SPAN: SET-PREDICTION ADAPTATION FOR CONTINUOUS BIOSIGNAL EVENT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous biosignal applications, including seizure monitoring, sleep analysis and motor-imagery decoding, require predictions with an accurate class, onset and offset. Yet pretrained biosignal foundation models are almost universally adapted through per-window classifiers, whose outputs must be thresholded, smoothed and merged into events after the fact. We show that this adaptation interface, and not only the pretrained representation, is a decisive factor in event-level quality. We introduce **SPAN**, a temporal set-prediction framework that maps tokens from heterogeneous biosignal backbones into a shared physical-time representation and directly predicts event intervals via multiscale temporal features, anchor-conditioned queries and duration-dependent attention. Using a paired protocol that fixes the pretrained initialization, training data and event scorer, we compare SPAN against dense per-token fine-tuning (DENSE-FT) across six public EEG foundation-model backbones, four tasks and three seeds. SPAN improves test event F1 at IoU> 0.5 in all 72 seed-matched comparisons, with average gains of 4.5–8.3 percentage points per task, and reduces onset error on matched seizure, motor-imagery and arousal events. A further comparison against a plain set decoder on identical backbone features shows that these gains come from SPAN's combined design rather than set prediction alone, and a component ablation study traces the gain to duration-dependent attention locality and query denoising. On densely overlapping artifact annotations, boundary localization and fragmentation remain difficult, pointing to a concrete direction for future work. Together, these results establish the adaptation interface as a substantial, previously underexamined lever for transferring pretrained biosignal representations to temporally localized event decoding, and position SPAN as a strong, broadly validated default for this setting.

1 INTRODUCTION

Continuous EEG applications require predictions with a class, an onset and an offset. Seizure monitoring locates episodes, sleep analysis identifies brief arousals, artifact detection marks contaminated intervals, and asynchronous motor-imagery decoding identifies periods of intended activity. In these cases, predictions must capture both what happens and when it happens, since accurate classification of individual windows alone does not mean a model can recover complete event instances.

EEG foundation models (FMs), including LaBraM, CBraMod, REVE, EEGPT, BIOT and LUNA learn representation from large collections of EEG, but their standard downstream interfaces emphasize classification and focus on window-level prediction (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025; Wang et al., 2024; Yang et al., 2023; Döner et al., 2025). To produce events, a dense classifier must convert its probability timeline into intervals through smoothing, thresholding, merging and duration filtering. These operations can change boundaries and combine or split events, so good sample-level performance therefore does not establish good event-level performance. Direct event prediction provides an alternative. DANCE (Lévy et al., 2026) showed that a DETR-style set predictor (Carion et al., 2020) trained on raw EEG can improve event detection over dense segmentation. These studies motivate a controlled question: for a fixed pretrained EEG backbone, how much does the adaptation interface affect continuous event decoding?

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.

We introduce SPAN, a temporal set-prediction framework for adapting EEG FMs to continuous event decoding. A common adapter attaches physical timing information to heterogeneous backbone tokens. A multiscale memory and anchor-conditioned queries then predict event classes and intervals, with duration-dependent attention, iterative boundary refinement and denoising supervision. The design draws on set prediction (Carion et al., 2020), anchor-based queries (Liu et al., 2022a) and query denoising (Li et al., 2022a; Zhang et al., 2023), integrating them into a shared interface for EEG backbones with different temporal resolutions.

We evaluate SPAN against dense per-token fine-tuning (DENSE-FT) in a paired protocol across six public backbones, four tasks and three seeds. Each pair starts from the same pretrained encoder checkpoint and uses the same data, adapter architecture and event scorer. Both encoders are fully fine-tuned, so the comparison measures the complete adaptation procedure, including the head and its training objective. Our contributions are:

- **A common event-decoding framework for heterogeneous EEG FMs.** SPAN expresses multiscale features and interval queries in physical time, supporting pretrained architectures through one decoder.
- **A paired evaluation across four continuous EEG tasks.** SPAN improves event F1 at IoU> 0.5 in all 72 seed-matched comparisons, with task-average gains of 4.5–8.3 percentage points.
- **An analysis of temporal quality and decoder design.** We compare a plain set decoder on the same adapted features, remove SPAN's locality, denoising and multiscale components one at a time, and examine boundary errors, fragmentation and sample-level scores. The results distinguish improved event detection from event coherence: onset errors decrease on three tasks, whereas overlapping artifacts remain challenging.

2 RELATED WORK

EEG representations and adaptation. Supervised EEG architectures such as DeepConvNet, EEGNet and EEG Conformer learn spatial and temporal features for decoding (Schirrmeister et al., 2017; Lawhern et al., 2018; Song et al., 2023). Self-supervised approaches extend this idea to reusable representations: temporal-context objectives (Banville et al., 2021), contrastive pretraining in BENDR (Kostas et al., 2021), and hierarchical time-series representations in TS2Vec (Yue et al., 2022) reduce dependence on task-specific labels. EEG FMs further explore tokenized pretraining (LaBraM), criss-cross modeling (CBraMod), flexible recording geometry (REVE), predictive representation learning (EEGPT), heterogeneous biosignal inputs (BIOT), and topology-agnostic latent representations (LUNA) (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025; Wang et al., 2024; Yang et al., 2023; Döner et al., 2025). We study how these pretrained features are adapted to structured temporal outputs. This complements work showing that downstream adaptation choices influence transfer (Kumar et al., 2022) and that pretrained transformer features can support detection through appropriate heads (Li et al., 2022b).

Continuous EEG events. Dense segmentation models such as U-Time (Perslev et al., 2019) assign labels along a signal, whereas DOSED (Chambon et al., 2019) predicts sleep micro-events as intervals. DETRtime applies set prediction to ocular events in EEG (Wolf et al., 2022), and DANCE jointly detects and classifies diverse EEG events (Lévy et al., 2026). Event-based evaluation of FMs has also been investigated for ICU burst suppression (Vasta et al., 2026). We hold the backbone fixed and vary only the interface, across six backbones and four tasks that include multi-class and overlapping events.

Temporal localization and set prediction. DETR introduced end-to-end detection with learned queries and bipartite matching (Carion et al., 2020). Deformable DETR and DAB-DETR improve spatial conditioning through reference points and anchors (Zhu et al., 2021; Liu et al., 2022a), while DN-DETR and DINO use denoising supervision (Li et al., 2022a; Zhang et al., 2023). In temporal action localization, TadTR uses a query-based set decoder and ActionFormer uses multiscale temporal features (Liu et al., 2022b; Zhang et al., 2022). We transfers these ideas to physical EEG time, where token rates differ between backbones and event durations range from seconds to minutes, and evaluates them against dense fine-tuning under a paired protocol.

3 METHOD

3.1 PROBLEM FORMULATION

Given a continuous EEG chunk $X \in \mathbb{R}^{C\times S}$ with C channels, S samples, and duration L seconds, the goal is to predict a set of events $\mathcal{Y} = \{(c_i, s_i, e_i)\}_{i=1}^{N}$, where $c_i \in \{1, \dots, K\}$ denotes the event class and $0 \le s_i < e_i \le L$ are its onset and offset. The number of events N varies across chunks. Our model produces Q candidate predictions $\widehat{\mathcal{Y}} = \{(\widehat{p}_j, \widehat{s}_j, \widehat{e}_j)\}_{j=1}^{Q}$, where $\widehat{p}_j$ is a probability distribution over the K event classes and a no-event class. Confidence filtering yields a variable-size event set, and predictions from successive chunks are stitched to form recording-level outputs.

3.2 A COMMON FOUNDATION-MODEL ADAPTER

The adapter is the only backbone-specific code. For each chunk, the adapter returns tokens $Z \in \mathbb{R}^{T\times W}$, token center times $t \in \mathbb{R}^{T}$ and token cell bounds cell $\in \mathbb{R}^{T\times 2}$ used to rasterize dense targets. Each FM processes its native 10-s windows at the input scale of its released code, and the per-window token grids are concatenated in time, so both methods see the same physical input. Channel tokens are reduced either by a mean over channels or by concatenating learned summary tokens, followed by LayerNorm and a linear projection $D \to W$; Appendix B gives the per-backbone rules and Table 4 lists the token layouts.

3.3 DENSE FINE-TUNING HEAD (DENSE-FT)

The dense head applies a two-layer MLP to every token, $\widehat{y}_k = \mathrm{MLP}(Z_k) \in \mathbb{R}^{K+1}$. After softmax, it produces probabilities $\widehat{q}_k$ representing the event classes and background. Labels $y_k \in \{0, \dots, K\}$ are rasterised from event intervals onto the token cells, with 0 denoting background. Training uses class-weighted cross-entropy:

$$\mathcal{L}_{\mathrm{dense}} = -\frac{\sum_{k\in\mathcal{V}} \omega_{y_k} \log \widehat{q}_k(y_k)}{\sum_{k\in\mathcal{V}} \omega_{y_k}}, \tag{1}$$

where $\mathcal{V}$ contains the valid tokens across the mini-batch and ω_c are class weights estimated from the training set. At inference, token probabilities are converted into events through moving-average smoothing, thresholding, merging of nearby runs and minimum-duration filtering. Post-processing parameters are selected on the validation split by event F1 at IoU>0.3 and fixed for test evaluation. The search grids and task-specific target construction are detailed in Appendix B.

3.4 SPAN: DIRECT INTERVAL-SET PREDICTION

Here we introduce SPAN, a framework for adapting pretrained EEG foundation models to continuous event decoding. As shown in Figure 1, it combines a multiscale temporal memory with an anchor-conditioned query decoder that iteratively refines candidate intervals. Denoising queries and an auxiliary dense head provide additional supervision during training.

Multiscale temporal memory. To represent events at different durations, we construct a temporal pyramid from the adapter tokens Z, which is average-pooled at physically matched strides of about $1, 2, 3, 6$ s, each level receiving its own projection, LayerNorm, level embedding and a Fourier encoding of the token centre time; all levels are concatenated as decoder memory with normalised centres $\tau_j = t_j/L$, so the decoder sees approximately matched physical resolutions across EEG FM's token rates.

