PHASE: A PHYSIOLOGY-GUIDED HIERARCHICAL FOUNDATION MODEL FOR INTRACRANIAL EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinicians and neuroscientists have long analyzed intracranial electroencephalography (iEEG) through directly measurable physiological characteristics, which carry much of the information that downstream tasks depend on. Recent iEEG foundation models learn by reconstructing or predicting their inputs, which leaves the retention of these characteristics implicit. They are also evaluated mainly on cognitive decoding and a narrow clinical task, i.e., seizure detection. On a broad, clinically relevant benchmark such as Omni-iEEG, they remain below task-specific models when used frozen. We introduce PHASE, a physiology-guided foundation model that makes these characteristics explicit learning targets, pairing them with masked latent prediction in a temporal stage (PHASE-T) within each channel and a spatiotemporal stage (PHASE-ST) across synchronized channels. PHASE is pretrained on heterogeneous recordings from 222 participants at nine clinical sites. On all five Omni-iEEG clinical tasks, frozen PHASE-T outperforms every evaluated foundation model by up to 31%, and fine-tuned PHASE-T surpasses the task-specific models, setting a new state of the art. PHASE-T benefits from physiological supervision, outperforming variants trained with latent prediction alone or auxiliary waveform reconstruction on every task in matched ablations. PHASE-T generalizes to unseen institutions, outperforming the compared models with few or no local labels. PHASE-ST further improves seizure-onset-zone identification over PHASE-T and, when frozen, decodes sound volume and pitch on BrainTreebank better than published models. Beyond task performance, PHASE learns to encapsulate the physiological characteristics clinicians recognize, from seizure onset and its propagation to anatomical region identity, even though its pretraining contains no ictal recordings or anatomical labels.

1 INTRODUCTION

Intracranial electroencephalography (iEEG) provides a direct view of brain activity for both clinical assessment and cognitive neuroscience. In clinical practice, it helps identify seizure-generating tissue and localize cortical regions supporting essential functions such as language and movement (Rosenow & Lüders, 2001; Jayakar et al., 2016; Kambara et al., 2018). In cognitive research, recordings during naturalistic listening reveal neural responses to speech and language (Wang et al., 2024; Zada et al., 2025). Supporting these applications calls for representations that preserve information about both pathological activity and normal brain function across diverse recording settings.

Measurable waveform and spectral properties have long guided the clinical interpretation and cognitive analysis of iEEG. Clinical criteria identify interictal discharges by waveform shape and describe seizure patterns through changes in frequency, amplitude, and morphology (Kane et al., 2017). These properties carry information across temporal scales: brief physiological and pathological high-frequency oscillations (HFOs) exhibit characteristic waveform and spectral features (Zhang et al., 2022a; 2025a; Daida et al., 2026), whereas ongoing rhythms vary with brain state and cortical region (Andrillon et al., 2011; Frauscher et al., 2018). Regional deviations from normative band power can help localize epileptogenic tissue (Taylor et al., 2022). In auditory cortex, high-frequency power indexes local population activity (Lachaux et al., 2012), and its amplitude tracks sentence onsets, intonation, and sound intensity (Hamilton et al., 2018; Tang et al., 2017; Potes et al., 2012). This motivates using directly measurable physiological properties to learn representations that preserve the information needed for diverse downstream iEEG tasks.

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.

Recent iEEG foundation models learn temporal and spectral structure from unlabeled recordings through reconstruction and prediction (Wang et al., 2023; Zhang et al., 2023; Carzaniga et al., 2026). These objectives leave the retention of this physiological information implicit. This information is expressed at several scales: brief events are interpreted within ongoing activity, and each channel in relation to the others (Frauscher et al., 2024). Electrode placement and coverage are also participant-specific (Jayakar et al., 2016), and acquisition practices and recording contexts differ across institutions, so representations are most broadly useful when they depend on neither a fixed channel layout nor electrode coordinates. The challenge is therefore to learn representations that capture these temporal and spatial relationships in heterogeneous recordings without relying on scarce expert labels.

To address these challenges, we introduce **PHASE** (**P**hysiology-Guided **H**ierarchical **A**rchitecture for **S**ignal Understanding in **E**lectrophysiology), a two-stage foundation model that makes measurable electrophysiology an explicit learning target across temporal and spatial scales. **PHASE-T** (PHASE-**T**emporal) pairs masked latent prediction, which predicts an exponential-moving-average (EMA) teacher's embeddings of the unmasked signal at masked positions (Assran et al., 2023), with physiological targets measured directly from the signal. Its fine–coarse–fine hierarchy connects local waveform structure with context from inputs of up to one minute. **PHASE-ST** (PHASE-**S**patio**T**emporal) extends these frozen temporal representations across synchronized channels, adding targets for measured cross-channel relationships, and requires neither electrode coordinates nor fixed channel identities. Our contributions can be summarized as follows:

1. **Physiological supervision leads to more informative representations across temporal and spatial scales.** PHASE turns measurable physiology into learning targets, within each channel and across synchronized channels, while connecting waveform structure to minute-scale context. In matched ablations, PHASE-T's physiological supervision outperforms waveform reconstruction on every task and raises HFO and sleep macro-F1 by 30% and 37% over latent prediction (Table 2).
2. **Diverse pretraining and superior generalization to unseen institutions.** PHASE is pretrained on a diverse cohort: 222 participants from nine sites, spanning clinical monitoring and cognitive tasks. PHASE-T generalizes to institutions it never saw: without local labels, it identifies pathological channels better than supervised models, and with a few labeled participants, exceeds the best published F1 by up to 83% (Table 3).
3. **PHASE achieves strong performance across clinical and cognitive applications.** On all Omni-iEEG clinical tasks (Duan et al., 2026), spanning brief events, brain states, and regional anatomy, PHASE-T outperforms every evaluated foundation model when frozen, with margins averaging 10.4% across the seven scores and up to 31.1% in anatomical localization, and sets a new state of the art when fine-tuned (Table 1), exceeding the performance of task-specific supervised models. In naturalistic listening on BrainTreebank (Wang et al., 2024), frozen PHASE-ST decodes volume and pitch better than published models (Table 4).
4. **PHASE encapsulates known physiological characteristics.** Although PHASE-T is pretrained without ictal recordings, electrode coordinates, or anatomical labels, retrospective analyses show that known physiological characteristics, including spatiotemporal patterns associated with seizure onset and propagation and regional anatomical organization, can be observed in the learned embedding space, with anatomical correspondence extending across institutions (Figure 2). PHASE-ST draws on the synchrony of a participant's concurrently recorded channels to identify the seizure-onset zone (SOZ), with relation targets accounting for most of the gain from concurrent context (Figure 3).

2 Related Work

Pretraining objectives. Pretraining targets determine which signal properties receive explicit supervision. BrainBERT reconstructs masked spectrograms (Wang et al., 2023), and Brant reconstructs waveform patches from temporal and band-power inputs (Zhang et al., 2023). MVPFormer instead contrastively predicts future wavelet embeddings (Carzaniga et al., 2026). Beyond the prediction objective, preprocessing determines which aspects of signal scale remain available for reconstruction. BrainBERT z-scores spectrograms within each frequency bin, emphasizing relative temporal variation rather than absolute spectral magnitude. Reconstructing these targets therefore does not require recovering the spectral scale removed by normalization. PHASE-T follows the joint-embedding predictive

architecture (JEPA) approach (Assran et al., 2023) and pairs latent prediction with a compact set of targets measured from the signal: multi-scale wavelet power, amplitude and local variation, and short-window high-frequency power. These targets are standardized using fixed training-set statistics, preserving a common reference across recordings.

Temporal coverage. Learning broadly useful iEEG representations requires both fine temporal detail and extended context. Existing models capture waveform structure and temporal dependencies at different scales: BrainBERT (Wang et al., 2023) and BaRISTA (Oganesian et al., 2025) encode 5-s and 3-s frames, whereas Brant models dependencies between 6-s patches (Zhang et al., 2023). These configurations resolve either fine structure within short windows or longer dependencies between coarse patches, but not both across a full minute. PHASE combines these two scales through a fine–coarse–fine hierarchy that produces dense, fine-resolution features over 60-s inputs.

Channel organization. Temporal learning and channel composition can be coupled within a model or separated into pretrained stages. Brant applies temporal and spatial attention sequentially (Zhang et al., 2023). BaRISTA structures encoding and masking by anatomical scales (Oganesian et al., 2025); NeuroCLUS learns functional channel groupings (Zheng & Wang, 2026). Closest to our design, PopT composes frozen temporal features with a coordinate-aware population Transformer (Chau et al., 2025). PHASE-ST also composes frozen temporal features, but needs no electrode coordinates and supervises the composition with measured cross-channel relationships (phase-locking value, envelope and broadband correlation, and signed lag) in addition to latent prediction.

Clinical and cognitive evaluation. Evidence for a shared backbone must span different prediction scales, functional demands, and recording populations. Task-specific models illustrate these varied requirements through HFO morphology (Zhang et al., 2025a), fast-ripple learning from detector agreement (Zhang et al., 2025b), channel and regional interactions (Chen et al., 2022), and speech decoding from electrocorticography (ECoG; Makin et al., 2020). Prior iEEG foundation models demonstrate complementary capabilities: BrainBERT is evaluated on auditory decoding (Wang et al., 2023), whereas Brant is evaluated on forecasting, imputation, and seizure detection (Zhang et al., 2023). MVPFormer combines auditory decoding with clinical transfer, but its clinical evaluation centers on seizure detection (Carzaniga et al., 2026). These evaluations do not establish whether a single frozen backbone supports the full range of prediction levels from transient-event classification through brain-state and channel characterization to participant-level outcomes, together with cross-institutional transfer. We evaluate one pretrained PHASE-T across these levels on Omni-iEEG, together with cross-institutional transfer and auditory decoding (Section 4.2).

3 METHOD

PHASE is pretrained in two stages (Figure 1). PHASE-T first encodes each channel independently, connecting local waveform structure with context spanning up to one minute. PHASE-ST then uses synchronized channels to update these representations while keeping the temporal encoder frozen. Both stages pair latent prediction with explicit targets for physiological characteristics measured from individual waveforms; PHASE-ST adds targets for measured relationships between channels.

3.1 A SHARED TEMPORAL REPRESENTATION ACROSS SCALES

PHASE-T brings brief waveform structure and sustained neural context into one dense temporal representation. Sharing the encoder across channels establishes a common feature space without electrode coordinates or fixed channel identities. The fine–coarse–fine hierarchy connects these temporal scales (Figure 1a). For the waveform x_c of channel c, the encoder produces $Z_c = f_\theta(x_c) \in \mathbb{R}^{N\times d}$ on a 32-ms grid, where N is the number of temporal tokens and $d = 768$ is the feature dimension. Shifted local attention encodes overlapping waveform patches into fine-scale features H_{fine}, while global attention over a pooled, coarser sequence spans inputs of up to one minute, yielding coarse contextual features H_{coarse}. The refinement decoder combines upsampled context with local features through a skip connection:

$$Z_c = \text{Decoder}(\text{Upsample}(H_{\text{coarse}}) + \text{Proj}(H_{\text{fine}})) \in \mathbb{R}^{N\times d}, \tag{1}$$

where Upsample learns to map coarse features back to the original token grid, Proj linearly projects local features to width d, and Decoder refines their sum. The final states serve as time-resolved features, and their temporal mean provides a summary for segment-level prediction.

