.3 Architecture and Optimization Details

The representation dimension is $d = 32$, the temporal patch length is 8 samples, and the temporal Event encoder contains two Transformer layers. The momentum teacher uses a coefficient of 0.99. The frequency-aware front-end models five bands: 0.5–4 Hz, 4–8 Hz, 8–12 Hz, 12–30 Hz, and 30–45 Hz. When a benchmark does not provide anatomically or functionally defined channel groups, individual EEG channels are treated as independent regions. The same representation architecture is used across benchmarks without dataset-specific architectural modifications.

C Additional Method Details

C.1 Context Isolation

During State-to-Event prediction, the contextual branch excludes the target Event's local temporal neighborhood to reduce trivial local information leakage. This mechanism is kept fixed across the main experiments and ablations and is treated as an implementation mechanism rather than an additional learning objective.

Context Isolation variants. For State-to-Event prediction, a contiguous temporal neighborhood around each target is blocked from the online context. Let g denote the isolation gap. The blocked set is denoted by $B(M, g)$, and Event tokens within this neighborhood are excluded from the prediction context.

We investigate two implementation variants that differ only in where this masking operation is applied. In the *original* variant, temporal Event encoding is performed first and Context Isolation is applied when constructing the State-to-Event context. In the *isolated* variant, the same blocked set is applied before temporal self-attention, preventing information from the target neighborhood from propagating through temporal refinement. Thus, the two variants isolate the effect of masking location while keeping the target sampling rule, teacher targets, and prediction objectives unchanged.

C.2 Perturbation Settings.

To evaluate robustness, we apply four controlled perturbations to the EEG test signals while keeping the trained model, labels, classifier, and test split unchanged. Let $\mathbf{x} \in \mathbb{R}^{C \times T}$ denote an EEG segment with C channels and T temporal samples.

Amplitude scaling is defined as

$$\tilde{\mathbf{x}}_{\text{amp}} = a\mathbf{x}, \qquad a \in \{0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}. \tag{20}$$

Non-circular temporal shifting is defined by

$$[\tilde{\mathbf{x}}_{\text{shift}}]_{c,t} = \begin{cases} \mathbf{x}_{c,t-\tau}, & 0 \le t-\tau < T, \\ 0, & \text{otherwise}, \end{cases} \qquad \tau \in \{-50, -25, 0, 25, 50\}. \tag{21}$$

Unlike circular shifting, samples shifted outside the temporal window are replaced with zeros.

DC offset perturbation is defined as

$$\tilde{\mathbf{x}}_{\text{DC}} = \mathbf{x} + b, \qquad b \in \{-10, -5, 0, 5, 10\}. \tag{22}$$

Additive Gaussian noise is defined as

$$\tilde{\mathbf{x}}_{\text{noise}} = \mathbf{x} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \sigma \in \{0, 0.05, 0.10, 0.15, 0.20\}. \tag{23}$$

For the combined perturbation condition, the four transformations are applied sequentially:

$$\tilde{\mathbf{x}}_{\text{combined}} = \mathcal{N}_\sigma\left(\mathcal{D}_b\left(\mathcal{S}_\tau\left(a\mathbf{x}\right)\right)\right), \tag{24}$$

where $a \sim \mathcal{U}(0.5, 2.0)$, $\tau \sim \mathcal{U}\{-50, \ldots, 50\}$, $b \sim \mathcal{U}(-10, 10)$, and $\sigma \sim \mathcal{U}(0, 0.2)$ are independently sampled for each test segment. All perturbations are generated deterministically using a fixed random seed of 2026.

“‘latex

C.3 Training Procedure

Pretraining Optimization. Before pretraining, the target encoder is initialized with the parameters of the online encoder and subsequently updated without backpropagation. For each mini-batch, the online encoder processes a stochastic view produced by Event- and State-level masking, whereas the target encoder processes the corresponding complete EEG segment without stochastic masking. The target representations are detached from the computational graph. The total pretraining objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{global}}\mathcal{L}_{\text{global}} + \lambda_{S\to E}\mathcal{L}_{S\to E} + \lambda_{E\to S}\mathcal{L}_{E\to S}, \tag{25}$$

where all four loss weights are set to 1.0 in the full model. The online encoder and prediction modules are optimized using RAdam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The target encoder is excluded from gradient-based optimization. After each optimizer step, its parameters are updated by exponential moving average:

$$\xi \leftarrow m\xi + (1-m)\theta, \tag{26}$$

where θ and ξ denote the online and target parameters, respectively, and $m = 0.99$. We use a constant learning rate without warm-up or learning-rate scheduling. Unless otherwise stated, pretraining is conducted for 100 epochs. The local Event, global State, State-to-Event, and Event-to-State losses are recorded separately during training.

Linear-Probing Evaluation. After pretraining, the online encoder is used to extract deterministic downstream representations. Event tokens are mean-pooled over regions and temporal positions, concatenated with the Global State representation, and normalized using LayerNorm. The encoder remains frozen, and only a logistic-regression classifier is fitted using the labeled training representations. Before classification, each representation dimension is standardized using statistics computed from the training representations. The classifier uses a one-versus-rest formulation and is optimized until convergence. The momentum encoder and all self-supervised prediction modules are excluded from linear-probing inference.

Full Fine-Tuning. For full fine-tuning, the pretrained online encoder and downstream classification head are optimized jointly using RAdam. We employ differential learning rates to allow the newly initialized downstream head to adapt faster than the pretrained encoder. Specifically, the encoder uses learning rate η_{ft} and weight decay ω_{ft}, while the representation-normalization layer and classification head use learning rate

$$\eta_{\mathrm{head}} = \gamma \eta_{\mathrm{ft}}, \tag{27}$$

with weight decay $0.1\omega_{\mathrm{ft}}$, where γ denotes the head learning-rate multiplier. Gradients are clipped to a maximum ℓ_2 norm of 4.0. The momentum encoder is frozen and is not used to produce downstream predictions.

Checkpoint and Threshold Selection. The fine-tuned checkpoint is selected exclusively according to performance on the validation set. AUROC is used as the selection metric for binary benchmarks, whereas balanced accuracy is used for TUEV. For binary classification, the decision threshold is determined after training using the validation set. We search 181 uniformly spaced threshold values from 0.05 to 0.95 and select the threshold that maximizes validation balanced accuracy. The selected checkpoint and threshold are then applied to the held-out test set without further adjustment. All reported results are averaged over five independent random seeds, and variability is reported using the sample standard deviation. "'

D Reproducibility and Statistical Reporting

All reported results are obtained using five independent random seeds. We report the mean and sample standard deviation across these runs. Dataset splitting is performed at the subject or recording level before temporal windowing whenever such grouping information is available, preventing windows from the same underlying recording from being distributed across training and evaluation sets.

For downstream experiments, hyperparameters and classification thresholds are determined using the validation set only. The test set is not used for model selection. For each method, the corresponding pretrained encoder is evaluated under the same downstream protocol, and the self-supervised prediction heads are removed before downstream classification.

Reporting Convention. All numerical results are reported with two decimal places. '$\pm$" denotes the sample standard deviation over five independent runs. Missing evaluations are denoted by '–" and are not treated as zero or excluded through imputation. For the main tables, comparisons are made independently for each dataset and metric because the benchmarks have different task definitions and class distributions.