Query initialisation. The slot contents start from a learned table $E \in \mathbb{R}^{Q\times W}$. To adapt it to the recording at hand, the mean $\bar{z}$ of the chunk tokens is mapped linearly to a feature-wise scale and shift, $q_i^{(0)} = E_i \odot (1+\gamma) + \beta$ with $(\gamma, \beta) = \mathrm{Linear}(\bar{z})$. The anchor is stored as two unconstrained coordinates $r_i^{(0)} \in \mathbb{R}^2$ and converted to a centre and a duration by a fixed squashing map, $(\mu_i, d_i) = g(r_i)$, given in Appendix B. The map places the duration on a logarithmic scale between one second and the chunk length, and restricts the centre so that the interval $[\mu_i - d_i/2,\ \mu_i + d_i/2]$ stays inside the chunk. Initial centres are evenly spaced over the chunk and initial durations are evenly spaced on the logarithmic scale.

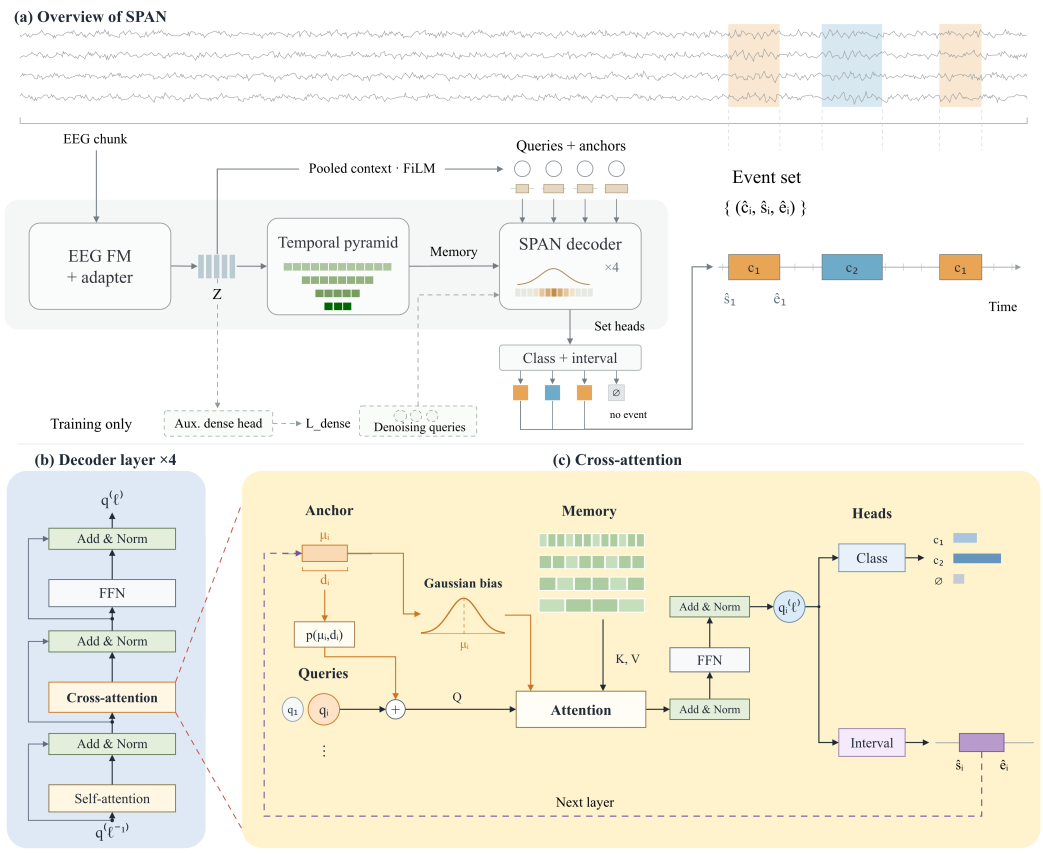

Figure 1: Overview of SPAN. **(a)** A foundation model and the common adapter produce tokens Z, which a temporal pyramid organises into a multiscale memory. Learned queries, modulated by a pooled chunk context and paired with interval anchors, pass through four decoder layers; set heads output a class and an interval per query, followed by confidence filtering and cross-chunk stitching. Dashed components are used only in training. **(b)** One decoder layer. **(c)** Cross-attention for query i: the anchor (μ_i, d_i) yields a positional embedding added to the attention query (Q) and a Gaussian bias added to the scores over the memory keys and values (K, V); the refined interval becomes the anchor of the next layer.

Decoder layer. Each of the four layers first updates the slot contents and then refines the anchors (Figure 1b,c). The contents pass through a standard Transformer decoder layer (Vaswani et al., 2017): self-attention among the slots, cross-attention to the memory, and a feed-forward network, each followed by a residual connection and LayerNorm. Self-attention lets every slot see which intervals the others currently claim, so one-to-one training can suppress duplicates without non-maximum suppression. Cross-attention gathers evidence from the memory, focused around the slot's anchor by the locality bias of Eq. 2. To tell both attentions where a slot is looking, its current anchor is embedded, inspired by DAB-DETR (Liu et al., 2022a), as $p_i = \mathrm{Linear}([\phi(\mu_i) \,\|\, \phi(d_i)]) \in \mathbb{R}^W$ with ϕ a Fourier encoding. This embedding is added to the attention queries, and in self-attention also to the keys, but never to the contents themselves, so the contents carry no position that would become outdated once the anchor moves. The decoder layer updates the reference coordinates and predicts class logits and boundaries $\widehat{b}_i^{(\ell)} = [\mu_i - d_i/2,\ \mu_i + d_i/2]$ with $(\mu_i, d_i) = g(r_i^{(\ell)})$.

Gaussian locality. Cross-attention receives an additive, per-layer bias

$$a_{ij}^{(\ell)} = \frac{q_i^\top k_j}{\sqrt{d_h}} - \frac{(\tau_j - \mu_i)^2}{2\,\sigma_{\ell i}^2}, \qquad \sigma_{\ell i} = \max(0.01, \mathrm{softplus}(\theta_\ell) d_i), \tag{2}$$

where $q_i, k_j \in \mathbb{R}^{d_h}$ with $d_h = W/8$, (μ_i, d_i) is the anchor entering layer ℓ, and θ_ℓ is a learnable scalar per decoder layer, shared by all queries and heads, that sets the window width as a multiple

of each query's anchor duration. The bias focuses each query on the neighborhood of its current anchor with a learned width but never masks distant tokens, so a badly placed anchor can still move. Because the width scales with the anchor duration, short and long events are attended at their own scale, and the lower bound prevents excessively narrow attention. Section 5.3 ablates this term.

Denoising queries and auxiliary dense head. Following DN-DETR and DINO (Li et al., 2022a; Zhang et al., 2023), we append two denoising groups during training. Each ground-truth event contributes a positive query with small, duration-scaled anchor perturbations and a negative query with a larger centre shift. Positive queries reconstruct the target class and interval, while negative queries predict no-event, isolated from the matching queries by attention masks. A linear auxiliary head on the adapter tokens Z is trained using the dense targets of Section 3.3 and supplies the scores used for per-second evaluation.

Objective. SPAN loss combines set prediction, denoising, and dense supervision:

$$\mathcal{L} = \sum_{\ell=1}^{4} w_\ell \left[\mathcal{L}_{\text{focal}}^{(\ell)} + \mathcal{L}_{\text{loc}}^{(\ell)}\right] + \mathcal{L}_{\text{DN}} + \mathcal{L}_{\text{dense}}, \tag{3}$$

where $\mathcal{L}_{\text{dense}}$ is defined in Eq. 1. The final layer has weight $w_4 = 1$, while the three intermediate losses are averaged and weighted by 0.5, equivalently $w_1 = w_2 = w_3 = 1/6$. At each decoder layer, Hungarian assignment (Kuhn, 1955) matches queries one-to-one to training targets using

$$\mathcal{C}_{ij} = -\widehat{p}_i(c_j) + \mathcal{R}(\widehat{b}_i, b_j), \qquad \mathcal{R}(\widehat{b}, b) = \|\widehat{b} - b\|_1 + 2\left[1 - \text{dIoU}(\widehat{b}, b)\right], \tag{4}$$

where layer indices are omitted and dIoU denotes the one-dimensional distance-IoU (Zheng et al., 2020). Let $\mathcal{P}_\ell$ and $\mathcal{U}_\ell$ denote matched and unmatched queries at layer ℓ, and let $(c_i^\star, b_i^\star)$ be the assigned target of a matched query. Using the focal penalty $f(p) = -(1-p)^2 \log p$ (Lin et al., 2017), we define

$$\mathcal{L}_{\text{focal}}^{(\ell)} = \left\langle f\left(\widehat{p}_i^{(\ell)}(c_i^\star)\right)\right\rangle_{\mathcal{P}_\ell} + 0.5\left\langle f\left(\widehat{p}_i^{(\ell)}(\varnothing)\right)\right\rangle_{\mathcal{U}_\ell}, \qquad \mathcal{L}_{\text{loc}}^{(\ell)} = \left\langle \mathcal{R}\left(\widehat{b}_i^{(\ell)}, b_i^\star\right)\right\rangle_{\mathcal{P}_\ell}. \tag{5}$$

Angle brackets denote means over the indicated queries across the mini-batch, with empty sets contributing zero. Thus, classification supervises all queries with separate matched and unmatched averages, whereas localisation supervises matched pairs only. Denoising uses the same loss components with known assignments:

$$\mathcal{L}_{\text{DN}} = \frac{1}{4}\sum_{\ell=1}^{4}\left[\mathcal{L}_{\text{focal,DN}}^{(\ell)} + \mathcal{L}_{\text{loc,DN}}^{(\ell)}\right]. \tag{6}$$

These terms follow Eq. 5, replacing matched and unmatched queries with valid positive and negative denoising queries, respectively. Localisation uses the clean intervals of positive queries only, without Hungarian assignment. We omit dense–event consistency regularisation from SPAN.

Inference. Only the final-layer matching queries are used at inference, without non-maximum suppression. For each query, the predicted class is the argmax over event classes and the score is its probability from that class. A query is retained if this score is at least the validation-selected threshold; no additional background-argmax rejection is applied. Same-class events touching adjacent chunk boundaries are stitched using task-specific gap and boundary margins: 10 s for TUSZ, 1 s for MI and TUAR, and 2 s for arousal. Dense predictions instead concatenate token probabilities across the recording before applying the validation-selected smoothing and event extraction.