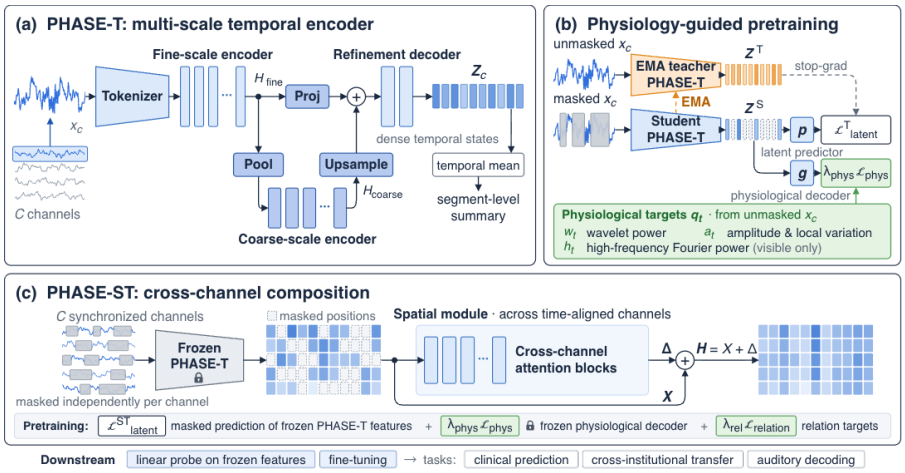

Figure 1: **PHASE separates shared temporal encoding from cross-channel composition.** **a**, PHASE-T, the temporal encoder shared across channels, whose refinement decoder combines fine- and coarse-scale features into temporal states Z_c. **b**, Pretraining with masked latent prediction and physiological targets. **c**, PHASE-ST, which updates frozen PHASE-T features across synchronized channels; rows are channels and columns are time steps. The bottom row shows downstream use.

3.2 Grounding Latent Prediction in Electrophysiology

PHASE-T combines two complementary objectives on the same dense temporal states (Figure 1b). Latent prediction learns dependencies across masked spans, while physiological supervision anchors these states to properties measured from the signal. For latent prediction, we follow JEPA (Assran et al., 2023): a student processes temporally masked inputs, and a predictor estimates the states of its unmasked EMA teacher. Mean squared error between feature-normalized predictions and teacher targets at masked positions defines $\mathcal{L}^{\mathrm{T}}_{\mathrm{latent}}$.

A physiological decoder supervises the same student states with electrophysiological properties measured directly from the signal. It predicts $q_t = [w_t; a_t; h_t]$ at each token position t, where w_t contains multi-scale wavelet power, a_t amplitude and local-variation measures, and h_t short-window high-frequency Fourier power. These targets are measured from unmasked waveforms and standardized using fixed training statistics shared across recordings. Wavelet and amplitude targets supervise masked prediction positions and fully visible positions. Local Fourier targets supervise fully visible patches only. Frequency-dependent targets contribute only within the recording's effective bandwidth. The physiological loss $\mathcal{L}_{\mathrm{phys}}$ averages Smooth L1 errors within each target group and then equally across active groups. With λ_{phys} weighting the physiological term, the student encoder, latent predictor, and physiological decoder are jointly optimized with

$$\mathcal{L}_{\mathrm{temporal}} = \mathcal{L}^{\mathrm{T}}_{\mathrm{latent}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}. \tag{2}$$

3.3 Composing Temporal Features across Channels

PHASE-ST represents each channel in the context of the participant's simultaneously recorded channels, without electrode coordinates and for montages of any size. It takes the frozen PHASE-T features of C synchronized channels, $X \in \mathbb{R}^{C\times N\times d}$, and applies self-attention across channels independently at each time position. Its Transformer blocks produce a residual update Δ in the temporal feature space, giving spatial states $H = X + \Delta$ (Figure 1c). Parameters are shared across channels and no channel identities or coordinates enter the model, so H is permutation equivariant and one set of parameters serves any number of channels.

PHASE-ST is pretrained with three objectives: latent prediction and physiological supervision on the spatial states, and relation supervision of channel pairs:

$$\mathcal{L}_{\mathrm{spatial}} = \mathcal{L}^{\mathrm{ST}}_{\mathrm{latent}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{relation}}. \tag{3}$$

For latent prediction, waveform spans are masked independently in each channel, and PHASE-ST is trained to predict the clean, frozen PHASE-T features at the masked positions, so the activity missing from one channel must be inferred from its own context and its synchronized peers. Physiological supervision applies PHASE-T's frozen physiological decoder to the spatial states with the targets defined above, so the states stay anchored to the signal properties, including at masked positions.

Relation supervision extends physiological guidance from single channels to their interactions. From the clean waveforms, we measure phase-locking value (Lachaux et al., 1999), envelope correlation (Bruns et al., 2000), broadband correlation, and signed lag between channel pairs, which characterize how activity couples and propagates across channels; relation heads predict them from temporal averages of the spatial states. Predicted phase locking and correlations do not depend on channel order, while lag changes sign when the channels are exchanged. PHASE-T stays frozen: only the spatial module, latent predictor, and relation heads are trained. Implementation details for both stages are given in Appendices B and C.

4 Experimental Setup

4.1 Pretraining Data and Training Stages

The pretraining corpus spans two recording contexts: clinical monitoring from Omni-iEEG (Duan et al., 2026) and Hatano Detroit (Hatano et al., 2026), and naturalistic listening from NYU Podcast ECoG (Zada et al., 2025) and BrainTreebank (Wang et al., 2024). Within each source, sampling is balanced across participants. No downstream evaluation data enter pretraining: Omni-iEEG test participants, the Tohoku, NCNP, MAYO, and FNUSA cohorts, and BrainTreebank's validation and test recordings are all excluded.

Both pretraining stages use 60-second inputs at 1 kHz. PHASE-T uses a 106.7M-parameter encoder and is evaluated after 60,000 pretraining updates (global batch of 192 crops; 2.8 GPU-days on four NVIDIA PRO 6000 GPUs). We then freeze PHASE-T and pretrain PHASE-ST on synchronized 64-channel sets from the same corpus. PHASE-ST adds a spatial module with 3.0M trainable parameters and is evaluated after 5,000 pretraining updates (global batch of 128 sets; approximately 1.9 GPU-days on two NVIDIA PRO 6000 GPUs). Source composition, data partitions, and full pretraining settings are detailed in Appendices A–C.

4.2 Downstream Evaluation

We evaluate PHASE-T on the five clinical tasks of Omni-iEEG (Duan et al., 2026), following its official task definitions, splits, evaluation units, and metrics. The five tasks, with the abbreviations used in Table 1, are pathological HFO classification (HFO), ictal period classification (Ictal), sleep–awake classification (Sleep), anatomical localization into five lobes (Lobe-5) or twelve regions (Region-12), and pathological brain region identification. Omni-iEEG evaluates this task using two criteria that a model should meet together (Duan et al., 2026): distinguishing SOZ channels from non-resected channels of seizure-free participants (Channel), and predicting post-operative seizure freedom from the resection ratio of the model's scores (Outcome). This ratio is the share of a participant's total pathology score that falls in resected channels. We compare frozen PHASE-T with the released foundation models BrainBERT (Wang et al., 2023), BaRISTA (Oganesian et al., 2025), and MVPFormer (Carzaniga et al., 2026), using linear probes. We compare fine-tuned PHASE-T with the benchmark's published models PyHFO-Omni (Zhang et al., 2024), CLAP (Wu et al., 2023b), TimeConv-CNN (Duan et al., 2026), and SEEG-NET (Wang et al., 2022).

We assess cross-institutional transfer by applying the Omni-iEEG-trained pathological-channel probe unchanged to Hatano's Tohoku and NCNP cohorts (Hatano et al., 2026). Following Omni-iEEG, we train CLAP and SEEG-NET on the same Omni-iEEG task and apply them without adaptation. We further evaluate cross-participant pathology classification at MAYO and FNUSA (Nejedly et al., 2020), following MVPFormer's data split: we fine-tune PHASE-T on four participants at each institution and test it on the remaining participants. We compare its scores with the MVPFormer and NeuroCLUS results reported by Zheng & Wang (2026).

For multichannel auditory decoding on BrainTreebank (Wang et al., 2024), we evaluate both frozen and fine-tuned PHASE-ST against the published results of PopT (with BrainBERT fea-

Table 1: **One representation across five clinical tasks, from HFO events to surgical outcome.** Results are on the official Omni-iEEG test participants. Frozen PHASE-T and the three foundation models are evaluated with linear probes; fine-tuned PHASE-T is compared with the specialist models published by Omni-iEEG. $C{=}1$ marks multichannel models run on one channel at a time. F1 is macro-F1 and AUC is ROC-AUC. Dashes denote tasks that a specialist model does not support. Bold marks the highest performance per column, and underline the highest among frozen encoders.

	Event	Window (60 s)		Anatomy		Pathol. brain region	
Model	HFO F1	Ictal F1	Sleep F1	Lobe-5 F1	Region-12 F1	Channel AUC	Outcome AUC
PyHFO-Omni	0.8061	–	–	–	–	0.7351	0.7438
CLAP	–	0.9245	0.7225	0.4750	0.3540	0.7684	0.6770
TimeConv-CNN	–	0.8533	0.7118	0.4788	0.3087	0.8061	0.7380
SEEG-NET	–	0.7526	0.6773	0.2520	0.1081	0.7850	0.5952
BrainBERT	0.7937	0.8323	0.6941	0.4054	0.2592	0.8055	0.6017
BaRISTA ($C{=}1$)	0.6050	0.7428	0.7213	0.3006	0.1658	0.6949	0.5739
MVPFormer ($C{=}1$)	0.4347	0.8311	0.5404	0.2423	0.1069	0.6604	0.5878
PHASE-T, frozen	<u>0.8138</u>	<u>0.8958</u>	<u>0.7413</u>	<u>0.4822</u>	<u>0.3398</u>	<u>0.8394</u>	<u>0.6369</u>
PHASE-T, fine-tuned	**0.8188**	**0.9327**	**0.8150**	**0.5134**	**0.3572**	**0.8502**	**0.7713**

tures), BaRISTA, MVPFormer, and NeuroCLUS, following BaRISTA's recording split and PopT's evaluation protocol (Chau et al., 2025). Literature results are reported at their published precision. Implementation details of downstream evaluation are provided in Appendix D.

5 Results

5.1 Clinical Reach across Temporal Scales and Institutions

On the Omni-iEEG test set, PHASE-T handles all five clinical tasks with linear probes on the same frozen model, from HFO events lasting tens of milliseconds to ictal periods and sleep–awake states in one-minute windows, and the anatomical origin of a channel (Table 1). In the frozen single-channel comparison, it leads three foundation-model baselines on every task, by the widest margins on Region-12, Lobe-5, and Ictal, which depend on regional physiology and minute-scale context. Its Region-12 macro-F1 is 31% higher than that of BrainBERT, the strongest of the three. Frozen PHASE-T outperforms specialist models on HFO, Sleep, Lobe-5, and Channel, although each is trained for particular tasks with fixed input formats and prediction targets. Fine-tuned PHASE-T sets a new state of the art on all tasks, leading on both Omni-iEEG's pathology criteria: its Outcome AUC of 0.7713 exceeds that of PyHFO-Omni, an event-based model built on expert HFO annotations.

Physiological supervision and fine-scale feature fusion improve downstream performance. We compare full PHASE-T with three pretraining variants (Table 2): Latent-only drops the physiological targets, Latent+Recon replaces them with Brant-style reconstruction (Zhang et al., 2023) of each masked token's clean 128-sample waveform patch, and No-skip removes the fine-stage skip connection. All four models are frozen after 60,000 pretraining updates and probed with the same protocol. Over Latent-only, physiological supervision adds the most macro-F1 on brief events (HFO, 0.63 to 0.81), sleep (0.54 to 0.74), and anatomical localization (Region-12, 0.20 to 0.34). Latent+Recon leaves HFO and Region-12 at the Latent-only level, and full PHASE-T outperforms it on all five tasks, so physiological supervision is a more useful pretraining signal for these tasks than sample-level waveform reconstruction. The fine-stage skip connection adds 0.15 macro-F1 on sleep and 0.13 on ictal over No-skip, showing that fine-scale features complement the coarse pathway's context.

