NEURODETR: MULTISCALE DISCRETE SPATIOTEMPORAL TOKENIZATION FOR ASYNCHRONOUS EEG EVENT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Asynchronous EEG event detection supports practical brain-computer interfaces by identifying event classes and temporal boundaries directly from continuous recordings without predefined event boundaries at inference time. The primary challenge lies in extracting event-related patterns embedded in ongoing background activity. To address this, we propose **NeuroDETR**, a discrete spatiotemporal multiscale framework that combines structured tokenization with event-aware feature refinement. Multiscale temporal codebooks capture brief responses and sustained dynamics, while Spatial-Pyramid VQ hierarchically encodes spatial structure across EEG channels. These complementary representations form a compact event memory for jointly classifying events and estimating their onsets and offsets. Before tokenization, we apply FOCAL, an event-aware sparse attention module that uses predicted event presence to guide temporal feature refinement and spatial aggregation, emphasizing likely event periods while preserving context across the recording. Experiments on eight EEG datasets that span diverse paradigms demonstrate state-of-the-art performance. These findings provide a basis for the development of BCI systems that recognize meaningful neural activity during continuous interaction and monitoring.

1 INTRODUCTION

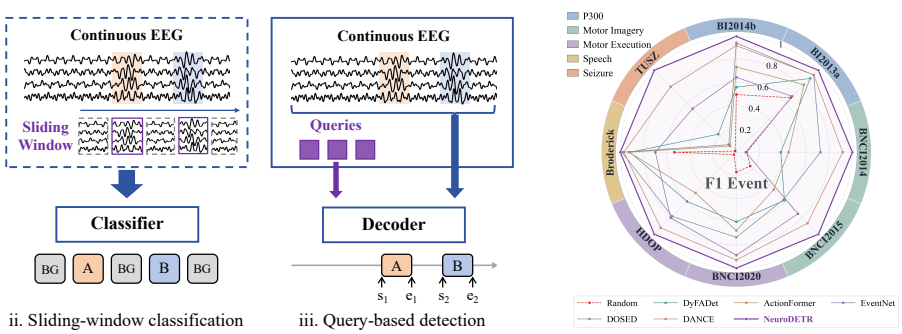
Brain–computer interfaces (BCIs) establish direct communication pathways between neural activity and external devices by translating neural signals into actionable information for device control, human–computer interaction, and clinical decision-making (Tang et al., 2023; Robinson et al., 2025). Achieving this goal depends on reliably identifying behaviorally or physiologically meaningful neural events in continuous electroencephalographic (EEG) recordings. Asynchronous EEG decoding addresses this challenge by jointly identifying event categories and localizing their onset and offset within the ongoing signal.

As shown in Figure 1a (i), a straightforward approach to asynchronous decoding applies sliding-window classifiers to continuous EEG, assigning each window an event or background label (Wu et al., 2024; Lopes-Dias et al., 2019). However, the mismatch between fixed-length windows and variable event durations and boundaries can fragment detections or merge adjacent events, making performance sensitive to window settings and post-processing rules (Seeuws et al., 2024). As illustrated in Figure 1a (ii), query-based detection addresses this mismatch by directly mapping continuous EEG to an unordered set of events, each characterized by its category, onset, and offset.

The effectiveness of this approach depends on how well the EEG features supplied to the decoder preserve event-specific information. Existing methods use cross-attention to aggregate information from long feature sequences into a fixed set of continuous latent tokens, but do not explicitly organize event cues across temporal scales and spatial regions (Jaegle et al., 2021; Lévy et al., 2026). Across EEG datasets and paradigms, events vary in duration, occurrence frequency, and spatial distribution. Their responses are also embedded in ongoing background activity, requiring feature aggregation to preserve weak or transient event cues.

In this study, we propose NeuroDETR, a multiscale discrete spatiotemporal framework for asynchronous EEG event detection. NeuroDETR uses vector quantization to organize continuous fea-

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.



(a) EEG decoding paradigms (b) Relative performance

Figure 1: EEG decoding paradigms and relative performance. (a) Sliding-window classification and query-based event detection. (b) Performance of NeuroDETR and six baselines across eight datasets. Scores on each dataset are normalized by NeuroDETR's score (NeuroDETR = 1).

tures into reusable learned codewords. To accommodate differences across datasets and paradigms, multiscale temporal VQ represents both brief responses and sustained dynamics, while Spatial-Pyramid VQ encodes channel information at complementary spatial levels. These representations form a compact event memory that supports joint localization and classification within a unified detection architecture.

To guide this representation toward event-related activity, we introduce FOCAL, an event-aware sparse attention module. FOCAL uses predicted event presence to guide temporal feature refinement before quantization and temporal pooling within the spatial branch, while retaining cross-window context. This shared event guidance connects the temporal and spatial representations, supplying the DETR decoder with event-aware multiscale information for predicting event classes and boundaries. We evaluate NeuroDETR on eight EEG datasets spanning P300, motor imagery, motor execution, speech, and seizure detection. NeuroDETR achieves the highest event-level F1 among the evaluated methods on all eight datasets, demonstrating its effectiveness across EEG paradigms with diverse event characteristics. By improving neural event detection in continuous EEG, NeuroDETR supports the development of practical BCI applications.

Our main contributions are summarized as follows:

- We propose **NeuroDETR**, which jointly predicts event classes, onsets, and offsets from continuous EEG without predefined event boundaries. At its core, vector quantization organizes EEG features into a compact event memory, with multiscale temporal and hierarchical spatial codebooks accommodating diverse event characteristics across paradigms.
- We introduce **FOCAL**, an event-aware sparse attention module that concentrates local feature refinement on likely event periods while preserving cross-window context. By sharing event-presence estimates with spatial aggregation, FOCAL directs both temporal and spatial tokenization toward event-relevant information.
- Experiments on eight EEG datasets spanning diverse paradigms demonstrate state-of-the-art performance. Ablation studies support the complementary contributions of spatiotemporal tokenization and event-aware refinement.

2 Related Work

2.1 Asynchronous EEG Decoding

Asynchronous EEG decoding identifies events in continuous signals without known onset times. Early systems estimate control states or classify sliding windows, using probabilistic aggregation or temporal decision rules (Zhang et al., 2008; Lopes-Dias et al., 2019; Wu et al., 2024). Recovering event instances typically requires temporal post-processing. DOSED and EventNet directly localize events through single-stage detection and center–duration regression, respectively (Chambon et al., 2019; Seeuws et al., 2024). DETRtime uses event queries and bipartite matching for set

prediction (Wolf et al., 2022), while a frozen foundation model has also been adapted for dense EEG segmentation (Ma & Ruotsalo, 2026). DANCE combines Perceiver-style resampling (Jaegle et al., 2021) with query decoding to predict event classes and temporal boundaries (Lévy et al., 2026).

2.2 Discrete Spatiotemporal EEG Representations

Discrete EEG tokens support EEG-to-text translation in DeWave (Duan et al., 2023) and pretraining or multimodal decoding in LaBraM, VQ-MTM, and NeuroLM (**?**Gui et al., 2024; Jiang et al., 2025). Recent tokenizers learn single-channel time–frequency motifs and decoupled temporal–spectral vocabularies (Pradeepkumar et al., 2026; Ma et al., 2026). Related approaches model EEG through electrode graphs, multiscale temporal filters, and regional and global spatial representations (Ding et al., 2023; Gao et al., 2024), or use sparse attention for motor imagery decoding (Liu et al., 2024; Miao et al., 2026). THD-BAR constructs hierarchical spatial tokens for autoregressive learning (Yang et al., 2025), while NeuroRVQ combines multiscale temporal convolutions with residual vector quantization for biosignal reconstruction and foundation-model training (Barmpas et al., 2026).

NeuroDETR uses shared event-presence estimates to route local temporal attention and weight temporal pooling of regional features before multiscale temporal and hierarchical spatial quantization. The resulting memory supports event-set prediction.

3 Method

3.1 Problem Formulation and Overview

We study asynchronous event detection from continuous EEG. Given $\mathbf{X} \in \mathbb{R}^{C\times T}$ with C channels and $T = f_s W$ samples, where f_s is the sampling rate and W the window duration, the target is an unordered event set:

$$\mathcal{Y} = \{(c_i, s_i, e_i)\}_{i=1}^{M}, \quad c_i \in \{1, \dots, K\}, \quad 0 \le s_i < e_i \le 1. \tag{1}$$

Here, M varies across windows, and boundaries are normalized by W. NeuroDETR adopts a DETR-style detection framework to predict events in parallel using N_e learnable event queries, with class 0 denoting no event.

As shown in Figure 2, a convolutional backbone extracts dense features, which FOCAL refines through event-guided local attention and cross-window interaction. Temporal VQ encodes dynamics at multiple resolutions, while spatial VQ retains regional electrode information before channel mixing. Query-based fusion combines both memories for event localization and classification (Carion et al., 2020).

3.2 Dense EEG Encoding

We adapt SimpleConv from Défossez et al. (Défossez et al., 2023). Spatial attention uses Fourier embeddings of normalized electrode coordinates $\mathbf{P} \in \mathbb{R}^{C\times 2}$ to mix the inputs into 270 virtual channels. A pointwise projection and ten dilated temporal convolutional stages yield

$$\mathbf{H} = f_{\text{enc}}(\mathbf{X}, \mathbf{P}) \in \mathbb{R}^{T\times D}, \qquad D = 128. \tag{2}$$

The backbone preserves temporal resolution, supplying features for FOCAL and auxiliary boundary supervision. Architectural details appear in Appendix E.

3.3 FOCAL: Event-Aware Sparse Attention

FOCAL combines event-guided local attention with axial interaction across windows.