4 Experimental setup

Datasets. We evaluate seizure detection on TUSZ (Shah et al., 2018), motor imagery (MI) on BNCI2014-001 (Tangermann et al., 2012), artifact detection on TUAR (Hamid et al., 2020), and arousal detection on the Human Sleep Project (HSP) (Li et al., 2026). TUSZ uses its official train/val/test split and a single seizure class. MI has four imagery classes and TUAR five artifact classes; both use five subject-grouped folds, pooled within each seed for evaluation. HSP uses 1,800 nights split into 1,200/300/300 for training/validation/testing. The evaluated recordings contain 467, 5,076, 40,247 and 41,006 reference events, respectively. We use 300-s chunks for TUSZ and arousal and 60-s chunks for MI and TUAR. Appendix A describes preprocessing and split construction.

Table 1: Test event F1 at IoU>0.5 (%), mean±SD over three seeds for every backbone. Counts are pooled over all five held-out folds for MI and TUAR. Bold indicates the higher mean.

	TUSZ		MI		TUAR		Arousal	
Backbone	DENSE-FT	SPAN	DENSE-FT	SPAN	DENSE-FT	SPAN	DENSE-FT	SPAN
LaBraM	42.3±2.5	**56.6±4.2**	18.7±0.6	**21.7±1.3**	14.6±0.2	**19.6±0.2**	24.9±0.2	**30.2±2.2**
CBraMod	41.7±1.4	**46.7±2.2**	17.6±0.5	**27.9±0.7**	14.8±0.4	**17.0±0.2**	23.9±1.1	**27.6±1.4**
REVE	48.1±0.8	**54.8±2.6**	27.7±0.8	**32.5±1.5**	12.4±1.2	**18.3±0.4**	24.0±0.3	**28.3±1.8**
EEGPT	36.2±3.5	**46.3±0.7**	12.9±0.1	**20.2±0.5**	13.2±0.7	**16.2±0.2**	12.9±0.5	**26.2±1.7**
BIOT	28.4±1.8	**38.2±0.6**	14.1±0.2	**28.5±0.1**	11.6±0.9	**17.3±0.0**	24.2±0.2	**27.7±2.6**
LUNA	29.7±2.2	**33.8±2.0**	26.5±1.0	**36.8±0.4**	14.5±0.3	**19.6±0.7**	24.6±0.4	**27.4±0.4**

Implementation details. The six backbones, in reporting order, are LaBraM, CBraMod, REVE, EEGPT, BIOT and LUNA. Seeds 7, 17 and 37 give 144 primary fine-tuning cells: six backbones, four tasks, two methods and three seeds, where a five-fold evaluation counts as one cell. The decoder comparison and the component ablation of Section 5.3 add LaBraM and REVE cells under the same protocol. Within a pair, the two methods use the same pretrained initialization and adapter architecture, with full encoder fine-tuning under shared task-specific AdamW settings (Loshchilov & Hutter, 2019) and training limits (Appendix B). SPAN checkpoints are selected by validation event F1 at IoU>0.3; the primary DENSE-FT checkpoint is selected by validation per-second AUPRC. We also evaluate the event-F1-selected dense checkpoint. For every checkpoint, inference settings are selected on validation event F1 at IoU>0.3 and fixed for all test metrics.

Metrics. Our primary metric is class-aware event F1 at IoU>0.5. Within each recording, confidence-sorted predictions are greedily matched one-to-one to same-class references using a strict IoU threshold. True positives, false positives and false negatives are summed over recordings and held-out folds before computing each seed's F1. We report the mean and population standard deviation across three seeds. For paired differences, we follow the repository's percentile bootstrap (Efron, 1979): 2,000 resamples of complete seed pairs, with random seed 0. Task summaries average all 18 backbone–seed differences and bootstrap those 18 pairs; these intervals describe variation within this evaluated collection. Appendix C specifies the resampling order and limitations. Secondary metrics are F1 at IoU>0.3, median absolute onset error among events matched at IoU>0.3, fragmentation, which is the fraction of reference events overlapped by at least two predictions, irrespective of class and per-second AUPRC. AUPRC complements event scoring under class imbalance, SPAN's sample scores come from its auxiliary dense head.

5 RESULTS

5.1 EVENT-LEVEL GAINS ACROSS SIX BACKBONES

Table 1 shows that SPAN improves event F1 at IoU>0.5 for every backbone on every task. All 72 seed-matched differences are positive. Averaging the 18 differences per task gives gains of 8.3 percentage points on TUSZ (95% bootstrap CI 6.4–10.3), 8.3 on MI (6.6–10.1), 4.5 on TUAR (3.7–5.2) and 5.5 on arousal (3.7–7.3). Figure 2 shows the paired effects and their uncertainty for each backbone.

The improvement depends on both task and backbone. LaBraM gains 14.3 points on TUSZ, BIOT gains 14.4 on MI, and EEGPT gains 13.4 on arousal. Backbone choice remains consequential: with dense adaptation, the range across backbone means is 19.6 points on TUSZ, 14.7 on MI, 3.2 on TUAR and 12.0 on arousal. Thus, adaptation produces substantial gains within backbones while preserving meaningful differences between them.

The result also holds when both methods select checkpoints by validation event F1. Against this dense variant, SPAN again improves all 72 pairs, with task-average gains of 8.4, 8.4, 3.9 and 5.7 points (Appendix E). At the less restrictive IoU>0.3 threshold, average gains remain positive on all tasks, although 61 of 72 individual differences are positive. The consistency is strongest when evaluation requires greater temporal overlap.

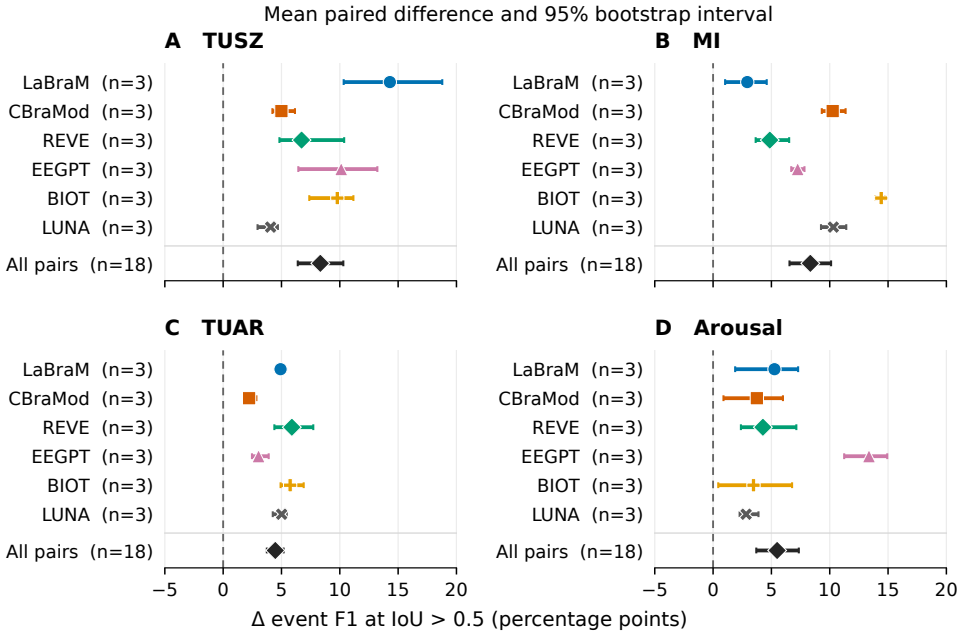

Figure 2: Paired test event-F1 differences, SPAN minus DENSE-FT, at IoU>0.5. Dots show the mean over three complete seed pairs per backbone; bars are 95% percentile bootstrap intervals from 2,000 resamples. "All pairs" averages and resamples the 18 backbone–seed differences within a task. The dashed line marks zero; all panels use the same $[-5, 20]$ percentage-point scale.

We can observe that, for event-level evaluation, SPAN is the better adaptation of every backbone on every task, and the margin is not an artifact of checkpoint selection. The dense head keeps the higher per-second AUPRC in 21 of 24 backbone–task means (Appendix D) yet yields fewer correct events: a good probability timeline is not the same as a good event decoder.

5.2 BOUNDARY LOCALIZATION AND EVENT COHERENCE

Event F1 rewards detection at a fixed overlap, it does not say whether boundaries are placed accurately or whether each reference is covered by a single prediction. We examine both, because onset statistics are conditional on matched events.

Onset errors decrease for all six backbones on TUSZ, MI and arousal (Table 6). For LaBraM, mean seed-level median absolute onset error falls from 9.66 to 3.24 s on TUSZ and from 0.752 to 0.412 s on MI, and from 1.00 to 0.864 s on arousal. This is not backbone dependent, on references matched by both methods, the error still falls from 9.62 to 5.66 s, from 0.78 to 0.31 s and from 1.05 to 0.83 s, in 53 of 54 backbone–seed cells. TUAR shows the opposite pattern: errors increase for all six backbones, from 0.45–0.51 s under DENSE-FT to 0.70–0.98 s under SPAN. Three observations locate the cause. The dense method matches only 14% of the references at IoU>0.3 (SPAN: 28%), and the references that only SPAN recovers carry larger errors (0.86 s), but the degradation persists on references matched by both methods (0.48 versus 0.73 s, in 18 of 18 cells). SPAN emits on average 2.9 same-class predictions overlapping each matched reference, against 1.02 for the dense method, and 72% of its matched references have more than one candidate. If each reference is paired with its best-aligned candidate instead of the most confident one, SPAN's onset error drops to 0.43 s, marginally below the dense value of 0.46 s (Table 13). SPAN's higher onset error on artifacts reflects repeated localization of the same event, rather than systematically less precise boundaries.

Higher event F1 does not imply less fragmentation. The fraction of references overlapped by two or more predictions increases for every backbone–task combination (Table 7). On TUAR it reaches 70.5–77.8% with SPAN, compared with 9.9–18.3% under DENSE-FT. Because the measure counts overlaps irrespective of class, simultaneous artifact classes contribute as well as repeated predictions.

