COMPACT SPATIOTEMPORAL REPRESENTATIONS VIA SUMMARY TOKENS FOR HUMAN INTRACRANIAL RECORDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracranial EEG (iEEG) provides high-resolution measurements of human brain activity, but building models that transfer across subjects and tasks is complicated by irregular electrode placement and by downstream tasks that rely on information at different spatial scales. Existing iEEG foundation models typically preserve channel-level structure during pretraining while leaving cross-channel aggregation to downstream adaptation, where labeled data are limited. We introduce a factorized spatiotemporal masked autoencoder that preserves channel-level tokens while learning two complementary summary representations: temporal summary tokens that aggregate information within each channel and spatial summary tokens that aggregate information across channels. During pretraining, an information bottleneck enforces reconstruction through these summary tokens, explicitly encouraging them to capture temporal and population-level structure. For downstream decoding, we use the learned summary representations across a range of iEEG tasks. Our results show that explicitly training the summary representations during pretraining improves downstream transfer across tasks with different spatial requirements. Further, we show that these representations capture task-relevant information and selectively aggregate signals from channels and time patches that contribute to downstream decoding.

1 INTRODUCTION

Intracranial EEG (iEEG) records neural activity from electrodes implanted inside or on the surface of the brain at millisecond temporal resolution and millimeter-scale spatial coverage across multiple brain regions (Lachaux et al., 2003; Parvizi & Kastner, 2018). As a result, iEEG offers a unique opportunity to study both local and distributed neural population dynamics across a variety of behaviors and cognitive processes (Johnson et al., 2020). Realizing this potential requires models that can preserve channel-level information and provide transferable summaries of population-level activity needed for decoding. Recent work has motivated pretraining iEEG foundation models with self-supervision to learn representations that can summarize multiregional activity while transferring across subjects, sessions, and downstream decoding tasks (Wang et al., 2023; Zhang et al., 2023; Yuan et al., 2025; Chau et al., 2025; Oganesian et al., 2025; Carzaniga et al., 2026; Han et al., 2026).

Despite early demonstrations of success, one key challenge for learning generalizable representations of iEEG is the irregular spatial coverage across subjects, as electrode placement is determined by clinical need rather than a standardized placement (Delgado-Garcia & Frauscher, 2022; Mentzelopoulos et al., 2024; Chau et al., 2025; Carzaniga et al., 2026). In that sense, individual electrodes are the fundamental units of measurement in iEEG: montages differ across subjects, spatial localization matters for both neuroscience and clinical readout, and any representation that discards channel identity upstream cannot recover it later. Thus, when building iEEG foundation models, channel-level tokenization is not an open design choice, but a constraint imposed by the data itself. Downstream decoding tasks, on the other hand, are typically defined over much coarser windows (e.g., hundreds of milliseconds to seconds), thereby necessitating information aggregation across irregular multichannel layouts and many atomic temporal patches within the window of interest (Nejedly et al., 2020; Peterson et al., 2022; Zahorodnii et al., 2026). Existing iEEG foundation models largely defer this aggregation to finetuning, attaching a pooling mechanism on top of individual channel tokens and

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.

letting the downstream task learn how to combine them (Wang et al., 2023; Zhang et al., 2023; Yuan et al., 2025; Oganesian et al., 2025; Carzaniga et al., 2026; Han et al., 2026). Addressing spatiotemporal pooling at the finetuning stage, however, places the burden of cross-channel summarization on the downstream task, which typically has access to less labeled iEEG data than pretraining. Given this constraint, we ask a different question: can self-supervised pretraining enable a channel-level model to form multi-channel structure atop atomic tokens and make the said structure available to downstream tasks—rather than learning it anew, data-hungrily, at finetuning time?

Standard masked reconstruction, the dominant self-supervised learning paradigm for neural foundation models (Zhang et al., 2023; Yuan et al., 2025; Oganesian et al., 2025; Han et al., 2026), does not by itself guarantee that cross-channel information is aggregated during pretraining. When the predictor can freely attend to unmasked channel tokens elsewhere in the sequence, it can often reconstruct a masked patch from that same channel's recent history or from spatially-adjacent (i.e., correlated) channels. Specifically, cross-channel aggregation becomes optional rather than necessary, and simply adding CLS or global query tokens does not guarantee population-level or temporal aggregation unless explicitly enforced by the pretraining objective or by restricting what information the predictor can access. As such, a pretraining objective should make both levels of structure available at finetuning time, rather than presuming in advance which one a given task will need.

Contributions. We introduce MATCHA (**M**asked **A**utoencoding **T**hrough **C**ompact summaries of **H**uman iEEG **A**ctivity), a factorized spatiotemporal masked autoencoder (MAE) for iEEG that learns compact temporal and spatial summaries alongside channel-level representations. During pretraining, we require the information used for masked reconstruction to pass through these summary tokens, explicitly training their aggregation mechanisms. The resulting model keeps both channel-level and population-level representations available within a single pretrained backbone, allowing downstream tasks to build on learned summaries rather than learning cross-channel spatiotemporal aggregation from scratch. Our contributions are as follows:

1. We introduce a self-supervised pretraining approach that forces dedicated summary tokens to aggregate temporal and population-level structure during pretraining. The encoder retains channel-level content tokens, while the predictor receives their information only through these summary streams, thereby explicitly training the summary tokens to carry reconstructive information without replacing the fine-grained channel representations.
2. We provide the capability to flexibly aggregate information using either the spatial or temporal summary tokens. Doing so enables dissociating and measuring the effect of different spatiotemporal features on each downstream task of interest.
3. In an attention-based interpretability analyses, we show that the learned spatial and temporal summary tokens learn functionally relevant spatiotemporal structure for each downstream task.
4. We evaluate MATCHA's representations on 8 iEEG datasets that span a diverse set of behavioral and cognitive downstream tasks, including spoken-word decoding, naturalistic-behavior state decoding, task-phase decoding in memory paradigms, and interictal epileptiform-spike detection.

2 Related work

Self-supervised pretraining for iEEG data. A growing body of work pursues self-supervised pretraining for iEEG (Wang et al., 2023; Zhang et al., 2023; Yuan et al., 2025; Chau et al., 2025; Oganesian et al., 2025; Carzaniga et al., 2026; Han et al., 2026). Across these works, Brant reconstructs masked neural waveforms (i.e., time-series) (Zhang et al., 2023) and BrainBERT reconstructs masked spectral representations (i.e., STFT) (Wang et al., 2023). Moreover, BaRISTA uses masked *latent* reconstruction with anatomically informed spatial encodings (Oganesian et al., 2025) and NeuroCLUS learns a functional context graph that guides soft clustering of channels into learned functional units (Zheng & Wang, 2026). Demonstrating a different pretraining paradigm, PopT learns population-level representations through self-supervised discriminative objectives (Chau et al., 2025). Here, we use masked reconstruction as our pretraining objective, but require reconstruction to proceed only through dedicated summary tokens rather than directly utilizing channel-level token information.

Cross-channel aggregation in iEEG models. iEEG models also differ in whether cross-channel aggregation is explicitly encouraged during pretraining or left for downstream adaptation. Brant, BrainWave, and seegnificant (Zhang et al., 2023; Yuan et al., 2025; Mentzelopoulos et al., 2024) separate temporal and spatial attention, while BaRISTA uses joint spatiotemporal attention (Oganesian et al., 2025). In all cases, a reconstruction objective can typically be satisfied from a channel's own history or from the activity of spatially-adjacent (i.e., correlated) channels. Thus, neither the pretraining objective nor the model formulation itself strictly enforce cross-channel information aggregation. In contrast, PopT learns a single, fixed global spatial aggregation over frozen BrainBERT channel embeddings using population-level discriminative objectives (Wang et al., 2023; Chau et al., 2025); the learned population-level CLS token is later also used for downstream decoding. Other approaches enforce aggregation architecturally. For example, Du-IN collapses within-region channels through convolutional layers tied to a fixed electrode layout. Although such an approach does not generalize to variable electrode montages, it can yield improved performance on downstream tasks that benefit from population-level representations (e.g., speech production) (Zheng et al., 2024). We instead ask whether pretraining can enforce aggregation without collapsing scale: whether a channel-level model can be learned while also building multi-channel summary structure on top of atomic tokens—thereby maintaining both channel-level and aggregated representations for downstream use.

3 METHODS

We introduce MATCHA, a self-supervised pretraining framework for iEEG that learns compact temporal and spatial summary representations while preserving channel-level information. We consider multivariate iEEG recordings across multiple subjects and recording sessions, with different temporal lengths, number of recorded channels, and electrode layouts. In our formulation, we first divide each iEEG segment into channel-wise temporal patches and map them to a spatiotemporal grid of patch tokens, augmented with dedicated temporal and spatial summary tokens (Fig. 1). During pretraining, a factorized spatiotemporal transformer encoder processes the unmasked patch tokens together with these summary tokens. Unlike a standard masked autoencoder, however, the predictor reconstructs masked patches using *only* the encoded summary representations, without direct access to the observed content tokens (Fig. 2). This restriction makes the summary representations an *information bottleneck*, explicitly training them to carry the information needed for reconstruction. Furthermore, for the downstream tasks we again only read out from either of these summary tokens flexibly, thus utilizing the cross-channel and time aggregation learned during pretraining. We detail this formulation below.

3.1 NEURAL SIGNAL TOKENIZATION

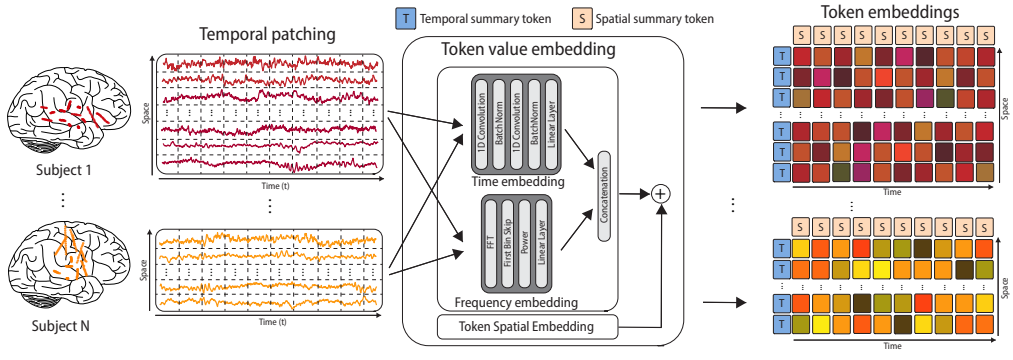

Figure 1: **Tokenization of iEEG signals.** Multivariate iEEG segments are first patchified into a spatiotemporal token sequence. For each patch, time and frequency value embeddings are computed and concatenated to form the token value embedding which is then combined via summation with a token spatial embedding. After prepending temporal ($\boldsymbol{T}$) and spatial ($\boldsymbol{S}$) summary tokens, the resulting token embeddings form the input to the transformer encoder.

Patching. We denote a multivariate iEEG data segment as $\boldsymbol{X}_i \in \mathbb{R}^{T_i \times C_i}$ where T_i and C_i represent number of timesteps and channels, respectively, for the ith data segment. We first form the token

sequence by creating non-overlapping temporal patches of length P (e.g., $P = 50$ samples, i.e., 100ms at a 500Hz sampling rate), resulting in $N_i = \lceil T_i/P \rceil$ patches per channel (padded when required) and a total of $N_i \times C_i$ atomic patches $\boldsymbol{x}^p_{n,c,i} \in \mathbb{R}^P$, for $n = 1, \ldots, N_i$ and $c = 1, \ldots, C_i$.