Event-guided routing. We partition $\mathbf{H}$ into $N_w = \lceil T/w \rceil$ windows of w tokens, padding the last window with zeros. A density head g_ϕ, comprising layer normalization, two linear layers, and GELU, predicts event presence from mean-pooled features:

$$\ell_n = g_\phi\left(\frac{1}{w}\sum_{t\in\mathcal{W}_n} \mathbf{h}_t\right), \qquad p_n = \sigma(\ell_n). \tag{3}$$

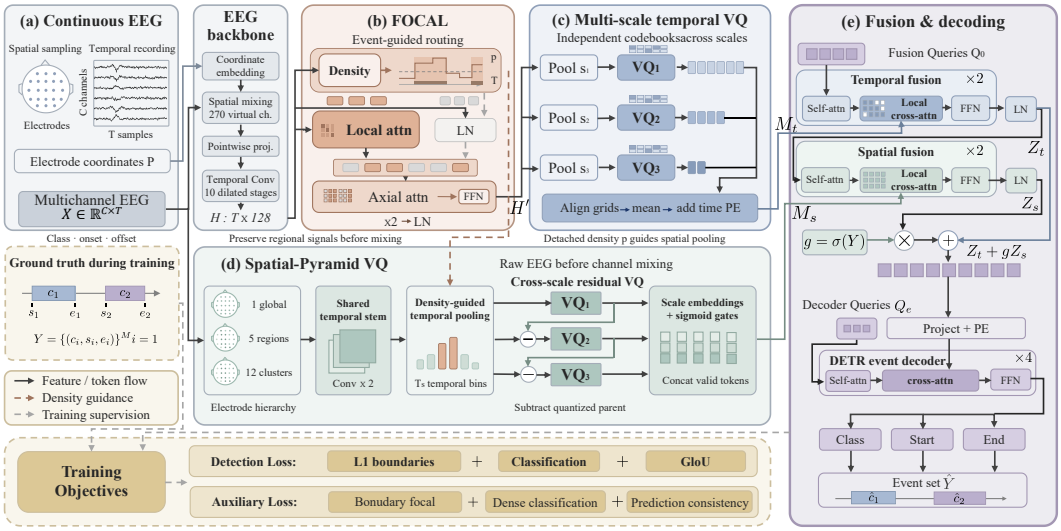


Figure 2: Overview of NeuroDETR. **(a)** Continuous EEG input. **(b)** FOCAL uses event-presence estimates to guide local attention. **(c)** Temporal features are quantized at multiple resolutions. **(d)** Spatial-Pyramid VQ quantizes EEG features at global, regional, and local spatial levels. **(e)** A query decoder predicts event classes and boundaries from fused spatiotemporal memory. Detection and auxiliary losses provide training supervision.

Window targets indicate overlap with an annotated event:

$$y_n = \mathbb{1}\left[\exists i:\ e_i > \frac{n}{N_w},\ s_i < \frac{n+1}{N_w}\right], \quad n = 0, \ldots, N_w - 1. \tag{4}$$

We train g_ϕ with binary cross-entropy:

$$\mathcal{L}_{\text{den}} = -\frac{1}{N_w} \sum_{n=0}^{N_w-1} \left[y_n \log p_n + (1 - y_n) \log(1 - p_n) \right]. \tag{5}$$

Windows are selected for local attention by

$$\mathcal{I}_{\text{evt}} = \{n : p_n > \tau\} \cup \text{TopK}(\boldsymbol{\ell}, k_{\min}), \quad m_n = \mathbb{1}[n \in \mathcal{I}_{\text{evt}}]. \tag{6}$$

The default uses $\tau = 0.4$ and $k_{\min} = 1$. The density map and routing mask are computed once and shared across both FOCAL blocks. Hard routing leaves g_ϕ supervised by $\mathcal{L}_{\text{den}}$.

Local–axial interaction. Selected windows receive local self-attention, and the others receive background normalization:

$$\widetilde{\mathbf{H}}_n = \begin{cases} \mathbf{H}_n + \text{MHSA}(\text{LN}(\mathbf{H}_n)), & m_n = 1, \\ \text{LN}_{\text{bg}}(\mathbf{H}_n), & m_n = 0, \end{cases} \tag{7}$$

For $\mathbf{H}_n \in \mathbb{R}^{w \times D}$, axial attention (Ho et al., 2019) connects corresponding token positions across all windows by reshaping

$$[B, N_w, w, D] \longrightarrow [Bw, N_w, D], \tag{8}$$

where B is the batch size. Each block applies axial attention and a feed-forward network with pre-normalization and residual connections. Two blocks, final normalization, and padding removal yield $\mathbf{H}' \in \mathbb{R}^{T \times D}$ for temporal VQ.

3.4 Spatiotemporal VQ Tokenization

Temporal and spatial branches use independent, scale-specific codebooks to encode temporal dynamics and channel-group structure.

Quantization and codebook learning. Following VQ representation learning (van den Oord et al., 2017), we assign $\mathbf{z}$ to its nearest codeword in $\mathcal{E} = \{\mathbf{e}_j\}_{j=1}^{V}$:

$$k^\star = \arg\min_j \|\mathbf{z} - \mathbf{e}_j\|_2^2, \qquad Q(\mathbf{z}) = \mathbf{z} + \mathrm{sg}[\mathbf{e}_{k^\star} - \mathbf{z}]. \tag{9}$$

$\mathrm{sg}[\cdot]$ denotes stop-gradient, so Q returns the codeword and passes gradients to $\mathbf{z}$. Each quantizer incurs a commitment penalty $\beta\,\mathrm{MSE}(\mathbf{z}, \mathrm{sg}[\mathbf{e}_{k^\star}])$. Codebooks are initialized from features and updated by exponential moving averages (EMA), with underused codes reinitialized. Commitment losses are averaged across scales within each branch, using only valid tokens for spatial VQ.

3.4.1 Multi-Scale Temporal VQ

At each scale $s \in \mathcal{S}_\mathrm{t}$, average pooling with kernel and stride s retains $\lceil T/s \rceil$ tokens, including the final partial bin:

$$\mathbf{H}_s = \mathrm{AvgPool}_s(\mathbf{H}'), \qquad \mathbf{q}_s^\mathrm{t} = Q^{\mathrm{t},s}(\mathbf{H}_s). \tag{10}$$

We linearly interpolate the codes to the finest pooled grid, $L_\mathrm{m} = \max_s \lceil T/s \rceil$, and average them before adding sinusoidal positional embeddings:

$$\mathbf{M}^\mathrm{t} = \frac{1}{|\mathcal{S}_\mathrm{t}|} \sum_{s \in \mathcal{S}_\mathrm{t}} \mathrm{Up}_{L_\mathrm{m}}(\mathbf{q}_s^\mathrm{t}) + \mathrm{PE}_{L_\mathrm{m}}\,. \tag{11}$$

A two-layer MLP r_ψ reconstructs dense features from the averaged codes resampled to length T:

$$\mathcal{L}_\mathrm{rec} = \mathrm{MSE}\left(r_\psi\left[\frac{1}{|\mathcal{S}_\mathrm{t}|} \sum_{s \in \mathcal{S}_\mathrm{t}} \mathrm{Up}_T(\mathbf{q}_s^\mathrm{t}) \right], \mathrm{sg}[\mathbf{H}'] \right). \tag{12}$$

Reconstruction uses codes before positional embedding and memory fusion.

3.4.2 Spatial-Pyramid VQ

Spatial-Pyramid VQ aggregates raw EEG channels before backbone mixing at three levels: one whole-head group, five coarse groups (frontal, temporal, central, parietal, and occipital), and twelve fine electrode clusters:

$$\mathbf{R}_\ell = \mathbf{A}_\ell \mathbf{X} \in \mathbb{R}^{N_\ell \times T}, \qquad (N_1, N_2, N_3) = (1, 5, 12). \tag{13}$$

Each nonempty row of $\mathbf{A}_\ell$ averages available channels assigned by channel name. A shared stem with two strided temporal convolutions and GELU produces $\mathbf{F}_{\ell,r,u} \in \mathbb{R}^D$, indexed by region r and stem time step u. Details of the hierarchical electrode grouping are provided in Appendix B.

Density-guided pooling. FOCAL probabilities $\mathbf{d} = \mathrm{sg}[\sigma(\boldsymbol{\ell})]$ are linearly interpolated to the stem grid as $\widetilde{d}_u$ and weight each group's features within T_s temporal bins:

$$\mathbf{E}_{\ell,r,b} = \frac{\sum_u \kappa(b,u)\widetilde{d}_u \mathbf{F}_{\ell,r,u}}{\max\left(\sum_u \kappa(b,u)\widetilde{d}_u, \epsilon\right)}. \tag{14}$$

Here, b indexes output bins, κ is a row-normalized triangular kernel centered on each bin, and $\epsilon = 10^{-6}$. The detached weights emphasize likely event periods without updating the density estimator.

Cross-scale residual quantization. The whole-head descriptor is quantized directly, while each finer level quantizes its residual relative to its parent's code:

$$\mathbf{u}_\ell = \begin{cases} \mathbf{E}_1, & \ell = 1, \\ \mathbf{E}_\ell - \mathrm{Broadcast}_\ell(\mathbf{q}_{\ell-1}^\mathrm{s}), & \ell > 1, \end{cases} \qquad \mathbf{q}_\ell^\mathrm{s} = Q^{\mathrm{s},\ell}(\mathbf{u}_\ell). \tag{15}$$

$\mathrm{Broadcast}_\ell$ assigns each child the immediately preceding level's quantized parent output at the same temporal bin. Learnable scale embeddings $\mathbf{a}_\ell$ and sigmoid gates $g_\ell = \sigma(\eta_\ell)$ form the spatial memory:

$$\mathbf{M}^\mathrm{s} = \Big\|_{\ell=1}^{3} \mathrm{Flatten}_\mathrm{valid}[g_\ell(\mathbf{q}_\ell^\mathrm{s} + \mathbf{a}_\ell)]\,. \tag{16}$$

$\mathrm{Flatten}_\mathrm{valid}$ retains all temporal bins of nonempty channel groups.

