Learning Time-Scale Invariant Population-Level Neural Representations

Eshani Patel[1] **Yisong Yue**[1] **Geeling Chau**[2]
Computing & Mathematical Sciences[1], Computation & Neural Systems[2]
California Institute of Technology , Pasadena, CA 91125
ejpatel@alumni.caltech.edu {yyue, gchau}@caltech.edu

Abstract

General-purpose foundation models for neural time series can help accelerate neuroscientific discoveries and enable applications such as brain computer interfaces (BCIs). A key component in scaling these models is population-level representation learning, which leverages information across channels to capture spatial as well as temporal structure. Population-level approaches have recently shown that such representations can be both efficient to learn on top of pretrained temporal encoders and produce useful representations for decoding a variety of downstream tasks. However, these models remain sensitive to mismatches in preprocessing, particularly on time-scales, between pretraining and downstream settings. We systematically examine how time-scale mismatches affects generalization and find that existing representations lack invariance. To address this, we introduce Time-scale Augmented Pretraining (TSAP), which consistently improves robustness to different time-scales across decoding tasks and builds invariance in the representation space. These results highlight handling preprocessing diversity as a key step toward building generalizable neural foundation models.

1 Introduction

Building general-purpose representations of neural time series data is a foundational goal for neuroscience research. High-fidelity neural recordings such as intracranial electroencephology (iEEG) capture complex activity patterns across multiple brain regions, but present significant modeling challenges due to inter-subject and session variability, and limited dataset sizes [Herff et al., 2020, Jiang et al., 2025]. As a result, many neuroscience studies and brain computer interface research use single-channel or subject-specific models, limiting their expressivity and generalizability [Willett et al., 2023, Wandelt et al., 2024, Kunz et al., 2024, Wang et al., 2024a].

Recent advances in self-supervised learning have inspired the development of foundation models for neural signals. Time-series foundation modeling and large-scale univariate pretraining provide robust and rich temporal embeddings for downstream decoding [Han et al., 2024, Ansari et al., 2024, Talukder et al., 2024, Wang et al., 2023, Liu et al., 2022]. However, one major limitation is a lack of learned spatial information at the level of populations of channels. Population-level pretraining on top of these temporal embeddings is one promising approach to provide the missing spatial component, demonstrating strong performance on downstream tasks while being computationally and sample-efficient to train [Chau et al., 2025, Liu et al., 2023]. Here we will focus on improving models that learn population-level representations from temporal encoders to enable efficient and generalizable scaling for neuroscience foundation models.

A key design limitation remains in this approach: the population-level layer is trained exclusively on the outputs of the temporal encoders. As a consequence, we empirically observe that these models are sensitive to preprocessing parameters used for input signals–such as the time-scales used for

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

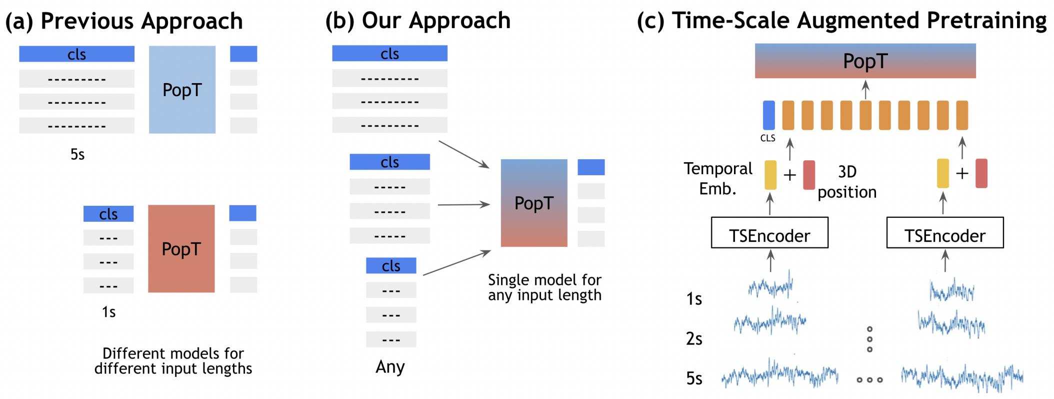

Figure 1: **Schematic of our approach**. (a) Previous population-level transformer (PopT) [Chau et al., 2025] pretrain on fixed temporal windows, which opens up the question of how sensitive these models are to inputs of different time-scales. (b) Our approach seeks to provide optimal performance for any input length. (c) Time-scale Augmented Pretraining (TSAP) use samples from varying input interval lengths to achieve invariance to input time-scales.