Token value embeddings. Given a patch, we encode its time-domain and frequency-domain content separately before combining them into a single value embedding. The time embedding $\boldsymbol{e}^{\text{time}}_{n,c,i} \in \mathbb{R}^{d_{\text{model}}/2}$ is computed by a small convolutional network applied directly to the patch's raw time-series signal, where d_{model} denotes the latent dimension (Fig. 1). The frequency embedding $\boldsymbol{e}^{\text{freq}}_{n,c,i} \in \mathbb{R}^{d_{model}/2}$ is computed by first applying a Hanning window to the patch to reduce spectral leakage, then taking its fast Fourier transform (FFT). The final value embedding $\boldsymbol{e}^v_{n,c,i} \in \mathbb{R}^{d_{\text{model}}}$ is formed by concatenating the two and applying a linear projection.

Token spatial embeddings. Since each electrode's position is fixed within a subject, the spatial embedding for a channel c depends only on c and the segment index i, not on the temporal patch index n. Specifically, we use each channel's MNI (Montreal Neurological Institute) coordinate to query the Cole-Anticevic Brain-wide Network Partition (CAB-NP) (Ji et al., 2019) for its corresponding assignment to one of 12 possible partitions. The assigned partition identifier is then mapped to a spatial embedding $\boldsymbol{e}^s_{c,i} \in \mathbb{R}^{d_{\text{model}}}$ through a learnable lookup table. The final token value embedding is $\boldsymbol{e}_{n,c,i} = \boldsymbol{e}^v_{n,c,i} + \boldsymbol{e}^s_{c,i}$ (Fig. 2). For all channels whose registration are designated as missing or unreliable, we utilize a single learnable unknown-position embedding, $\boldsymbol{e}^s_{\text{unk}} \in \mathbb{R}^{d_{\text{model}}}$ (see details in Appendix A.2).

Temporal and spatial summary tokens. After value and spatial embeddings are formed for the atomic channel–time patches, we augment the token grid with two learnable summary streams that act as dedicated aggregation sites for the subsequent transformer (Fig. 1). Let L and M denote the numbers of temporal and spatial summary tokens, respectively. First, for each channel c, we attach L learnable *temporal summary* tokens along the temporal axis, denoted per channel as $\boldsymbol{T}_c \in \mathbb{R}^{L \times d_{\text{model}}}$ and jointly across channels as $\boldsymbol{T} = \{\boldsymbol{T}_c\}_{c=1}^{C_i}$ (Fig. 1). These tokens share the channel's spatial embedding but have no associated waveform content; their role is to provide a per-channel temporal summary of the visible patches. Second, we prepend M learnable *spatial summary* tokens along the channel axis, denoted per time patch as $\boldsymbol{S}_n \in \mathbb{R}^{M \times d_{\text{model}}}$ and jointly across time patches as $\boldsymbol{S} = \{\boldsymbol{S}_n\}_{n=1}^{N_i}$. These tokens are shared across the channels and are replicated at every temporal position, yielding a cross-channel summary stream at each patch time. The resulting token tensor therefore is composed of spatiotemporal patch embeddings together with per-channel temporal summaries and population-level spatial summaries. As detailed in Section 3.2, these summary streams are the carriers of the *information bottleneck*. Finally, the augmented token sequences along each axis are:

$$[\,\boldsymbol{T}_c;\, \boldsymbol{e}_{1,c,i};\, \ldots;\, \boldsymbol{e}_{N_i,c,i}\,] \qquad \text{for channel } c \text{ in segment } i, \tag{1}$$

$$[\,\boldsymbol{S}_n;\, \boldsymbol{e}_{n,1,i};\, \ldots;\, \boldsymbol{e}_{n,C_i,i}\,] \qquad \text{for temporal patch } n \text{ in segment } i. \tag{2}$$

3.2 Pretraining and finetuning with information bottleneck

As shown in Fig. 2, after forming the spatiotemporal token sequence, we next encode it with a factorized spatiotemporal transformer encoder. We pretrain our model using a masked reconstruction objective with an *information bottleneck* wherein masked patches are reconstructed using the temporal and spatial summary streams only. Below we describe the transformer encoder, the reconstruction objective, and the bottleneck restriction.

Factorized spatiotemporal transformer encoder. The tokenizer output is processed by a stack of N_L factorized transformer layers (e.g., N_L=10, with hidden dimension d_{model}=256). Within each layer, attention is applied sequentially along the two axes of the spatiotemporal token grid in an interleaved manner (Fig. 2). First, *temporal attention* is computed independently within each channel, allowing patch tokens to exchange information along time. In this block, instead of absolute temporal position embeddings, we use rotary positional embeddings (RoPE) (Su et al., 2023) to encode relative temporal position. Second, *spatial attention* is computed independently at each temporal position, allowing patch tokens across channels—including the spatial summary tokens—to

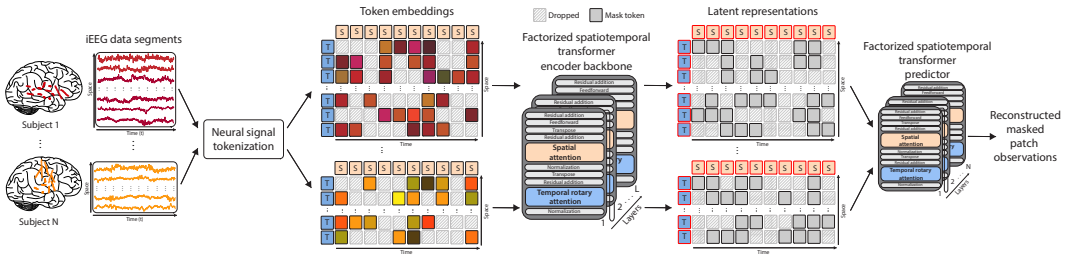

Figure 2: **Masked reconstruction with an information bottleneck.** Token embeddings from the tokenization stage (Fig. 1) are masked with contiguous temporal spans drawn independently per channel; masked patches (hatched) are dropped and never seen by the encoder, while summary tokens remain visible throughout. The encoder processes the remaining tokens with temporal attention within each channel, then spatial attention across channels at each time step, producing latent representations for all visible patches and the summary tokens. The predictor then reconstructs each masked patch's observations using only the summary tokens' encoder embeddings (outlined in red) together with a shared learnable mask token at each masked position, indicating which patches it must reconstruct; embeddings of visible (unmasked) tokens are withheld entirely.

exchange information. We do not use RoPE in this block as there is no clear way to indicate channel ordering. Coordinate-based relative encodings are not always available since spatial metadata can be missing for some channels or datasets (Section 3.1), and we do not want to bias the attention mechanism toward requiring it. In addition, encoding relative position via channel index is arbitrary and would break the desired permutation equivariance of spatial attention across channels. We instead convey spatial position through the absolute token spatial embeddings (Fig. 1), which supports a fallback mechanism to an unknown-position embedding when spatial metadata is unavailable (see *Electrode localization* in Appendix A.2).

Masked autoencoding. We pretrain with a masked autoencoder (MAE) in which a fraction of the content patches is held out from the encoder and must be reconstructed by a lightweight predictor. Masking is performed with contiguous temporal spans drawn independently per channel: span lengths follow a geometric distribution with mean 4 patches (e.g., 400ms at 100ms patches), and span placement is adjusted so that exactly a target fraction p (e.g., 0.7) of each channel's temporal positions are masked. Here, importantly, summary tokens are never masked or used as reconstruction targets; they always remain available to the encoder. The encoder therefore sees only the tokenized unmasked content patches together with the metadata and summary tokens. A factorized predictor of the same spatiotemporal design, but with fewer layers ($N_p = 4$) and a reduced hidden dimension ($d_{\text{predictor}} = 192$), receives the summary tokens ($\boldsymbol{T}$, $\boldsymbol{S}$) and the shared learnable mask token ($\boldsymbol{e}^{s}_{\text{mask}}$) at the reconstruction target positions. In subsequent sections we use the index sets $\mathcal{V}_i$ and $\overline{\mathcal{V}}_i$ to denote unmasked and masked content patches, respectively, across all patches $\boldsymbol{x}^{p}_{n,c}$ in the ith segment.

Reconstruction through *information bottleneck*. Under standard masked autoencoding, a masked patch can typically be reconstructed from temporally or spatially adjacent unmasked content tokens due to correlated activity. Thus even when summary tokens are present nothing forces them to carry useful information. We address this with an *information bottleneck*: after a single encoder forward pass over the unmasked tokens, the predictor is given access only to the temporal and spatial summary token embeddings whereas all content token embeddings are omitted from the predictor transformer's attention mechanism, thereby forcing masked patches to be reconstructed from the summary token streams alone. Formally, the encoder E_θ processes the visible content tokens together with the summary tokens

$$(\boldsymbol{H}_{e_{\mathcal{V}_i}}, \boldsymbol{H}_{\boldsymbol{T}}, \boldsymbol{H}_{\boldsymbol{S}}) = E_\theta\big(\{\boldsymbol{e}_{n,c,i}\}_{(n,c)\in\mathcal{V}_i}, \boldsymbol{T}, \boldsymbol{S}\big). \tag{3}$$

where $\boldsymbol{H}_{e_{\mathcal{V}_i}}$ denotes the encoded content tokens, and $\boldsymbol{H}_{\boldsymbol{T}} \in \mathbb{R}^{LC_i \times d_{\text{model}}}$ and $\boldsymbol{H}_{\boldsymbol{S}} \in \mathbb{R}^{MN_i \times d_{\text{model}}}$ denote the encoded temporal and spatial summary tokens, respectively. The predictor D_ψ then receives only the summaries and a shared learnable mask token $\boldsymbol{e}^{\text{mask}}$ at each masked position—never $\boldsymbol{H}_{e_{\mathcal{V}_i}}$, unlike a standard MAE decoder. We define the output of the decoder as

$$\{\hat{\boldsymbol{x}}^{p}_{n,c,i}\}_{(n,c)\in\overline{\mathcal{V}}_i} = D_\psi\big(\boldsymbol{e}^{\text{mask}}, \boldsymbol{H}_{\boldsymbol{T}}, \boldsymbol{H}_{\boldsymbol{S}}\big) \tag{4}$$

where $\hat{\boldsymbol{x}}^{p}_{n,c,i}$ denotes reconstructed content patch waveforms for the ith segment.

It is important to note that equation 4 in our formulation explicitly ties reconstruction quality to how well each summary has aggregated the structure it is positioned to capture: cross-channel dependencies for the spatial stream and within-channel dynamics for the temporal stream. The result is an encoder that produces both fine-grained, channel-level representations, as well as summary representations that are pressured, by construction, to learn informative spatiotemporal features from multiregional neural data. Finally, as an auxiliary loss the short-time Fourier transform of the full segment is also decoded from the temporal-summary tokens.

Finetuning with summary tokens. During finetuning we perform decoding using the summary streams trained under the bottleneck, rather than pooling over individual channel-level representations directly—as was done in prior works (Zhang et al., 2023; Oganesian et al., 2025; Chau et al., 2025). Moreover, by using the summary streams for downstream evaluations, our approach offers the additional capability to interrogate whether it is more important to aggregate spatial or temporal information (or the combination of both) for each target task considered. We pool either temporal summaries, spatial summaries, or their combination with a small cross-attention readout into a fixed-size representation, which is passed to a linear task head for downstream decoding (see Appendix Finetuning readout formulation, for the full formulation).

Although in equation 5 we combine the temporal and spatial summary streams prior to downstream decoding, our approach gives us the capability to consider the impacts of spatial summarization and temporal summarization in isolation, that is, using only the temporal or only the spatial summary stream as input to the readout. This flexibility enables dissociating and evaluating the relative importance of population-level spatial interactions vs. individual-channel temporal dynamics across a variety of downstream tasks.