PHASE-T transfers to institutions held out from pretraining. Generalization to new institutions is important for the clinical utility of pathology models, as recording setups differ and expert labels are scarce. Fitted only on Omni-iEEG, the PHASE-T probe identifies pathological channels at Tohoku and NCNP better than every supervised model trained on the same data, most clearly at NCNP (ROC-AUC 0.6292 against 0.5528; Table 3). For cross-participant pathology classification, fine-tuned

Table 2: **Ablation Studies.** Full PHASE-T is the unablated model, and each variant changes one component. All rows are evaluated under the same pretraining and linear probe conditions; the Full PHASE-T row repeats the frozen scores of Table 1. F1 is macro-F1 and AUC is ROC-AUC. Bold marks the highest performance per column.

	Event	Window (60 s)		Anatomy		Pathol. brain region	
Pretraining variant	HFO F1	Ictal F1	Sleep F1	Lobe-5 F1	Region-12 F1	Channel AUC	Outcome AUC
Full PHASE-T	**0.8138**	**0.8958**	**0.7413**	**0.4822**	**0.3398**	**0.8394**	**0.6369**
Latent-only	0.6276	0.8597	0.5398	0.3271	0.1996	0.7570	0.6285
Latent+Recon	0.6298	0.8532	0.6405	0.3126	0.1954	0.7745	0.6156
No-skip	0.7093	0.7684	0.5904	0.3640	0.2371	0.7415	0.6354

Table 3: **Cross-institution generalization with no local labels or labels from four participants.** **(a)** Pathological-channel identification at Tohoku and NCNP: the PHASE-T probe and supervised comparators are trained on Omni-iEEG and applied unchanged. **(b)** Pathology classification at MAYO and FNUSA after fine-tuning on four local participants; comparator values are published results. Bold marks the highest performance per column.

(a) Cross-institution transfer (ROC-AUC)

Model	Tohoku	NCNP
CLAP	0.7663	0.5021
SEEG-NET	0.7636	0.5528
PHASE-T, frozen	**0.7775**	**0.6292**

(b) Cross-participant fine-tuning (binary F1)

Model	MAYO	FNUSA
MVPFormer	0.36	0.46
NeuroCLUS	0.40	0.51
PHASE-T, fine-tuned	**0.7337**	**0.6656**

PHASE-T nearly doubles the best published F1 at MAYO (0.7337 against 0.40; Zheng & Wang 2026) and exceeds it at FNUSA.

5.2 Label-Free Seizure Dynamics and Regional Identity

Task scores measure how useful PHASE-T's embeddings are, not what they encode. We therefore analyze the frozen embeddings directly, without updating the encoder or fitting a linear probe, and find that they encode two kinds of structure PHASE-T was never explicitly taught: how seizures unfold, and what distinguishes brain regions across institutions (Figure 2).

PHASE-T captures how seizures unfold across space and time. We analyze 200 seizures from 53 participants recorded at the Hospital of the University of Pennsylvania (HUP; Bernabei et al., 2023) within Omni-iEEG, using the release's annotations of ictal period and SOZ channels. For each channel, we measure the distance of its frozen embedding from its own pre-ictal baseline (up to 90 s ending 10 s before onset) in a space of up to 32 whitened principal components, and express this distance as a robust z-score against the baseline; a channel departs when it first stays above $z = 5$ for five consecutive seconds. When an ictal period begins, channels in the clinically marked seizure-onset zone move farther from baseline and peak earlier than the rest (Figure 2b), although PHASE-T's pretraining data contain no ictal recordings. PHASE-T encodes each channel independently and receives no electrode positions, yet departure times derived from its embeddings are ordered along the array: channels farther from the first departing channel depart later (Figure 2a; median Spearman correlation 0.60 over 706 stereo-EEG shaft–seizure pairs at HUP and 0.66 over 75 ECoG strip–seizure pairs from three other institutions (Li et al., 2021); 0.34 and 0.39 with departure times shuffled within each array). PHASE-T captures the distinct dynamics of SOZ channels and how seizures unfold along the electrode array, reflecting the spatiotemporal organization of human seizures (Martinet et al., 2017).

PHASE-T learns a map of the cortex that holds across institutions. Averaged by brain region, PHASE-T's frozen embeddings of 16,444 ECoG channels in the Omni-iEEG anatomy recordings line up across institutions: all eight Region-12 regions we compare at UCLA (32 participants) are

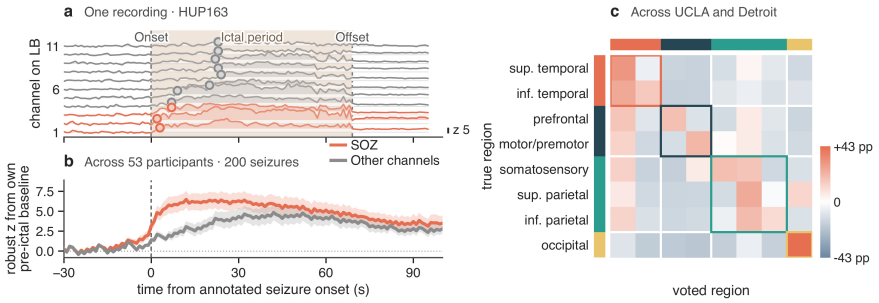

Figure 2: **Frozen PHASE-T embeddings carry seizure dynamics and regional identity across institutions. a**, Embedding distance from the pre-ictal baseline for each channel on one stereo-EEG shaft before, during, and after one ictal period; orange traces are SOZ channels, and circles mark departures. **b**, The same distance for SOZ and other channels: the per-seizure median over channels, averaged over each participant's seizures and then over participants; shading shows 95% participant-bootstrap confidence intervals. **c**, The proportion of each true region's channels (rows) voted into each region (columns), pooled over both directions between UCLA and Detroit and shown relative to chance (red, above; blue, below). Each channel is voted on by its ten most similar participants at the other institution; boxes group regions by lobe.

most similar to their own counterparts at Detroit (135 participants), and six of the eight Detroit regions to theirs at UCLA. These 14 matches out of 16 are far above chance: about seven times the 2 expected at random, and a count that none of 2,000 shuffles of the region labels reached (permutation test, $p = 5 \times 10^{-4}$). The different regions whose averages lie closest together are neighbors in the brain: motor and somatosensory cortex, parietal and occipital cortex, and the two temporal regions (Appendix D.7). The map reaches single channels: voted on by their ten most similar participants at the other institution, the channels of every region land in their own region more often than chance, most often in their own region or, for inferior temporal and inferior parietal cortex, in the superior region of the same lobe (Figure 2c). PHASE-T's lead in anatomical localization (Table 1) thus reflects a map of the cortex that its frozen embeddings carry consistently across institutions.

5.3 Context from the Concurrent Montage

PHASE-ST interprets each channel in the context of the other channels recorded at the same time. In naturalistic listening on BrainTreebank, its frozen features decode volume and pitch better than published models (Table 4). In clinical recordings, it draws on the synchrony of a participant's concurrently recorded channels to identify SOZ channels more accurately than PHASE-T; relation targets account for most of the gain from this concurrent context (Figure 3).

PHASE-ST decodes acoustic properties during naturalistic listening. We evaluate PHASE-ST on four auditory decoding tasks across seven BrainTreebank participants (Table 4). Frozen PHASE-ST reaches mean ROC-AUCs of 0.9450 for volume and 0.8822 for pitch, above the strongest published results of 0.92 and 0.83 (Zheng & Wang, 2026). Fine-tuning raises the scores on all four tasks (volume to 0.9515, pitch to 0.8886).

PHASE-ST draws on synchrony with concurrently recorded channels to identify the SOZ. With the participant's own channels from the same minute, in panels of up to 64, PHASE-ST reaches a pooled ROC-AUC of 0.862, compared with 0.843 when each channel is processed alone and 0.841 for PHASE-T on the same channels. The added context yields a mean per-participant gain of $+0.048$ across 52 held-out Omni-iEEG participants (Figure 3a). This advantage depends on synchrony with the target channel. Taking the same companion channels from another minute preserves the participant and recording setup and keeps the companions mutually synchronized, while breaking their alignment with the unchanged target waveform. With linear probes fitted to their respective conditions, the mean gain falls to $+0.015$, significantly below the same-minute gain (two-sided sign test, $p = 0.003$). Replacing the companions with another participant's channels instead gives a mean per-participant change of -0.019 under the fixed same-minute probe.

Table 4: **Auditory decoding on BrainTreebank.** Results are mean test ROC-AUC across the seven evaluation participants. Literature rows are published results. Bold marks the highest and underline the second-highest performance per column.

Method	Sentence onset	Speech	Volume	Pitch
PopT + BrainBERT	0.90	0.93	0.87	0.74
BaRISTA	0.91	0.92	0.85	0.73
MVPFormer	0.87	0.90	0.88	0.83
NeuroCLUS	**0.96**	**0.99**	0.92	0.83
PHASE-ST, frozen	0.9176	0.9703	<u>0.9450</u>	<u>0.8822</u>
PHASE-ST, fine-tuned	<u>0.9245</u>	<u>0.9786</u>	**0.9515**	**0.8886**

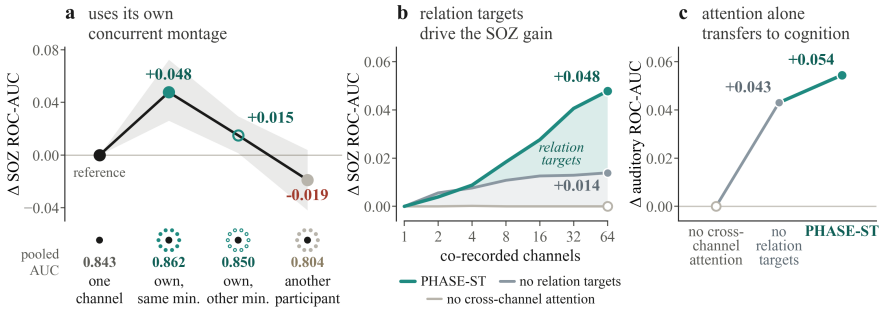

Figure 3: **PHASE-ST uses a participant's concurrent montage, learns this from relation targets, and carries context into auditory decoding. a**, Per-participant mean change in SOZ ROC-AUC relative to scoring the target channel alone, for each set of companion channels; shading shows the 95% participant-bootstrap confidence interval. Glyphs depict the target channel (black) and its companions, with each condition's pooled ROC-AUC below. **b**, The same change as a function of the number of co-recorded channels. **c**, Change in auditory ROC-AUC on BrainTreebank relative to the stage without cross-channel attention, averaged over four tasks and seven participants.

Relation targets drive the SOZ context gain, and cross-channel attention carries context into auditory decoding. We compare PHASE-ST with two variants trained identically except for one component: one without relation targets and one without cross-channel attention. PHASE-ST's SOZ context gain grows with the number of co-recorded channels to $+0.048$ (95% CI $[+0.026, +0.071]$) at the largest panels of up to 64 channels; without relation targets it reaches only $+0.014$, so $+0.034$ of the gain comes from the relation targets, and without cross-channel attention it stays at zero by construction (Figure 3b). On BrainTreebank, PHASE-ST improves decoding over the stage without cross-channel attention for all seven participants, by a mean $+0.054$ ROC-AUC. The variant without relation targets has a mean gain of $+0.043$ over the same stage (Figure 3c). Appendix E details the context probe and spatial-stage ablations.

6 Conclusion and Limitations

PHASE uses physiology-guided pretraining to learn temporal and cross-channel representations of iEEG. PHASE-T leads the evaluated foundation models on all five Omni-iEEG clinical tasks with frozen representations and sets a new state of the art after fine-tuning, while PHASE-ST achieves the best volume and pitch decoding among compared methods. PHASE-T's embeddings also capture seizure dynamics and consistent regional organization across institutions. Together, these results support PHASE as a physiology-guided foundation for clinical and cognitive iEEG analysis. The framework uses separate temporal and spatial pretraining stages and context limited to one minute. Future work will explore joint pretraining, extend context toward hour-long recordings, and expand both patient cohorts and downstream evaluations to assess broader generalization (Appendix F).