training. In practice, neural recordings vary widely in length across datasets and tasks [Peterson et al., 2022, Wang et al., 2024a, Gwilliams et al., 2025], so building invariance into our models along this dimension is important. In this work, we quantify the degradation we get from ignoring these preprocessing mismatches, and explore techniques to remedy this. In particular, we propose a new strategy: Time-scale Augmented Pretraining (TSAP). Applied to iEEG decoding tasks, TSAP consistently improves generalization across seen and unseen time-scales, outperforming interval-specific baselines (Figure 1b). Our analysis further demonstrates that TSAP reduces time-scale clustering in embedding space, leading to more invariant and transferable representations. Addressing this limitation is critical for realizing the promise of foundation models that are applicable across a wide range of tasks and experimental settings.

Our contributions are:

1. A comparative evaluation of decoding performance on mismatched time-scales, confirming that pretraining and finetuning on the same interval length leads to better performance.
2. A novel Time-scale Augmented Pretraining (TSAP) strategy that exposes the model to a spectrum of interval lengths, improving generalization, even to unseen lengths.
3. Analysis of the embeddings from different time-scales across pretrained models.

2 Methods

Population Transformer Framework. To investigate the time-scale invariance of spatial aggregation pretrained transformers, we adopt the Population Transformer (PopT) [Chau et al., 2025] as our core framework, using architectural parameters and training configuration described in the original work.

At a high level, for each electrode channel, c, a given interval i of length l (where l can be varied) is pushed through a frozen temporal encoder, in this case we use BrainBERT [Wang et al., 2023]. These temporal embeddings are summed with a positional embedding derived from its 3D electrode coordinates (Figure 1c), before being passed through the transformer encoder, yielding spatially contextualized channel representations and an aggregated `[CLS]` output. This `[CLS]` is then projected with a linear layer for downstream decoding. We adjust the PopT pretraining strategy below.

Augmented Data and Training. To adapt the PopT framework for temporal invariance, we modified the data generation pipeline to expose the model to iEEG signals at multiple interval lengths. Specifically, we sampled recording segments of lengths $l \in 1, 2, 4, 5$ seconds, with a fixed gap g between consecutive windows for each input channel. Each interval was independently encoded into BrainBERT embeddings, yielding distinct temporal representations for the same underlying signal at different scales. These embeddings include overlapping windows across interval lengths, which the temporal encoder maps into non-identical representations. This augmentation strategy, TSAP, encourages the model to generalize across multiple time-scales.

Embedding analysis. To better understand the distributions of temporal embeddings and PopT representations, we perform 2D PCA on the embedding spaces. We take 100 samples from a specific subject-session from the Word Onset task, across the time-scales 1, 2, 3, 4, and 5 seconds. For temporal embedding PCA analysis, each sample is represented as a concatenation of all the channels embeddings, producing high-d $n_{chan} * h_{dim}$ vectors. For pretrained PopT `[CLS]` token analysis, each sample is pushed through pretrained 5s and TSAP PopT models, producing the processed `[CLS]` token representations. To analyze the clustering, we performed K-Means clustering with 5 means, aligned the clusters to the true classes based on the mode interval of points assigned to the cluster, and plotted the confusion matrices of cluster assignments to true intervals.