3.5 Event Memory Fusion and Set Prediction

Temporal fusion and gated spatial integration. Starting from N_q learnable queries initialized with sinusoidal embeddings, temporal fusion alternates global self-attention, sliding-window cross-attention (SWA), and feed-forward refinement:

$$\mathbf{Q} \leftarrow \mathbf{Q} + \mathrm{MHSA}(\mathrm{LN}(\mathbf{Q})), \quad \mathbf{Q} \leftarrow \mathbf{Q} + \mathrm{SWA}(\mathrm{LN}(\mathbf{Q}), \mathbf{M}^{\mathrm{t}}), \quad \mathbf{Q} \leftarrow \mathbf{Q} + \mathrm{FFN}(\mathrm{LN}(\mathbf{Q})). \tag{17}$$

Cross-attention spans are centered proportionally along the memory. The normalized output $\mathbf{Z}_{\mathrm{t}} = \mathrm{LN}[\mathcal{F}_{\mathrm{t}}(\mathbf{Q}_0, \mathbf{M}^{\mathrm{t}})]$ initializes an independently parameterized spatial Transformer with the same structure, whose output enters through a gated residual:

$$\mathbf{Z} = \mathbf{Z}_{\mathrm{t}} + g\,\mathrm{LN}[\mathcal{F}_{\mathrm{s}}(\mathbf{Z}_{\mathrm{t}}, \mathbf{M}^{\mathrm{s}})], \qquad g = \sigma(\gamma). \tag{18}$$

Each stream has two layers and four heads, yielding memory $\mathbf{Z} \in \mathbb{R}^{N_q \times D}$. The spatial gate is initialized at $g = 0.02$.

Set-based event decoding. A four-layer, four-head DETR decoder (Carion et al., 2020) applies N_e event queries to the fused memory, projected from 128 to 256 dimensions and augmented with sinusoidal embeddings. Query features $\mathbf{o}_j$ predict classes and boundaries:

$$\widehat{\mathbf{p}}_j = \mathrm{softmax}(\mathrm{Cls}(\mathbf{o}_j)), \quad \hat{s}_j = \sigma(\mathrm{Start}(\mathbf{o}_j)), \quad \hat{e}_j = \sigma(\mathrm{End}(\mathbf{o}_j)). \tag{19}$$

We discard queries whose highest-probability class is background and multiply the remaining boundaries by W. Here, N_q controls memory length and N_e the candidate-event budget.

3.6 Training Objectives

Detection Loss. Hungarian matching (Kuhn, 1955) assigns ground-truth events to predicted queries using a matching cost based on class probabilities and temporal IoU. The classification loss $\mathcal{L}_{\mathrm{cls}}$ applies cross-entropy supervision to all queries, assigning event labels to matched queries and the background label to unmatched queries. For matched queries, endpoint regression loss $\mathcal{L}_{\ell_1}$ penalizes absolute onset and offset errors. Temporal generalized IoU loss $\mathcal{L}_{\mathrm{GIoU}}$ (Rezatofighi et al., 2019) complements endpoint regression by supervising interval overlap and relative position. The detection objective is

$$\mathcal{L}_{\mathrm{det}} = \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{giou}}\mathcal{L}_{\mathrm{GIoU}}. \tag{20}$$

Auxiliary Loss. We complement set prediction with full-resolution boundary supervision, dense classification, and prediction consistency. A boundary head applied to the backbone features $\mathbf{H}$ predicts class-agnostic onset and offset heatmaps. The targets are pointwise maxima of Gaussians centered at annotated boundaries. Sigmoid focal loss $\mathcal{L}_{\mathrm{bnd}}$ supervises these predictions while down-weighting easy non-boundary positions.

For dense classification and prediction consistency, we follow the auxiliary-supervision strategy of DANCE (Lévy et al., 2026). Dense cross-entropy loss $\mathcal{L}_{\mathrm{dense}}$ supervises class predictions at each position of the fused memory $\mathbf{Z}$. To construct the set-derived predictions, we accumulate query class probabilities over their discretized predicted intervals and normalize the accumulated scores at each covered grid position. Uncovered positions are assigned zero vectors. We apply componentwise clamping to both predictions without renormalization and compute a KL-style regularizer $\mathcal{L}_{\mathrm{cons}}$. At covered grid positions, this term encourages agreement between dense and set-derived predictions. At uncovered positions, the clamped zero vectors induce an entropy regularization effect on the dense predictions.

The complete event-supervision objective is

$$\mathcal{L}_{\mathrm{event}} = \boldsymbol{\mathcal{L}_{\mathbf{det}}} + \lambda_{\mathrm{bnd}}\mathcal{L}_{\mathrm{bnd}} + \lambda_{\mathrm{dense}}\mathcal{L}_{\mathrm{dense}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}. \tag{21}$$

More details on the training objectives are provided in the appendix.

4 Experiments

4.1 Datasets

We evaluate our method on eight EEG datasets spanning five task domains with diverse temporal characteristics. These datasets cover densely occurring and potentially overlapping P300 events

Table 1: Results on eight EEG datasets: mean ± SD over five cross-validation folds, except TUSZ, which uses a single result on its predefined split. Best and second-best scores are **bold** and underlined, respectively.

	Method	P300		Motor Imagery		Motor Execution		Speech	Seizure
		BI2014b	BI2013a	BNCI2014	BNCI2015	BNCI2020	HDOP	Broderick	TUSZ
F1 Sample	Random	$35.46_{\pm 0.60}$	$40.40_{\pm 0.36}$	$20.06_{\pm 0.54}$	$33.75_{\pm 1.12}$	$33.50_{\pm 0.50}$	$25.11_{\pm 0.39}$	$48.94_{\pm 0.15}$	19.67
	DyFADet	$34.84_{\pm 9.46}$	$\underline{55.08}_{\pm 2.22}$	$34.58_{\pm 20.67}$	$63.68_{\pm 3.52}$	$45.99_{\pm 22.05}$	$51.40_{\pm 1.23}$	$47.78_{\pm 2.66}$	25.60
	ActionFormer	$52.18_{\pm 1.65}$	$54.71_{\pm 2.29}$	$42.38_{\pm 4.85}$	$53.04_{\pm 1.29}$	$53.50_{\pm 2.16}$	$46.63_{\pm 1.80}$	$42.75_{\pm 1.17}$	22.56
	EventNet	$44.04_{\pm 2.74}$	$51.61_{\pm 5.88}$	$\underline{53.64}_{\pm 3.50}$	$\underline{73.94}_{\pm 4.02}$	$\mathbf{59.91}_{\pm 2.94}$	$57.52_{\pm 3.82}$	$\underline{52.18}_{\pm 0.61}$	24.52
	DOSED	$38.56_{\pm 2.30}$	$44.75_{\pm 3.01}$	$46.06_{\pm 2.98}$	$60.85_{\pm 3.55}$	$55.36_{\pm 2.93}$	$\underline{64.15}_{\pm 1.69}$	$43.30_{\pm 0.44}$	18.27
	DANCE	$\underline{56.24}_{\pm 3.04}$	$54.73_{\pm 2.41}$	$53.46_{\pm 6.44}$	$73.36_{\pm 2.54}$	$55.79_{\pm 2.56}$	$61.74_{\pm 3.37}$	$50.31_{\pm 2.71}$	$\underline{27.78}$
	NeuroDETR	$\mathbf{59.61}_{\pm 3.12}$	$\mathbf{57.48}_{\pm 4.02}$	$\mathbf{55.63}_{\pm 5.17}$	$\mathbf{75.82}_{\pm 3.30}$	$\underline{59.05}_{\pm 2.80}$	$\mathbf{66.32}_{\pm 3.35}$	$\mathbf{53.80}_{\pm 0.68}$	**32.26**
F1 Event	Random	$22.58_{\pm 0.58}$	$37.55_{\pm 0.50}$	$4.05_{\pm 0.56}$	$11.44_{\pm 0.76}$	$7.21_{\pm 0.54}$	$1.51_{\pm 0.55}$	$20.69_{\pm 0.06}$	0.35
	DyFADet	$25.38_{\pm 2.58}$	$50.86_{\pm 1.86}$	$17.40_{\pm 15.51}$	$39.76_{\pm 3.95}$	$25.10_{\pm 12.94}$	$31.80_{\pm 1.14}$	$35.49_{\pm 0.43}$	5.51
	ActionFormer	$33.29_{\pm 1.72}$	$46.53_{\pm 1.34}$	$20.47_{\pm 6.14}$	$30.88_{\pm 2.11}$	$28.28_{\pm 1.92}$	$26.30_{\pm 1.30}$	$35.86_{\pm 0.28}$	1.89
	EventNet	$29.22_{\pm 0.59}$	$38.52_{\pm 1.44}$	$30.67_{\pm 4.63}$	$50.86_{\pm 5.24}$	$37.21_{\pm 2.21}$	$42.34_{\pm 4.42}$	$27.05_{\pm 0.76}$	13.33
	DOSED	$\underline{42.62}_{\pm 2.64}$	$\underline{53.28}_{\pm 0.88}$	$32.92_{\pm 3.75}$	$39.08_{\pm 0.93}$	$30.71_{\pm 2.98}$	$41.37_{\pm 2.16}$	$\underline{38.22}_{\pm 0.5}$	2.31
	DANCE	$41.63_{\pm 2.12}$	$53.26_{\pm 0.75}$	$\underline{41.65}_{\pm 6.96}$	$\underline{58.78}_{\pm 3.37}$	$\underline{38.92}_{\pm 3.64}$	$\underline{47.48}_{\pm 4.46}$	$37.55_{\pm 0.83}$	$\underline{20.03}$
	NeuroDETR	$\mathbf{45.23}_{\pm 1.22}$	$\mathbf{56.50}_{\pm 0.97}$	$\mathbf{45.55}_{\pm 4.96}$	$\mathbf{67.77}_{\pm 3.99}$	$\mathbf{41.85}_{\pm 3.83}$	$\mathbf{52.91}_{\pm 3.33}$	$\mathbf{38.76}_{\pm 0.57}$	**25.06**