AI USE STATEMENT

In this work, we used generative AI tools for [design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis]. We have not used generative AI tools for [Generate synthetic data sets, propose or refine hypotheses, interpret results], and [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. The authors reviewed and revised all AI-assisted text, checking technical descriptions and interpretations against the methods and experimental results. AI-assisted code was reviewed and tested by the authors before use in the reported experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing open-sourced iEEG datasets and involves no new participant recruitment or data collection. PHASE is intended for research; its clinical use would require further validation.

REPRODUCIBILITY STATEMENT

We document the data sources, pretraining and evaluation partitions, and participant sampling in Appendix A.1. Signal preprocessing, model architectures, target construction, and optimization settings are provided in Appendix B, with preprocessing details in Appendix A.2. Downstream evaluation protocols, including readout training, fine-tuning, and prediction aggregation, are specified in Appendix D. PHASE implementation and evaluation details are provided in Appendix C. We will release the implementation and pretrained checkpoints upon publication.

REFERENCES

Thomas Andrillon, Yuval Nir, Richard J Staba, Fabio Ferrarelli, Chiara Cirelli, Giulio Tononi, and Itzhak Fried. Sleep spindles in humans: insights from intracranial eeg and unit recordings. *Journal of Neuroscience*, 31(49):17821–17834, 2011.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

John M Bernabei, Adam Li, Andrew Y Revell, Rachel J Smith, Kristin M Gunnarsdottir, Ian Z Ong, Kathryn A Davis, Nishant Sinha, Sridevi Sarma, and Brian Litt. Quantitative approaches to guide epilepsy surgery from intracranial EEG. *Brain*, 146(6):2248–2258, 2023. doi: 10.1093/brain/awad007.

Andreas Bruns, Reinhard Eckhorn, Hennric Jokeit, and Alois Ebner. Amplitude envelope correlation detects coupling among incoherent brain signals. *NeuroReport*, 11(7):1509–1514, 2000. doi: 10.1097/00001756-200005150-00029.

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.

Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. BrainNet: Epileptic wave detection from SEEG with hierarchical graph diffusion learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2741–2751, 2022. doi: 10.1145/3534678.3539178.

Atsuro Daida, Sotaro Kanai, Yipeng Zhang, Chenda Duan, Naoto Kuroda, Tonmoy Monsoor, Yuanyi Ding, Kikuko Kaneko, Joe Qiao, Shaun A. Hussain, Aria Fallah, Noriko Salamon, Raman Sankar, Richard J. Staba, Jerome Engel, William Speier, Eishi Asano, Vwani Roychowdhury, and Hiroki Nariai. Developmental profile of physiological high-frequency oscillations in the human brain. *NeuroImage*, 336:122017, 2026. doi: 10.1016/j.neuroimage.2026.122017.

Yuanyi Ding, Yipeng Zhang, Chenda Duan, Atsuro Daida, Yun Zhang, Sotaro Kanai, Mingjian Lu, Shaun Hussain, Richard J Staba, Hiroki Nariai, and Vwani Roychowdhury. PyHFO 2.0: an open-source platform for deep learning–based clinical high-frequency oscillations analysis. *Journal of Neural Engineering*, 22(5):056040, 2025. doi: 10.1088/1741-2552/ae10e0.

Chenda Duan, Yipeng Zhang, Sotaro Kanai, Yuanyi Ding, Atsuro Daida, Pengyue Yu, Tiancheng Zheng, Naoto Kuroda, Shaun A. Hussain, Eishi Asano, Hiroki Nariai, and Vwani Roychowdhury. Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rv9lQpY5cG.

Rosana Esteller, Javier Echauz, Tung Tcheng, Brian Litt, and Benjamin Pless. Line length: an efficient feature for seizure onset detection. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pp. 1707–1710, 2001.

B Frauscher, D Mansilla, C Abdallah, A Astner-Rohracher, S Beniczky, Milan Brázdil, V Gnatkovsky, J Jacobs, G Kalamangalam, P Perucca, et al. Learn how to interpret and use intracranial eeg findings. *Epileptic Disorders*, 26(1):1–59, 2024. doi: 10.1002/epd2.20190.

Birgit Frauscher, Nicolas von Ellenrieder, Rina Zelmann, Irena Doležalová, Lorella Minotti, André Olivier, Jeffery Hall, Dominique Hoffmann, Dang Khoa Nguyen, Philippe Kahane, François Dubeau, and Jean Gotman. Atlas of the normal intracranial electroencephalogram: Neurophysiological awake activity in different cortical areas. *Brain*, 141(4):1130–1144, 2018. doi: 10.1093/brain/awy035.

Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, 28(12):1860–1871.e4, 2018. doi: 10.1016/j.cub.2018.04.033.

Keisuke Hatano, Naoto Kuroda, Hiroshi Uda, Kazuki Sakakura, Michael J. Cools, Aimee F. Luat, Shin-Ichiro Osawa, Hitoshi Nemoto, Kazushi Ukishiro, Hidenori Endo, Nobukazu Nakasato, Yutaro Takayama, Keiya Iijima, Masaki Iwasaki, and Eishi Asano. Internal and external validation of comprehensive high-frequency activity biomarkers for epilepsy surgery. *Clinical Neurophysiology*, 190:2111981, 2026. doi: 10.1016/j.clinph.2026.2111981.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Zhentao Huang, Yahong Ma, Rongrong Wang, Weisu Li, and Yongsheng Dai. A model for EEG-based emotion recognition: CNN-Bi-LSTM with attention mechanism. *Electronics*, 12(14):3188, 2023. doi: 10.3390/electronics12143188.

Prasanna Jayakar, Jean Gotman, A Simon Harvey, André Palmini, Laura Tassi, Donald Schomer, Francois Dubeau, Fabrice Bartolomei, Alice Yu, Pavel Kršek, et al. Diagnostic utility of invasive eeg for epilepsy surgery: indications, modalities, and techniques. *Epilepsia*, 57(11):1735–1747, 2016.

Toshimune Kambara, Sandeep Sood, Zahraa Alqatan, Christine Klingert, Diksha Ratnam, Akane Hayakawa, Yasuo Nakai, Aimee F Luat, Rajkumar Agarwal, Robert Rothermel, et al. Presurgical language mapping using event-related high-gamma activity: The detroit procedure. *Clinical Neurophysiology*, 129(1):145–154, 2018.

Nick Kane, Jayant Acharya, Sandor Beniczky, Luis Caboclo, Simon Finnigan, Peter W Kaplan, Hiroshi Shibasaki, Ronit Pressler, and Michel JAM Van Putten. A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the eeg findings. revision 2017. *Clinical neurophysiology practice*, 2:170–185, 2017. doi: 10.1016/j.cnp.2017.07.002.

Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J. Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208, 1999.

Jean-Philippe Lachaux, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E. Crone. High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, 98(3):279–301, 2012. doi: 10.1016/j.pneurobio.2012.06.008.

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

Adam Li, Chester Huynh, Zachary Fitzgerald, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel O. Claudio, Andres M. Kanner, Jennifer Hopp, Stephanie Chen, et al. Neural fragility as an EEG marker of the seizure onset zone. *Nature Neuroscience*, 24(10):1465–1474, 2021. doi: 10.1038/s41593-021-00901-w.

Joseph G. Makin, David A. Moses, and Edward F. Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582, 2020. doi: 10.1038/s41593-020-0608-8.

Louis-Emmanuel Martinet, Grant Fiddyment, Joseph R Madsen, Emad N Eskandar, Wilson Truccolo, Uri T Eden, Sydney S Cash, and Mark A Kramer. Human seizures couple across spatial scales through travelling wave dynamics. *Nature communications*, 8(1):14896, 2017. doi: 10.1038/ncomms14896.

Tonmoy Monsoor, Yipeng Zhang, Atsuro Daida, Shingo Oana, Qiujing Lu, Shaun A. Hussain, Aria Fallah, Raman Sankar, Richard J. Staba, William Speier, Vwani Roychowdhury, and Hiroki Nariai. Optimizing detection and deep learning-based classification of pathological high-frequency oscillations in epilepsy. *Clinical Neurophysiology*, 154:129–140, 2023. doi: 10.1016/j.clinph.2023.07.012.

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, 2020. doi: 10.1038/s41597-020-0532-5.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Lucine L. Oganesian, Saba Hashemi, and Maryam M. Shanechi. BaRISTA: Brain scale informed spatiotemporal representation of human intracranial neural activity. In *Advances in Neural Information Processing Systems*, 2025.

Cristhian Potes, Aysegul Gunduz, Peter Brunner, and Gerwin Schalk. Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. *NeuroImage*, 61(4):841–848, 2012. doi: 10.1016/j.neuroimage.2012.04.022.

Felix Rosenow and Hans Lüders. Presurgical evaluation of epilepsy. *Brain*, 124(9):1683–1700, 2001. doi: 10.1093/brain/124.9.1683.

Claire Tang, Liberty S Hamilton, and Edward F Chang. Intonational speech prosody encoding in the human auditory cortex. *Science*, 357(6353):797–801, 2017. doi: 10.1126/science.aam8577.

Peter N Taylor, Christoforos A Papasavvas, Thomas W Owen, Gabrielle M Schroeder, Frances E Hutchings, Fahmida A Chowdhury, Beate Diehl, John S Duncan, Andrew W McEvoy, Anna Miserocchi, Jane de Tisi, Sjoerd B Vos, Matthew C Walker, and Yujiang Wang. Normative brain mapping of interictal intracranial EEG to localize epileptogenic tissue. *Brain*, 145(3):939–949, 2022. doi: 10.1093/brain/awab380.

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 Yaari, Aaditya Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph Madsen, Scellig Stone, Gabriel Kreiman, et al. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. *Advances in Neural Information Processing Systems*, 37:96505–96540, 2024.

Yiping Wang, Yanfeng Yang, Gongpeng Cao, Jinjie Guo, Penghu Wei, Tao Feng, Yang Dai, Jinguo Huang, Guixia Kang, and Guoguang Zhao. SEEG-Net: An explainable and deep learning-based cross-subject pathological activity detection method for drug-resistant epilepsy. *Computers in Biology and Medicine*, 148:105703, 2022. doi: 10.1016/j.compbiomed.2022.105703.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Zaid Zada, Samuel A. Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Sasha Devore, Adeen Flinker, Orrin Devinsky, Ariel Goldstein, and Uri Hasson. The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135, 2025. doi: 10.1038/s41597-025-05462-2.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26304–26321, 2023. doi: 10.52202/075280-1144.

Yipeng Zhang, Hoyoung Chung, Jacquline P. Ngo, Tonmoy Monsoor, Shaun A. Hussain, Joyce H. Matsumoto, Patricia D. Walshaw, Aria Fallah, Myung Shin Sim, Eishi Asano, Raman Sankar, Richard J. Staba, Jerome Engel, William Speier, Vwani Roychowdhury, and Hiroki Nariai. Characterizing physiological high-frequency oscillations using deep learning. *Journal of Neural Engineering*, 19(6):066027, 2022a. doi: 10.1088/1741-2552/aca4fa.

Yipeng Zhang, Qiujing Lu, Tonmoy Monsoor, Shaun A. Hussain, Joe X. Qiao, Noriko Salamon, Aria Fallah, Myung Shin Sim, Eishi Asano, Raman Sankar, Richard J. Staba, Jerome Engel, William Speier, Vwani Roychowdhury, and Hiroki Nariai. Refining epileptogenic high-frequency oscillations using deep learning: a reverse engineering approach. *Brain Communications*, 4(1): fcab267, 2022b. doi: 10.1093/braincomms/fcab267.