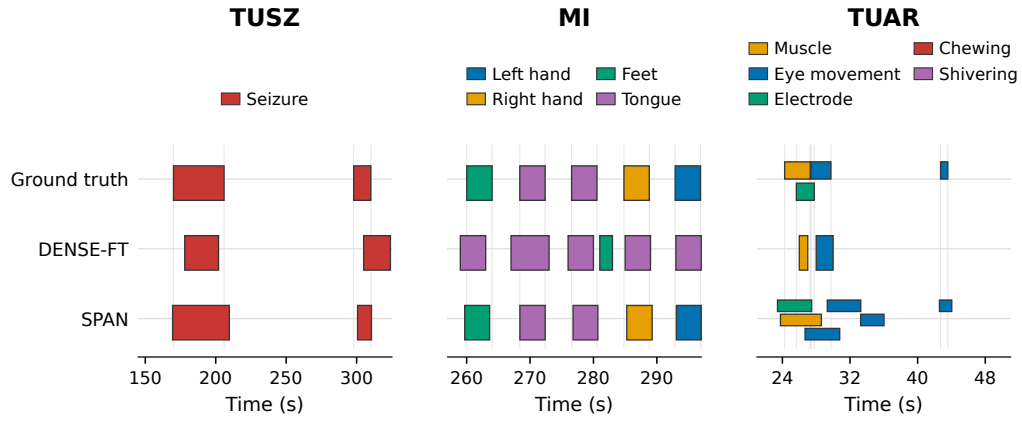

Figure 3: Illustrative test segments from LaBraM, seed 7, intentionally selected to show favourable SPAN predictions; they are not statistically representative. Both methods use the same recording and displayed interval within each panel, with the original validation-selected settings. TUSZ illustrates closer seizure boundaries; MI shows improved event classes; TUAR shows recovery of overlapping and missed artifacts, alongside an extra eye-movement prediction. Colours denote event classes and overlapping bars occupy separate lanes. Boundary markers indicate intervals continuing outside the window.

This identifies event coherence in overlapping annotations as a remaining weakness. Per-second AUPRC is also lower for SPAN in 21 of 24 backbone–task means (Table 10), showing that better interval detection need not accompany a better auxiliary probability timeline.

Figure 3 provides a qualitative view of these behaviours. SPAN more closely follows seizure boundaries in the selected TUSZ segment and recovers the four reference classes in the MI segment, where the dense output includes class errors. In TUAR, it recovers an electrode artifact and a later eye-movement event missed by the dense method, but also predicts an extra eye-movement interval.

We can observe that, where events are isolated, SPAN places boundaries more accurately than the dense pipeline, and the gain survives restriction to commonly detected events. While annotations are dense and overlapping, its remaining weakness is coherence.

5.3 Decoder comparison and component ablation

To test whether a plain set decoder obtains the same benefit, we replace SPAN's head and loss with a DANCE-style decoder on LaBraM and REVE (Table 2; Appendix F). SPAN performs better in all 24 seed pairs. Mean gains over the six pairs per task are 6.6 points on TUSZ, 10.4 on MI, 8.6 on TUAR and 24.4 on arousal. The plain decoder improves over dense adaptation on TUSZ but underperforms on the other tasks. The end-to-end DANCE reference is lower still, reaching 39.0% on TUSZ, 16.2% on MI, 9.2% on TUAR and 3.1% on arousal.

We can observe that, pretrained features lift a plain set decoder above the end-to-end model on three of four tasks, and SPAN's decoder and training design lift it further on all four. Because the plain-set comparison changes several components at once, we therefore remove one component of SPAN at a time and retrain LaBraM and REVE with three seeds under the unchanged protocol: the Gaussian locality, the denoising queries and loss, or the multiscale pyramid. We report TUAR and arousal, the two tasks with the most reference events (Table 3; Appendix G). On TUAR, removing locality lowers event F1 at IoU>0.5 by 3.6 points over the six seed pairs, removing denoising lowers it by 2.4 points, and multiscale pyramid lowers it by 0.4 points, all in all six pairs. Without locality the median onset error also rises, from 0.70 to 0.96 s for LaBraM and from 0.79 to 1.43 s for REVE. On arousal, no single removal changes event F1 beyond seed variation, and every interval includes zero. However, removing locality and denoising sill slightly lower the performance based on the mean event F1.

Table 2: Decoder comparison on LaBraM and REVE. Test event F1 at IoU>0.5 (%), mean$\pm$SD over three seeds. The last column gives the paired SPAN minus plain-set difference in percentage points with its 95% bootstrap CI. The plain set decoder applies a DANCE-style head and objective to the adapted FM features; the last row of each block is a DANCE trained end-to-end from scratch as in their paper, without a pretrained backbone, for reference.

Task	Backbone	DENSE-FT	Plain set	SPAN	Δ vs. plain set
TUSZ	LaBraM	42.3±2.5	48.6±0.7	56.6±4.2	+8.0 [+2.8, +11.2]
	REVE	48.1±0.8	49.6±1.9	54.8±2.6	+5.2 [+0.2, +10.8]
	DANCE		39.0±2.0		
MI	LaBraM	18.7±0.6	16.0±1.2	21.7±1.3	+5.7 [+3.0, +8.2]
	REVE	27.7±0.8	17.4±1.1	32.5±1.5	+15.1 [+11.7, +17.1]
	DANCE		16.2±1.7		
TUAR	LaBraM	14.6±0.2	10.4±0.3	19.6±0.2	+9.2 [+8.7, +9.9]
	REVE	12.4±1.2	10.3±0.1	18.3±0.4	+8.0 [+7.2, +8.5]
	DANCE		9.2±0.2		
Arousal	LaBraM	24.9±0.2	5.7±2.5	30.2±2.2	+24.4 [+17.8, +27.9]
	REVE	24.0±0.3	3.9±1.1	28.3±1.8	+24.4 [+22.0, +27.9]
	DANCE		3.1±0.3		

Table 3: Component ablation on LaBraM and REVE. Test event F1 at IoU>0.5 (%), mean$\pm$SD over three seeds. Δ rows give the seed-paired difference from full SPAN over the six backbone with its 95% bootstrap CI; bold intervals exclude zero.

Task	Backbone	Full SPAN	− Gaussian locality	− Denoising	− Multiscale pyramid
TUAR	LaBraM	19.6±0.2	17.4±1.3	17.1±0.2	19.4±0.2
	REVE	18.3±0.4	13.2±0.3	15.9±0.5	18.0±0.3
	Δ, six pairs		**−3.6 [−4.9, −2.2]**	**−2.4 [−2.8, −1.9]**	**−0.4 [−0.6, −0.1]**
Arousal	LaBraM	30.2±2.2	29.3±2.5	29.1±2.1	29.4±1.1
	REVE	28.3±1.8	28.1±1.8	27.6±1.3	29.5±0.8
	Δ, six pairs		−0.5 [−2.8, +1.6]	−0.9 [−3.1, +1.7]	+0.3 [−2.0, +2.6]

6 DISCUSSION AND CONCLUSION

SPAN adapts six representative EEG FMs to continuous event decoding through one interface, and in every one of 72 paired comparisons it detects events better than dense fine-tuning of the same backbone. The effect is substantial: task-average gains of 4.5–8.3 points are of the same order as the differences between backbones under dense adaptation, and they survive event-based checkpoint selection for the baseline and a plain set decoder on the same features. The results support treating the adaptation head and objective as substantive choices when transferring pretrained EEG representations to event detection.

addressing the limitation, better detection did not bring better coherence. SPAN places boundaries more accurately when events are isolated, but on densely annotated artifacts it often produces multiple overlapping predictions for the same event. The one-to-one scorer counts these duplicates as fragmentation and can also assign a reference to a less well-aligned duplicate, increasing the measured onset error even when the best-aligned prediction is as accurate as that of the dense baseline. Future work should address overlapping event identities and prediction coherence while preserving the gains in event recall and overlap. Also, our study evaluates full fine-tuning on four fixed datasets and three seeds. It does not establish the benefit of pretraining over random initialization of the same backbone or the behavior of frozen encoders. Within this scope, the results provide consistent evidence that event-oriented adaptation improves the utility of pretrained EEG models for temporally localized prediction.

AI USE STATEMENT

We used generative AI tools to assist with coding, the preparation of tables and figures from saved experimental outputs, polishing the writing, and drafting sections of the paper. All AI-drafted or AI-polished text was reviewed and edited by the authors, and all AI-assisted code was reviewed by the authors before use; for example, the DANCE reimplementation was checked term by term against its published description. All research questions, experimental design, model implementation, training runs and measurements are the authors' own. Every number reported in this paper was produced by the authors' code and verified by the authors against the corresponding result files on disk. We did not use generative AI tools to generate research ideas or to develop theoretical models or conceptual frameworks, and no AI system generated experimental results, citations or claims that the authors did not independently verify. No AI system is an author of this work, and the authors take full responsibility for the final content of this submission, including any text that originated as an AI-assisted suggestion.

REPRODUCIBILITY STATEMENT

Section 3 specifies the shared representation, objectives and inference procedures. Appendix A describes preprocessing and splits, Appendix B gives implementation settings, and Appendix C defines aggregation and uncertainty. Tables and figures are generated from the evaluation outputs, with source paths, checkpoint selection and operating points recorded in the accompanying provenance files.

REFERENCES

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 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. *Journal of Neuroscience Methods*, 321:64–78, 2019.

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

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ahmad Hamid, Kaitlyn Gagliano, Sadia Rahman, Nicolas Tulin, Vivian Tchiong, Iyad Obeid, and Joseph Picone. The Temple University artifact corpus: An annotated corpus of EEG artifacts. In *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020.

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

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. doi: 10.3389/fnhum.2021.653659.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Jarod Lévy, Hubert Banville, Jérémy Rapin, Jean-Remi King, Thomas Moreau, and Stéphane d'Ascoli. DANCE: Detect and classify events in EEG. *arXiv preprint arXiv:2605.10688*, 2026.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Qichen Li, Shenghan Wen, Haoqi Sun, Wolfgang Ganglberger, Ayush Tripathi, Niels Turley, Samuel Waters, Arnav Gupta, Aditya Gupta, Manohar Ghanta, Bruce Nearing, Han Wu, Katie L. Stone, Chad Robichaux, Zhiyong Zhang, Qiao Li, Gauri Ganjoo, Christine Tsien Silvers, Bharath Gunapati, Kiran Maski, Samaneh Nasiri, Dennis Hwang, Lynn Marie Trotti, Umakanth Katwa, Gari D. Clifford, Emmanuel Mignot, Robert J. Thomas, and Michael Brandon Westover. The human sleep project: A multi-center clinical polysomnography dataset across the human lifespan. *Sleep*, pp. zsag215, 2026. doi: 10.1093/sleep/zsag215.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision (ECCV)*, 2022b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations (ICLR)*, 2022a.

Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 2022b.

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

Mathias Perslev, Michael Hejselbak Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-Time: A fully convolutional network for time series segmentation applied to sleep staging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38:5391–5420, 2017. doi: 10.1002/hbm.23730.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

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

Elisa Vasta, Thorir Mar Ingolfsson, Andrea Cossettini, Luca Benini, Tilman Beck, Emanuela Keller, and Una Pale. Evaluation of EEG foundation models for event-based burst-suppression detection in ICU. *arXiv preprint arXiv:2606.20074*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ICLR)*, 2025.

Lukas Wolf, Ard Kastrati, Martyna Beata Płomecka, Jie-Ming Li, Dustin Klebe, Alexander Veicht, Roger Wattenhofer, and Nicolas Langer. A deep learning approach for the segmentation of electroencephalography data in eye tracking applications. In *International Conference on Machine Learning*, 2022.

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

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022. doi: 10.1609/aaai.v36i8.20881.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, 2020.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

A Data processing and event construction

All tasks use caches of 10-s windows at 200 Hz together with a recording index, per-second labels and an exact interval index in seconds. Before windowing, each channel is band-pass filtered using a Butterworth filter (0.5–$70, \mathrm{Hz}$), notch-filtered at the relevant line frequency (50 or $60, \mathrm{Hz}$), and resampled using polyphase filtering to a common cache rate of $200, \mathrm{Hz}$. Set targets keep the original interval boundaries, whereas dense targets are rasterized onto token cells (Appendix B). Windows are assembled into chunks within each recording, with padding and a validity mask for the final chunk. EEGPT and LUNA inputs are resampled to their native 256 Hz inside the adapters.

TUSZ. We use the official v2.0.3 train/val/evaluation partition and 18 bipolar EEG derivations, and merge all seizure subtypes into one class. The evaluation set contains 467 events over 125.5 h. Both methods sample training chunks with weight $0.01 + 0.99 \min(1, 5r)$, where r is the fraction of the chunk occupied by seizures.

Motor imagery. BNCI2014-001 contains nine subjects and 22 EEG channels. The four-second imagery period of each trial defines one of four event classes: left hand, right hand, feet or tongue. Each run is a separate recording, and cue markers are hidden at inference. The held-out subject groups are (A01,A02), (A03,A04), (A05,A06), (A07,A08) and (A09); the first subject of the next group, with cyclic indexing, supplies the validation data. The five test folds cover 108 recordings and 5,076 events, and the scorer reports 11.4 h of valid evaluated time.

Artifacts. TUAR v3.0.1 uses 20 channels and five subject-grouped folds over 213 subjects and 310 recordings. Channel annotations are consolidated by class and onset, keeping the latest offset within each group; near-duplicate annotations with different onsets remain distinct, and classes may overlap. The five classes are muscle, eye movement, electrode artifact, chewing and shivering. The pooled test data contain 40,247 events over 99.7 h. Training chunks with more than 80 events keep only the 80 longest set targets; dense targets are unaffected.

Arousals. The HSP subset contains 1,800 nights from a fixed stratified subject split: 1,200 for training, 300 for val and 300 for testing. We use six EEG derivations and the expert arousal intervals. The test set contains 41,006 events over 2,268.1 h. Seconds without a scored sleep stage are excluded from dense training and from sample metrics, and for both methods a predicted event whose center falls in an unscored second is removed before event scoring. Boundary statistics are recomputed from the scored recording export rather than from an unfiltered query export. Per-epoch monitoring uses 60 validation nights; final operating-point selection uses all 300.

B Implementation and inference settings

Adapters. Table 4 lists the temporal layouts entering the common projection. Because the decoder works in physical time rather than a fixed token index, the native backbone rates may differ. All adapters output width 256 after aggregation, LayerNorm and projection.

Table 4: Backbone adapter outputs for a 10-s input. C denotes channels; D is the width before projection to 256. Rates are nominal tokens per second.

Backbone	Temporal aggregation	D	Tokens/s
LaBraM	$[C, 10, 200]$, mean over channels	200	1.0
CBraMod	$[C, 10, 200]$, mean over channels	200	1.0
REVE	$[C, 11, 512]$, mean over channels	512	1.1
EEGPT	$[40, 4, 512]$, concatenate summaries	2048	4.0
BIOT	$[C, 19, 256]$, mean over channels	256	1.9
LUNA	$[64, 4, 64]$, concatenate summaries	256	6.4

Channel identity. TUSZ and TUAR use bipolar montages of 18 and 20 pairs, MI uses 22 referential electrodes, and arousal uses six mastoid-referenced electrodes. LaBraM indexes its pretrained electrode embeddings by 10-20 name (TUH T3/T4/T5/T6 mapped to T7/T8/P7/P8) and averages the two embeddings of a bipolar or mastoid-referenced channel; the result stays trainable. REVE averages the 3-D coordinates of the two electrodes of such a channel, following its official position rule. LUNA places a bipolar pair at the mean of its two standard-1005 electrode positions and a mastoid-referenced channel at its electrode's position; its channel-name embedding table, used only by the pretraining decoder, is dropped. EEGPT uses its pretrained channel embeddings directly for the MI electrodes and for the arousal electrodes after removing the mastoid reference, and maps the TUSZ and TUAR bipolar inputs onto 20 named virtual channels with a 1×1 max-norm convolution. BIOT keeps the pretrained token of every channel in its 18-channel pretraining montage (16 TCP pairs plus C3-A2 and C4-A1, with T3/T5/T4/T6 and M1/M2 aliases) and appends a fresh trainable token for each other channel: FZ-CZ and CZ-PZ on TUSZ, the four T3-C3, C3-CZ, CZ-C4 and C4-T4 pairs on TUAR, all 22 MI channels, and the F3, F4, O1 and O2 derivations on arousal. CBraMod has no channel vocabulary; channel position enters only through its convolutional positional encoding over the channel-by-patch grid.

Dense targets. Dense labels are rasterized from the interval index onto the token cells of each backbone. For MI, TUAR and arousal, a cell receives the event class with the greatest coverage if that coverage exceeds half of the cell duration, and background otherwise; TUSZ applies the same threshold to the summed coverage of all seizure intervals. The dense target therefore assigns one class per cell even where reference events overlap, whereas set targets keep every interval. On arousal, cells without a scored sleep stage are excluded from the dense loss.

Training. Table 5 gives the task-specific settings, which are identical for the two methods of a pair. All runs use AdamW with head learning rate 10^{-4}, encoder learning rate 10^{-5}, weight decay 10^{-4}, gradient clipping at 1 and bf16 autocast. A warm-up precedes cosine decay to 5% of the peak rate, where progress is the larger of the step fraction and the elapsed-budget fraction. Runs use one NVIDIA A30 (24 GB) or RTX 6000 Ada (48 GB); memory-dependent batch sizes are combined with gradient accumulation to an effective batch of eight. Seeds 7, 17 and 37 control initialization, data order and denoising noise. Training caps are 12 h for TUSZ and arousal, 2 h per MI fold and TUAR fold, matched between the two methods within every backbone–seed pair. The saved training arguments record all settings.

Table 5: Task-specific configuration. The gap and margin apply to set-prediction stitching at adjacent chunk boundaries. Training may stop earlier through early stopping on the development split.

	TUSZ	MI	TUAR	Arousal
Chunk duration (s)	300	60	60	300
Event classes	1	4	5	1
Queries Q	10	16	80	16
Maximum epochs	30	100	100	30
Patience (epochs)	6	15	15	6
Warm-up (steps)	500	100	100	500
Stitch gap / margin (s)	10 / 10	1 / 1	1 / 1	2 / 2

SPAN decoder. The four-layer decoder has width 256, eight attention heads, feed-forward width 512 and dropout 0.1. Pyramid strides in tokens are $(1, 2, 3, 6)$ for LaBraM, CBraMod and REVE, $(4, 8, 12, 24)$ for EEGPT, $(2, 4, 6, 11)$ for BIOT and $(6, 13, 19, 38)$ for LUNA, which approximate common physical scales at each backbone's native rate. Given raw reference coordinates $r = (r_0, r_1)$, let $\bar{r} = 8 \tanh(r/8)$ and $d_{\min} = 1/L$. The interval geometry is

$$d = \exp[\log d_{\min} + \mathrm{sigmoid}(\bar{r}_1)(-\log d_{\min})], \qquad \mu = d/2 + \mathrm{sigmoid}(\bar{r}_0)(1 - d), \tag{7}$$

so durations are at least one second and boundaries stay inside the chunk; decoder offsets refine the raw coordinates. Training uses two denoising groups with noise scale 0.4. The focal exponent is 2, the no-event weight 0.5, the endpoint L1 weight 1 and the distance-IoU weight 2. The intermediate

decoder losses have total weight 0.5, and denoising and auxiliary dense supervision each have weight 1. Dense–event consistency regularization is disabled for SPAN.

Development-selected inference. The dense sweep varies the smoothing width (1, 3 or 7 tokens), the confidence threshold, the merge gap and the minimum duration. TUSZ uses merge gaps of 0, 1, 2, 5, 10 or 30 s and minimum durations of 0, 1, 2, 5 or 10 s. MI and TUAR use merge gaps of 0, 0.5, 1 or 2 s, arousal uses 0, 1 or 2 s, and these three tasks use minimum durations of 0, 1, 2 or 3 s. The threshold grid runs from 0.05 to 0.95 in steps of 0.05, followed by 0.96, 0.97, 0.98, 0.985, 0.99, 0.9925, 0.995, 0.9975, 0.9985, 0.999, 0.9995 and 0.9999. SPAN sweeps only the confidence threshold, with its fixed stitching settings. No suppression or additional merging is applied in the main results; Appendix H evaluates suppression separately. The provenance files record each selected checkpoint and operating point, including the corrected EEGPT arousal rerun.

C AGGREGATION AND UNCERTAINTY

For each method, backbone and seed, we pool event counts over all test recordings, and over all five folds for MI and TUAR, before computing F1. Onset errors are absolute differences between the original, unclipped onsets of pairs matched at IoU>0.3; we pool these errors across folds, take one median per seed and average the three medians. These values are conditional on matched events, and the methods can match different subsets of the references.