3 Experiments

Data. We used iEEG, a type of neural time-series data collected via probes implanted within the brain to record local electrical signals at high temporal resolution and spatial precision. We used the publicly available BrainTreeBank dataset [Wang et al., 2024a]. Data was collected from 10 subjects (total 1,688 electrodes, mean 167 electrodes per subject) who watched 26 movies (19 for pretraining and 7 for downstream decoding), while intracranial probes recorded their neural activity. We evaluate on two auditory-linguistic classification tasks from BrainTreeBank: (1) determining whether any speech at all is occurring (word onset), and (2) determining whether the beginning of a sentence is occurring (sentence onset). For each subject, we randomly select 90 electrodes to use for fine-tuning.

Baselines and Methods Compared. For our baselines, we compared against the non-pretrained PopT to assess the impact of pretraining, as well as the original PopT pretraining formulation as an optimal matched preprocessing baseline. To study the effect of interval-specific pretraining, we pretrained PopT on individual interval lengths of 1, 2, 3, 4, or 5 seconds to produce an "optimal" baseline model (5 seconds is the original PopT formulation). Each variant was trained following the same setup as the PopT paper, with the exception of using a learning rate of 1×10^{-4} to improve training stability. These experiments serve to test whether models pretrained and finetuned on the same interval length outperform non-corresponding counterparts across tasks and interval lengths.

Evaluating TSAP. To understand how TSAP performs on a variety of downstream time-scales, we pretrain PopT with TSAP as described in Section 2 with interval lengths of 1, 2, 4, and 5 seconds (note that 3 seconds is held out). We doubled the number of pretraining steps from 500,000 to 1,000,000, while keeping the learning rate fixed at 1×10^{-4} to allow the model to process the larger augmented dataset. For all pretraining runs, we selected the best-performing model checkpoint based on validation loss. After pretraining, we evaluated the TSAP model on all downstream interval lengths to compare its performance with the optimal baselines for held-in (1, 2, 4, and 5 seconds) and held-out (3 seconds) interval lengths.

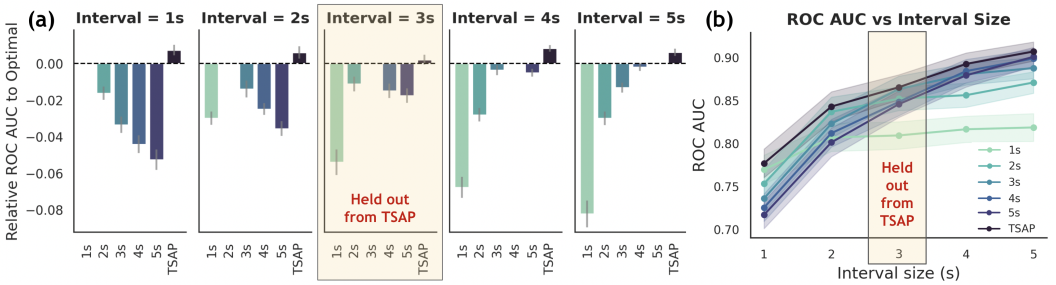

Figure 2: **Performance drop from mismatch in input time-scales is recovered by TSAP.** (a) Compared to the optimal (dotted line), models (x-axis) trained with mismatched time-scales perform much worse (below the line), while TSAP (dark blue) generally improves upon the optimal baseline. Shown are the Word Onset ROC AUC difference means and standard error across subjects and 5 seeds. (b) We see TSAP (dark blue) match or outperform all other models across all input lengths. Shown are the Word Onset ROC AUC mean and standard error across subjects and 5 seeds.