Yipeng Zhang, Lawrence Liu, Yuanyi Ding, Xin Chen, Tonmoy Monsoor, Atsuro Daida, Shingo Oana, Shaun Hussain, Raman Sankar, Aria Fallah, et al. PyHFO: Lightweight deep learning-powered end-to-end high-frequency oscillations analysis application. *Journal of Neural Engineering*, 21(3): 036023, 2024.

Yipeng Zhang, Atsuro Daida, Lawrence Liu, Naoto Kuroda, Yuanyi Ding, Shingo Oana, Sotaro Kanai, Tonmoy Monsoor, Chenda Duan, Shaun A. Hussain, Joe X. Qiao, Noriko Salamon, Aria Fallah, Myung Shin Sim, Raman Sankar, Richard J. Staba, Jerome Engel, Eishi Asano, Vwani Roychowdhury, and Hiroki Nariai. Self-supervised data-driven approach defines pathological high-frequency oscillations in epilepsy. *Epilepsia*, 66(11):4434–4450, 2025a. doi: 10.1111/epi.18545.

Yipeng Zhang, Yuanyi Ding, Chenda Duan, Atsuro Daida, Hiroki Nariai, and Vwani Roychowdhury. Self-supervised distillation of legacy rule-based methods for enhanced EEG-based decision-making. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15967 of *Lecture Notes in Computer Science*, pp. 477–486. Springer Nature Switzerland, 2025b. doi: 10.1007/978-3-032-04984-1_46.

Hui Zheng and Haiteng Wang. NeuroCLUS: A foundation model with functional clustering for intracranial neural decoding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pFweJM4Uw8.

Maeike Zijlmans, Premysl Jiruska, Rina Zelmann, Frans S. S. Leijten, John G. R. Jefferys, and Jean Gotman. High-frequency oscillations as a new biomarker in epilepsy. *Annals of Neurology*, 71(2): 169–178, 2012. doi: 10.1002/ana.22548.

Appendix: Supplementary Material

CONTENTS OF APPENDIX

A Data and Signal Preparation **16**
A.1 Pretraining Data and Partitions 16
A.2 Data Preparation and Signal Preprocessing 16

B PHASE-T Architecture and Pretraining **17**
B.1 Temporal Architecture 17
B.2 Physiological Target Construction 17
B.3 Masking and Temporal Objectives 18
B.4 Pretraining Settings 19

C PHASE-ST Architecture and Pretraining **20**
C.1 Spatial Residual Architecture 20
C.2 Latent Prediction and Physiological Supervision 21
C.3 Pairwise Relation Targets and Loss 21
C.4 Pretraining and Inference Settings 22

D Evaluation Protocols and Extended Results **23**
D.1 Shared Encoder Settings 23
D.2 Clinical Prediction across Scales 23
D.2.1 Linear Probing and Task-Specific Fine-Tuning 23
D.2.2 Released Foundation Checkpoints and External Adapters 23
D.2.3 Direct Physiological-Target Control 24
D.3 Cross-Institution Transfer 25
D.4 Adaptation from Four Participants 25
D.5 Multichannel Auditory Decoding 26
D.6 Seizure Trajectory Extraction 26
D.7 Cross-Institution Regional Anatomical Correspondence 26

E Spatial Context Analysis and Ablations **27**
E.1 Concurrent-Channel Context Probe 27
E.2 Spatial-Stage Pretraining Ablations 28
E.3 Single-Channel Auditory Probe 28

F Limitations and Future Directions **29**

A Data and Signal Preparation

A.1 Pretraining Data and Partitions

PHASE-T uses 568 recordings from Omni-iEEG (Duan et al., 2026), the Hatano Detroit release (Hatano et al., 2026), NYU Podcast ECoG (Zada et al., 2025), and BrainTreebank (Wang et al., 2024) (Table 5). Sampling is balanced across participants within each source, not by duration.

Pretraining uses the official Omni-iEEG training partition, the Hatano Detroit non-REM recordings, the NYU Podcast ECoG recordings, and BrainTreebank's pretraining split. No downstream evaluation data enter pretraining. Omni-iEEG test participants are excluded from both clinical sources. Pretraining also contains no ictal recordings: all ictal recordings in Omni-iEEG, including every seizure analyzed in Section 5.2, are excluded, and the remaining sources contribute interictal clinical monitoring or naturalistic listening sessions. The Tohoku and NCNP cohorts of the Hatano release and the MAYO and FNUSA cohorts are used only for evaluation. For BrainTreebank, we follow BaRISTA's recording split (Oganesian et al., 2025): 17 recordings for pretraining, 2 for validation, and 7 for testing.

Table 5: **Pretraining sources.** Recs, recordings. Sampling shares are proportional to each source's participant count, and channel-hours are counted before artifact rejection. Omni-iEEG and Hatano share 76 participants, so the corpus includes 222 distinct participants.

Source	Participants	Recs	Channel-hours	Sampling	Recording context	Evaluation use
Omni-iEEG	151	399	5,235	50.67%	Clinical monitoring; official training partition	Test participants and tasks
Hatano	129	144	10,315	43.29%	Clinical monitoring; Detroit non-REM sleep	Tohoku and NCNP cohorts
NYU Podcast	8	8	632	2.68%	Naturalistic listening	Pretraining only
BrainTreebank	10	17	5,434	3.36%	Naturalistic listening; BaRISTA pretraining split	7 test recordings

A.2 Data Preparation and Signal Preprocessing

Recordings are converted into single-channel waveforms in microvolts on a common 1 kHz grid, retaining their native clinical referencing, acquisition bandwidth, and per-sample validity masks.

The deterministic filtering in Table 6 is applied when estimating target statistics, during pretraining, and at inference; the table's bandwidth augmentation is used only during pretraining.

Table 6: **Shared preprocessing for training and inference.**

Parameter	Setting
Waveform units and grid	Microvolts at 1 kHz; released reference convention retained
Pretraining crop	60 seconds (60,000 samples)
High-pass filter	Second-order Butterworth, 2-Hz corner
Mains-notch filters	50 and 60 Hz, each with quality factor $Q = 30$
Deterministic filtering	Zero-phase product of the squared-magnitude filter responses, applied in the frequency domain after reflection padding
Input amplitude transform	$u_t = \operatorname{asinh}(x_t/s_0)$ for filtered waveforms with scale parameter $s_0 = 20\,\mu\mathrm{V}$, preserving absolute signal amplitudes
Bandwidth augmentation	Probability 0.25 for crops with usable acquisition bandwidth of at least 240 Hz
Augmentation low-pass	Cutoff sampled uniformly from 200–220 Hz, followed by a 20-Hz raised-cosine transition; reflection padding of 1 second on each side
Augmented views and targets	Teacher, student, and physiological targets use the same bandwidth-reduced waveform; the sampled cutoff becomes its effective target bandwidth
Frequency eligibility	Wavelet center frequency or Fourier-band upper edge must not exceed the effective acquisition bandwidth

B PHASE-T Architecture and Pretraining

B.1 Temporal Architecture

For 60-second, 1 kHz crops, a convolutional tokenizer embeds overlapping 128-sample patches at a 32-sample stride, with 48 zero-padded samples on each side. GELU, linear projection, and RMS normalization produce $N = 1{,}875$ tokens of width 384, spaced 32 ms apart. Eight fine-scale Transformer blocks use six heads and 256-token local attention windows, shifted by 128 tokens in alternating blocks, yielding $H_{\mathrm{fine}} \in \mathbb{R}^{1,875\times 384}$.

Convolutional pooling with kernel eight and stride four maps the fine sequence to 469 tokens of width 768. For the 1,875-token input, pooling reflect-pads two tokens on the left and three on the right. Ten coarse blocks with 12 heads use full bidirectional attention, yielding $H_{\mathrm{coarse}} \in \mathbb{R}^{469\times 768}$. A learned transposed convolution with the same kernel and stride returns these tokens to the fine grid; the asymmetric padding is cropped away. The result is added to a 384-to-768 projection of H_{fine} and refined by two 12-head blocks with 256-token windows, the second shifted by 128 tokens. This is the Z_c of Equation 1, with $N = 1{,}875$ and $d = 768$.

All Transformer blocks use pre-normalized residual connections, RMS normalization, rotary position embeddings, and SwiGLU feed-forward layers. The SwiGLU hidden dimension is $8/3$ times the block dimension: 1,024 in the fine-scale encoder and 2,048 in the coarse-scale encoder and refinement decoder. The encoder comprises 106.7M parameters. The latent predictor and physiological decoder are two-layer GELU networks on Z_c.

B.2 Physiological Target Construction

At each token position t, we construct an 84-dimensional physiological target comprising wavelet power, amplitude and local variation, and short-window high-frequency power:

$$q_t = [w_t; a_t; h_t] \in \mathbb{R}^{84}, \qquad w_t \in \mathbb{R}^{64}, \quad a_t \in \mathbb{R}^{4}, \quad h_t \in \mathbb{R}^{16}.$$

All targets are computed from unmasked waveforms in microvolts after signal preprocessing and any pretraining bandwidth augmentation, but before the asinh input amplitude transform in Table 6.

Wavelet power. The spectral target w_t describes power at 64 center frequencies: 24 logarithmically spaced frequencies from 2 to 80 Hz and 40 linearly spaced frequencies above 80 up to 450 Hz. We use complex Morlet wavelets, each with a Gaussian frequency response and temporal standard deviation $\sigma(f) = 3/f$. To obtain an analytic response with amplitude-preserving normalization, the frequency-domain filter is set to zero at negative frequencies and multiplied by two at positive frequencies. Reflection padding extends the waveform by 6 seconds on each side, corresponding to four temporal standard deviations of the lowest-frequency wavelet.

For each frequency, the squared magnitude of the wavelet coefficients is averaged within consecutive 32-sample windows to align the power measurements with the token grid. The resulting power P is log-transformed as

$$\widetilde{P} = \log\left(1 + \frac{P}{P_0}\right), \qquad P_0 = (1\,\mu\mathrm{V})^2.$$

Amplitude and local variation. The target a_t contains four waveform descriptors: centered root-mean-square amplitude, mean absolute deviation, peak-to-peak amplitude, and mean absolute successive difference (line length; Esteller et al., 2001). These descriptors are computed over 128-sample patches using the same stride and zero padding as the tokenizer. Each descriptor a is log-transformed as

$$\widetilde{a} = \log\left(1 + \frac{a}{s_0}\right), \qquad s_0 = 20\,\mu\mathrm{V}.$$

Local high-frequency power. The target h_t captures spectral content using the same 128-sample patches. Each patch is multiplied by a symmetric Hann window before Fourier transformation. Squared Fourier magnitudes are normalized by the window energy, and positive-frequency bins other than the Nyquist bin are doubled to obtain a one-sided power spectrum; the DC bin is unchanged. The spectrum is summarized by averaging power across bins within each of 16 bands: eight equal-width bands over 80–250 Hz and eight over 250–450 Hz, in the ripple and fast-ripple ranges (Zijlmans et al., 2012). Each band-power measurement P is log-transformed as $\log(1 + P/s_0^2)$.

Target standardization. Each log-transformed target feature is standardized using a fixed mean and standard deviation estimated from the participant-balanced pretraining mixture. These statistics are computed using the same deterministic signal-filtering pipeline used by the encoder and remain fixed throughout pretraining.

B.3 Masking and Temporal Objectives

Temporal pretraining combines latent prediction at masked positions with physiological supervision at positions selected according to target type and waveform visibility.

Temporal masking. For each microbatch, we sample temporal spans covering 55% of token positions, denoted by $\mathcal{M}$. Each span length is sampled uniformly from one of three integer ranges, with range-selection probabilities specified in Table 7. Spans may overlap, and the final span may be shortened to reach the specified number of prediction positions.

To prevent overlapping patches from exposing masked waveform samples, we first zero all samples covered by the tokenizer patches at positions in $\mathcal{M}$. Every token whose patch overlaps these zeroed samples is replaced with a learned mask embedding. We denote the fully visible positions by $\mathcal{V}$. The positions outside $\mathcal{M} \cup \mathcal{V}$ correspond to partially masked patches adjacent to the prediction spans.