(BI2013 (Congedo et al., 2011) and BI2014b (Korczowski et al., 2019)), fine-grained high-frequency speech-related events (Broderick (Broderick et al., 2018; 2019)), sparse and long-duration motor imagery events (BNCI2014 (Tangermann et al., 2012) and BNCI2015 (Faller et al., 2012)), motor execution events (BNCI2020 (Schwarz et al., 2020) and HDOP (Lomele et al., 2026)), and rare epileptic seizures with durations ranging from several seconds to several minutes (TUSZ (v2.0.6) Shah et al. (2018)). This diversity enables a comprehensive evaluation of whether a unified set-prediction framework can generalize across variations in event frequency, duration, category complexity, and recording configurations. Further details on the datasets are provided in the Appendix **??**.

4.2 Baselines

We compare NeuroDETR with four dense temporal detection frameworks, EventNet (Seeuws et al., 2024), ActionFormer (Zhang et al., 2022), DyFADet (Yang et al., 2024), and DOSED (Chambon et al., 2019), and the query-based set-prediction framework DANCE (Lévy et al., 2026). We also include a dataset-aware Random event generator as a lower bound. For each learned baseline, we adapt the input representation, classification head, and temporal decoding procedure to EEG event detection while retaining its original detection mechanism. All methods use the same data splits and event-level evaluation protocol. For all dense detectors, we apply a confidence threshold of 0.05 and non-maximum suppression (NMS) with a temporal IoU threshold of 0.5, retaining at most N_e predictions per window. Further implementation details appear in Appendix D.

4.3 Evaluation Metrics

We report $F1_{\text{sample}}$ and $F1_{\text{event}}$. The former measures sample-wise classification quality through macro-averaged F1, while the latter jointly evaluates event detection, classification, and localization. Under one-to-one matching, a predicted event counts as a true positive when

$$\text{IoU}(\hat{I}_j, I_i) > \tau_{\text{IoU}} \quad \text{and} \quad \hat{c}_j = c_i, \tag{22}$$

where $\hat{I}_j$ and I_i denote the predicted and ground-truth intervals, and $\hat{c}_j$ and c_i their respective classes. We set $\tau_{\text{IoU}} = 0.5$ for the main evaluation. The two metrics are computed as

$$\begin{aligned} F1_{\text{sample}} &= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2TP_c^{\text{s}}}{2TP_c^{\text{s}} + FP_c^{\text{s}} + FN_c^{\text{s}}}, \\ F1_{\text{event}} &= \frac{2TP^{\text{e}}}{2TP^{\text{e}} + FP^{\text{e}} + FN^{\text{e}}}. \end{aligned} \tag{23}$$

Here, $\mathcal{C}$ is the evaluated class set, and superscripts s and e denote counts over time samples and event instances, respectively.

Table 2: Core component ablations: $F1_{\text{event}}$ on eight EEG datasets (mean $\pm$ SD over five cross-validation folds; TUSZ: a single result on its predefined split). Best and second-best scores are **bold** and underlined, respectively.

Variant	P300		Motor Imagery		Motor Execution		Speech	Seizure
	BI2014b	BI2013a	BNCI2014	BNCI2015	BNCI2020	HDOP	Broderick	TUSZ
M1	$43.03_{\pm 1.75}$	$52.77_{\pm 1.22}$	$43.11_{\pm 6.70}$	$65.17_{\pm 5.33}$	$39.49_{\pm 3.13}$	$48.12_{\pm 3.84}$	$38.02_{\pm 0.43}$	$\underline{21.32}$
M2	$43.88_{\pm 2.46}$	$53.46_{\pm 0.99}$	$43.38_{\pm 5.52}$	$\underline{65.67}_{\pm 4.30}$	$\underline{40.92}_{\pm 4.04}$	$48.87_{\pm 5.05}$	$\underline{38.73}_{\pm 0.57}$	18.96
M3	$\underline{44.54}_{\pm 2.32}$	$\underline{53.89}_{\pm 0.86}$	$\underline{43.72}_{\pm 7.79}$	$65.44_{\pm 3.02}$	$40.57_{\pm 7.63}$	$\underline{48.97}_{\pm 3.17}$	$38.71_{\pm 0.43}$	19.69
M4	$43.46_{\pm 2.43}$	$52.44_{\pm 1.22}$	$42.39_{\pm 7.18}$	$64.93_{\pm 4.11}$	$40.49_{\pm 3.06}$	$47.96_{\pm 3.22}$	$38.12_{\pm 0.62}$	16.30
M5	$42.83_{\pm 2.08}$	$52.25_{\pm 1.61}$	$42.83_{\pm 5.19}$	$64.23_{\pm 4.53}$	$28.67_{\pm 1.29}$	$46.77_{\pm 3.38}$	$37.87_{\pm 0.66}$	7.98
Full	$\mathbf{45.23}_{\pm 1.22}$	$\mathbf{56.50}_{\pm 0.97}$	$\mathbf{45.55}_{\pm 4.96}$	$\mathbf{67.77}_{\pm 3.99}$	$\mathbf{41.85}_{\pm 3.83}$	$\mathbf{52.91}_{\pm 3.33}$	$\mathbf{38.76}_{\pm 0.57}$	**25.06**

Table 3: Quantization and FOCAL ablation: $F1_{\text{event}}$ (%, five-fold mean $\pm$ SD; TUSZ: a single result).

Configuration		P300		Motor Imagery		Motor Execution		Speech	Seizure
Quantization	FOCAL	BI2014b	BI2013a	BNCI2014	BNCI2015	BNCI2020	HDOP	Broderick	TUSZ
✗	✗	$42.02_{\pm 1.38}$	$50.64_{\pm 0.65}$	$39.44_{\pm 3.27}$	$63.05_{\pm 3.50}$	$36.73_{\pm 3.61}$	$46.34_{\pm 2.79}$	$37.15_{\pm 0.30}$	12.24
✓	✗	$43.50_{\pm 2.74}$	$51.11_{\pm 1.32}$	$41.23_{\pm 6.30}$	$64.15_{\pm 5.11}$	$37.28_{\pm 4.54}$	$46.50_{\pm 1.43}$	$\underline{37.98}_{\pm 0.44}$	16.32
✗	✓	$\underline{43.98}_{\pm 2.64}$	$\underline{51.99}_{\pm 1.33}$	$\underline{42.76}_{\pm 5.88}$	$\underline{65.34}_{\pm 2.88}$	$\underline{37.79}_{\pm 3.70}$	$\underline{47.98}_{\pm 3.04}$	$37.94_{\pm 0.51}$	$\underline{20.14}$
✓	✓	$\mathbf{45.23}_{\pm 1.22}$	$\mathbf{56.50}_{\pm 0.97}$	$\mathbf{45.55}_{\pm 4.96}$	$\mathbf{67.77}_{\pm 3.99}$	$\mathbf{41.85}_{\pm 3.83}$	$\mathbf{52.91}_{\pm 3.33}$	$\mathbf{38.76}_{\pm 0.57}$	**25.06**

4.4 Main Results

As shown in Table 1, NeuroDETR ranks first on most datasets in $F1_{\text{sample}}$. It achieves an average of 57.50, outperforming DANCE, the strongest baseline on average, by 3.32 percentage points and demonstrating consistently strong sample-wise discrimination.

For $F1_{\text{event}}$, NeuroDETR achieves the best performance on all eight datasets, averaging 46.70 and surpassing DANCE by 4.29 percentage points. Its improvement over the strongest baseline reaches 8.99 percentage points on BNCI2015, highlighting its effectiveness in jointly detecting, classifying, and localizing EEG events.

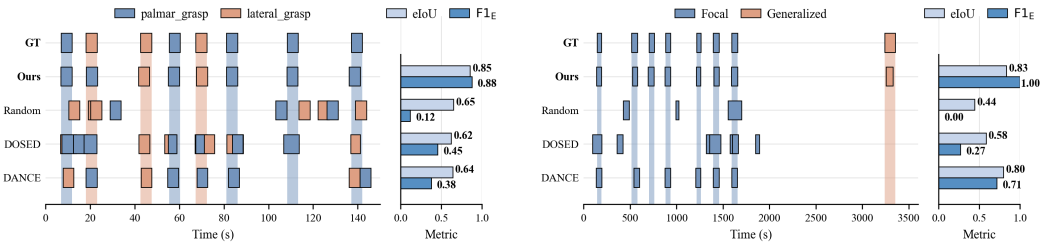
Some baselines achieve competitive sample-wise classification, yet this advantage does not consistently translate into better event detection. NeuroDETR achieves the best average performance on both metrics, combining strong sample-wise classification with effective event-instance recovery. Its consistent gains across P300, motor imagery, motor execution, speech, and seizure tasks support its broad applicability across EEG datasets and paradigms.

4.5 Ablation Studies

We evaluate the contributions of NeuroDETR's core components and examine the relationship between spatiotemporal quantization and event guidance.