4 Results

4.1 Dataset and evaluation details

We performed pretraining using the Brain Treebank dataset (Wang et al., 2024), and evaluated downstream decoding performance on 8 publicly available iEEG datasets spanning a variety of domains including spoken-word decoding, naturalistic-behavior state decoding, task-phase decoding in memory paradigms, and interictal epileptiform-spike detection: Du-IN (Zheng et al., 2024), AJILE12 (Peterson et al., 2022), three publicly availabe datasets from the OpenNeuro repository [1] (`ds004809`, `ds005523`, and `ds004752`), MAYO and FNUSA (Nejedly et al., 2020), and Brain Treebank sessions excluded from pretraining (Zahorodnii et al., 2026). Datasets and preprocessing details are provided in Appendix A.1. For each downstream task, we used the same training, validation, and test splits across all evaluated models. Validation data were used for model selection, and final performance was measured on the held-out test data using relevant metrics for each dataset. We compare against three pretrained iEEG foundation models: BaRISTA (Oganesian et al., 2025), Brant (Zhang et al., 2023), and PopT (Chau et al., 2025) – adapting each to every downstream dataset in this evaluation – as well as a linear baseline using power features. Full details of their implementation and adaptation are reported in Appendix A.3.

4.2 Pretrained summary tokens support transfer across iEEG datasets

We evaluated our pretrained model by finetuning on each downstream dataset. We performed decoding using only the compact summary streams, the spatial and temporal summary embeddings, which we pool into a single embedding via a learned attention pooling mechanism prior to read-out (see *Finetuning with summary tokens* in Section 3.2). Architecture and optimization details are in Appendix A.2.

As shown in Table 1, MATCHA achieves strong performance across all evaluated downstream dataset, with improvements spanning spoken-word classification (Du-IN), task-phase decoding in memory paradigms (`ds004809`, `ds005523`, `ds004752`), naturalistic-behavior decoding (AJILE12), and interictal epileptiform spike detection (MAYO and FNUSA). Notably, even though PopT also learns

[1] `https://openneuro.org/`

a population-level CLS token during pretraining, it achieves substantially lower accuracy than MATCHA on Du-IN. Unlike PopT's single global summary token, MATCHA maintains separate temporal and spatial summary streams, capturing both within-channel dynamics and cross-channel, population-level information. Moreover, in most datasets the three readouts options perform closely, indicating that the *information bottleneck*'s compression is robust regardless of which summary stream is used as readout during finetuning (see Table 9). Nevertheless, on the Du-IN dataset there is a larger gap across readout options: the spatial readout substantially outperforms both the temporal and combined readouts, consistent with this task's dependence on population-level activity rather than channel-level temporal structure (Zheng et al., 2024) (see Section 4.3 for details).

Taken together, our results demonstrate that summaries learned on one recording corpus (e.g., Brain Treebank) support adaptation to distinct downstream iEEG decoding tasks or domains. MATCHA learns its summary representations through bottlenecked masked autoencoding, forcing reconstruction through these summaries during pretraining. In contrast, BaRISTA and Brant—which also pretrain with masked reconstruction but do not impose similar information summarization constraints—underperform MATCHA, showing that having summary tokens that aggregate population-level and cross-time structure during pretraining drives the improvement over these baselines. PopT provides a complementary comparison: although it also learns a population-level CLS token during pretraining, it compresses information into a single global summary, whereas MATCHA maintains separate spatial and temporal summary streams at finer granularity. The stronger transfer of MATCHA on several downstream tasks therefore suggests that not only learning summary representations, but also the granularity at which information is summarized across space and time, is important for downstream decoding.

	Downstream dataset						
Model	Du-IN	AJILE12	ds004809	ds005523	ds004752	MAYO	FNUSA
Random	0.015 ± 0.001	0.295 ± 0.053	0.329 ± 0.002	0.487 ± 0.014	0.245 ± 0.052	0.109	0.293
Power features	0.058 ± 0.016	0.829 ± 0.083	0.741 ± 0.054	0.762 ± 0.063	0.675 ± 0.128	0.291	0.482
Brant	0.064 ± 0.097	0.614 ± 0.202	0.452 ± 0.079	0.609 ± 0.112	0.452 ± 0.102	0.591	0.638
BaRISTA	0.235 ± 0.195	0.888 ± 0.096	0.865 ± 0.104	$\mathbf{0.897 \pm 0.058}$	0.732 ± 0.099	0.644	0.607
PopT	0.188 ± 0.092	0.845 ± 0.129	0.769 ± 0.087	0.776 ± 0.047	0.725 ± 0.010	0.622	0.656
MATCHA	$\mathbf{0.663 \pm 0.169}$	$\mathbf{0.928 \pm 0.078}$	$\mathbf{0.894 \pm 0.110}$	$\mathbf{0.896 \pm 0.058}$	$\mathbf{0.854 \pm 0.056}$	**0.667**	**0.695**

Table 1: **Pretrained summary streams transfer across downstream iEEG datasets.** Comparison with baseline models in macro-F1 across datasets (mean $\pm$ std across sessions or subjects where applicable); for MAYO/FNUSA (interictal epileptiform-spike detection), values are positive-class F1. Tasks and labels differ across datasets, so values are not comparable across columns.

4.3 Information bottleneck pretraining improves downstream decoding

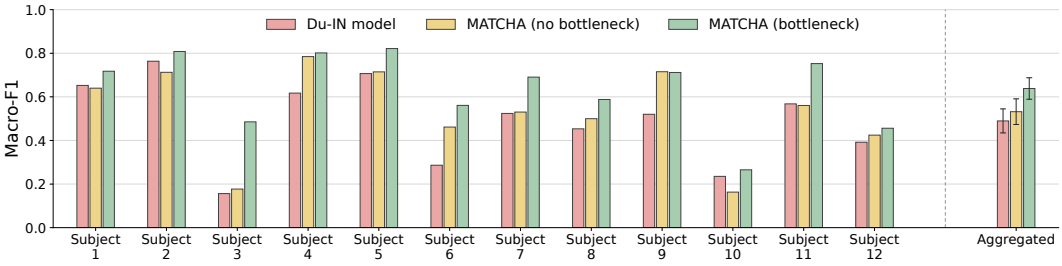

Figure 3: **Summary pretraining improves Du-IN speech decoding.** Per subject (individual bars) and averaged across 12 subjects macro-f1 on DU-IN (mean $\pm$ std). The bottleneck model outperforms both the MATCHA non-bottleneck and Du-IN baseline models in every subject (two-sided Wilcoxon signed-rank, Holm-corrected, $n = 12$, $p = 1.95 \times 10^{-3}$).

We next tested whether requiring masked reconstruction through summaries improves downstream performance. To do this, we pretrained MATCHA both with and without *information bottleneck*, specifically the MAE predictor either has its input restricted to the summary streams (Section 3.2) or may also attend to the unmasked content tokens, as in a standard MAE decoder. We used the combined summary token readout for both models, additionally evaluating the non-bottleneck model with an all-tokens readout (content plus summary tokens) for a fairer comparision and reporting whichever performed better for it. For this evaluation, we focused on Du-IN spoken-word classification, where

the authors show this task depends on population-level information and multichannel integration, along with fine temporal signals, to decode the spoken word (Zheng et al., 2024). This provides a relevant test of whether bottleneck-trained summaries support effective multichannel decoding while retaining channel-level representations. For a complete comparison, we also evaluated against the Du-IN model originally proposed for this dataset.

Fig. 3 illustrates the advantage offered by the summary streams for individual subjects on Du-IN spoken-word classification. Requiring reconstruction through the summaries increased mean Du-IN macro-F1 from 0.532 to 0.638 with improvement in 11/12 subjects. Notably, MATCHA also outperformed the Du-IN baseline model, with significant improvements observed in all 12 subjects compared to ($n = 12$ subjects; unadjusted $p = 9.77 \times 10^{-4}$, Holm-adjusted $p = 1.95 \times 10^{-3}$ for each comparison). Unlike Du-IN's montage-specific input fusion, MATCHA retains channel-level tokens alongside learned summaries, demonstrating that summarization and channel-level preservation are not mutually exclusive. Interestingly, MATCHA without the bottleneck was statistically indistinguishable from the Du-IN model (two-sided Wilcoxon signed-rank, Holm-corrected, $n = 12$, $p = 0.233$), further reinforcing that channel-level representations do not preclude competitive population-level decoding. Bottleneck pretraining also improved performance across the other evaluated datasets (Table 10), supporting its contribution across different decoding tasks.

Together, the results in Table 10 and Fig. 3 show that the benefit of training through summary tokens is not confined to a single dataset or task, and that preserving channel-level representations is compatible with population-level aggregation for effective multichannel decoding across a variety of downstream domains.

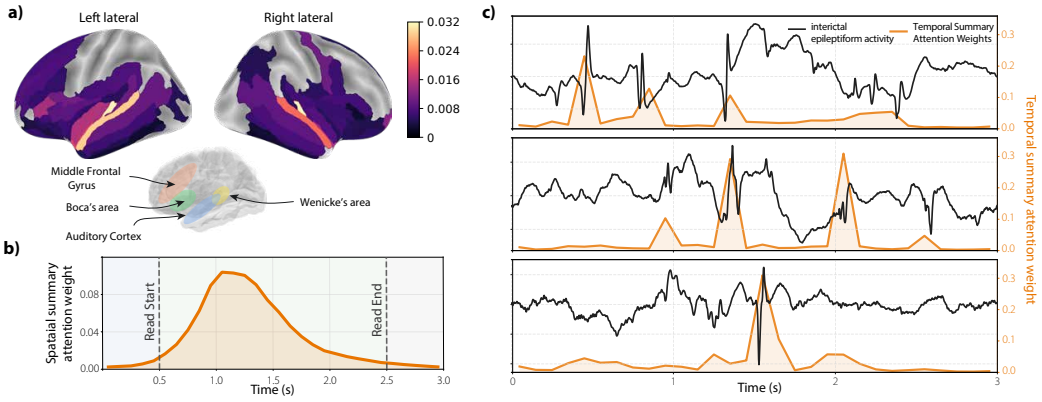

Figure 4: **Summary representations reflect task-related channel and temporal structure.** (a) Per-channel temporal summary contributions for speech vs. non-speech decoding in Brain Treebank, averaged onto anatomical parcels; activation is highest in auditory and language regions. (b) The readout's attention to per-time-step spatial summary tokens during Du-IN spoken-word classification, concentrating in the period when the word was spoken (dashed lines). (c) Temporal summary tokens' attention to time patches for example interictal epileptiform-spike segments (orange), concentrating around periods of high-frequency epileptiform spiking.

4.4 Summary tokens aggregate task-related signals

To test whether the temporal and spatial summary tokens learned during pretraining remain interpretable once applied to downstream tasks, we finetuned the pretrained model separately on three tasks – speech-versus-non-speech decoding, cued word production, and interictal epileptiform-spike detection – and, in each case, examined which channels or time patches the model relies on most for downstream decoding by measuring how each summary stream's activations or attention weights are distributed across the channels or time patches it summarizes (Fig. 4; computation and visualization details in Appendix A.4).