Latent prediction. The masked student and unmasked teacher produce representations Z^{S} and Z^{T}, respectively. Both views use the same deterministic preprocessing and the same bandwidth augmentation as the physiological targets. The teacher is evaluated in inference mode, and its parameters are updated by an exponential moving average of the student.

At each position in $\mathcal{M}$, the latent predictor p_ϕ maps the student representation to the corresponding teacher target. The prediction loss is

$$\mathcal{L}^{\mathrm{T}}_{\mathrm{latent}} = \frac{1}{d|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left\| \mathrm{LN}(p_\phi(Z^{\mathrm{S}}_t)) - \mathrm{sg}[\mathrm{LN}(Z^{\mathrm{T}}_t)] \right\|_2^2, \tag{4}$$

where d is the embedding dimension, LN denotes feature normalization without learned affine parameters, and sg denotes stop-gradient. The teacher update schedule is specified below.

Physiological supervision and target visibility. The physiological decoder g_ψ maps the same student representations to predictions of $q_t = [w_t; a_t; h_t]$. Wavelet-power and amplitude losses are

evaluated at both masked prediction positions and fully visible positions, $t \in \mathcal{M} \cup \mathcal{V}$. The local Fourier-power loss is evaluated only at fully visible positions, $t \in \mathcal{V}$.

Partially masked patches outside $\mathcal{M}$ are excluded from the physiological loss.

Bandwidth-dependent target selection. Frequency-dependent targets contribute to the loss only when supported by the recording's effective bandwidth after any bandwidth augmentation. A wavelet-power target is included when its center frequency lies within this bandwidth; a Fourier-power band is included when its upper edge lies within it. This selection prevents supervision at frequencies unavailable in the original recording or removed by augmentation.

Physiological loss and joint optimization. For each target group $b \in \{w, a, h\}$, let $\mathcal{I}_b$ contain the token–feature pairs satisfying the visibility and bandwidth conditions above, and let $\mathcal{B}$ denote the groups with at least one valid pair. We first average the loss within each group and then average across groups, giving each active target group equal weight:

$$\mathcal{L}_{\text{phys}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{1}{|\mathcal{I}_b|} \sum_{(t,k) \in \mathcal{I}_b} \rho\big(g_\psi(Z_t^{\text{S}})_{b,k} - q_{t,b,k}\big)\,, \tag{5}$$

where ρ is the Smooth L1 loss with transition parameter one. The student encoder, latent predictor, and physiological decoder are optimized jointly using $\mathcal{L}_{\text{latent}}^{\text{T}} + \lambda_{\text{phys}} \mathcal{L}_{\text{phys}}$, as defined in Equation 2.

Ablation variants. We compare PHASE-T with three variants in Table 2, each pretrained with the settings in Table 7 apart from the changes described below. Latent-only sets $\lambda_{\text{phys}} = 0$. Latent+Recon replaces physiological supervision with Brant-style reconstruction (Zhang et al., 2023) of clean 128-sample waveform patches aligned to the token grid. Targets use the preprocessed, bandwidth-augmented waveform before masking and the input amplitude transform, standardized by a fixed scalar mean and standard deviation estimated from the training samples. A two-layer GELU decoder predicts each patch from its student token. As in masked autoencoders (He et al., 2022), mean squared error is computed only at the requested masked positions $\mathcal{M}$, averaged over in-crop samples, and added to $\mathcal{L}_{\text{latent}}^{\text{T}}$ with unit weights on both losses. Both physiological and reconstruction targets are standardized using fixed training-set statistics, and both auxiliary losses enter the latent-prediction objective with weight one. The comparison shares the encoder architecture, pretraining settings, and downstream probing protocol, while each auxiliary objective follows its specified target definition, supervision locations, and loss function. No-skip removes $\text{Proj}(H_{\text{fine}})$ from Equation 1, leaving the refinement decoder to receive only upsampled coarse features.

B.4 Pretraining Settings

We optimize the student encoder and prediction heads with AdamW, applying weight decay to matrix weights while excluding biases, normalization gains, and the learned mask embedding.

After each successful optimizer update s, the teacher parameters $\bar{\theta}$ are updated from the student parameters θ:

$$\bar{\theta} \leftarrow m(s)\bar{\theta} + [1 - m(s)]\theta.$$

The EMA momentum increases from m_{base} to m_{final} on a cosine schedule.

All downstream evaluations use the EMA teacher checkpoint at 60,000 updates, corresponding to 11.52 million sampled training crops, as either a frozen encoder or the initialization for task-specific fine-tuning. Table 7 summarizes the optimization and masking settings.

Table 7: **Temporal pretraining with multiscale masking and a moving-average teacher.** Mask-span lengths are in fine-grid tokens; schedules are in optimizer updates.

Parameter	Setting
Requested mask fraction	55% of fine-grid tokens
Short mask spans	2–4 tokens, probability 0.35
Medium mask spans	16–62 tokens, probability 0.45
Long mask spans	156–312 tokens, probability 0.20
Optimizer	AdamW; $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-10}$
Weight decay	0.1 on matrix weights
Learning rate	Peak 3×10^{-4}
Learning-rate schedule	1,000 linear warm-up updates, then cosine decay to 7.7×10^{-5} at 60,000 updates
Global batch	192 crops (48 crops per GPU across 4 GPUs)
Training hardware	4× NVIDIA PRO 6000; 2.8 GPU-days for 60K updates
Numerical precision	Bfloat16 autocast
Gradient clipping	Global gradient norm of 1.0
EMA momentum	Cosine from $m_{\text{base}} = 0.996$ to $m_{\text{final}} = 0.9989$ over 60,000 updates
Temporal objective weights	Latent: 1; physiological: $\lambda_{\text{phys}} = 1$

C PHASE-ST Architecture and Pretraining

C.1 Spatial Residual Architecture

Frozen temporal interface. For a synchronized set of C channels, the frozen PHASE-T teacher encodes each waveform independently. Stacking its dense outputs gives $X \in \mathbb{R}^{C \times N \times 768}$, with $N = 1{,}875$ for a 60-second input and $N = 156$ for a five-second input at 1 kHz. Both training and inference use PHASE-T's waveform preprocessing. Within the spatial branch, we subtract a fixed training-mean template $\mu^{(L)} \in \mathbb{R}^{N \times 768}$ matched to the input duration L. This position-dependent template is shared across channels and retained with the model; one template per input duration is estimated once with frozen PHASE-T from training windows (two per participant, up to eight channels each), weighted as in pretraining sampling. The residual branch is centered without feature rescaling; the original tokens X remain the skip input and clean prediction target.

A bias-free projection $W_\downarrow$ maps centered tokens from width 768 to 256. Four Transformer blocks then apply self-attention across the C channels independently at each time position. The same block parameters are used at all positions. Each block has pre-attention RMS normalization, four attention heads of width 64, and a pre-normalized SwiGLU feed-forward layer with hidden dimension 512. Both sublayers have residual connections; dropout is zero. Writing F_ℓ for spatial block ℓ, the complete update is

$$\begin{aligned} U^{(0)}_{:,t} &= W_\downarrow (X_{:,t} - \mu^{(L)}_t), \\ U^{(\ell)}_{:,t} &= F_\ell(U^{(\ell-1)}_{:,t}), \qquad \ell = 1, \ldots, 4, \\ H_{:,t} &= X_{:,t} + W_\uparrow \big(U^{(4)}_{:,t} - U^{(0)}_{:,t}\big), \end{aligned} \tag{6}$$

where the bias-free $W_\uparrow$ maps width 256 back to 768. Attention uses the current states of all channels, including the query channel, as keys and values at every layer. The module adds no temporal attention, positional rotation, channel-index embedding, or electrode-coordinate input; temporal context is already carried by PHASE-T. Consequently, permuting the channels permutes H in the same way, and the number of channels can vary without changing the parameters.

Attention-output and feed-forward-output weights start at zero, giving $U^{(4)} = U^{(0)}$ and $H = X$ at initialization. The spatial module contains 3,016,704 trainable parameters, including its two projections. The latent predictor and relation heads bring the pretraining total to 4,260,560; the frozen temporal encoder, frozen physiological decoder, and EMA copy are excluded from this count.

C.2 Latent Prediction and Physiological Supervision

Latent prediction of fixed temporal targets. Independent multiscale masks are sampled for each channel, using the 55% requested fraction and span distribution in Table 7. Let $\mathcal{M}$ contain the requested channel–time positions. After PHASE-T's waveform preprocessing, we zero waveform samples covered by those tokens' receptive fields and mark every overlapping tokenizer patch as masked. Frozen PHASE-T encodes both the clean and masked waveforms. Its masked outputs $\widetilde{X}$ retain the contextual states at all positions, including the masked positions, and feed the spatial module to produce $\widetilde{H}$. The retained within-channel context distinguishes channels while spatial attention supplies synchronized cross-channel information.

A two-layer predictor p (768–768–768, with GELU between the linear layers) maps $\widetilde{H}$ to clean PHASE-T targets. With non-affine, per-token feature normalization LN, the loss is

$$\mathcal{L}_{\text{latent}}^{\text{ST}} = \frac{1}{768|\mathcal{M}|} \sum_{(c,t)\in\mathcal{M}} \left\| \text{LN}(p(\widetilde{H}_{c,t})) - \text{sg}[\text{LN}(X_{c,t})] \right\|_2^2. \tag{7}$$

The target is fixed by the frozen temporal encoder. Only the requested mask positions are scored; the additional overlap halo is excluded.

Frozen physiological supervision. The 84-output physiological decoder saved with PHASE-T is applied directly to $\widetilde{H}$. Its weights and target statistics remain fixed, while gradients through the decoder train the spatial module. Targets are computed from the clean, preprocessed waveforms using the definitions, transforms, and fixed standardization in Appendix B.2; supervision follows the visibility, bandwidth, and loss rules in Appendix B.3.

C.3 Pairwise Relation Targets and Loss

Pairwise relation targets. Targets are measured from the continuous clean waveforms after PHASE-T's waveform preprocessing and temporal mean removal. Fourth-order Butterworth band-pass filters applied forward and backward, followed by the Hilbert transform, define analytic signals

$$z_{c,b}(t) = x_{c,b}(t) + i\,\mathcal{H}\{x_{c,b}(t)\} = A_{c,b}(t)e^{i\phi_{c,b}(t)}. \tag{8}$$

For bands 1–4, 4–8, 8–13, 13–30, 30–80, 80–250, and 250–500 Hz, we measure the phase-locking value $|\langle e^{i(\phi_{i,b}-\phi_{j,b})}\rangle_t|$ and the Pearson correlation of the amplitude envelopes. The filter's upper edge is capped at 99% of Nyquist. Three broadband targets give the signed correlation at the largest absolute cross-correlation peak within ± 100 ms, zero-lag correlation, and the peak's signed lag. A band contributes only when both channels' acquisition bandwidths reach its nominal upper edge.

Phase-locking values are clipped to $[0.001, 0.999]$ and logit transformed; correlations are clipped to $[-0.999, 0.999]$ and Fisher transformed. Lag is divided by 50 ms. Each coordinate is then standardized using fixed training moments.

Pairwise relation heads and loss. Relation prediction uses the same masked spatial output $\widetilde{H}$. For each channel, an affine projection maps its temporal mean to $r_c \in \mathbb{R}^{64}$. A symmetric head takes $[r_i + r_j;\, r_i \odot r_j]$ through a 128–64–16 GELU network. A separate bias-free 64–64–1 network predicts lag from the difference:

$$\widehat{y}_{ij}^{\text{sym}} = h_\eta([r_i + r_j;\, r_i \odot r_j]), \qquad a_{ij} = W_2 \tanh(W_1(r_i - r_j)). \tag{9}$$

The symmetric outputs are invariant to channel exchange. For lag, the standardized prediction is $a_{ij} - \mu_\tau/\sigma_\tau$, where μ_τ, σ_τ are the fixed training moments of the scaled lag target. Undoing this standardization gives $\sigma_\tau a_{ij}$, which changes sign when channels are exchanged. Thus antisymmetry holds in the underlying lag units even when its training mean is nonzero. We sample 64 distinct unordered pairs per window and randomize their orientation. Smooth L1 errors are averaged over valid pairs separately for each window and target coordinate, then over nonempty window–coordinate groups to give $\mathcal{L}_{\text{relation}}$.