Ablation of Core Components. We compare five variants with the full model. M1 and M2 replace the temporal and spatial quantized representations with continuous features, respectively. M3 retains only the finest temporal scale while preserving the memory grid and fusion output length. M4 retains only fine electrode clusters, removing higher spatial levels and cross-level residual connections. M5 removes FOCAL and event-presence supervision, feeds backbone features directly into temporal tokenization, and replaces event-guided spatial pooling with adaptive average pooling.

Table 2 shows that the full model outperforms all five variants in event F1 on each of the eight datasets. The largest drops under M5 occur on BNCI2020, HDOP, and TUSZ, reaching 13.18, 6.14, and 17.08 points, respectively. These datasets have event occupancies of 27.03%, 6.15%, and 3.18%, respectively. On Broderick, M5 reduces event F1 by 0.89 points. These results show that the combined contribution of FOCAL-associated processing and supervision varies across datasets, with larger gains in these low-occupancy settings. M4 also reduces event F1 by 4.95 points on HDOP and 8.76 points on TUSZ, supporting the contribution of the multilevel spatial representation and its cross-level residual connections.



(a) BNCI2020 (150 s)

(b) TUSZ (3600 s)

Figure 3: Temporal decoding examples from (a) BNCI2020 and (b) TUSZ. Colored intervals indicate event classes and temporal extents; GT denotes ground truth. Side panels report metrics computed on the displayed segments: eIoU is the mean temporal IoU over overlapping, same-class one-to-one matches, and event-level F1 ($F1_{\mathrm{E}}$) is evaluated at an IoU threshold of 0.5. N/A indicates that no matches are available for computing eIoU.

Quantization and Event Guidance. We evaluate continuous and quantized representations with and without event guidance in a 2×2 study. Continuous variants bypass both quantizers while preserving pooling, hierarchy, and fusion. Disabling guidance replaces event-based routing with event-independent window selection and uses uniform weights for regional feature pooling. The attention architecture, event-presence head, and its supervision are retained across configurations.

For each dataset, let $F_{q,g}$ denote the reported event F1, where $q, g \in \{0, 1\}$ indicate whether quantization and event guidance are enabled. We measure their interaction as

$$\Delta_{\mathrm{int}} = (F_{1,1} - F_{0,1}) - (F_{1,0} - F_{0,0}).$$

A positive value indicates quantization provides a larger gain with event guidance than without it.

Table 3 shows that the joint configuration achieves the highest event F1 on all eight datasets. Across the seven cross-validation datasets, the average quantization gain increases from 0.91 points without guidance to 2.97 points with guidance, yielding an average interaction contrast of 2.06 points. The interaction contrasts computed from the reported means are largest on HDOP (4.77 points), BI2013a (4.04), and BNCI2020 (3.51). BI2014b shows a negative contrast (-0.23), while Broderick is close to zero (-0.01). These point estimates indicate that event guidance increases the benefit of quantization on average, with interactions that vary across datasets.

4.6 Visualization of decoded events

Figure 3 compares NeuroDETR with Random, DOSED, and DANCE on a 150 s EEG segment from BNCI2020 and a 3600 s segment from TUSZ. The aligned timelines show predicted event classes and temporal boundaries alongside the ground-truth annotations, allowing inspection of boundary deviations, missed events, and false positives. The accompanying metrics summarize detection performance within each displayed segment.

5 Conclusion

We presented NeuroDETR, a framework for asynchronous EEG event detection that jointly predicts event classes, onsets, and offsets from continuous recordings. Multiscale temporal VQ and Spatial-Pyramid VQ organize temporal dynamics and hierarchical spatial information into a compact event memory. FOCAL uses predicted event presence to guide temporal feature refinement and spatial pooling while preserving cross-window context. Across eight datasets spanning P300, motor imagery, motor execution, speech, and seizure detection, NeuroDETR achieves the highest event-level F1 among the compared methods, improving the average over DANCE by 4.29 percentage points. Ablation studies support the complementary contributions of discrete spatiotemporal representations, multiscale structure, and event-aware refinement. These results demonstrate the value of organizing event-relevant information across temporal scales and spatial regions for joint event classification and localization in continuous EEG.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and language polishing, focusing on improving clarity and readability. We take full responsibility for the final content of this work, including all AI-assisted text. All AI-assisted text was reviewed and revised by the authors.

REPRODUCIBILITY STATEMENT

The main text describes the architecture and training objectives of NeuroDETR. Dataset statistics, input window lengths, and per-GPU batch sizes are summarized in Table 6. Section F specifies the preprocessing, data splits, model hyperparameters, optimization settings, random seed, and validation-based checkpoint selection procedure.

REFERENCES

Konstantinos Barmpas, Na Lee, Dimitrios Chalatsis, William Raftery, Yannis Panagakis, Dimitrios A. Adamos, Nikolaos Laskaris, et al. NeuroRVQ: Multi-scale biosignal tokenization for generative foundation models, 2026.

Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.*, 28(5):803–809.e3, 2018.

Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. Data from: Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Dryad, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pp. 213–229, 2020.

Stanislas Chambon, Valentin Thorey, Pierrick J. Arnal, Emmanuel Mignot, and Alexandre Gramfort. DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal. *J. Neurosci. Methods*, 321:64–78, 2019.

Marco Congedo, Matthieu Goyat, Nicolas Tarrin, Gelu Ionescu, Léo Varnet, Bertrand Rivet, Ronald Phlypo, et al. "Brain Invaders": A prototype of an open-source P300-based video game working with the OpenViBE platform. In *Int. BCI Conf.*, pp. 280–283, 2011.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.*, 5(10):1097–1107, 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 Trans. Affect. Comput.*, 14(3):2238–2250, 2023.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *NeurIPS*, volume 36, pp. 9907–9918, 2023.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 20(3):313–319, 2012.

Hongxiang Gao, Xingyao Wang, Zhenghua Chen, Min Wu, Zhipeng Cai, Lulu Zhao, Jianqing Li, et al. Graph convolutional network with connectivity uncertainty for EEG-based emotion recognition. *IEEE J. Biomed. Health Inform.*, 28(10):5917–5928, 2024.

Haokun Gui, Xiucheng Li, and Xinyang Chen. Vector quantization pretraining for EEG time series with random projection and phase alignment. In *ICML*, pp. 16731–16750, 2024.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, pp. 4651–4664, 2021.

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 *ICLR*, pp. 55436–55457, 2025.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luis Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain Invaders solo versus collaboration: Multi-user P300-based brain-computer interface dataset (bi2014b). Zenodo, 2019.

H. W. Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2(1–2): 83–97, 1955.

Jarod Lévy, Hubert Banville, Jérémy Rapin, Jean-Remi King, Thomas Moreau, and Stéphane d'Ascoli. DANCE: Detect and classify events in EEG, 2026.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017.

Ming Liu, Yanbing Liu, Weiyou Shi, Yitai Lou, Yuan Sun, Qi Meng, Dezheng Wang, et al. EMPT: A sparsity transformer for EEG-based motor imagery recognition. *Front. Neurosci.*, 18:1366294, 2024.

G. Lomele, T. Lencioni, S. D'Ambrosio, A. Comanducci, F. Lucchetti, A. Marzegan, C. Derchi, et al. High-density EEG and multi-muscle EMG dataset during object prehension with a sensorized grasping box in humans. *Sci. Data*, 13(1):932, 2026.

Catarina Lopes-Dias, Andreea I. Sburlea, and Gernot R. Müller-Putz. Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Sci. Rep.*, 9(1):17596, 2019.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *ICLR*, pp. 83526–83574, 2026.

Jun Ma and Tuukka Ruotsalo. Adapting frozen foundation models for montage-agnostic high-resolution EEG event segmentation. *J. Neural Eng.*, 23(3):036020, 2026.

Tianhua Miao, Liansen Sha, Kun Huang, Yongbin Li, and Bin Liu. SATrans-Net: Sparse attention transformer for EEG-based motor imagery decoding. *Sci. Rep.*, 16:1192, 2026.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel EEG with time-frequency motif learning. In *ICLR*, pp. 36611–36637, 2026.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pp. 658–666, 2019.

Jacob T. Robinson, Sumner L. Norman, Matthew R. Angle, Timothy G. Constandinou, Timothy Denison, John P. Donoghue, Ryan M. Field, et al. An application-based taxonomy for brain–computer interfaces. *Nat. Biomed. Eng.*, 9(6):789–791, 2025.

Andreas Schwarz, Carlos Escolano, Luis Montesano, and Gernot R. Müller-Putz. Analyzing and decoding natural reach-and-grasp actions using gel, water and dry EEG systems. *Front. Neurosci.*, 14:849, 2020.

Nick Seeuws, Maarten De Vos, and Alexander Bertrand. Avoiding post-processing with event-based detection in biomedical signals. *IEEE Trans. Biomed. Eng.*, 71(8):2442–2453, 2024.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, et al. The Temple University Hospital Seizure Detection Corpus. *Front. Neuroinform.*, 12:83, 2018.

Xin Tang, Hao Shen, Siyuan Zhao, Na Li, and Jia Liu. Flexible brain–computer interfaces. *Nat. Electron.*, 6(2):109–118, 2023.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, et al. Review of the BCI Competition IV. *Front. Neurosci.*, 6:55, 2012.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, volume 30, pp. 6306–6315, 2017.

Lukas Wolf, Ard Kastrati, Martyna Beata Płomecka, Jie-Ming Li, Dustin Klebe, Alexander Veicht, Roger Wattenhofer, et al. A deep learning approach for the segmentation of electroencephalography data in eye tracking applications. In *ICML*, pp. 23912–23932, 2022.

Huanyu Wu, Siyang Li, and Dongrui Wu. Motor imagery classification for asynchronous EEG-based brain–computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 32:527–536, 2024.

Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. DyFADet: Dynamic feature aggregation for temporal action detection. In *ECCV*, pp. 305–322, 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 *NeurIPS*, volume 38, pp. 150290–150314, 2025.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pp. 492–510, 2022.

Haihong Zhang, Cuntai Guan, and Chuanchu Wang. Asynchronous P300-based brain–computer interfaces: A computational approach with statistical models. *IEEE Trans. Biomed. Eng.*, 55(6): 1754–1763, 2008.

Appendix

A Additional Results

A.1 Event-Level F1 at Different Temporal IoU Thresholds

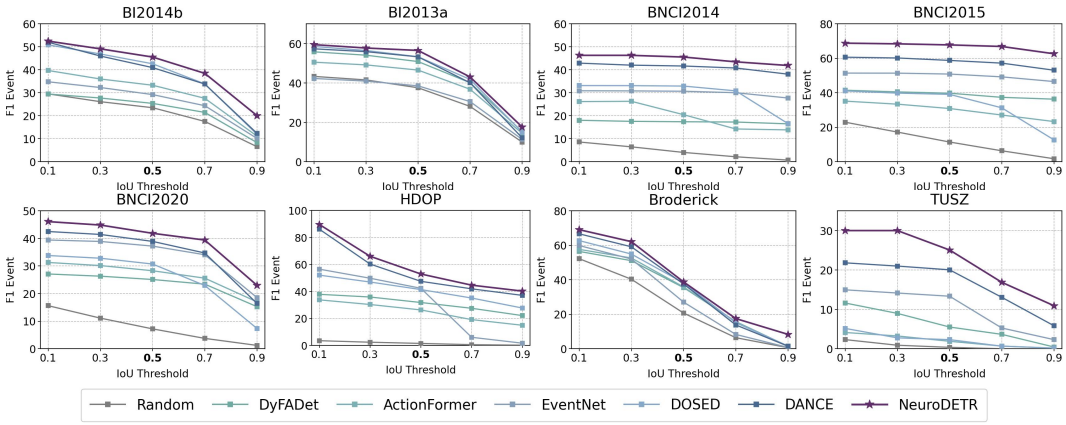


Figure 4: Event-level F1 of NeuroDETR and baseline methods at different temporal IoU thresholds. Higher thresholds impose stricter temporal localization requirements.

As shown in Figure 4, we compare NeuroDETR with the baselines in $F1_{\text{event}}$ at temporal IoU thresholds $\tau_{\text{IoU}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to assess detection performance under increasingly strict localization requirements. Lower thresholds allow greater boundary deviations, whereas higher thresholds require closer alignment between predicted and annotated intervals.

NeuroDETR achieves the highest event-level F1 among the compared methods at every evaluated IoU threshold, demonstrating its effectiveness in joint event classification and temporal localization even under stringent boundary-matching requirements.

A.2 Detection Error Analysis

To distinguish missed events from incorrect detections, we evaluate the event-level false discovery rate (FDR) and false negative rate (FNR). Predictions are matched one-to-one to ground-truth events in descending confidence order, with a true positive requiring the correct class and temporal IoU > 0.5. The metrics are defined as

$$\text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \times 100\%, \qquad \text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \times 100\%. \tag{24}$$

Here, TP counts correct matches, FP counts incorrect predictions, and FN counts ground-truth events without a correct match. FDR measures the proportion of incorrect detections among predictions, while FNR measures the proportion of missed events among annotations.

As shown in Figures 5 and 6, NeuroDETR achieves the lowest FDR and FNR on six of the eight datasets each, while attaining the highest F1_{event} on all eight. Where competing methods achieve a lower individual error rate, this advantage is accompanied by a higher complementary error rate: lower FDR coincides with more missed events, whereas lower FNR coincides with a larger proportion of incorrect detections. Such single-metric advantages therefore do not translate into better overall detection performance. NeuroDETR achieves the highest F1_{event} across all eight datasets, indicating superior overall event detection performance.

A.3 Sensitivity to the Number of Event Queries

The number of decoder queries N_e bounds the number of event predictions within an input window. Table 4 reports event F1 at five query counts for each dataset to assess sensitivity to this prediction capacity.

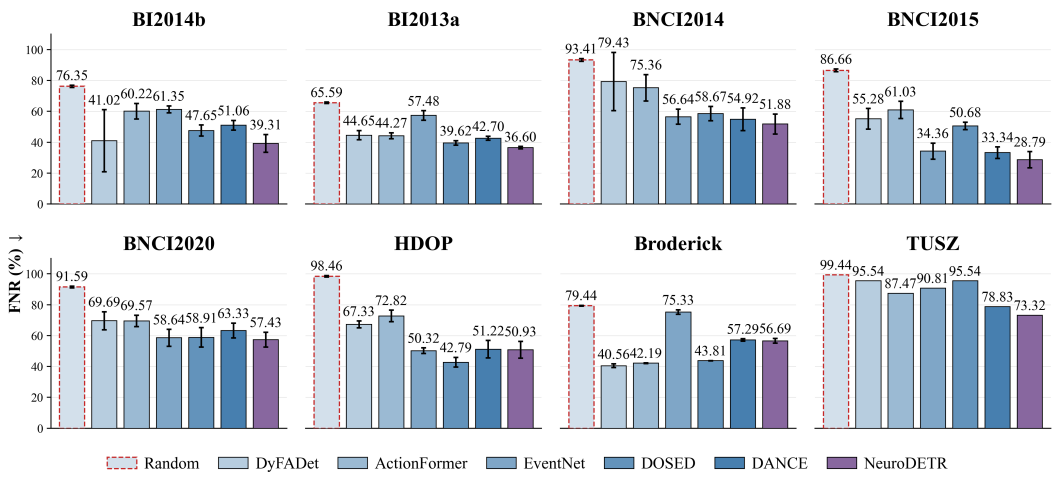


Figure 5: Event-level false negative rate (FNR) across eight datasets. Lower values indicate fewer missed events. NeuroDETR achieves the lowest FNR on six datasets. Error bars denote standard deviations and are omitted for TUSZ, where standard deviations are unavailable.

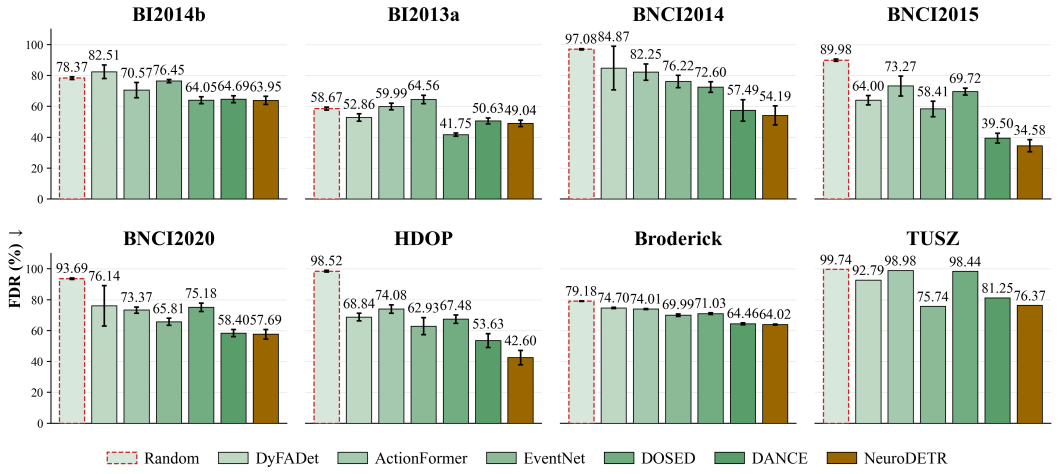


Figure 6: Event-level false discovery rate (FDR) across eight datasets. Lower values indicate a smaller proportion of incorrect detections among predictions. NeuroDETR achieves the lowest FDR on six datasets. Error bars denote standard deviations and are omitted for TUSZ, where standard deviations are unavailable.

Several datasets show small differences over a range of query counts. On Broderick, event F1 ranges from 38.65 to 38.76 for N_e between 50 and 200, before decreasing to 37.05 at 400 queries. BNCI2014 yields scores of 45.55 and 45.44 with 5 and 10 queries, respectively, while TUSZ achieves 25.06 and 25.00 with 8 and 16 queries.

Increasing the query count can also produce substantial, non-monotonic changes. On BNCI2015, event F1 decreases from 67.77 to 64.12 when N_e increases from 40 to 80. HDOP shows a sharper change: increasing N_e from 16 to 24 reduces event F1 from 52.91 to 41.17, followed by a recovery to 51.94 at 32 queries.

Overall, query-count sensitivity depends on the dataset: some query ranges yield similar scores, while individual settings can produce substantial drops. Larger query budgets do not consistently improve event detection. These findings motivate selecting N_e separately for each dataset using validation performance.

Table 4: Sensitivity to decoder query count N_e. For each dataset, the upper and lower rows report N_e and $F1_{\text{event}}$, respectively. **Bold** indicates the selected query count and its corresponding score.

Dataset	Decoder queries / Event F1				
BI2014b	50	100	**150**	200	300
	44.00	45.13	**45.23**	44.20	43.20
BI2013a	50	100	**150**	200	300
	55.65	52.71	**56.50**	52.46	54.28
BNCI2014	1	3	**5**	10	20
	42.55	43.10	**45.55**	45.44	42.77
BNCI2015	10	20	**40**	60	80
	66.85	67.44	**67.77**	67.49	64.12
BNCI2020	10	20	**40**	60	80
	40.03	41.32	**41.85**	40.99	39.80
HDOP	4	8	**16**	24	32
	51.52	51.78	**52.91**	41.17	51.94
Broderick	50	100	**150**	200	400
	38.65	38.73	**38.76**	38.71	37.05
TUSZ	2	4	**8**	16	32
	23.65	24.09	**25.06**	25.00	24.18

B Hierarchical Brain-Region Grouping

NeuroDETR uses a fixed three-level hierarchy of scalp-electrode groups: one global group (S_1), five coarse regions (S_2), and twelve fine clusters (S_3). The global group aggregates all electrodes assigned to the hierarchy, while the coarse-to-fine organization is summarized in Table 5.