First, in the speech versus non-speech decoding task in Brain Treebank, per-channel temporal summary tokens' activations were highest in the annotated auditory and language regions, consistent with their established role in speech processing (Fig. 4a). Similarly, an inference-time occlusion analysis showed that removing the lateral superior temporal gyrus reduced decoding performance

more than removing the same number of randomly chosen electrodes, in speech versus non-speech and sentence onset tasks (all $p \leq 0.009$; Appendix A.4). Moreover, in Du-IN word classification, the downstream readout's attention to the per-time-step spatial summary tokens, averaged across all segments, concentrated around the interval in which the cued word was spoken (Fig. 4b), showing that the model draws on this population-level structure specifically when it is needed to classify the correct spoken word. Finally, in the MAYO interictal epileptiform-spike detection task, the temporal summary tokens' attention weights to content patches increased around time patches with more high-frequency epileptiform activity (Fig. 4c), showing that the temporal summary stream weights the specific time patches most predictive of the task. Computation and visualization details in Appendix A.4.

Together, across these three analyses – spanning a speech-decoding task drawn from the same corpus the model was pretrained on, and two independent downstream tasks – the summary representations learned during pretraining remain informative and interpretable under downstream transfer, capturing task-relevant population-level and temporal signals. These results point to the broader potential of our pretraining approach, and the temporal and spatial summary tokens it produces, as a tool for decoding and understanding task-relevant neural signals in various downstream tasks.

4.5 Benchmark evaluation on Brain Treebank

Finally, we evaluated MATCHA on a benchmark comprising 15 auditory, visual, and linguistic decoding tasks from 12 Brain Treebank movie-watching sessions across 6 subjects. These sessions were withheld from pretraining, and we finetuned the model using summary readouts under the benchmark's within-session evaluation protocol (Dataset in Appendix A.1). Using both spatial and temporal summary tokens achieved the highest mean AUROC over all 15 tasks on this benchmark, outperforming baselines including DIVER-1 and PopT. Notably, during pretraining PopT uses Brain Treebank recordings (similar to our model) and learns a population-level CLS token that is used for downstream decoding (Chau et al., 2025). However, MATCHA's reconstruction-trained summaries at a finer temporal resolution support stronger decoding performance on this benchmark. These results demonstrate that finetuning MATCHA's pretrained summaries supports effective decoding of audiovisual and linguistic information. Full per-task results are reported in Tables 3 and 4.

Model	AUROC
Linear, spectrogram*	0.660
CNN, spectrogram*	0.669
BrainBERT, frozen*	0.626
PopT*	0.670
DIVER-1, frozen*	<u>0.678</u>
MATCHA frozen	0.686
MATCHA finetuning	**0.697**

Table 2: **Evaluation on Neuroprobe benchmark.** Mean AUROC across 15 decoding tasks. Best previously reported result underlined. (Reported from the benchmark leaderboard on September 5, 2026.)

5 Discussion

Our results show that useful summary representations of iEEG activity can be learned directly during self-supervised pretraining rather than introduced only as downstream pooling. By requiring masked reconstruction to proceed through temporal and spatial summary tokens, the information bottleneck explicitly trains these representations to aggregate structure across time and channels. We demonstrated that the resulting summaries support effective transfer and capture task-relevant information across diverse downstream tasks, and that reading out through the spatial, temporal, or combined summary streams offers a way to investigate which axis of aggregation matters most for a given task. Their ability to selectively aggregate informative channels further suggests that they learn structured representations rather than simple averages of the underlying tokens.

Although we evaluated MATCHA across a diverse set of iEEG downstream tasks and datasets, future work can extend evaluation to additional downstream datasets and tasks, and can pretrain on a larger, more heterogeneous cohort spanning a wider range of spatial coverage, acquisition, and preprocessing conditions. Future work can also explore incorporating spectral information, such as the Short-Time Fourier Transform, as model input rather than only as a reconstruction loss, to better capture frequency content and potentially improve downstream performance. Overall, our findings demonstrate that explicitly training temporal and spatial aggregation through an information bottleneck produces informative representations that transfer across diverse iEEG decoding tasks.

AI USE STATEMENT

In this work, we used generative AI tools solely to edit and polish the language and presentation of the main text and appendix. We did not use generative AI tools for developing the methodology, designing experiments, processing data, or interpreting results. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final content.

REFERENCES

Jeffrey R Binder, Julie A Frost, Thomas A Hammeke, Patrick SF Bellgowan, Jane A Springer, Jackie N Kaufman, and Edward T Possing. Human temporal lobe activation by speech and nonspeech sounds. *Cerebral cortex*, 10(5):512–528, 2000.

Francesco S. Carzaniga, Michael Hersche, Abu Sebastian, Kaspar Schindler, and Abbas Rahimi. A foundation model with multi-variate parallel attention to generate neuronal activity. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5M1YOW3bRq.

Geeling Chau, Christopher Wang, Sabera J Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FVuqJt3c4L.

Alain de Cheveigné. ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage*, 207:116356, February 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116356. URL https://www.sciencedirect.com/science/article/pii/S1053811919309474.

Guillermo Delgado-Garcia and Birgit Frauscher. Future of Neurology & Technology: Stereoelectroencephalography in Presurgical Epilepsy Evaluation. *Neurology*, 98(4):e437–e440, January 2022. doi: 10.1212/WNL.0000000000013088. URL https://www.neurology.org/doi/10.1212/WNL.0000000000013088.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15, 2010a.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010b. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.06.010. URL https://www.sciencedirect.com/science/article/pii/S1053811910008542.

Vasileios Dimakopoulos, Pierre Mégevand, Lennart H Stieglitz, Lukas Imbach, and Johannes Sarnthein. Information flows from hippocampus to auditory cortex during replay of verbal working memory items. *elife*, 11:e78677, 2022.

Vasileios Dimakopoulos, Lennart Stieglitz, Lukas Imbach, and Johannes Sarnthein. Dataset of intracranial EEG, scalp EEG, and beamforming sources from epilepsy patients performing a verbal working memory task. OpenNeuro, 2023. Version 1.0.1. doi:10.18112/openneuro.ds004752.v1.0.1.

Danny Dongyeop Han, Yonghyeon Gwon, Ahhyun Lucy Lee, Taeyang Lee, Seong Jin Lee, Jubin Choi, Sebin Lee, Jihyun Bang, Seungju Lee, David Keetae Park, Shinjae Yoo, Chun Kee Chung, and Jiook Cha. DIVER-1: Scaling Intracranial EEG Foundation Models for Transferable Representations, May 2026. URL http://arxiv.org/abs/2512.19097. arXiv:2512.19097 [cs.LG].

Haydn G. Herrema and Michael J. Kahana. Categorized free recall: Delayed free recall of word lists organized by semantic categories. OpenNeuro, 2024a. Version 2.2.0. doi:10.18112/openneuro.ds004809.v2.2.0.

Haydn G. Herrema and Michael J. Kahana. Spatial memory of object locations with open-loop stimulation at encoding. OpenNeuro, 2024b. Version 1.0.1. doi:10.18112/openneuro.ds005523.v1.0.1.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Jie Lisa Ji, Marjolein Spronk, Kaustubh Kulkarni, Grega Repovš, Alan Anticevic, and Michael W. Cole. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, 185:35–57, January 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.10.006. URL https://www.sciencedirect.com/science/article/pii/S1053811918319657.

Elizabeth L Johnson, Julia W Y Kam, Athina Tzovara, and Robert T Knight. Insights into human cognition from intracranial EEG: A review of audition, memory, internal cognition, and causality. *Journal of Neural Engineering*, 17(5):051001, October 2020. ISSN 1741-2552. doi: 10.1088/1741-2552/abb7a5. URL https://doi.org/10.1088/1741-2552/abb7a5.

J.Ph Lachaux, D Rudrauf, and P Kahane. Intracranial eeg and human brain mapping. *Journal of Physiology-Paris*, 97(4):613–628, 2003. ISSN 0928-4257. doi: https://doi.org/10.1016/j.jphysparis.2004.01.018. URL https://www.sciencedirect.com/science/article/pii/S0928425704000348. Neuroscience and Computation.

Georgios Mentzelopoulos, Evangelos Chatzipantazis, Ashwin G Ramayya, Michelle Hedlund, Vivek Buch, Kostas Daniilidis, Konrad Kording, and Flavia Vitale. Neural decoding from stereotactic EEG: accounting for electrode variability across subjects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LR1nnsD7H0.

Petr Nejedly, Vaclav Kremen, Vladimir Sladky, Jan Cimbalnik, Petr Klimes, Filip Plesinger, Filip Mivalt, Vojtech Travnicek, Ivo Viscor, Martin Pail, Josef Halamek, Benjamin H. Brinkmann, Milan Brazdil, Pavel Jurak, and Gregory Worrell. Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals. *Scientific Data*, 7(1):179, June 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0532-5. URL https://www.nature.com/articles/s41597-020-0532-5.

Lucine L Oganesian, Saba Hashemi, and Maryam M. Shanechi. BaRISTA: Brain scale informed spatiotemporal representation of human intracranial neural activity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=LDjBDk3Czb.

Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21(4):474–483, April 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0108-2. URL https://www.nature.com/articles/s41593-018-0108-2.

Eshani Patel, Yisong Yue, and Geeling Chau. Learning time-scale invariant population-level neural representations, 2025. URL https://arxiv.org/abs/2511.13022.

Steven M. Peterson, Satpreet H. Singh, Benjamin Dichter, Michael Scheid, Rajesh P. N. Rao, and Bingni W. Brunton. AJILE12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific Data*, 9(1):184, April 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01280-y. URL https://www.nature.com/articles/s41597-022-01280-y.

Saul Sternberg. High-speed scanning in human memory. *science*, 153(3736):652–654, 1966.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, November 2023. URL http://arxiv.org/abs/2104.09864. arXiv:2104.09864 [cs].

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Christopher Wang, Adam Uri Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=KZlJF8kguO.

Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. BrainWave: A Brain Signal Foundation Model for Clinical Applications, September 2025. URL http://arxiv.org/abs/2402.10251. arXiv:2402.10251 [q-bio.NC] version: 7.

Andrii Zahorodnii, Christopher Wang, Geeling Chau, Bennett Stankovits, Charikleia Moraitaki, Eli Gross, Alexander Brady, Andrei Barbu, Boris Katz, and Ila R. Fiete. Neuroprobe: Evaluating Intracranial Brain Responses to Naturalistic Stimuli, May 2026. URL http://arxiv.org/abs/2509.21671. arXiv:2509.21671 [cs.LG].

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DDkl9vaJyE.

Hui Zheng and Haiteng Wang. NeuroCLUS: A Foundation Model with Functional Clustering for Intracranial Neural Decoding. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pFweJM4Uw8.

Hui Zheng, Haiteng Wang, Weibang Jiang, Zhongtao Chen, Li He, Peiyang Lin, Penghu Wei, Guoguang Zhao, and Yunzhe Liu. Du-IN: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=uyLtEFnpQP.

A EXPERIMENTS

A.1 DATASET AND PREPROCESSING DETAILS

We pretrain MATCHA on the Brain Treebank dataset (Wang et al., 2024) and evaluate its representations across eight downstream iEEG datasets spanning spoken-word decoding, naturalistic-behavior state decoding, task-phase decoding in memory paradigms, and clinical seizure detection. These datasets vary substantially in experimental paradigm, electrode coverage, and availability of spatial metadata, providing a diverse evaluation of transfer across recording settings and task domains. We first describe the preprocessing pipeline shared across datasets, followed by dataset-specific details, including the downstream decoding task and any additional preprocessing or evaluation procedures.

Dataset preprocessing. Raw recording sessions are first Zapline-filtered (de Cheveigné, 2020) and notch filtered to remove line-noise harmonics (60 Hz in the US; 50 Hz in Europe and Asia), then bandpass filtered between 0.5 Hz and 250 Hz. This is followed by a channel-quality check with two criteria based on robust (median/MAD) z-scores computed across 10-second chunks: channels are rejected for anomalously high variance ($z > 10$), and for weak spatial correlations, computed as the maximum absolute correlation with any neighbor, flagged either by a low z-score relative to other channels ($z < -3$) or by falling below an absolute correlation floor of 0.1.