C.4 Pretraining and Inference Settings

Channel-set pretraining (PHASE-ST). Training uses 64-channel, 60-second sets sampled from the four-source training corpus: each source contributes a fixed quota of sets, and within a source a participant is drawn uniformly before one of their windows. The global batch is 128 window sets. AdamW uses learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 0.05. Learning rate warms up linearly for 1,000 updates and then remains constant. The spatial module is trained for 5,000 updates, taking approximately 1.9 GPU-days on two NVIDIA PRO 6000 GPUs. Training uses bfloat16 autocast, float32 residual states, and gradient clipping at norm 1.0. Equation 3 uses weights $1 : 1 : 0.2$ for latent prediction, physiological supervision, and relation losses. All three losses share the masked spatial output. PHASE-T and its physiological decoder remain frozen throughout.

An EMA of the spatial module, with momentum rising from 0.996 to 0.9966 along a cosine ramp over the 5,000 updates, is updated after each successful optimizer step and supplies the exported spatial representation; the prediction targets remain the clean frozen PHASE-T tokens. Both reported auditory rows use the resulting step-5,000 spatial EMA checkpoint together with the same step-60,000 temporal teacher.

Inference and exported representations. For PHASE-ST, inference uses unmasked inputs and an EMA of the spatial module, with pretraining heads removed. PHASE-T supplies the temporal tokens. The exported H retains every channel and time position. Temporal averaging yields one vector per channel. Linear probing and fine-tuning for auditory decoding are described in Appendix D.5.

D Evaluation Protocols and Extended Results

D.1 Shared Encoder Settings

All PHASE-T evaluations use the same EMA-teacher checkpoint, saved at step 60,000 of participant-balanced pretraining. This checkpoint serves as the frozen feature extractor for task-specific linear probes and the initialization for jointly fine-tuning the encoder and prediction head. Both settings retain the preprocessing associated with the pretrained model.

D.2 Clinical Prediction across Scales

We evaluate clinical prediction following the official Omni-iEEG benchmark formulations, participant-level partitions, and evaluation units (Duan et al., 2026), spanning four temporal and clinical prediction scales:

- **Pathological HFO (event level, 570 ms)** classifies 570-ms crops of candidate events, detected with PyHFO (Ding et al., 2025) and annotated by experts in Omni-iEEG, into artifact, non-spike HFO, or spike-associated HFO, evaluated by macro-F1 across 14,141 held-out test events from 20 participants. Final token representations are averaged across each crop.
- **Ictal period and sleep–awake classification (window level, 60 s)** evaluate 60-second recording windows by averaging channel decision logits within each window. Following the official Omni-iEEG evaluator, macro-F1 is computed at the threshold maximizing Youden's J.
- **Anatomical localization (channel level)** classifies channels into Lobe-5 and Region-12 anatomical targets using standard 60-second segments, evaluated by macro-F1.
- **Pathological channel identification and surgical outcome (channel and participant levels)** follow Omni-iEEG's official pathological-brain-region formulation using 60-second windows and its benchmark-defined evaluators at two complementary clinical levels:
  - *Channel-level pathology (ROC-AUC)* assesses discriminative identification of seizure-onset-zone (SOZ) channels across 102 test participants (807 SOZ and 7,297 normal channels), using normal channels from seizure-free participants as controls.
  - *Participant-level surgical outcome (ROC-AUC)* follows the benchmark's official resection-based evaluation of post-operative seizure freedom (Engel Class I vs II–IV), which builds on the resection-ratio analyses of prior HFO studies (Zhang et al., 2022b; Monsoor et al., 2023).

D.2.1 Linear Probing and Task-Specific Fine-Tuning

Following Omni-iEEG's official split, linear probes are fitted on frozen training-set representations with a fixed ℓ_2 penalty weight $\lambda = 0.01$ on the mean cross-entropy, without cross-validation. Frozen and fine-tuned evaluations, and the ablation probes in Table 2, use the same training rows for each task; the source probe in Table 10 is fitted on the complete official training set. For task-specific fine-tuning, models are trained for three epochs using AdamW (encoder learning rate 10^{-5}, head 10^{-4}, weight decay 0.01, effective batch size 32, bfloat16 precision). Learning rates use a 5% warm-up, with the first epoch updating only the head. Pathological-channel training uses class-margin cross-entropy with margin $m = 2\log(n_{\text{normal}}/n_{\text{pathological}})$ on the normal class. Checkpoint selection uses the benchmark validation partitions.

D.2.2 Released Foundation Checkpoints and External Adapters

We evaluate three released foundation model checkpoints, BrainBERT STFT-Large (Wang et al., 2023), BaRISTA `parcels_chans` (Oganesian et al., 2025), and MVPFormer-M base (Carzaniga et al., 2026). Omni-iEEG evaluates channel-level predictions or their task-specific aggregates. For the frozen comparison with PHASE-T, we adapt the released foundation models to a single-channel input interface following each model's official protocol, and evaluate their frozen representations with linear probes. The benchmark provides no electrode coordinates, and its anatomical labels are themselves prediction targets; each model encodes one channel at a time ($C = 1$), and BaRISTA receives its released `UNKNOWN` parcel. BrainBERT pools the central ten tokens of five-second STFT frames, BaRISTA encodes three-second frames, and MVPFormer filters inputs at 0.5–120 Hz, resamples them to 512 Hz, and uses the representation of the last 2,560-sample segment; this band-pass removes

most of the 80–500 Hz HFO band. All evaluations follow the benchmark's official train/test rows, task-native aggregation, and scoring workflows. For 60-second rows, BrainBERT and BaRISTA cover the complete waveform with consecutive frames before averaging their frame or token representations, and MVPFormer encodes the complete waveform as twelve consecutive segments, the last of which attends to all twelve. For HFO events, each event is centered in a single frame, and BaRISTA pools its two central tokens.

Table 8 gives the complete clinical comparison, including every task-specific specialist published by Omni-iEEG: LSTM+Attention, adapted from Huang et al. (2023); PatchTST (Nie et al., 2023); TimesNet (Wu et al., 2023a); eHFO (Monsoor et al., 2023); PyHFO spkHFO and PyHFO-Omni (Zhang et al., 2024); CLAP (Wu et al., 2023b); TimeConv-CNN; SEEG-NET (Wang et al., 2022); and EEGNet (Lawhern et al., 2018).

Table 8: **Clinical prediction from events to participant outcomes: full comparison.** Rows above the rule are specialist models published by Omni-iEEG; dashes denote tasks that a specialist does not support. F1 is macro-F1 and AUC is ROC-AUC. Bold marks the highest performance per column, and underline the highest among frozen encoders.

	Event	Window (60 s)		Anatomy		Pathol. brain region	
Model	HFO F1	Ictal F1	Sleep F1	Lobe-5 F1	Region-12 F1	Channel AUC	Outcome AUC
LSTM+Attention	0.7338	–	–	–	–	–	–
TimesNet	0.7652	–	–	–	–	–	–
PatchTST	0.7726	–	–	–	–	–	–
eHFO	–	–	–	–	–	0.6611	0.4521
PyHFO spkHFO	–	–	–	–	–	0.6557	0.4972
PyHFO-Omni	0.8061	–	–	–	–	0.7351	0.7438
CLAP	–	0.9245	0.7225	0.4750	0.3540	0.7684	0.6770
TimeConv-CNN	–	0.8533	0.7118	0.4788	0.3087	0.8061	0.7380
SEEG-NET	–	0.7526	0.6773	0.2520	0.1081	0.7850	0.5952
EEGNet (adapted)	–	–	–	–	–	0.7468	0.5942
BrainBERT STFT-Large	0.7937	0.8323	0.6941	0.4054	0.2592	0.8055	0.6017
BaRISTA `parcels_chans`, $C{=}1$	0.6050	0.7428	0.7213	0.3006	0.1658	0.6949	0.5739
MVPFormer-M, $C{=}1$	0.4347	0.8311	0.5404	0.2423	0.1069	0.6604	0.5878
PHASE-T, frozen	<u>0.8138</u>	<u>0.8958</u>	<u>0.7413</u>	<u>0.4822</u>	<u>0.3398</u>	<u>0.8394</u>	<u>0.6369</u>
PHASE-T, fine-tuned	**0.8188**	**0.9327**	**0.8150**	**0.5134**	**0.3572**	**0.8502**	**0.7713**

D.2.3 Direct Physiological-Target Control

Table 9 compares full PHASE-T with a direct physiological-target shortcut control that uses no encoder, under the task-specific fitting and evaluation protocols of the ablations in Table 2; Full PHASE-T repeats the frozen scores of Table 1, and Outcome is the participant-level aggregate of the pathological-channel probe. The control time-averages the complete 84-dimensional physiological target for the four 60-s tasks and fits task-matched linear probes. For HFO, whose 570-ms crops are shorter than the support of the low-frequency wavelets, the control uses the duration-matched 20-dimensional patch-local target; a 60-dimensional high-frequency variant reaches 0.6531. Full PHASE-T outperforms the encoder-free control on all five tasks.

Table 9: **Direct physiological-target control across five clinical tasks.** F1 is macro-F1 and AUC is ROC-AUC. *Duration-matched patch-local target. Bold marks the highest performance per column.

	Event	Window (60 s)		Anatomy		Pathol. brain region	
Representation	HFO F1	Ictal F1	Sleep F1	Lobe-5 F1	Region-12 F1	Channel AUC	Outcome AUC
Full PHASE-T	**0.8138**	**0.8958**	**0.7413**	**0.4822**	**0.3398**	**0.8394**	**0.6369**
Direct targets (no encoder)	0.6223*	0.8112	0.3089	0.2431	0.1425	0.7293	0.5962

D.3 Cross-Institution Transfer

For transfer from Omni-iEEG to Hatano (Hatano et al., 2026), we evaluate pathological-channel identification using 60-second window representations. Feature standardization and a logistic probe with ℓ_2 penalty weight $\lambda = 0.01$ are fitted on the Omni-iEEG training windows. The fitted probe is then applied without adaptation to Tohoku and NCNP. Target-window probabilities are averaged within recording-channel units, yielding 915 channels from 26 participants at Tohoku and 1,000 channels from 30 participants at NCNP. Tohoku and NCNP are absent from both pretraining and probe fitting. Threshold-independent ROC-AUC serves as the evaluation metric.

For comparison, we follow Omni-iEEG (Duan et al., 2026) in training TimeConv-CNN, CLAP (Wu et al., 2023b), and SEEG-NET (Wang et al., 2022) on the same pathological-channel identification task. Each model is then applied to the target institutions using its training-time preprocessing. The full transfer comparison appears in Table 10.

Table 10: **Cross-institution transfer without local labels: full comparison.** Pathological-channel ROC-AUC at Tohoku and NCNP. Bold marks the highest performance per column.

Model	Training regime	Tohoku	NCNP
TimeConv-CNN	Supervised source model	0.7434	0.4716
CLAP	Supervised source model	0.7663	0.5021
SEEG-NET	Supervised source model	0.7636	0.5528
PHASE-T	Frozen encoder + source probe	**0.7775**	**0.6292**

D.4 Adaptation from Four Participants

We follow MVPFormer's evaluation protocol (Carzaniga et al., 2026) on the MAYO and FNUSA release (Nejedly et al., 2020). At each institution, the first four participants (release IDs 0–3) are used for training and the rest for testing; pathological clips are classified against noise and physiological clips, using all clips; and binary F1 is pooled across test clips. This yields 21,665 training clips and 133,517 test clips from 20 held-out participants at MAYO, and 33,805 training clips and 159,313 test clips from 10 held-out participants at FNUSA. The three-second clips are resampled to 1 kHz.