Table 5: Coarse-to-fine spatial grouping.

Coarse region (S_2)	**Fine clusters** (S_3)	**Count**
Frontal	Left, midline, and right frontal	3
Temporal	Left and right temporal	2
Central	Left, midline, and right central	3
Parietal	Left, midline, and right parietal	3
Occipital	Undivided occipital group	1
Total		12

Each fine cluster belongs to one coarse region, and all coarse regions share the global group as their parent. The two temporal clusters are merged at the coarse level, whereas the occipital group remains undivided. This hierarchy provides a spatial prior over scalp electrodes and is instantiated using the available mapped channels of each dataset.

C Details of Event Supervision

This appendix specifies the set-prediction loss and auxiliary supervision in Eq. 21. For each of the N_e queries, $\hat{\mathbf{p}}_j$ denotes the class probabilities over $\{0, \dots, K\}$ and $\hat{I}_j = [\hat{s}_j, \hat{e}_j]$ the predicted interval. Ground-truth event i has class c_i and interval $I_i = [s_i, e_i]$, with M events in the input window. Class 0 denotes background, and all interval boundaries are normalized to the input-window duration.

C.1 Matching and Set-Prediction Loss

Hungarian matching. Following set-based detection (Carion et al., 2020), Hungarian matching produces a one-to-one assignment $\mathcal{M}$ by minimizing the total matching cost, with pairwise cost

$$\mathcal{C}_{ji} = -\lambda_{\text{mcls}} \log \hat{p}_j(c_i) + 1 - \text{IoU}(\hat{I}_j, I_i). \tag{25}$$

The assignment determines the classification targets and the query–event pairs used for localization supervision.

Endpoint regression. For matched queries, the endpoint loss penalizes absolute errors in normalized onset and offset predictions:

$$\mathcal{L}_{\ell_1} = \frac{1}{N_e} \sum_{(j,i)\in\mathcal{M}} \left(|\hat{s}_j - s_i| + |\hat{e}_j - e_i|\right). \tag{26}$$

Event classification. Matched queries receive their assigned event labels, while unmatched queries receive the background label. Cross-entropy supervision is applied to all queries:

$$c_j^\star = \begin{cases} c_i, & (j,i)\in\mathcal{M}, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathcal{L}_{\text{cls}} = -\frac{1}{N_e}\sum_{j=1}^{N_e} \log \hat{p}_j(c_j^\star). \tag{27}$$

Interval localization. Temporal generalized IoU (Rezatofighi et al., 2019) complements endpoint regression with interval-level supervision. For intervals A and B, let $H(A,B)$ denote their smallest enclosing interval and $|\cdot|$ denote temporal length. We define

$$\begin{aligned} \text{IoU}(A,B) &= \frac{|A\cap B|}{|A\cup B|}, \\ \text{GIoU}(A,B) &= \text{IoU}(A,B) - \frac{|H(A,B)| - |A\cup B|}{|H(A,B)|}, \\ \mathcal{L}_{\text{GIoU}} &= \frac{1}{N_e}\sum_{(j,i)\in\mathcal{M}} \Big[1 - \text{GIoU}(\hat{I}_j, I_i)\Big]. \end{aligned} \tag{28}$$

The enclosing-interval term provides localization supervision even when the predicted and target intervals do not overlap.

Both localization losses are normalized by the full query count N_e, and unmatched queries receive no localization gradients. For windows without annotated events, all classification targets are background and both localization losses are zero. The set-prediction objective is

$$\mathcal{L}_{\text{det}} = \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{giou}}\mathcal{L}_{\text{GIoU}}. \tag{29}$$

C.2 Auxiliary Supervision

We complement set prediction with full-resolution boundary supervision, dense classification, and prediction consistency.

Boundary supervision. A linear head on the backbone features $\mathbf{H}$ predicts class-agnostic onset and offset logits b_t^u, where $u \in \{\text{on}, \text{off}\}$. Let μ_i^u denote the corresponding annotated boundary in sample coordinates. For windows containing events, the target heatmaps are

$$Y_t^u = \max_{i=1,\dots,M} \exp\left(-\frac{(t-\mu_i^u)^2}{2\sigma_{\text{bnd}}^2}\right), \qquad t = 1,\dots,T, \tag{30}$$

where σ_{bnd} controls the Gaussian width. For windows without events, $Y_t^u = 0$ for both boundary types. Sigmoid focal loss (Lin et al., 2017) supervises the heatmaps while downweighting easy non-boundary positions:

$$\mathcal{L}_{\text{bnd}} = \frac{1}{2T}\sum_{u\in\{\text{on},\text{off}\}}\sum_{t=1}^{T} \text{FL}_{\text{sigmoid}}(b_t^u, Y_t^u). \tag{31}$$

Dense classification. A classifier on the fused memory $\mathbf{Z}$ produces class probabilities $\mathbf{P}_{\text{dense}} \in \mathbb{R}^{N_q\times(K+1)}$. Let y_r denote the annotation label at grid position r, with 0 indicating background. The dense classification loss is

$$\mathcal{L}_{\text{dense}} = -\frac{1}{N_q}\sum_{r=1}^{N_q} \log P_{\text{dense}}(r, y_r). \tag{32}$$

Prediction consistency. We project query predictions onto the same N_q-position grid as the dense predictions. Let $a_{rj} \in \{0, 1\}$ indicate whether position r lies within query j's discretized predicted interval. We accumulate class probabilities from all queries, including the background class, and normalize the scores at covered positions:

$$S_{r,c} = \sum_{j=1}^{N_e} a_{rj} \hat{p}_j(c),$$
$$P_{\text{set}}(r, c) = \begin{cases} \dfrac{S_{r,c}}{\sum_{k=0}^{K} S_{r,k}}, & \sum_{k=0}^{K} S_{r,k} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

Overlapping predictions contribute jointly, while uncovered positions are assigned zero vectors. Let $\widetilde{P}$ denote componentwise clamping for numerical stability, without subsequent renormalization, and let $\epsilon_{\text{clamp}} > 0$ denote its lower bound. The KL-style consistency penalty is averaged over the full grid:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N_q} \sum_{r=1}^{N_q} \sum_{c=0}^{K} \widetilde{P}_{\text{dense}}(r, c) \log \frac{\widetilde{P}_{\text{dense}}(r, c)}{\widetilde{P}_{\text{set}}(r, c)}. \tag{34}$$

At covered positions, this penalty encourages agreement between the dense and set-derived predictions. At uncovered positions, all components of the set-derived vector are clamped to ϵ_{clamp}. When clamping leaves the dense probabilities unchanged, the contribution at an uncovered position, before averaging over the grid, reduces to

$$\sum_{c=0}^{K} P_{\text{dense}}(r, c) \log \frac{P_{\text{dense}}(r, c)}{\epsilon_{\text{clamp}}} = \log \frac{1}{\epsilon_{\text{clamp}}} - H(P_{\text{dense}}(r, \cdot)),$$

where H denotes Shannon entropy. This term therefore favors higher-entropy dense predictions at uncovered positions, while $\mathcal{L}_{\text{dense}}$ provides supervision from the annotation labels.

At covered positions, gradients propagate to both classification heads through probability entries unaffected by clamping. At uncovered positions, the set-derived vector is constant, so this term backpropagates only through the dense predictions. The discrete interval masks block gradients to the predicted boundaries through this projection.

The event-supervision objective combines the set-prediction loss with the three auxiliary terms:

$$\mathcal{L}_{\text{event}} = \mathcal{L}_{\text{det}} + \lambda_{\text{bnd}} \mathcal{L}_{\text{bnd}} + \lambda_{\text{dense}} \mathcal{L}_{\text{dense}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}. \tag{35}$$

Table 6: Dataset statistics and training configurations.

Dataset	Domain	Subjects	Channels	Event (h)	Non-event (h)	Mean event duration (s)	Event coverage (%)	Split	Window (s)	Batch size per GPU
Korczowski2014b	P300	38	32	1.0	1.3	1.0	44	CV	8	32
Vaineau2018Brain	P300	24	16	4.4	1.1	1.0	80	CV	32	8
Tangermann2012	Motor imagery	9	22	5.8	5.8	4.0	50	CV	8	32
Faller2012	Motor imagery	12	13	7.8	8.9	5.0	47	CV	32	4
BNCI2020	Motor execution	45	12/32/58	10	27	5.0	27	CV	32	4
HDOP	Motor execution	14	62	0.24	3.7	0.20	6.1	CV	9	8
Broderick	Speech	19	128	14	6.5	0.23	69	CV	8	32
TUSZ (v2.0.6)	Seizure	675	17–25	47	1.4×10^3	63	3.2	Fixed	256	4†

†TUSZ uses eight GPUs with a batch size of four per GPU, giving an effective batch size of 32 (8×4).

Table 6 summarizes the eight EEG datasets used in our experiments, including their domains, subject and channel counts, event and non-event recording times, mean event durations, event coverage, data splits, and input window lengths. For each recording, event time is the temporal union of the configured event annotations, and non-event time is the remaining recording time. Summing these quantities across recordings gives E and N, reported in hours, with event coverage defined as $E/(E + N) \times 100\%$. Overlapping intervals contribute only once to E.