Segment generation. For pretraining, we apply session-wide normalization, segment recordings into 6-second windows, and apply a second quality check on the z-scored segments. Values are first clipped to ± 10 standard deviations. On the clipped data, channels are flagged within each segment for railing (staying at the clip bound for more than 5% of the segment), excessive kurtosis (> 20), excessive peak-to-peak amplitude (> 20 standard deviations), or near-zero variance (flat channels, std < 0.01). Flagged channels are removed from that segment while the remaining channels are retained. Dataset-specific segment generation and splitting procedures for downstream evaluation are described with each dataset below.

Brain Treebank dataset (Wang et al., 2024). This is a large-scale stereo-EEG (sEEG) dataset collected from 10 pediatric epilepsy patients at Boston Children's Hospital while they watched Hollywood movies. The dataset comprises roughly 43.5 hours of naturalistic viewing across 26 films and 1,688 electrodes (106–246 per subject, sampled at 2048 Hz), accompanied by multimodal audio-visual annotations. Electrode positions were localized and standardized to MNI coordinate space, which allows for obtaining spatial information required by our model's spatial encoding. Because electrode placement was driven entirely by clinical seizure-localization needs rather than experimental design, the dataset offers broad, naturalistic coverage of language- and stimulus-related neural activity across cortex, making it well suited as a pretraining corpus for learning general-purpose sEEG representations prior to finetuning on downstream decoding tasks. 19 recording sessions are used during **pretraining** ($\approx$29 hours of recording), using the segmentation procedure described above, and the remaining sessions are left-out for downstream evaluations using Neuroprobe benchmark (Zahorodnii et al., 2026).

Du-IN dataset (Zheng et al., 2024). In this dataset, sEEG recordings are collected from 12 subjects with pharmacologically intractable epilepsy during Chinese word-reading. Subjects were implanted with 7–13 sEEG probes (72–158 channels total at 1000 Hz) and performed a cued word-reading task, speaking aloud individual words from a 61-word set (each word repeated twice per block across 25 blocks, yielding 122 trials per block) while sEEG and voice were recorded simultaneously – totaling roughly 3 hours of task recordings per subject. Each trial follows a prepare–read–rest structure (0.5s cue, 2s speech production, 0.5s rest). The channel selection for this dataset was performed by the original authors based on individual channels' contribution to decoding performance, resulting in 10 channels for every subject. The spatial information for these channels are not available, so pretrained models leveraged the learned unknown spatial embedding during **finetuning**. We segment the recordings into 3-second windows corresponding to individual trials. Following the original work, we apply subject-level normalization before assigning random training, validation, and test splits. For this dataset, the downstream decoding task is 61-class word classification.

AJILE12 (Peterson et al., 2022). AJILE12 (Annotated Joints in Long-term Electrocorticography for 12 human participants) contains multi-day, naturalistic ECoG recordings from 12 epilepsy patients undergoing clinical monitoring, together with electrode positions in MNI space, and coarse behavioral annotations of each recording epoch. Because participants were not instructed to perform any task, the recordings capture unconstrained daily behavior, making the dataset a test of whether pretrained representations transfer to spontaneous, non-trial-structured activity. We define a multilabel activity-recognition task over five annotated activities—sleep/rest, talking, watching TV, using a computer or phone, and eating—excluding the "inactive" and "other activity" labels. Annotated epochs are tiled into non-overlapping 1-second windows, each inheriting the (possibly multiple) activity labels of its epoch; windows that span a gap in the recording are discarded. Because activity durations are highly imbalanced, we undersample epochs within each session so that each activity is represented by a comparable number of epochs, retaining approximately one hour of data per session. We finetune on 12 daily sessions from subjects 01, 03, and 04 (approximately 291 hours of recording, from which roughly one hour of labeled windows per session is retained after balancing, giving 43,118 one-second windows or 12.0 hours in total), and split windows at random into training, validation, and test sets (80/10/10%) within each session, stratified by label combination.

Mayo and FNUSA (Nejedly et al., 2020). These two datasets comprise expert-annotated, single-channel iEEG clips from patients with drug-resistant epilepsy recorded at Mayo Clinic (Rochester, MN, USA; 24 patients, 155,182 clips totaling 129 hours, depth and strip electrodes) and St. Anne's University Hospital (FNUSA; Brno, Czech Republic; 14 patients, 193,118 clips totaling 161 hours, depth electrodes). Each clip is 3 seconds long, sampled at 5 kHz, and labeled into one of four classes: power-line noise, other artifacts, physiological activity, or pathological (interictal epileptiform) activity. All clips are retained without removing noisy ones, and we define a binary task that separates pathological activity (positive class) from all other clips, merging physiological activity, artifacts, and power-line noise into the negative class. Since MNI coordinates are not available, all channels use the learned unknown-position embedding. We finetune on the first four patients of each dataset, holding out 10% of their clips for validation, and test on all clips from the remaining patients, so that evaluation is entirely out-of-patient. Because many test patients contain no pathological activity, we report the F1-score computed over the pooled test clips of all patients rather than averaged per patient, together with the average sensitivity and specificity.

`ds004809`: Categorized free recall (Herrema & Kahana, 2024a). This OpenNeuro dataset was collected by the Computational Memory Lab at the University of Pennsylvania across multiple clinical sites, and contains iEEG from epilepsy patients performing a categorized delayed free-recall task. In each list, 12 words drawn from three of 25 semantic categories are presented visually one at a time (encoding), followed by an arithmetic distractor task and a free-recall period in which participants verbally recall as many words from the list as possible in any order. MNI coordinates and anatomical annotations are provided where available. We define a 3-class task-phase classification: encoding, distractor, and recall. Phase boundaries are taken from the behavioral event markers (encoding start/end, distractor start/end, and recall start/end, with fallbacks to adjacent markers for sessions in which a marker is missing); each phase block is then chunked into non-overlapping 1-second segments. We finetune on five sessions from four subjects (R1004D, R1013E, R1015J, and R1016M) which comprise approximately 6.3 hours of recording and yield 2.8 hours of labeled task-phase segments. Segments are split at random into training, validation, and test sets (80/10/10%), and the training set is class-balanced by undersampling the majority classes; validation and test sets retain their natural class proportions.

`ds005523`: Spatial navigation memory with open-loop stimulation (Herrema & Kahana, 2024b). This OpenNeuro dataset contains iEEG recordings from patients performing a virtual spatial-navigation memory task. Each experimental trial consists of two guided learning trials in which the participant is automatically turned toward and driven to an object's location, followed by a test trial in which the object is invisible and the participant navigates freely to where they believe it was located. Open-loop electrical stimulation is delivered to a single hippocampal or entorhinal electrode during the learning trials of half of the experimental trials, and never during test. We define a binary navigation-phase classification (learning vs. test), with phase blocks delimited by consecutive learning- and test-onset markers and chunked into non-overlapping 1-second segments. Because stimulation artifacts are confined to the learning phase and could otherwise be exploited by

the classifier, stimulation periods are blanked and interpolated during preprocessing. We finetune on 17 sessions from five subjects (R1006P, R1013E, R1015J, R1017J, and R1019J), which comprise approximately 14.6 hours of recording and yield 12.9 hours of labeled navigation-phase segments, and use the same random split and training-set class balancing as for `ds004809`.

`ds004752`: Verbal working memory (Dimakopoulos et al., 2022; 2023). This OpenNeuro dataset contains simultaneous iEEG and scalp EEG from 15 epilepsy patients at the Swiss Epilepsy Center (Zurich) performing a modified Sternberg (Sternberg, 1966) task with temporally separated encoding, maintenance, and retrieval periods. Each stored trial epoch lasts 8 seconds and consists of four stages. A 1-second fixation period is followed by a 2-second encoding period, during which 4, 6, or 8 consonants are shown on screen while their corresponding letter names are played aloud; unused positions in the eight-letter display are filled with Xs. This is followed by a 3-second maintenance period, during which participants rehearse the letters subvocally, and the first 2 seconds of the retrieval period, beginning with a probe letter, during which participants indicate by button press whether the probe appeared in the encoded string or not. We randomly assign trials as a whole to the training, validation, and test sets, keeping the four segments of a given trial in the same split. Following the original study, we keep only the correctly answered, artifact-free trials to focus decoding on successful task performance, and finetune the models for each participant by pooling their sessions to increase the number of training trials.

The release contains 68 sessions from 15 subjects, with 2–8 sessions and up to 50 trials per session, totaling 3,353 trials and approximately 7.5 hours of stored trial data. Of these, 168 trials are flagged as artifactual and 3,045 were answered correctly. Depth electrodes were implanted in the hippocampus of every patient, and four patients additionally received subdural grid or strip electrodes, giving 32–80 intracranial channels per session. Released intracranial recordings were made at 2, 4, or 4.096 kHz depending on the session, and MNI coordinates are provided for the intracranial electrodes. We use only the intracranial channels and discard the simultaneously recorded scalp EEG and the beamforming-source derivatives.

Benchmark Evaluation on Brain Treebank (Zahorodnii et al., 2026). Neuroprobe is a standardized multitask benchmark built on Brain Treebank dataset for evaluating iEEG foundation models. It defines 15 binary decoding tasks spanning auditory (e.g., speech, volume, pitch), visual (e.g., frame brightness, optical flow, face count), and language (e.g., GPT-2 surprisal, part-of-speech, word position) features aligned with naturalistic movie stimuli. Each task takes as input a 1 second window of neural activity starting at the event onset and is evaluated using 2-fold cross-validation within a recording session, where temporally contiguous train and test segments are separated to prevent leakage from temporal autocorrelation. In our experiments, we follow the benchmark protocol and report performance on the within-session split using AUROC averaged across the 15 decoding tasks, 2 evaluation folds, and 12 recording sessions. For MATCHA, we read out from all summary tokens and concatenate them rather than pooling with cross-attention, as concatenation yielded better validation performance on this benchmark. We also adhered to the benchmark protocols of the benchmark by not including any evaluation session in the pretraining corpus.

Model	Sentence onset	Speech	Volume	Delta volume	Voice pitch	Word position	Inter-word gap	GPT-2 surprisal
Linear band power	0.895	0.881	0.738	0.701	0.601	0.605	0.540	0.581
Linear + spectrogram*	0.888	0.882	0.717	0.756	0.584	0.740	0.613	0.608
CNN + spectrogram*	0.886	0.888	0.757	0.762	<u>0.618</u>	0.734	0.583	0.597
BrainBERT (frozen)*	0.883	0.887	0.723	0.677	0.564	0.620	0.526	0.557
PopT*	<u>0.930</u>	<u>0.919</u>	<u>0.766</u>	0.781	0.596	0.758	0.620	0.603
DIVER-1 (frozen)*	<u>0.930</u>	0.897	0.719	<u>0.803</u>	0.592	<u>0.798</u>	<u>0.636</u>	<u>0.629</u>
MATCHA, frozen	0.936	0.917	0.761	0.802	0.614	0.784	0.615	0.619
MATCHA, finetuning	**0.948**	**0.934**	**0.767**	**0.826**	**0.625**	**0.815**	**0.639**	**0.633**

Table 3: Within-session AUROC on tasks 1–8, averaged across folds and sessions. We report our three finetuning readouts alongside the best leaderboard results for each baseline. In each column, bold marks the best result in the table and underline marks the best result among the benchmark baselines.