	1s Interval	**2s Interval**	**3s Interval**	**4s Interval**	**5s Interval**
Word Onset:					
Non-Pretrained	0.645 ± 0.015	0.665 ± 0.027	0.663 ± 0.018	0.671 ± 0.019	0.678 ± 0.018
1s Interval	$\underline{0.770} \pm \underline{0.017}$	0.807 ± 0.016	0.809 ± 0.016	0.817 ± 0.016	0.819 ± 0.016
2s Interval	0.753 ± 0.017	$\underline{0.837} \pm \underline{0.017}$	0.852 ± 0.014	0.856 ± 0.014	0.871 ± 0.013
3s Interval	0.736 ± 0.016	0.823 ± 0.016	$\underline{0.863} \pm \underline{0.015}$	0.881 ± 0.013	0.888 ± 0.012
4s Interval	0.725 ± 0.016	0.812 ± 0.016	0.849 ± 0.016	$\underline{0.884} \pm \underline{0.014}$	0.899 ± 0.010
5s Interval	0.717 ± 0.016	0.801 ± 0.017	0.846 ± 0.015	0.879 ± 0.014	$\underline{0.901} \pm \underline{0.011}$
TSAP	$\mathbf{0.777 \pm 0.017}^{*}$	$\mathbf{0.843 \pm 0.017}$	$\mathbf{0.866 \pm 0.015}$	$\mathbf{0.893 \pm 0.013}^{*}$	$\mathbf{0.907 \pm 0.011}^{*}$
Sentence Onset:					
Non-Pretrained	0.715 ± 0.021	0.740 ± 0.018	0.731 ± 0.018	0.717 ± 0.017	0.710 ± 0.018
1s Interval	$\underline{0.790} \pm \underline{0.016}$	0.798 ± 0.015	0.785 ± 0.015	0.778 ± 0.015	0.776 ± 0.014
2s Interval	0.771 ± 0.018	$\underline{0.837} \pm \underline{0.014}$	0.831 ± 0.013	0.828 ± 0.014	0.829 ± 0.011
3s Interval	0.764 ± 0.016	0.822 ± 0.015	$\mathbf{0.846 \pm 0.013}$	$\underline{0.852} \pm \underline{0.012}$	0.853 ± 0.011
4s Interval	0.760 ± 0.017	0.803 ± 0.016	0.833 ± 0.013	0.851 ± 0.013	0.851 ± 0.011
5s Interval	0.760 ± 0.016	0.798 ± 0.016	0.821 ± 0.014	0.847 ± 0.012	$\underline{0.860} \pm \underline{0.011}$
TSAP	$\mathbf{0.802 \pm 0.015}^{*}$	$\mathbf{0.841 \pm 0.014}$	$\underline{0.843} \pm \underline{0.013}$	$\mathbf{0.855 \pm 0.012}$	$\mathbf{0.865 \pm 0.010}$

Table 1: **Results across models and decoding tasks.** For each interval length (columns), we evaluate on models non-pretrained, pretrained with specific interval lengths (1s, 2s, 3s, 4s, and 5s), and pretrained on using TSAP (3s interval is heldout) (rows). We show two different downstream tasks: Word Onset and Sentence Onset (sections). Shown are the test ROC-AUC mean and standard error across subjects and 5 seeds per subject. Best per task is bolded, second best is underlined. Asterisks denote significant improvement compared to the optimal with paired t-test (Table 2).

Finetuning. For finetuning, embeddings were generated at interval lengths of 1, 2, 3, 4, or 5 seconds, with each sample labeled as positive or negative depending on whether it was centered around a word or sentence onset (depending on the task). Each finetuning experiment was performed on a single subject and interval length, with models evaluated across every subject-interval combination. The 3-second interval dataset served as the held-out interval set, as TSAP was not pretrained on 3-second interval data. To improve robustness, each experiment was repeated five times with different random seeds. For each seed, we selected the best model checkpoint based on validation ROC-AUC, and reported test performance on that model.

4 Results

Decoding Performance. We sought to understand how much of a performance decrease we get from using mismatched input lengths between pretraining and finetuning. We find dramatic reductions in performance when these are mismatched (Figure 2a). When we introduce TSAP, we recover and occasionally exceed the performance achieved by the optimal baseline models (Figure 2). This is consistent across downstream time-scales, even for unseen time-scales such as the 3-second interval shown in Figure 2b. We also see the same patterns across additional decoding tasks Table 1. By augmenting the pretraining task with additional time-scales, we obtain a model that is able to match or exceed the expected performance for each time-scale.

Embedding Space Analysis. We hypothesized that the temporal encoding representations of trials cropped to different time-scales would be represented drastically differently, despite containing overlapping information. To see if this is true, we project our BrainBERT encoded samples into the top 2 PCA component space, and find strong clustering by interval length (Figure 3a).