Paired intervals use 2,000 percentile-bootstrap resamples of the seed-level differences, with the 2.5th and 97.5th percentiles as endpoints. For each task, NumPy's default random generator starts at seed 0 and advances through LaBraM, CBraMod, REVE, EEGPT, BIOT and LUNA in that order; the task summary then resamples the 18 backbone–seed differences as a flat collection. This is not a hierarchical subject or backbone bootstrap. With only three seeds per backbone, the intervals give a limited description of run-to-run variation rather than population-level clinical uncertainty. All 72 pairs are complete.

The diagnostics in Appendices D and H recompute the scorer's class-aware greedy matching from the saved test events, pooling the five folds within a cell before any statistic is taken. Unless stated otherwise, appendix tables report means over the 18 backbone–seed cells of a task.

D TEMPORAL AND SAMPLE-LEVEL DIAGNOSTICS

Boundary and fragmentation tables. Table 6 gives the boundary diagnostic needed to interpret the event-F1 gains: onset errors improve on TUSZ, MI and arousal and degrade on TUAR. Table 7 complements it with the fraction of references that have positive overlap with at least two predictions, irrespective of class. Its denominator includes all reference events, including undetected ones, and distinct classes overlapping the same reference are counted, so it is not a duplicate-only measure.

Table 6: Mean of seed-level median absolute onset errors (s) for events matched at IoU>0.3. Each cell is DENSE-FT $\rightarrow$ SPAN; lower is better. All six backbones have three complete seed pairs.

Backbone	TUSZ	MI	TUAR	Arousal
LaBraM	$9.66 \rightarrow 3.24$	$0.75 \rightarrow 0.41$	$0.45 \rightarrow 0.70$	$1.00 \rightarrow 0.86$
CBraMod	$10.03 \rightarrow 4.17$	$0.77 \rightarrow 0.36$	$0.45 \rightarrow 0.87$	$1.00 \rightarrow 0.87$
REVE	$8.11 \rightarrow 3.80$	$0.58 \rightarrow 0.31$	$0.51 \rightarrow 0.79$	$1.00 \rightarrow 0.84$
EEGPT	$8.22 \rightarrow 4.78$	$1.34 \rightarrow 0.55$	$0.45 \rightarrow 0.98$	$1.00 \rightarrow 0.86$
BIOT	$12.20 \rightarrow 6.28$	$1.21 \rightarrow 0.39$	$0.51 \rightarrow 0.83$	$1.08 \rightarrow 0.96$
LUNA	$11.06 \rightarrow 5.84$	$0.25 \rightarrow 0.11$	$0.45 \rightarrow 0.70$	$1.16 \rightarrow 0.86$

Detection statistics. Table 8 decomposes event F1 into precision and recall and adds the number of predictions per reference and the false-positive rate. The SPAN gains come mainly from recall: at IoU>0.3, recall rises on every task (TUSZ 44.0 to 50.6%, MI 23.1 to 32.6%, TUAR 14.1 to 28.3%,

Table 7: Fragmentation rate (%), averaged over three seeds. Each cell is DENSE-FT → SPAN. Higher values mean that more references are overlapped by multiple predictions.

Backbone	TUSZ	MI	TUAR	Arousal
LaBraM	0.9 → 5.1	3.3 → 19.6	10.8 → 73.1	0.9 → 11.1
CBraMod	1.1 → 4.6	3.1 → 11.9	11.3 → 77.8	0.7 → 7.0
REVE	0.9 → 1.1	1.4 → 11.3	9.9 → 72.5	0.8 → 11.5
EEGPT	0.9 → 3.0	8.9 → 40.0	18.3 → 75.1	1.0 → 12.1
BIOT	1.3 → 1.4	7.6 → 11.2	11.4 → 72.8	0.8 → 11.2
LUNA	1.3 → 3.9	1.4 → 7.8	11.0 → 70.5	1.0 → 9.4

arousal 36.4 to 43.2%), while precision is similar on TUSZ, MI and arousal. TUAR is the exception: the dense method predicts only 0.37 events per reference and reaches a precision of 38.8% on the artifacts it does report, whereas SPAN predicts 1.14 events per reference at 25.1% precision, with 8,293 instead of 2,238 false positives per 24 h. The dense method under-predicts on all tasks except arousal (0.37–0.88 predictions per reference), which is the behavior expected from a threshold on a smoothed probability timeline.

Table 8: Precision and recall (%) and predictions per reference of the primary checkpoints at their development-selected operating points, pooled over folds and averaged over 18 backbone–seed cells per task.

Task	Method	IoU>0.3		IoU>0.5		Pred./ref.
		Precision	Recall	Precision	Recall	
TUSZ	DENSE-FT	53.2	44.0	42.4	35.0	0.88
	SPAN	51.2	50.6	46.6	46.0	1.01
MI	DENSE-FT	26.1	23.1	20.8	18.5	0.88
	SPAN	28.6	32.6	26.3	29.9	1.16
TUAR	DENSE-FT	38.8	14.1	25.5	9.3	0.37
	SPAN	25.1	28.3	17.0	19.2	1.14
Arousal	DENSE-FT	32.6	36.4	21.4	23.8	1.16
	SPAN	33.6	43.2	25.0	32.1	1.31

Per-class event F1. Table 9 reports class-wise event F1 at IoU>0.5 for the two multi-class tasks, with per-class counts pooled over the five folds of each cell. On MI, SPAN improves every class, by 5.9 (tongue) to 13.8 points (feet); the dense baseline is weakest on feet (12.6%), the class that the example in Figure 3 shows it confusing with tongue. On TUAR, the three frequent classes improve (muscle $+5.4$, eye movement $+3.1$ and electrode $+4.0$ points, the first and third in all 18 cells), whereas chewing (868 test events) is marginally worse and shivering (41 test events) is detected by neither method. The class-aware gain on TUAR is therefore carried by the classes with thousands of test events, and the rare classes remain unsolved by both interfaces.

Sample-level scores. Table 10 measures per-second ranking separately from interval detection, using the dense classifier or SPAN's auxiliary dense head. For MI and TUAR, per-second labels and probabilities are pooled across the five folds before class-wise average precision is computed and macro-averaged over event classes; these values are not averages of fold-level AUPRC. Unscored arousal seconds are excluded.

Matched-subset decomposition. Onset statistics are conditional on matched events, and the two methods match different references. We therefore split the matched references of every cell into those matched by both methods and those matched by SPAN only. Table 11 reports recall at IoU>0.3, the median absolute onset error on all matched references, on the shared subset and on the SPAN-only subset, and the reference-duration distribution. On TUSZ, MI and arousal the shared-

Table 9: Per-class test event F1 at IoU>0.5 (%) for the multi-class tasks. Counts are pooled over the five folds within each backbone–seed cell; values are means over the 18 cells, and Δ is the mean paired difference. The test-event column counts references per class.

Task	Class	Test events	DENSE-FT	SPAN	Δ
MI	Left hand	1,244	22.5	32.4	+9.8
	Right hand	1,294	16.9	26.1	+9.2
	Feet	1,261	12.6	26.4	+13.8
	Tongue	1,277	20.6	26.5	+5.9
	All classes	5,076	19.6	27.9	+8.3
TUAR	Muscle	14,840	14.0	19.4	+5.4
	Eye movement	15,817	17.1	20.2	+3.1
	Electrode	8,681	6.4	10.4	+4.0
	Chewing	868	3.5	2.8	-0.7
	Shivering	41	0.0	0.0	+0.0
	All classes	40,247	13.5	18.0	+4.5

Table 10: Per-second AUPRC (%), mean±SD over three seeds. Multi-class tasks use macro-average precision over event classes.

	TUSZ		MI		TUAR		Arousal	
Backbone	DENSE-FT	SPAN	DENSE-FT	SPAN	DENSE-FT	SPAN	DENSE-FT	SPAN
LaBraM	65.7±3.6	58.5±2.1	18.6±0.4	15.5±0.3	35.9±3.5	36.2±0.8	26.5±0.2	26.6±1.0
CBraMod	65.2±5.2	60.1±3.9	17.6±0.4	15.7±0.2	31.9±0.3	29.5±0.9	25.4±0.3	21.2±0.8
REVE	75.3±0.4	61.3±3.2	25.2±0.7	24.9±1.2	35.8±3.3	34.7±0.6	25.6±0.4	25.4±0.3
EEGPT	48.2±5.0	50.0±5.1	15.6±0.1	14.8±0.1	34.8±0.5	34.4±0.8	11.0±0.2	8.3±0.2
BIOT	49.6±2.2	43.8±1.2	16.1±0.2	15.7±0.1	30.6±1.1	29.7±0.2	25.5±0.1	22.6±1.5
LUNA	51.0±1.0	38.4±0.5	24.6±0.3	22.8±0.1	34.2±0.6	34.1±0.6	27.2±0.5	21.4±0.2

subset error is lower for SPAN in 18, 18 and 17 of 18 cells; on TUAR it is higher in all 18. The median signed onset error and the ratio of predicted to reference duration explain the direction. On TUSZ the dense pipeline starts 7.50 s late with a length ratio of 0.78 (SPAN: +0.80 s and 0.98). On MI it starts 0.44 s early with a ratio of 1.17 (SPAN: 0.00 s and 1.02). On arousal both start early, with ratios 1.35 and 1.28. On TUAR both start slightly early (-0.08 and -0.14 s), SPAN ends later ($+0.15$ s), and the ratios are 1.03 and 1.12.

Table 11: Matched-subset decomposition of onset errors. Recall is the fraction of references matched at IoU>0.3. Onset errors are median absolute onset errors (s) over all matched references, over references matched by both methods, and over references matched by SPAN only. Arrows read DENSE-FT $\rightarrow$ SPAN. The last column gives the median reference duration and the fraction of references shorter than 2 s. Means over 18 backbone–seed cells.