The datasets span P300, motor imagery, motor execution, speech, and clinical seizure detection. Korczowski2014b (BI2014b) and Vaineau2018Brain (BI2013a) contain visual oddball stimulation,

with both target and non-target stimuli represented as event classes using 1-s annotation intervals. Tangermann2012 (BNCI2014-001) and Faller2012 (BNCI2015-001) provide four-class and two-class motor-imagery tasks, respectively, with annotated intervals of 4 and 5 s. BNCI2020 contains palmar and lateral grasp events annotated with 5-s intervals, while rest is treated as background. HDOP represents LED onset, object contact, and object lift as three event classes, each assigned a 0.2-s interval. Broderick contains continuous EEG recorded during naturalistic speech listening, with word-level temporal annotations. TUSZ (v2.0.6) provides clinical seizure annotations mapped to four configured categories: focal, generalized, absence, and tonic-clonic. Duplicate seizure annotations across channels are removed before computing the statistics.

Channel counts refer to EEG channels retained by the dataset loaders. BNCI2020 uses 12, 32, and 58 channels in its dry-, water-, and gel-electrode groups, respectively; the dry group includes the A2 reference channel. HDOP retains 62 EEG channels after excluding VEOG and HEOG, while TUSZ retains 17–25 EEG channels depending on the recording montage.

For the seven non-clinical datasets, we use five-fold cross-validation (CV) grouped by subject. One fold is held out for testing, and 10% of the remaining subjects are reserved for validation, rounded up to the next integer. TUSZ follows its official train/dev/eval partition, comprising 579, 53, and 43 subjects, respectively. Input windows range from 8 to 32 s for the non-clinical datasets and span 256 s for TUSZ.

D BASELINE

D.1 BASELINE IMPLEMENTATIONS

We compare NeuroDETR with six event-level baselines: a dataset-aware random predictor, two biomedical event detectors, two temporal action detection models, and DANCE. The comparison excludes sliding-window classifiers because evaluating densely overlapping windows repeatedly processes the same portions of long EEG recordings.

All methods use the same dataset-specific EEG channels, preprocessing, event annotations, class mappings, temporal units, and subject-level training, validation, and test partitions. Model selection is performed exclusively on the validation set. All methods share the event-level evaluation protocol described below.

For all dense detectors (EventNet, ActionFormer, DyFADet, and DOSED), we apply a confidence threshold of 0.05 and class-wise non-maximum suppression (NMS) with a temporal IoU threshold of 0.5. After NMS, we retain at most N_e predictions per window, ranked by confidence.

Random. We construct a dataset-aware random predictor to provide a non-trivial lower bound. For each training fold, the event-class distribution, event-duration distribution, and event occurrence rate are estimated using only the training annotations. The number of predicted events in each evaluation segment is sampled according to the estimated occurrence rate. Their classes and durations are sampled from the corresponding training distributions, while their onset locations are sampled uniformly within the segment. Predicted intervals are clipped to the valid temporal range and evaluated using the same event-level protocol as the learned models.

EventNet. EventNet denotes our implementation of the event-based modeling framework introduced by Seeuws et al. (Seeuws et al., 2024). The original method represents each event using its temporal center and duration. A sequence-to-sequence backbone produces a center-confidence signal and a duration estimate at every temporal position. During inference, local maxima in the center-confidence signal are selected as event centers, and the corresponding duration predictions are converted into temporal intervals.

We extend the original binary formulation with class-specific center and duration heads for multi-class EEG event detection. Candidate intervals are decoded independently for each class, allowing events of different classes to overlap. The decoded intervals undergo the shared post-processing described above.

ActionFormer. ActionFormer (Zhang et al., 2022) is a single-stage, anchor-free temporal detector originally developed for temporal action localization in untrimmed videos. It constructs a multiscale temporal feature pyramid, uses local self-attention to model temporal context, and predicts an event class together with the distances from each positive temporal location to the left and right event boundaries.

For EEG adaptation, we replace the video feature extractor with a one-dimensional EEG encoder. Its output is arranged as a temporal feature sequence and processed by the original multiscale detection architecture. The temporal strides and regression ranges of the feature pyramid are adjusted according to the EEG sampling resolution and the dataset-specific event-duration distribution. Predicted boundary distances are converted into onset–offset intervals, followed by the shared post-processing.

DyFADet. DyFADet (Yang et al., 2024) extends temporal action detection with dynamic feature aggregation. Its dynamic encoder adjusts the aggregation weights and temporal receptive fields according to the input features. The associated dynamic detection head further adapts its parameters across different feature-pyramid levels, allowing the model to handle events with substantially different durations.

We use the same EEG encoder, input resolution, and temporal feature interface as in the ActionFormer adaptation. The video feature sequence is replaced by the EEG temporal representation, while the dynamic encoder, multiscale feature pyramid, and boundary-regression mechanism are retained. The supported regression ranges are rescaled to the event durations of each EEG dataset. The output intervals are decoded in seconds and processed using the shared post-processing.

DOSED. DOSED (Chambon et al., 2019) is a one-dimensional, anchor-based detector originally designed for detecting multiple sleep micro-events in EEG. The model associates a collection of default temporal events with different positions and durations. For every default event, it jointly predicts a class score and localization offsets. Training matches ground-truth events to default events according to temporal overlap and optimizes the classification and localization objectives jointly.

We adapt DOSED to the event categories of each dataset using its default-event detection formulation. Default-event durations are estimated using only the training annotations of each fold. Predicted offsets are decoded into onset–offset intervals, clipped to the input duration, and processed using the shared post-processing.

DANCE. DANCE (Lévy et al., 2026) is a query-based model for detecting and classifying temporal events in EEG recordings. An EEG encoder first maps the input signal into a temporal representation. A fixed set of learnable queries then interacts with the encoded features, with each query predicting an event category and its temporal extent. During training, bipartite matching establishes a one-to-one assignment between predicted and annotated events, while unmatched predictions are assigned to the no-event category.

In the query-sensitivity analysis, we vary only the number of decoder queries. The capacity used to encode ground-truth events remains fixed across the sweep, and all query settings use identical training and evaluation annotations.

Unified evaluation. All methods ultimately output a collection of $(t_{\text{start}}, t_{\text{end}}, c, s)$ tuples, where c denotes the event class and s is the confidence score. Predictions are sorted by confidence and greedily assigned to unmatched ground-truth intervals according to temporal IoU. A matched pair contributes a true positive only when its class is correct and its IoU satisfies the evaluation threshold. We report event-level F1 at a temporal IoU threshold of 0.5.

E Backbone Architecture

We use a SimpleConv backbone adapted from (Défossez et al., 2023) to extract spatial and temporal features from EEG recordings. The backbone first performs spatial channel mixing, where Fourier embeddings of electrode coordinates determine learnable attention weights over the recording channels. This operation maps the input signals to a fixed set of virtual channels, incorporating electrode positions into the spatial representation. A pointwise convolution then projects the mixed signals into the feature space used by the temporal stack.

Temporal features are extracted through ten convolutional stages with increasing dilation rates, progressively expanding the temporal receptive field. The first nine stages use 512 hidden channels with batch normalization, GELU activation, and dropout, while the final stage produces 128-dimensional features. All temporal convolutions preserve the sequence length, maintaining alignment between the extracted features and the input EEG samples for subsequent event localization and classification. Table 7 summarizes the module configurations and input–output dimensions, where B, C, and T denote the batch size, number of EEG channels, and sequence length, respectively.

Table 7: Backbone architecture and hyperparameters. The temporal resolution is preserved throughout the backbone.

Module	Input shape	Output shape	Configuration
Spatial channel mixing	$B \times C \times T$	$B \times 270 \times T$	Fourier embedding dimension: 2048; spatial dropout radius: 0.2
Pointwise projection	$B \times 270 \times T$	$B \times 256 \times T$	Conv1d, kernel size: 1
Input dropout	$B \times 256 \times T$	$B \times 256 \times T$	Dropout probability: 0.1
Temporal stage 1	$B \times 256 \times T$	$B \times 512 \times T$	Kernel size: 9; dilation: 1
Temporal stages 2–9	$B \times 512 \times T$	$B \times 512 \times T$	Kernel size: 9; dilation: $d_\ell = \lfloor 2.5^{\ell-1} \rfloor$
Temporal stage 10	$B \times 512 \times T$	$B \times 128 \times T$	Kernel size: 9; dilation: 3814; no normalization or activation
Transpose	$B \times 128 \times T$	$B \times T \times 128$	Feature sequence for downstream modules

All temporal convolutions use stride 1 and symmetric padding $4d_\ell$. Stages 1–9 each apply Conv1d, BatchNorm1d, GELU, dropout ($p = 0.1$), and a pointwise convolution followed by ReLU. Stage 10 contains only the temporal convolution.

F Implementation Details

We implement NeuroDETR using PyTorch and PyTorch Lightning. EEG signals are resampled to 128 Hz and normalized using robust scaling. For the seven cross-validation datasets, we perform five-fold subject-wise cross-validation. Within each fold, 10% of the remaining subjects are reserved for validation. TUSZ uses its predefined training, development, and evaluation splits. The random seed is fixed to 42. We train with AdamW and a OneCycle learning-rate schedule, using the first 10% of training steps for learning-rate increase. Unless otherwise specified, the peak learning rate is 10^{-4}, weight decay is 10^{-3}, and the maximum gradient norm is 1. Training runs for up to 100 epochs with an early-stopping patience of 20 epochs. Checkpoint selection and early stopping are based on validation $F1_{\text{event}}$ at a temporal IoU threshold of 0.5. All experiments use FP32 precision. The seven cross-validation datasets are trained on a single GPU. TUSZ uses distributed data parallel training on eight GPUs, with four windows per GPU and an effective batch size of 32.