Model	Head-word position	Part of speech	Word length	Global optical flow	Local optical flow	Frame brightness	Number of faces
Linear band power	0.590	0.593	0.609	0.628	0.619	0.500	**0.541**
Linear + spectrogram*	0.603	0.601	0.617	0.630	0.615	0.528	0.519
CNN + spectrogram*	0.614	0.621	0.636	0.638	0.630	**0.536**	0.530
BrainBERT (frozen)*	0.567	0.576	0.586	0.609	0.599	0.497	0.516
PopT*	0.601	0.604	0.622	0.621	0.618	0.496	0.517
DIVER-1 (frozen)*	0.623	0.627	0.646	0.628	0.616	0.495	0.525
MATCHA, frozen	0.627	0.638	0.654	**0.643**	**0.647**	0.495	0.537
MATCHA, finetuning	**0.633**	**0.648**	**0.674**	**0.643**	0.632	0.507	0.528

Table 4: Within-session AUROC on tasks 9–15. Models, averaging, and bold/underline conventions follow Table 3. Mean AUROC across all 15 tasks is reported in Table 2.

A.2 Model architecture and training details

Encoder. Segments are resampled to 500Hz and split into non-overlapping 50-sample (100ms) patches, one per channel. Each patch is embedded from its time-domain waveform and its FFT magnitude (DC bin excluded, since it encodes the patch's mean amplitude rather than oscillatory content), and the channel's spatial embedding is added to every patch token (Fig. 1). Channels without usable coordinates use a shared learnable unknown-position embedding. During training, 10% of training segments have the spatial embedding of every channel replaced by this unknown embedding, mimicking recordings that carry no spatial metadata at all. To condition the encoder on systematic differences in recording and preprocessing, we augment each channel's temporal token sequence with two learned metadata tokens. The modality token indicates whether the channel was recorded using stereo-electroencephalography (sEEG) or electrocorticography (ECoG). The re-referencing token indicates the reference applied to the signal, such as bipolar, Laplacian, or common-average referencing, with a separate embedding used when this information is unavailable. These tokens are concatenated along the temporal axis of each channel and therefore participate in temporal attention alongside the patch tokens. The encoder is a 10-layer factorized spatiotemporal transformer (d_{model}=256) with rotary position embeddings in the temporal attention and no positional encoding in the spatial attention. It has 10.72 M parameters (0.19 M tokenizer, 10.53 M transformer); the pretraining heads (2.49 M for the MAE predictor, 3.62 M for STFT reconstruction) are discarded after pretraining.

Summary tokens. Each channel carries L=7 temporal summary tokens, and a single (M=1) spatial summary row spans the full time axis. Because the temporal-summary tokens are attached before the spatial-summary row, the row also crosses the 7 temporal-summary columns; these $M \times L$=7 positions form a third, *spacetime* summary stream. The three streams are disjoint, and the bottleneck predictor reads all three. The 7 temporal-summary slots of each channel are also the decoding sites of the seven STFT bands (one band per slot).

Pretraining. We pretrain on 19 Brain Treebank sessions from 10 subjects with masked autoencoding: contiguous temporal spans (geometric length, mean 4 patches) are masked independently per channel until 70% of each channel's patches are hidden. Under the bottleneck, a 4-layer factorized predictor (d=192) reconstructs the masked waveforms from the summary tokens alone. The loss sums three terms, each with weight 1.0: the waveform MSE, the MSE on the log-power spectrum $\log(1+|\mathrm{rFFT}|^2)$ of each masked patch (26 bins), and the band-wise STFT log-power reconstruction. We keep the checkpoint with the lowest validation MAE loss.

Finetuning. Each downstream session is adapted independently, with the encoder optimization mode selected using validation performance. The newly initialized readout and classifier are trained at a learning rate of 6.25×10^{-4}, while the pretrained encoder is either updated end-to-end at a smaller learning rate (10^{-5}) or held fixed. The readout is a 4-query, 8-head cross-attention pool over the chosen summary tokens, producing a 1024-dimensional vector (4×256) regardless of how many tokens it reads, followed by LayerNorm, dropout and a linear classifier. We use either spatial or temporal summary tokens or their combination for the readout in all downstream evaluations. Early stopping and checkpoint selection both use validation F1 as logged during training: macro-averaged

over classes for multiclass and multilabel tasks, and the F1 of the positive class for the binary tasks. Reported test scores are macro-F1 over the classes present in each session's test split.

Electrode localization. We note that for datasets that provide only native subject-space coordinates alongside a structural MRI, we first register the subject MRI to MNI space using Statistical Parametric Mapping (SPM)[2], and then use the registration to transform the electrode coordinates accordingly. We visually inspect each transformed coordinate set against the registered anatomy and treat coordinates from unreliable registrations as missing. Channels without usable spatial metadata are assigned the unknown-position embedding, $\boldsymbol{e}^s_{\text{unk}}$. Finally, we also apply dropout at the level of individual channels during training: with some probability, a channel's spatial embedding is replaced with the same learnable unknown-position embedding, $\boldsymbol{e}^s_{\text{unk}}$. This exposes the model to missing spatial information during pretraining and allows it to remain functional at inference for channels, subjects, or datasets that lack spatial metadata entirely.

Finetuning readout formulation During finetuning we apply the pretrained encoder, as described in equation 3, to all content patches in a segment and concatenate the encoded temporal and spatial summary representations along the token dimension as such,

$$\boldsymbol{Z} = [\, \boldsymbol{H}_T;\, \boldsymbol{H}_S \,] \in \mathbb{R}^{K \times d_{\text{model}}}, \qquad K = LC_i + MN_i. \tag{5}$$

We then use a single multi-head cross-attention layer with a set of learnable query vectors, $\boldsymbol{Q} \in \mathbb{R}^{Q \times d_{\text{model}}}$, to attend over $\boldsymbol{Z}$ and generate a more compact representation $\boldsymbol{U}$

$$\boldsymbol{U} = \text{MHA}\left(\text{LN}(\boldsymbol{Q}), \text{LN}(\boldsymbol{Z}), \text{LN}(\boldsymbol{Z})\right) \in \mathbb{R}^{Q \times d_{\text{model}}}, \tag{6}$$

where MHA denotes multi-head attention and LN denotes layer norm. Each query therefore produces an attention-weighted pool of the encoded summary representations. Because the number of queries Q is fixed (e.g., $Q = 4$), $\boldsymbol{U}$ has a fixed size ($\mathbb{R}^{Q \times d_{\text{model}}}$) regardless of the number of channels or temporal patches in the input. This key design decision makes our approach robust to situations such as channel dropout—a common occurrence during iEEG data collection. Finally, we flatten and pass the Q pooled representations to a linear task head for downstream decoding,

$$\hat{\boldsymbol{y}} = \boldsymbol{W}_o \,\text{Drop}(\text{LN}\left([\, \boldsymbol{u}_1; \ldots; \boldsymbol{u}_Q \,]\right)) + \boldsymbol{b}_o, \tag{7}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{d_{\text{out}} \times Q d_{\text{model}}}$ and d_{out} denotes the number of task outputs. Unless otherwise stated, in all of our analyses the backbone model is finetuned end-to-end with the cross-attention and task heads following a smaller learning rate.

A.3 Baseline Details

Band-power features. As a classical baseline, we decode each downstream task from spectral power using the same data splits as the other models. For each segment, we estimate the power spectral density of every channel using Welch's method with a Hann window of $\min(f_s, T)$ samples (up to one second at f_s=500 Hz) and 50% overlap. We average the PSD within six canonical bands: theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), low gamma (30–55 Hz), gamma (55–90 Hz), and high gamma (90–200 Hz), and apply $\log(1 + x)$, yielding six features per channel.

Because per-trial bad-channel rejection produces variable channel sets, we define a fixed electrode ordering for each session and place each segment's features into the corresponding slots. Missing-channel entries are imputed using training-split means, which become zero after standardization. This produces a fixed $6N_{\text{ch}}$-dimensional vector per segment without discarding segments; only segments with remaining non-finite features are removed.

Features are standardized using statistics estimated from the training split. A one-vs-rest ℓ_2-regularized logistic regression is then fit separately for each session, with $C \in \{0.01, 0.1, 1, 10\}$ selected by validation macro-F1 and the selected model evaluated on the test split.

Brant. The original Brant model (Zhang et al., 2023) was pretrained on recordings downsampled to 250 Hz and uses a sequence of 15 six-second patches, each containing 1,500 samples per channel.

[2] https://www.fil.ion.ucl.ac.uk/spm/

Hyperparameter	Value
Tokenization	
Sampling rate	500 Hz
Temporal patch length	50 samples (100 ms), non-overlapping
Value embedding	time-domain conv. (128 ch.) $\oplus$ FFT magnitude (DC excluded)
Spatial encoding	12 partitions (CAB-NP), added to tokens
Unknown-position embedding	learnable; replaces all channels' spatial embeddings in 10% of training samples
Metadata tokens	modality, re-referencing (1 concatenated token each)
Metadata ID dropout	0.1 (modality and re-referencing IDs, per channel)
Temporal summary tokens L	7 per channel
Spatial summary tokens M	1
Encoder	
Architecture	factorized spatiotemporal transformer
Layers / d_{model} / heads	10 / 256 / 8
MLP ratio / activation	4 / GELU
Normalization	RMSNorm, pre-norm
Temporal / spatial attention	RoPE (max. position 2048) / no positional encoding
Dropout	0.1
Parameters	10.72 M
Pretraining objective	
Masking	per-channel temporal spans, ratio 0.7, mean length 4 patches
Predictor	factorized transformer, 4 layers, d=192, 8 heads
Information bottleneck	predictor sees temporal and spatial summary tokens only
Reconstruction losses	waveform MSE + patch log-power spectrum MSE (each weight 1.0)
STFT reconstruction	7 bands: 0–4, 4–8, 8–13, 13–30, 30–55, 55–90, 90–200 Hz
STFT parameters	n_{fft}=500, window 500, hop 125; DC and Nyquist excluded
STFT loss weight	1.0
Pretraining optimization	
Optimizer	AdamW
Peak learning rate	6.25×10^{-4}
Learning-rate schedule	linear warm-up from $0.3\times$ over 1,500 steps, then cosine to 10^{-6} over 50,000 steps
Weight decay	0.1 $\rightarrow$ 0.4, linear over 50,000 steps
Epochs	up to 100; early stopping (patience 50) on validation MAE loss
Selected checkpoint	lowest validation MAE loss (epoch 90, step 53,417)
Pretraining data	
Dataset	Brain Treebank, 19 sessions, 10 subjects
Segments	6 s
Preprocessing	0.5 Hz high-pass, 250 Hz low-pass, Laplacian re-referencing, per-session z-scoring

Table 5: Architecture and pretraining configuration of the pretrained MATCHA checkpoint used for all downstream results.

Hyperparameter	Value
Readout	4-query, 8-head cross-attention pool $\rightarrow 4\times256 = 1024$ dims
Classifier	LayerNorm $\rightarrow$ dropout $\rightarrow$ linear
Optimizer	AdamW
Learning rate	10^{-5} pretrained parameters; 6.25×10^{-4} new parameters
Model selection	validation F1 (early stopping and checkpoint)

Table 6: Finetuning configuration, shared by all readouts and downstream datasets in Table 1.

This results in a 90-second input window with a total of $15 \times 1{,}500 = 22{,}500$ samples. Following the approach in the authors' released preprocessing for the 3-second segments of the MAYO/FNUSA datasets, we resize our shorter downstream segments with Fourier resampling to 22,500 samples, divided into 15 patches per channel to match Brant's input.