Task	Recall@0.3 (%)	Onset error (s)			Ref. duration (s)
	DENSE-FT $\rightarrow$ SPAN	all	both	SPAN only	median (% $<$2 s)
TUSZ	44.0 $\rightarrow$ 50.6	9.88 $\rightarrow$ 4.68	9.62 $\rightarrow$ 5.66	2.91	34.9 (0%)
MI	23.1 $\rightarrow$ 32.6	0.82 $\rightarrow$ 0.35	0.78 $\rightarrow$ 0.31	0.38	4.0 (0%)
TUAR	14.1 $\rightarrow$ 28.3	0.47 $\rightarrow$ 0.81	0.48 $\rightarrow$ 0.73	0.86	2.5 (43%)
Arousal	36.4 $\rightarrow$ 43.2	1.04 $\rightarrow$ 0.88	1.05 $\rightarrow$ 0.83	1.03	3.0 (0%)

TUAR by duration and by assignment. Table 12 splits the TUAR references by duration: 43% are shorter than 2 s and 22% shorter than 1 s, which is the minimum duration SPAN can predict in a 60-s chunk. SPAN matches about twice as many references in every bin, and its onset error is larger in every bin, including references longer than 5 s, where the minimum duration is irrelevant. The

error also increases in every artifact class with matched events (muscle 0.60 to 0.96 s, eye movement 0.37 to 0.68 s, electrode 0.88 to 1.09 s; chewing is matched only about 25 times per cell and also degrades). Table 13 isolates the assignment. For each matched reference we count the same-class predictions with IoU>0.3 and record the onset error of the best-aligned one. The dense method has essentially one candidate per reference, so its matched and best-aligned errors coincide. SPAN has 2.91 candidates on average, more than one for 72% of matched references, and its best-aligned candidate (0.43 s) is slightly more accurate than the dense prediction (0.46 s); its matched prediction (0.81 s) is worse only because the greedy scorer takes the most confident duplicate. The TUAR onset degradation is therefore a consequence of duplicate overlapping predictions, the same behavior that raises the fragmentation rate. Appendix H shows that overlap-based suppression does not remove these duplicates without cost.

Table 12: TUAR references by duration. Columns give the share of references in each bin, the fraction matched at IoU>0.3 by each method, and the median absolute onset error (s) of the matched references. Means over 18 backbone–seed cells.

Duration (s)	References (%)	Matched (%)		Onset error (s)	
		DENSE-FT	SPAN	DENSE-FT	SPAN
<1	22	6	9	0.25	0.43
1–2	21	15	27	0.27	0.40
2–5	28	17	36	0.44	0.64
≥ 5	30	17	37	1.44	2.36

Table 13: TUAR onset error under the scorer's confidence-ordered matching versus the best-aligned same-class candidate with IoU>0.3, with the number of such candidates per matched reference. Means over 18 backbone–seed cells.

	DENSE-FT	SPAN
Onset error of the matched prediction (s)	0.47	0.81
Onset error of the best-aligned candidate (s)	0.46	0.43
Same-class candidates per matched reference	1.02	2.91
References with more than one candidate (%)	2	72

E SENSITIVITY TO CHECKPOINT SELECTION AND OVERLAP THRESHOLD

Dense checkpoint selection. The primary DENSE-FT checkpoint is selected by validation per-second AUPRC. Table 14 reports the absolute test event F1 of this checkpoint next to the checkpoint selected by validation event F1 at IoU>0.3, evaluated with the same scorer and the same validation-selected post-processing. Event-F1 selection changes the dense F1 at IoU>0.5 by between -0.24 and $+0.55$ points on average per task and gives the higher value in 34 of 72 cells, so neither criterion systematically favors the dense baseline; the largest single change is REVE on TUSZ, from 48.1% to 43.3%. Table 15 then gives the paired differences between SPAN and the event-selected dense checkpoint; all 72 differences remain positive.

Overlap threshold. Table 16 evaluates the primary checkpoints at IoU>0.3 with the same operating points as the main results; positive differences occur in 14/18 TUSZ pairs, 16/18 MI pairs, 18/18 TUAR pairs and 13/18 arousal pairs. The primary result therefore persists under event-oriented checkpoint selection and is more consistent at the stricter overlap threshold.

F PLAIN SET DECODER AND END-TO-END REFERENCE

Table 17 lists the three decoder configurations compared in Section 5.3. The plain set decoder is our DANCE-style implementation applied to the adapted LaBraM and REVE features: adapter to-

Table 14: Test event F1 (%) of the two DENSE-FT checkpoints, selected by validation per-second AUPRC (the primary checkpoint) or by validation event F1 at IoU>0.3; mean±SD over three seeds.

	TUSZ		MI		TUAR		Arousal	
Backbone	AUPRC	event F1	AUPRC	event F1	AUPRC	event F1	AUPRC	event F1
Event F1 at IoU>0.5								
LaBraM	42.3±2.5	42.4±2.4	18.7±0.6	18.8±0.2	14.6±0.2	15.5±0.4	24.9±0.2	24.6±0.3
CBraMod	41.7±1.4	43.3±0.6	17.6±0.5	17.3±1.3	14.8±0.4	14.6±0.3	23.9±1.1	23.4±1.0
REVE	48.1±0.8	43.3±2.0	27.7±0.8	28.0±1.9	12.4±1.2	14.3±0.2	24.0±0.3	23.5±0.6
EEGPT	36.2±3.5	38.3±0.9	12.9±0.1	13.3±0.7	13.2±0.7	13.0±0.2	12.9±0.5	13.0±0.4
BIOT	28.4±1.8	27.5±0.4	14.1±0.2	13.9±0.6	11.6±0.9	12.1±0.3	24.2±0.2	23.6±0.3
LUNA	29.7±2.2	30.8±0.9	26.5±1.0	25.8±1.0	14.5±0.3	14.9±0.4	24.6±0.4	25.0±1.1
Event F1 at IoU>0.3								
LaBraM	51.2±1.5	51.1±1.6	24.2±0.7	23.7±0.3	22.1±0.3	23.5±0.4	36.8±0.6	37.0±0.4
CBraMod	51.6±0.6	51.7±0.9	22.5±0.7	22.5±0.8	22.3±0.6	22.4±0.4	35.9±0.4	35.6±0.5
REVE	56.9±0.6	53.6±1.3	33.4±0.6	33.4±2.0	19.2±1.9	21.9±0.3	37.2±0.4	37.1±0.2
EEGPT	44.1±4.5	44.8±0.6	18.0±0.2	18.6±0.4	20.6±0.5	20.9±0.4	21.6±0.4	21.8±0.3
BIOT	39.4±0.7	38.9±1.1	19.7±0.2	19.2±1.0	17.9±1.3	18.5±0.5	36.3±0.2	36.1±0.3
LUNA	41.2±0.7	42.1±1.2	29.2±1.2	28.2±0.9	21.5±0.3	22.1±0.6	37.4±0.4	37.6±0.5

Table 15: Paired event-F1 differences at IoU>0.5: SPAN minus event-F1-selected DENSE-FT, in percentage points with 95% bootstrap intervals. Each backbone uses three seed pairs; the final row uses 18.

Backbone	TUSZ	MI	TUAR	Arousal
LaBraM	+14.1 [+10.3, +18.3]	+2.9 [+0.8, +4.1]	+4.1 [+3.7, +4.4]	+5.6 [+2.8, +7.3]
CBraMod	+3.4 [+0.6, +5.9]	+10.6 [+9.1, +13.4]	+2.5 [+2.2, +2.7]	+4.3 [+0.9, +6.0]
REVE	+11.5 [+10.7, +12.6]	+4.5 [+0.5, +8.6]	+4.0 [+3.2, +4.7]	+4.9 [+3.9, +6.6]
EEGPT	+8.0 [+6.8, +8.8]	+6.9 [+5.4, +8.4]	+3.2 [+2.7, +3.5]	+13.2 [+11.1, +14.7]
BIOT	+10.7 [+10.0, +11.0]	+14.6 [+13.7, +15.3]	+5.2 [+5.0, +5.6]	+4.0 [+0.9, +6.8]
LUNA	+2.9 [+1.5, +4.3]	+11.0 [+10.0, +11.6]	+4.6 [+3.3, +5.7]	+2.5 [+0.9, +4.2]
All pairs	+8.4 [+6.4, +10.6]	+8.4 [+6.3, +10.4]	+3.9 [+3.4, +4.4]	+5.7 [+4.0, +7.5]

kens are averaged into one-second bins, projected from 256 to 128 dimensions and given sinusoidal position codes, and a plain DETR decoder with learned content queries and direct sigmoid endpoint regression predicts the event set. Training combines cross-entropy, a $5(1-\mathrm{IoU})$ localization loss, auxiliary dense supervision from a linear head on the one-second memory and dense–event consistency with weight 0.5. It omits SPAN's pyramid, anchors, Gaussian bias, denoising and intermediate decoder losses. The comparison preserves the pretrained initialization, splits, query budgets and event-based selection rule but changes the decoder, objective and memory construction together, so it evaluates the complete adaptation design under our training protocol.

The end-to-end reference replaces the pretrained backbone and adapter by a convolutional encoder trained from scratch: a 1×1 channel mixer, five residual dilated convolution blocks (kernel 9, dilations 1 to 16, batch normalization, GELU, dropout 0.2), a projection to 128 dimensions and a Perceiver resampler with one latent per second (six cross- and self-attention blocks, two heads, feed-forward width 512). The same decoder, dense head and objective as the plain set decoder are applied to the latents. It uses our chunking, optimizer and training caps, and one fixed montage per task with the 1×1 mixer in place of DANCE's sensor-position attention. DANCE has public code; our implementation and protocol are not an exact reproduction of its reported benchmark scores. Its three-seed mean event F1 at IoU>0.5 is 39.0% on TUSZ, 16.2% on MI, 9.2% on TUAR and 3.1% on arousal, and it is excluded from the six-backbone paired aggregate. On arousal, the plain-set and end-to-end runs used the 11 h training cap, so they are budget-matched with each other but not with the 23 h cap of the primary LaBraM and REVE pairs.

Table 16: Test event F1 at IoU>0.3 (%), mean over three seeds. Each cell is DENSE-FT $\rightarrow$ SPAN, using the primary checkpoints and their original validation-selected operating points.

Backbone	TUSZ	MI	TUAR	Arousal
LaBraM	51.2 $\rightarrow$ 60.3	24.2 $\rightarrow$ 24.8	22.1 $\rightarrow$ 28.6	36.8 $\rightarrow$ 38.8
CBraMod	51.6 $\rightarrow$ 50.6	22.5 $\rightarrow$ 30.2	22.3 $\rightarrow$ 25.2	35.9 $\rightarrow$ 38.3
REVE	56.9 $\rightarrow$ 58.7	33.4 $\rightarrow$ 35.0	19.2 $\rightarrow$ 27.0	37.2 $\rightarrow$ 37.5
EEGPT	44.1 $\rightarrow$ 51.0	18.0 $\rightarrow$ 23.6	20.6 $\rightarrow$ 25.0	21.6 $\rightarrow$ 36.5
BIOT	39.4 $\rightarrow$ 44.5	19.7 $\rightarrow$ 30.8	17.9 $\rightarrow$ 25.4	36.3 $\rightarrow$ 37.7
LUNA	41.2 $\rightarrow$ 38.7	29.2 $\rightarrow$ 38.0	21.5 $\rightarrow$ 27.7	37.4 $\rightarrow$ 36.8

Table 17: Decoder configurations compared in Section 5.3. The plain set decoder and the end-to-end model share DANCE's decoder and objective; the end-to-end model additionally replaces the pretrained backbone and adapter by an encoder trained from scratch. Q is the task query budget of Table 5.