A linear two-class head is initialized by fitting a logistic probe to training-participant embeddings with ℓ_2 penalty weight $\lambda = 0.01$. We then train the head for one epoch and jointly update the encoder and head for five epochs. AdamW uses encoder and head learning rates of 3×10^{-5} and 10^{-4}, respectively, weight decay 0.01, an effective batch size of 32, and 5% warm-up followed by cosine decay. We use unweighted cross-entropy, seed 0, and no augmentation. Because released MAYO and FNUSA clips are standardized to unit variance rather than calibrated microvolts, inputs are scaled to match the nominal amplitude regime ($s_0 = 20\,\mu\mathrm{V}$), with checkpoint-owned waveform preprocessing applied identically during initialization, training, and evaluation.

The final epoch is evaluated with the trained head at probability threshold 0.5.

We compare against PopT, BaRISTA, MVPFormer, and NeuroCLUS as reported in Table 1 of Zheng & Wang (2026). MAYO and FNUSA are absent from PHASE pretraining. Tables 11 and 12 give the adaptation comparators and held-out cohort metrics (ROC-AUC, precision, recall, and specificity).

Table 11: **Four-participant adaptation: full comparison.** Binary F1 on held-out participants at MAYO and FNUSA. Bold marks the highest performance per column.

Model	MAYO	FNUSA
NeuroCLUS	0.40	0.51
MVPFormer	0.36	0.46
PopT	0.34	0.43
BaRISTA	0.30	0.45
PHASE-T	**0.7337**	**0.6656**

Table 12: **Four-participant adaptation: held-out cohort metrics.**

Quantity	MAYO	FNUSA
Training participants	4	4
Training clips	21,665	33,805
Test participants	20	10
Test clips	133,517	159,313
Binary F1	0.7337	0.6656
ROC-AUC	0.9571	0.8625
Precision	0.7023	0.5600
Recall	0.7681	0.8201
Specificity	0.9666	0.7780

D.5 Multichannel Auditory Decoding

We follow BaRISTA's recording split (Oganesian et al., 2025) and PopT's evaluation protocol (Chau et al., 2025): four binary tasks over five-second windows (sentence onset, speech detection, volume, and pitch), within-participant train/validation/test splits, 90 channels per participant and task chosen on validation data, and ROC-AUC averaged over the seven evaluation participants. Signals follow the benchmark's Laplacian referencing scheme and are resampled to 1 kHz. Both PHASE-ST rows use the spatial-module checkpoint after 5,000 pretraining updates over the frozen PHASE-T teacher at step 60,000.

As in PopT, the frozen row trains only a linear classifier, and the fine-tuned row also updates the spatial module while the temporal encoder stays frozen. The classifier's inverse regularization strength is 0.01. Fine-tuning runs for eight epochs with batch size 32 and AdamW (learning rates 3×10^{-4} for the spatial module and 10^{-5} for the classifier, 5% warm-up, cosine decay), and the state with the highest validation ROC-AUC is evaluated on test. Comparator rows report published results.

D.6 Seizure Trajectory Extraction

For the retrospective seizure analysis, each channel is encoded independently in 60-s windows with a 10-s stride. Only tokens centered in the middle 10 s of each window are retained and averaged within each second. Each time-resolved embedding therefore includes approximately 25–35 s of waveform context on either side of its timestamp. The trajectories in Figure 2a–b are aligned to the clinically annotated seizure onset.

D.7 Cross-Institution Regional Anatomical Correspondence

We assess anatomical correspondence across institutions using 16,444 ECoG channels from 135 Detroit participants and 32 UCLA participants in the eight Region-12 classes with at least 30 ECoG channels at each institution. For each channel, embeddings are averaged across recording windows, centered within participants, standardized across features, and normalized to unit length.

For region-level matching, channels are averaged by anatomical region within each institution, and the resulting regional vectors are normalized to unit length. Each region is compared by cosine similarity with every region at the other institution. The same region is the most similar for all 8 UCLA regions and 6 of 8 Detroit regions, and it ranks first or second in all 16 comparisons. Permuting anatomical labels within each institution yields an average of 1.0 such match per direction; none of 2,000 permutations reaches the observed 14 ($p = 5 \times 10^{-4}$). The eight largest similarities between different regions, from 0.20 to 0.44, all join anatomical neighbors: motor and somatosensory cortex, superior and inferior parietal cortex, parietal and occipital cortex, and the two temporal regions.

For channel-level voting (Figure 2c), each of the ten participants at the other institution with the most similar channels votes for the region of its most similar channel, weighted by that similarity, and each region's vote total is divided by the region's frequency at the institution searched. Every region's channels are voted into their own region more often than chance; the balanced accuracy is 0.271 against 0.125 ($p = 5 \times 10^{-4}$).

E Spatial Context Analysis and Ablations

E.1 Concurrent-Channel Context Probe

Figure 3a probes whether the frozen spatial stage uses synchronized channels from the same participant for SOZ identification. Unlike the single-channel comparison in Table 1, this analysis constructs channel panels. Linear probes for the panel-size comparison are fitted and scored on matching panels; replacement controls also test the fixed same-minute probe. Table 13 lists the conditions, and Figure 4 shows the pooled ROC-AUC at every panel size.

Panels. For this analysis, every training recording (292 recordings, 137 participants) and every test recording with at least eight labeled channels in a synchronized window (154 of 237 recordings, 76 participants) is given two 60-second windows at the labeled rows' own starts, holding every channel the release marks good; labels are the released labels (SOZ positive). Training holds 26,286 labeled channel-windows (2,704 SOZ), test 16,040 (1,446 SOZ). Each channel is encoded once by frozen PHASE-T. For each panel size $k \in \{1, 2, 4, 8, 16, 32, 64\}$ the channels of a window are partitioned at random into balanced groups of at most k (at $k = 64$ the mean group holds 46 channels), and the spatial module mixes each group on its own; three partitions are drawn per window. Substitution arms at the largest size replace a channel's group-mates by the same channels taken from the recording's other window (60 seconds apart) or by 63 channels of another test participant's window; a mixed arm joins each own group of at most 32 channels with 32 channels of another test participant. In the other-minute arm, hardware and reference are unchanged and the companions remain mutually synchronized, but their activity is no longer time-aligned with the target channel.

Probes and statistics. One logistic probe per size, with ℓ_2 penalty weight $\lambda = 0.01$ on standardized time-averaged features, is fitted on all three training partitions and scores the first test partition. The other-minute arm is scored both by a probe fitted on other-minute training panels and by the $k = 64$ same-minute probe. The mixed and other-participant arms also use the $k = 64$ same-minute probe. Per-participant ROC-AUC is defined for the 52 test participants with both classes. Paired comparisons use the exact two-sided sign test on nonzero paired differences and a 10,000-resample participant bootstrap of the mean and median paired difference.

Results. The gain grows with the number of own-participant channels (Figure 4). It depends on time-aligned activity: with matched probes, the same minute exceeds another minute by a mean $+0.033$ [$+0.017$, $+0.051$] per participant (median $+0.006$ [$+0.003$, $+0.030$]; 36 of 50 participants; $p = 0.003$), and the other-minute panels score alike under their own probe and the same-minute probe (mean per-participant difference, own probe minus same-minute probe, -0.004 [-0.014, $+0.004$]). With the same-minute probe held fixed, the participant's own channels exceed 63 channels of another participant by $+0.067$ [$+0.042$, $+0.093$] (median $+0.032$ [$+0.016$, $+0.056$]; 45 of 51; $p = 2 \times 10^{-8}$).

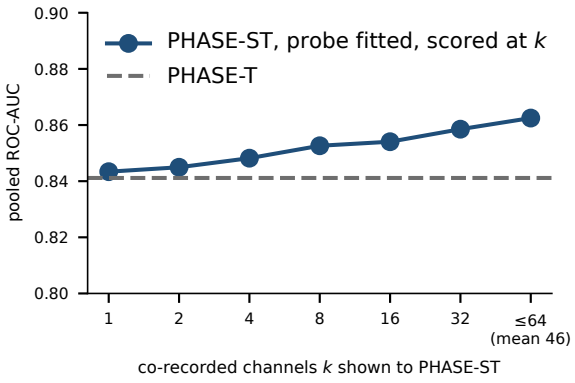

Figure 4: **SOZ identification against panel size.** Pooled SOZ ROC-AUC as a function of the number of co-recorded channels from the same participant; the dashed line is PHASE-T.

Table 13: **SOZ identification under different contexts.** Pooled ROC-AUC and paired per-participant differences (Δ) from the one-channel condition; the $\leq$32-own-plus-32-other-participant row is compared with the $\leq$32-channel same-minute condition. Brackets are 95% participant-bootstrap confidence intervals.

context given to each channel	probe fitted on	pooled	mean Δ [95%]	median Δ
PHASE-T, no context	—	0.841	—	—
1 channel	same	0.843	0	0
$\leq$8 channels, same minute	same	0.853	$+0.018$ [$+0.007$, $+0.032$]	$+0.003$
$\leq$32 channels, same minute	same	0.859	$+0.041$ [$+0.022$, $+0.061$]	$+0.013$
$\leq$64 channels, same minute	same	0.862	$+0.048$ [$+0.026$, $+0.071$]	$+0.016$
same channels, another minute	another minute	0.850	$+0.015$ [$+0.002$, $+0.029$]	$+0.001$
same channels, another minute	same minute	0.841	$+0.019$ [$+0.000$, $+0.040$]	$+0.003$
$\leq$32 own + 32 other-participant	same minute	0.855	-0.009 [-0.021, $+0.002$]	$+0.001$
1 own + 63 other-participant	same minute	0.804	-0.019 [-0.042, $+0.004$]	-0.015

E.2 Spatial-Stage Pretraining Ablations

Both ablated stages in Figure 3b and c are trained exactly as PHASE-ST (Appendix C.4), except that one sets the relation weight to zero and the other restricts each channel's attention to itself, so that a channel's output depends on its own PHASE-T features alone. For the stage without cross-channel attention, the gain from adding companions is therefore zero by construction. Panel b repeats the panel-size comparison above for each stage. At the largest panels, the stage without relation targets gains $+0.014$ [$+0.003$, $+0.025$] per participant; PHASE-ST's context gain is larger by $+0.034$ [$+0.016$, $+0.054$] (95% intervals from 10,000-resample paired participant bootstraps).

E.3 Single-Channel Auditory Probe

Figure 3c uses the windows, splits, labels, and 90-channel panel of the auditory protocol (Appendix D.5). The spatial stage sees all 90 channels; following BrainBERT (Wang et al., 2023), a linear probe is fitted to each channel's output. Each stage's change is taken against the same channel's output from the stage without cross-channel attention, averaged over the 90 channels and four tasks within a participant. PHASE-ST exceeds that stage for all seven participants (mean $+0.054$; smallest difference $+0.038$), and the stage without relation targets exceeds it by a mean $+0.043$.

F LIMITATIONS AND FUTURE DIRECTIONS

PHASE-T and PHASE-ST are pretrained sequentially, with the temporal model frozen during spatial pretraining. Joint single- and multichannel pretraining is a direction for future study.

Temporal context spans up to one minute. Longer recordings can be processed in separate windows, but dependencies across windows are not explicitly modeled. Extending context toward hour-long recordings will require efficient sequence modeling and evaluation on tasks that depend on longer-term changes in neural activity.

Our evaluation covers clinical prediction, cross-institution transfer, and naturalistic auditory decoding. Future work will expand pretraining to additional participants and institutions and evaluate a wider range of recording contexts and downstream tasks.