Brant also uses frequency encodings derived from eight spectral bands spanning 4–128 Hz. Computing these features from the resized patches distorts their physical frequency interpretation and does not align with the model's pretraining. While correcting for the sampling rate could be a workaround, dividing the short 1-second segments into 15 patches means each patch spans approximately 67 ms (originally sampled at 250Hz), leading to a frequency-bin spacing of about 15 Hz. This lies above both the theta (4–8 Hz) and alpha (8–13 Hz) bands. To resolve this, we compute spectral-band features for each channel's complete segment at its original sampling rate before resizing, and share these features across the channel's 15 temporal patches. For a 1-second segment (our shortest downstream segment), this results in a 1 Hz frequency-bin spacing, correctly capturing the low-frequency bands of the signal.

Finally, to obtain a single prediction for each segment, we use cross-attention pooling over all 2048-dimensional channel-time embeddings. We use a single learned query with four attention heads, followed by a linear classifier, and define the segment-level loss function as cross-entropy for multiclass tasks and binary cross-entropy for binary and multilabel tasks. The encoders and the readout are finetuned with learning rates of 5×10^{-6} and 10^{-4}, respectively. The finetuning lasts for 150 epochs, with checkpoint selection and early stopping based on validation F1, with a patience of 30 epochs.

Our Brant configuration achieved a higher mean macro-F1 than the original configuration, which uses a token-wise MLP classifier, token-wise cross-entropy loss, linear interpolation instead of Fourier resampling, and the released per-patch spectral computation, on the evaluated datasets (Table 7).

Model	Du-IN	AJILE12	`ds004809`	`ds005523`	`ds004752`
Brant (original)	0.0047 ± 0.0024	0.3904 ± 0.1042	0.3742 ± 0.0583	0.5956 ± 0.0929	0.3420 ± 0.0747
Brant (adapted)	$\mathbf{0.0638 \pm 0.0969}$	$\mathbf{0.6144 \pm 0.2016}$	$\mathbf{0.4521 \pm 0.0789}$	$\mathbf{0.6093 \pm 0.1118}$	$\mathbf{0.4524 \pm 0.1018}$

Table 7: Downstream performance of the original and adapted Brant, in macro-F1 (mean $\pm$ standard deviation across session-specific finetuned models; per-subject models for Du-IN and `ds004752`).

Population Transformer (PopT). The original PopT model (Chau et al., 2025) uses a frozen BrainBERT encoder to process each electrode independently and was pretrained on 5-second neural segments sampled every 0.2 seconds. BrainBERT maps each electrode's STFT representation to a 768-dimensional embedding, which is then projected into the 512-dimensional PopT transformer space. We use the released checkpoint for BrainBERT [3], the same checkpoint PopT used to extract its pretraining features, together with the released PopT upstream weights.[4]

In our initial transfer experiments, this released configuration performed poorly when downstream segment durations differed from the 5-second pretraining duration. This sensitivity is consistent with Patel et al. (2025), who report that population-level representations of PopT are not invariant to time-scale mismatches between pretraining and downstream settings, and who address it by pretraining across a range of segment lengths. Motivated by this, we adapted the input encoder and PopT's pretraining segment length to our downstream setting, where most datasets use 1-second segments and Du-IN, MAYO, and FNUSA use 3-second segments. We replaced BrainBERT with a learned linear projection over flattened STFT features and pretrained PopT from scratch on 1-second Brain Treebank segments—the shortest segment length across datasets in our evaluation, so that no padding or resampling is required when transferring to downstream datasets. Preprocessing and pretraining sessions match MATCHA.

Apart from the input encoder and segment duration, we retained the original PopT transformer backbone and pretraining hyperparameters.[5] The linear input encoder uses log-power STFT features computed at 500 Hz using Welch's method with a 125-sample Hann window, 92-sample overlap, and 50 frequency bins. Each channel's features are flattened into a 650-dimensional vector and projected into the 512-dimensional PopT representation.

For downstream datasets, with 1-second segments, we used the same finetuning recipe for decoding as the original PopT. For downstream trials longer than 1 second, we divided each trial into overlapping

[3] `https://drive.google.com/file/d/14ZBOafR7RJ4A6TsurOXjFVMXiVH6Kd_Q/view`
[4] `https://huggingface.co/PopulationTransformer/popt_brainbert_stft`
[5] `https://github.com/czlwang/PopulationTransformer`

1-second crops sampled every 0.2 seconds, each retaining all electrodes. PopT processes each crop independently and produces one CLS representation per crop. For example, in the Du-IN dataset a 3-second trial yields 21 crops with a 100 ms stride, matching the temporal patch resolution of MATCHA. We then aggregate the resulting CLS summary representations from PopT across all crops using cross-attention pooling with learned queries, then concatenate the pooled query representations and pass them to the classifier, following the same readout setup as in our downstream decoder.

Linear+PopT outperformed the original BrainBERT-based configuration on the evaluated datasets (Table 8). On Du-IN, macro-F1 increased from 0.0225 to 0.1876.

Model	Du-IN	AJILE12	`ds004809`	`ds005523`	`ds004752`
BrainBERT+PopT (original)	0.0225 ± 0.0080	0.4390 ± 0.1756	0.4212 ± 0.0462	0.6632 ± 0.0562	0.4390 ± 0.0919
Linear+PopT	$\mathbf{0.1876 \pm 0.0918}$	$\mathbf{0.8450 \pm 0.1292}$	$\mathbf{0.7691 \pm 0.0869}$	$\mathbf{0.7756 \pm 0.0468}$	$\mathbf{0.7254 \pm 0.0995}$

Table 8: Downstream performance of the original and adapted PopT, in macro-F1 (mean $\pm$ standard deviation across session-specific finetuned models; per-subject models for Du-IN and `ds004752`).

BaRISTA. For BaRISTA (Oganesian et al., 2025), we finetuned the publicly released parcel checkpoint, which achieved the best reported performance among the checkpoints provided by the authors. This checkpoint uses regions from the Destrieux atlas for spatial encoding.

We made several adaptations for our downstream setting. Because our segments are shorter than those used to develop BaRISTA for most datasets, and some datasets contain only one channel, the model receives relatively few tokens per sample. We therefore used the existing `temporal_subsegment_step` configuration in the public repository to generate patches with 75% overlap, while retaining the original 250 ms patch size. This modification only affects token generation within each segment, and the input segments and train/validation/test splits remain unchanged. We also disabled the tokenizer's per-segment, channel-wise z-scoring, since all inputs were already robust-scaled per channel over the full recording session.

Because BaRISTA relies on electrode locations and the Du-IN dataset provides no electrode coordinates, all of its channels are assigned to the same unknown region. We therefore added a learned 64-dimensional embedding per electrode, initialized from $\mathcal{N}(0, 0.02^2)$ and added to each corresponding token after the tokenizer and before the transformer. For consistency, we used this adaptation across all datasets, except MAYO and FNUSA, which only had one channel. We kept the tokenizer frozen and finetuned the transformer. The output tokens were aggregated using attention pooling with a single learned query, four attention heads, and a dropout rate of 0.1, then passed to a linear classifier shared across all sessions within each model. We used attention pooling because it improved performance in preliminary experiments.

We finetuned the model using AdamW, with a learning rate of 10^{-4} for pretrained weights and 10^{-3} for newly introduced parameters, a weight decay of 0.001, and a batch size of 32. For FNUSA, we increased the learning rates for pretrained and new parameters to 10^{-3} and 10^{-2}, respectively.

Models were trained for up to 300 epochs with 60 epoch early-stopping patience on validation F1, and the checkpoint with the highest validation F1 was used for final evaluation on the test set.

Du-IN. Du-IN (Zheng et al., 2024) is pretrained separately for every subject in two stages. A vector-quantized autoencoder is first trained on that subject's recordings to produce a discrete neural codebook, after which a masked autoencoder is trained to predict the codebook indices of masked temporal patches. Both stages operate on a per-subject spatial encoder that collapses a fixed set of ten selected channels—chosen for that subject according to their contribution to decoding performance—into a single regional representation through a linear layer followed by depthwise temporal convolutions, before a shared temporal transformer. The channel-fusion weights are therefore tied to one subject's montage, channel count, and channel ordering.

On the Du-IN dataset, we use the per-subject pretrained weights the authors share on Hugging Face,[6] and finetune each subject's model on the same train, validation, and test splits used for every other model in this work. To keep the comparison matched, we replace the model's native data pipeline

[6] `https://huggingface.co/datasets/liulab-repository/Du-IN`

with our own segment loader, so that the baseline sees exactly the same segments and labels as every other model. Finetuning then initializes the tokenizer and temporal encoder from that subject's masked-autoencoder checkpoint and trains a newly initialized classification head, for 200 epochs with batch size 32 and a cosine learning-rate schedule from 5×10^{-6} to 2×10^{-4} with 20 warmup epochs, which are the finetuning settings specified in Du-IN's public codebase.[7] For each subject, we then select the epoch with the highest validation accuracy and report test performance at that epoch.

Here, we evaluate Du-IN model only on Du-IN dataset. The reason is that its pretraining is defined per subject: the spatial encoder's input dimension and channel ordering are fixed to a single montage, it encodes neither electrode coordinates nor anatomical identity, and the ten channels it operates on are selected using labeled data from that subject. Transferring to a new subject or electrode implant therefore requires repeating the channel-selection procedure and both pretraining stages from scratch, at a reported cost of approximately 1.2 V100-days per subject (Zheng et al., 2024). Consequently, there is no single pretrained Du-IN model that could be transferred across datasets, as the evaluation setting of our work requires. We therefore restrict this baseline to the Du-IN dataset, for which the authors provide pretrained checkpoints for each of its 12 subjects.

	Downstream dataset						
Readout	Du-IN	AJILE12	ds004809	ds005523	ds004752	MAYO	FNUSA
Spatial summary	$\underline{\mathbf{0.663 \pm 0.169}}$	$\underline{\mathbf{0.928 \pm 0.078}}$	$\underline{\mathbf{0.894 \pm 0.110}}$	$\underline{\mathbf{0.896 \pm 0.058}}$	$\underline{\mathbf{0.854 \pm 0.056}}$	0.675	$\underline{0.695}$
Temporal summary	0.472 ± 0.240	0.908 ± 0.092	0.864 ± 0.087	0.886 ± 0.054	0.834 ± 0.078	$\mathbf{0.689}$	0.691
Combined (all)	0.638 ± 0.171	0.902 ± 0.101	0.877 ± 0.085	0.880 ± 0.061	0.832 ± 0.081	$\underline{0.667}$	$\mathbf{0.700}$

Table 9: **Per-dataset performance for each summary-token readout option.** Spatial, temporal, and combined readouts compared across all seven downstream datasets. Bold marks the best test performance per dataset; underline marks the readout selected by validation performance, which is the one reported in Table 1.

	Downstream dataset						
Model	Du-IN	AJILE12	ds004809	ds005523	ds004752	MAYO	FNUSA
MATCHA w/o bottleneck	0.532 ± 0.204	$\mathbf{0.923 \pm 0.073}$	0.829 ± 0.093	0.864 ± 0.061	$\mathbf{0.855 \pm 0.058}$	0.657	0.688
MATCHA w/ bottleneck	$\mathbf{0.638 \pm 0.171}$	0.902 ± 0.101	$\mathbf{0.877 \pm 0.085}$	$\mathbf{0.880 \pm 0.061}$	0.832 ± 0.081	$\mathbf{0.667}$	$\mathbf{0.700}$

Table 10: **Pretraining with an information-bottleneck provides consistent improvement across downstream datasets.** Comparison of MATCHA with and without the information bottleneck, using both spatial and temporal summary tokens with attention-pooling readout. Dataset metrics and reporting conventions follow Table 1; higher values indicate better performance. mean $\pm$ std across subjects

A.4 Summary-token interpretability analyses

We examined finetuned models using three complementary analyses: channel contributions in Brain Treebank (Figure 4a), downstream attention to spatial summaries in Du-IN (Figure 4b), and temporal-summary attention to content patches in MAYO (Figure 4c).