To check what was happening to the representations by doing TSAP versus training on only one length, we projected 100 samples with their `[CLS]` token representations from the two respective models. We find that the model trained only with 5-second time-scales still exhibits strong clustering of samples within their time-scales (Figure 3b), while the TSAP model has much more overlap of its clusters suggesting greater time-scale invariance in its representations (Figure 3c). The confusion matrices following K-means clustering of the `[CLS]` tokens further show how the 5-second Pretrained PopT produces extremely time-scale dependent clusters while there is more confusion with the cluster assignments with the TSAP model Figure 5.

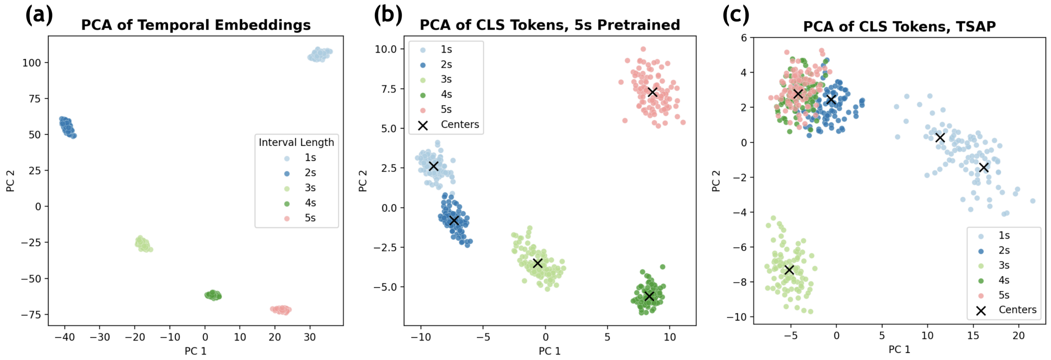

Figure 3: **PCA Analysis of Raw Embeddings and CLS tokens**. (a) PCA projection of temporal embeddings taken from different time-scales (colors) from 1 subject and 1 session from the Word Onset task. Temporal Embeddings tend to cluster by interval length despite them being from the same 100 samples. (b) PCA of CLS token after training with 5-second intervals only. We again see strong clustering by time-scale, with K-Means clusters identified for each ("X" marks). (c) PCA of CLS token after training with TSAP. We see that TSAP CLS tokens across several time-scales are clustered more closely together, with confused K-Means cluster ("X" marks).

5 Discussion

Our exploration into the sensitivity of variable time-scale inputs reveals that current models are overfit to specific time-scale and preprocessing styles, preventing optimal performance across arbitrary time-scales (Figure 2a). We find that pretraining with the same interval length provides optimal performance for evaluating on that time-scale. However, model performance drops as we evaluate on interval lengths different from the pretraining time-scale. Despite this, none of the pretrained models do as poorly as the non-pretrained performances, suggesting that there is still some valuable information being learned in pretraining despite mismatched time-scales (Table 1). To amend this sensitivity to input length, we tested an augmentation strategy that involves pretraining with samples of different interval lengths, and find that it is incredibly effective at recovering lost performance (Figure 2) across all time-scale inputs. In fact, we find that our augmented pretraining strategy allows the resulting model to consistently outperform the optimal models in many cases, even performing better than most on held-out time-scale lengths. Models that can achieve optimal performance across time-scales can allow neuroscience and BCI research to use neural foundation models more effectively out of the box, providing optimal performance no matter the time-scale input size of the experimental paradigm.

Limitations and Future Work. We identify a gap in generalizability of current population-level foundation models to perform well across input time-scales, and explore a solution to remedy this. However, it remains to be seen how this solution compares with or augments other approaches, such as building invariance into the temporal encoders themselves [Zhang et al., 2022, Liu et al., 2023, Somaiya et al., 2022], or leveraging smaller fixed patches and learning temporal and spatial components together [Wang et al., 2024b, Jiang et al., 2024, Zhang et al., 2024, Talukder et al., 2024]. Future work could experiment with combinations of these approaches and evaluate which scales efficiently and provides best generalization to input time-scales.