	SPAN	Plain set decoder	End-to-end
Input	adapter tokens Z (256-d) at the native token rate	adapter tokens Z averaged into 1-s bins	raw EEG: 1×1 channel mixer, five dilated conv blocks, Perceiver with one latent per second
Decoder memory	masked pyramid at strides $\approx$1, 2, 3, 6 s with level and Fourier time codes	linear 256$\rightarrow$128 and sinusoidal position code	128-d latents and sinusoidal position code
Decoder	4 layers, 8 heads, width 256, FF 512	4 layers, 4 heads, width 128, FF 512	as plain set
Queries	Q content queries with chunk-conditioned scale and shift; anchors (μ, d) refined at every layer	Q learned queries	Q learned queries
Attention bias	Gaussian locality with learned per-layer width	none	none
Interval output	anchor plus offsets, log-duration, $d_{\min} = 1$ s	direct sigmoid endpoints, $d_{\min} = 1$ s	as plain set
Matching cost	$-\widehat{p}(c) + \|\widehat{b} - b\|_1 + 2\,(1 - \mathrm{dIoU})$ at every layer	$-\widehat{p}(c) + 5\,(1 - \mathrm{IoU})$ at the final layer	as plain set
Classification loss	focal ($\gamma = 2$), no-event weight 0.5	cross-entropy	cross-entropy
Localization loss	L1 (weight 1) and dIoU (weight 2)	$5\,(1 - \mathrm{IoU})$	as plain set
Auxiliary terms	intermediate layers (0.5), denoising (1), dense head (1)	dense head (1), dense–event consistency (0.5)	as plain set
Denoising queries	two groups, noise scale 0.4	none	none
Trained parts	backbone, adapter and head	backbone, adapter and head	encoder and head, from scratch

G COMPONENT ABLATION

Each arm of Section 5.3 differs from full SPAN in one setting. *Without Gaussian locality*, cross-attention receives no additive bias, so Eq. 2 reduces to scaled dot-product attention; the anchors, their Fourier conditioning of self- and cross-attention and the iterative refinement are unchanged. *Without denoising*, no denoising queries are appended and $\mathcal{L}_{\mathrm{DN}}$ is dropped from Eq. 3. *Without the multiscale pyramid*, the decoder memory is the single finest level, that is the adapter tokens with their projection, LayerNorm, level embedding and Fourier time code, which removes about 0.2 M head parameters. Backbone, adapter, optimizer, query budget, training caps, checkpoint selection and the validation-selected confidence threshold follow the primary SPAN runs of the same backbone and

seed, which serve as the full arm. The ablation adds 36 fine-tuning cells: two backbones, two tasks, three arms and three seeds.

Table 18 gives the per-backbone results. On TUAR, the locality and denoising effects are negative with intervals excluding zero for both backbones, and the locality effect is larger for REVE (-5.1 points) than for LaBraM (-2.2). Removing either component also raises the median onset error, while fragmentation stays near 72–75% in every arm, so none of the three components is the source of the duplicate predictions discussed in Section 5.2. On arousal, no per-backbone interval excludes zero; removing locality lowers fragmentation, from 11.1 to 8.2% for LaBraM and from 11.5 to 9.2% for REVE, without a reliable change in event F1. Intervals resample three seed pairs per backbone and are therefore wide; they describe variation within the evaluated runs.

Table 18: Component ablation per backbone. Test event F1 (%), median absolute onset error among events matched at IoU>0.3 and fragmentation, mean$\pm$SD over three seeds. ΔF1@0.5 is the seed-paired difference from full SPAN with its 95% bootstrap CI over three pairs; bold marks intervals that exclude zero.

Task / backbone	Decoder	F1@0.5	ΔF1@0.5 [95% CI]	F1@0.3	Onset (s)	Frag. (%)
TUAR / LaBraM	Full SPAN	19.6±0.2		28.6±0.3	0.70±0.03	73.1±2.6
	− Gaussian locality	17.4±1.3	**−2.2 [−3.6, −0.8]**	26.6±1.4	0.96±0.16	73.0±0.5
	− Denoising	17.1±0.2	**−2.5 [−2.8, −2.1]**	25.9±0.2	0.92±0.02	73.2±1.8
	− Multiscale pyramid	19.4±0.2	**−0.2 [−0.6, +0.0]**	28.6±0.4	0.70±0.02	75.0±3.3
TUAR / REVE	Full SPAN	18.3±0.4		27.0±0.4	0.79±0.03	72.5±3.0
	− Gaussian locality	13.2±0.3	**−5.1 [−5.2, −5.0]**	21.8±0.3	1.43±0.05	72.2±2.8
	− Denoising	15.9±0.5	**−2.4 [−3.1, −1.4]**	24.7±0.2	0.97±0.06	73.2±1.5
	− Multiscale pyramid	18.0±0.3	**−0.3 [−0.7, −0.1]**	27.1±0.5	0.75±0.03	71.7±2.0
Arousal / LaBraM	Full SPAN	30.2±2.2		38.8±1.1	0.86±0.10	11.1±1.6
	− Gaussian locality	29.3±2.5	−0.9 [−1.5, +0.1]	39.1±0.1	0.91±0.12	8.2±2.4
	− Denoising	29.1±2.1	−1.1 [−5.1, +4.9]	38.0±0.9	0.88±0.07	11.4±3.6
	− Multiscale pyramid	29.4±1.1	−0.7 [−3.7, +3.9]	38.3±0.7	0.86±0.00	10.6±2.4
Arousal / REVE	Full SPAN	28.3±1.8		37.5±1.1	0.84±0.07	11.5±4.8
	− Gaussian locality	28.1±1.8	−0.2 [−5.1, +2.3]	37.4±1.4	0.85±0.01	9.2±0.6
	− Denoising	27.6±1.3	−0.7 [−1.3, +0.4]	36.2±1.1	0.86±0.08	14.4±4.0
	− Multiscale pyramid	29.5±0.8	+1.2 [−2.1, +3.1]	37.2±0.8	0.82±0.03	11.3±3.9

H Duplicate suppression at inference

The main results apply no suppression. To test whether the duplicates identified in Appendix D can be removed at inference, we re-select the operating point of stored predictions on the validation split over a joint grid of the confidence threshold, a suppression mode and an overlap threshold in $\{0, 0.1, 0.3, 0.5\}$, keep the configuration with the highest validation event F1 at IoU>0.3, and score the test split once. The modes are class-agnostic one-dimensional NMS, class-aware NMS, in which only predictions of the same class suppress each other, and merging, which replaces overlapping predictions by their union with the class of the most confident member. The no-suppression grid point reproduces the main operating point. Sweeps exist for SPAN on LaBraM, CBraMod, REVE and EEGPT, for the plain set decoder on LaBraM and REVE, and for the end-to-end model; Table 19 averages them per head.

Where duplicates are the main source of fragmentation, suppression removes them at no cost: on arousal, SPAN's fragmentation falls from 10.4% to 0.7% and event F1 at IoU>0.5 rises by 0.9 to 1.2 points on all four backbones; on TUSZ, fragmentation halves and F1 changes by between -0.6 and $+0.7$ points. On MI, merging is selected in most cells, fragmentation falls from 20.7% to 11.6%, and F1 at IoU>0.5 changes by between -1.4 and $+0.5$ points. On TUAR, class-aware NMS is selected in every cell; F1 at IoU>0.3 rises from 26.4% to 29.1% and predictions per reference fall from 1.18 to 0.79, but fragmentation stays at 70.5% and F1 at IoU>0.5 decreases on all four backbones, by 0.3 to 1.6 points. The duplicates that inflate the TUAR onset error therefore cannot be removed by overlap-based suppression without also removing correct predictions of co-

occurring artifacts. This is why Section 5.2 attributes the TUAR degradation to coherence rather than precision, and why we regard suppression as a validated but insufficient remedy.

Table 19: validation-selected duplicate suppression applied at inference to stored test predictions. Each cell reads without → with suppression, where the suppressed variant re-selects the confidence threshold jointly with the suppression mode and overlap threshold. Values are means over three seeds and over the backbones with stored sweeps (SPAN: LaBraM, CBraMod, REVE and EEGPT; plain set decoder: LaBraM and REVE).

Task	Head	Backbones	F1@0.3 (%)	F1@0.5 (%)	Fragmentation (%)	Pred./ref.
TUSZ	SPAN	4	55.2 → 55.4	51.1 → 51.2	3.4 → 1.8	1.00 → 1.01
	Plain set decoder	2	53.8 → 54.7	49.1 → 49.6	3.0 → 1.1	0.85 → 0.86
	DANCE	1	43.0 → 43.9	39.0 → 39.2	4.2 → 1.3	1.02 → 1.13
MI	SPAN	4	28.4 → 29.6	25.6 → 25.4	20.7 → 11.6	1.21 → 1.08
	Plain set decoder	2	21.5 → 26.5	16.7 → 16.0	67.4 → 33.0	1.75 → 1.14
	DANCE	1	21.8 → 25.7	16.2 → 15.3	80.3 → 48.8	1.91 → 1.32
TUAR	SPAN	4	26.4 → 29.1	17.8 → 17.1	74.6 → 70.5	1.18 → 0.79
	Plain set decoder	2	18.1 → 20.5	10.4 → 10.1	79.2 → 79.4	1.41 → 1.08
	DANCE	1	16.7 → 18.6	9.2 → 9.1	79.0 → 79.3	1.16 → 0.92
Arousal	SPAN	4	37.8 → 39.8	28.1 → 29.1	10.4 → 0.7	1.31 → 1.17
	Plain set decoder	2	11.8 → 11.9	4.8 → 4.8	8.8 → 6.7	1.96 → 1.90
	DANCE	1	8.5 → 8.6	3.1 → 3.2	6.9 → 5.7	3.12 → 2.98