Speech-versus-non-speech decoding (Brain Treebank). We analyzed a model finetuned using combined summary streams. For each test segment, we applied the trained classification head's LayerNorm to the complete concatenation of all summary streams (dropout disabled at evaluation), then extracted the normalized features corresponding to the seven final-layer time-summary-token representations associated with each channel. Each normalized feature was multiplied by its corresponding linear-classifier weight, and these signed contributions were summed across embedding dimensions and all seven time-summary tokens for each channel; we then took the absolute value of this sum and normalized the resulting scores across channels within each segment. Channel importance was averaged separately within each test class, then equally across classes and folds of the evaluated speech-vs-non-speech task. Channels were assigned to Destrieux parcels (Destrieux et al., 2010a) using Brain Treebank electrode-localization labels, and parcel scores were obtained by averaging channels within each session, sessions within each subject, and subjects with coverage in that parcel; only cortical parcels were visualized.

[7] https://github.com/liulab-repository/Du-IN

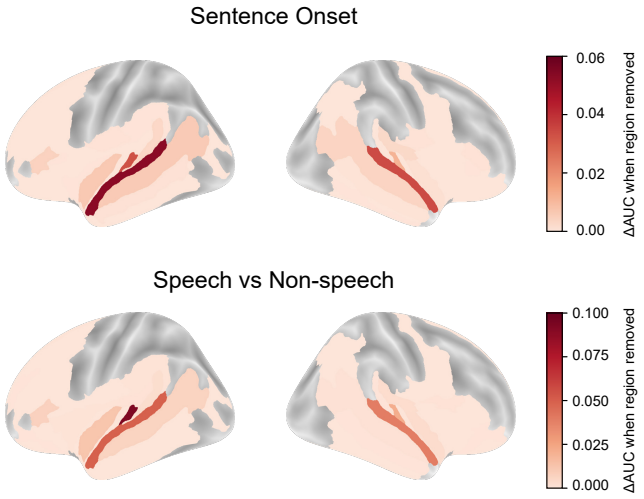

Figure 5: **Regional occlusion reveals preferential dependence on superior temporal recording sites.** Decrease in test AUROC (ΔAUROC) after inference-time occlusion of the electrodes in each Destrieux region, for sentence-onset (top) and speech-versus-non-speech (bottom) decoding. For each region, the content tokens and temporal-summary tokens of all its electrodes were removed. Values are averaged over sessions and folds within each subject and then across subjects covering the region. Regions without electrodes are shown in gray.

Recording-Site Dependence of Summary Representations We next asked whether the learned summary representations depend preferentially on recording sites in particular anatomical regions. We performed inference-only regional occlusion on the Neuroprobe speech and sentence-onset tasks (12 sessions from 6 subjects $\times$ 2 folds per task). For each region of the Destrieux atlas (Destrieux et al., 2010b), we removed all electrodes assigned to that region by excluding both their content tokens and their temporal summary tokens from attention, and measured the resulting decrease in test AUROC (ΔAUROC).

Model dependence was concentrated in superior temporal cortex (Fig. 5). The largest effect was observed in the left transverse temporal sulcus for speech decoding ($\Delta\text{AUROC} = 0.092$); the same region also showed an effect for sentence onset (0.033). The lateral superior temporal gyrus (STG) showed consistent dependence across both tasks and hemispheres, with ΔAUROC values of 0.050 and 0.053 in the left hemisphere and 0.039 and 0.034 in the right hemisphere for speech and sentence onset, respectively. Outside superior temporal cortex, no region exceeded $\Delta\text{AUROC} = 0.013$.

Because larger regions contain more electrodes and may therefore produce larger drops simply through the removal of more input channels, we compared each regional occlusion with 20 random electrode sets of the same size from the same session and fold. Removing these size-matched random sets reduced AUROC by at most 0.005 for the regions above. We tested whether each region's ΔAUROC differed from that of the size-matched random sets using a rank test combined across sessions, for the 47 regions sampled in at least two subjects. Lateral STG exceeded size-matched random removal in both hemispheres and both tasks (speech: $p = 0.009$ left, $p = 0.003$ right; sentence onset: $p = 0.001$ left, $p < 0.001$ right). The concentration of effects in superior temporal cortex is consistent with its role in auditory processing (Binder et al., 2000; Hickok & Poeppel, 2007), providing evidence that the learned representations capture functionally relevant spatial structure aligned with known neuroanatomy.

Spoken-word classification (Du-IN). We analyzed a model finetuned using both temporal and spatial summary tokens. For this analysis, we retained the trained readout's cross-attention mechanism (Section 3.2) but focused on attention from $Q = 4$ learned queries to the spatial summaries. For each attention head, this produces

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_h}}\right) \in \mathbb{R}^{Q\times N}, \tag{8}$$

where $\boldsymbol{K}$ are the projected keys of the N spatial summary tokens. Entry $A_{q,n}$ is the attention from readout query q to the spatial summary token at time patch n: it measures how much that time patch's cross-channel content contributes to the pooled representation query q builds, which is what the linear classifier ultimately acts on. Because this pooled representation is exactly what the model uses to make its word-classification decision, $\boldsymbol{A}$ tells us which time steps within a spoken-word segment actually drive the downstream prediction. We average $\boldsymbol{A}$ across attention heads and across the Q queries giving one attention value per time step; these values sum to one across time patches and form a probability distribution. We report the mean of these per-time-step values across all evaluated Du-IN test segments and spoken words. These weights describe how the downstream readout uses the time-resolved spatial summaries.

Interictal epileptiform-spike detection (MAYO). We analyzed a model finetuned using the combined summary readout for this task (Table 9). Here, we instead examine the backbone encoder's last temporal attention layer. This attention governs how each temporal summary token aggregates content-patch information before that representation reaches the readout for decoding. Similar to the spoken-word decoding procedure above, we calculate attention from the temporal summary tokens to the time patches, but here from the backbone encoder's own last self-attention layer rather than the finetuning readout's cross-attention. We extracted last-layer backbone attention from the seven temporal summary tokens to the 30 content patches (100 ms resolution over a 3-second segment). For each temporal summary token, we averaged post-softmax attention across heads, retained only the content-patch weights, and renormalized them to sum to one; averaging the resulting seven distributions gave one temporal attention profile per segment. With a_{qht} denoting attention from temporal summary token q through head h to content patch t,

$$\bar{a}_{qt} = \frac{1}{H}\sum_{h=1}^{H} a_{qht}, \qquad p_t = \frac{1}{7}\sum_{q=1}^{7} \frac{\bar{a}_{qt}}{\sum_{u=1}^{30} \bar{a}_{qu}}. \tag{9}$$

Heads are averaged before renormalization, and each of the seven temporal summary tokens contributes equally to the final profile p_t, which sums to one across the 30 content patches. Example profiles are plotted alongside their corresponding iEEG signals in Figure 4c. p_t therefore reflects how much attention the temporal summary tokens pay to each time patch when predicting interictal epileptiform spikes in the MAYO dataset.

A.5 Held-out-session reconstruction through the summary-token bottleneck

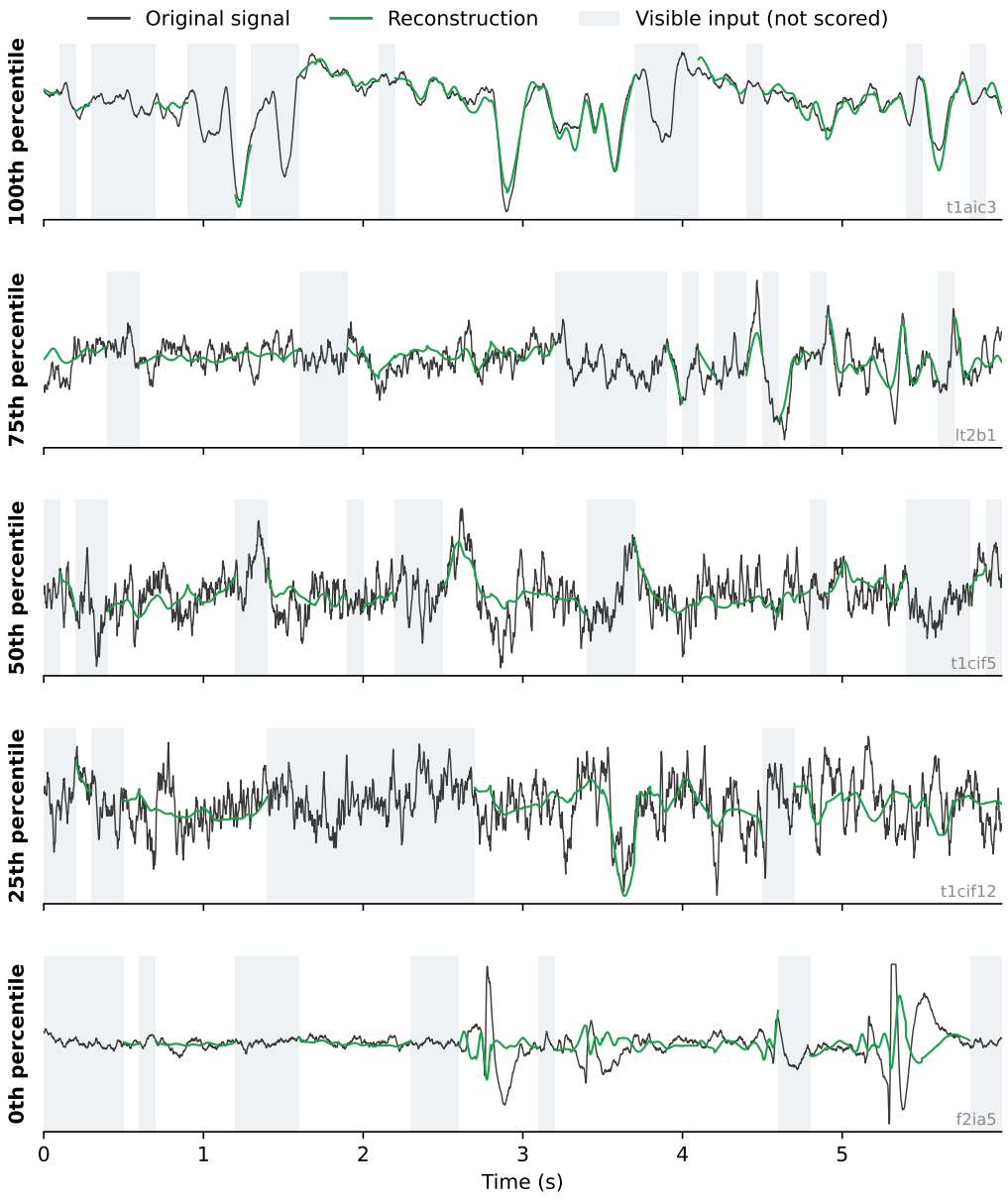

Figure 6: **Masked waveforms reconstructed through the summary-token bottleneck.** We evaluated masked reconstruction on Brain Treebank sessions withheld from pretraining. For each evaluated segment, the encoder processed the visible content patches together with the temporal and spatial summary tokens. The pretrained predictor then reconstructed the masked waveforms using both encoded summary streams, without access to the encoded content tokens. These predictions are compared with the original signals; shaded intervals indicate visible patches supplied to the encoder and excluded from reconstruction.