6 Conclusion

Here we focused on improving population-level foundation models by addressing a key limitation in generalizability to time-scales. We quantify the drop in performance due to mismatches in preprocessing, and show that such performance drop can be easily recovered with our proposed TSAP technique. We further investigate the nature of the embeddings spaces and generalizability of this approach to unseen time-scales to demonstrate that it is an effective strategy for foundation models to adopt in order to gain generalizbility along this critical dimension.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. Mbrain: A multi-channel self-supervised learning framework for brain signals. KDD '23, page 130–141, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599426. URL `https://doi.org/10.1145/3580305.3599426`.

Geeling Chau, Yujin An, Ahamed Raffey Iqbal, Soon-Jo Chung, Yisong Yue, and Sabera Talukder. Generalizability under sensor failure: Tokenization+ transformers enable more robust latent spaces. *arXiv preprint arXiv:2402.18546*, 2024.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. *ArXiv*, pages arXiv–2406, 2025.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M Sandino, and Joseph Y Cheng. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *bioRxiv*, pages 2024–04, 2025.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7129–7142, 2024.

Christian Herff, Dean J Krusienski, and Pieter Kubben. The potential of stereotactic-eeg for brain-computer interfaces: current progress and future directions. *Frontiers in neuroscience*, 14:123, 2020.

Brian Kenji Iwana and Seiichi Uchida. Time series data augmentation for neural networks by time warping with a discriminative teacher. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3558–3565. IEEE, 2021.

Linxing Preston Jiang, Shirui Chen, Emmanuel Tanumihardja, Xiaochuang Han, Weijia Shi, Eric Shea-Brown, and Rajesh PN Rao. Data heterogeneity limits the scaling effect of pretraining neural data transformers. *bioRxiv*, pages 2025–05, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

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

Erin M Kunz, Benyamin Meschede-Krasa, Foram Kamdar, Donald Avansino, Samuel R Nason-Tomaszewski, Nicholas S Card, Brandon Jacques, Payton Bechefsky, Nick Hahn, Carrina Iacobacci, et al. Representation of verbal thought in motor cortex and implications for speech neuroprostheses. *bioRxiv*, pages 2024–10, 2024.

Ran Liu, Mehdi Azabou, Max Dabagia, Jingyun Xiao, and Eva Dyer. Seeing the forest and the tree: Building representations of both individual and collective dynamics with transformers. *Advances in neural information processing systems*, 35:2377–2391, 2022.

Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

Steven M Peterson, Satpreet H Singh, Benjamin Dichter, Michael Scheid, Rajesh PN Rao, and Bingni W Brunton. Ajile12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific data*, 9(1):184, 2022.

Pratik Somaiya, Harit Pandya, Riccardo Polvara, Marc Hanheide, and Grzegorz Cielniak. Ts-rep: Self-supervised time series representation learning from robot sensor data. 2022.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*, 2024.

Sarah K Wandelt, David A Bjånes, Kelsie Pejsa, Brian Lee, Charles Liu, and Richard A Andersen. Representation of internal speech by single neurons in human supramarginal gyrus. *Nature human behaviour*, 8(6):1136–1149, 2024.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

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, 2024a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024b.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36, 2024.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8980–8987, 2022.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36, 2024.

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

A Related Works

Foundation Models for Neural Signals. Recent work on neural time-series—particularly EEG and iEEG—has leveraged foundation-model approaches to learn generalizable representations across individuals, tasks, and recording setups. Channel-independent pretraining has shown promise for spiking data [Liu et al., 2022], electrophysiological recordings [Wang et al., 2023, Talukder et al., 2024, Chau et al., 2024], and general time-series [Talukder et al., 2024], while related models have also been explored for EEG [Chien et al., 2022, Kostas et al., 2021, Yi et al., 2023]. However, these methods often focus on single channels or assume fixed sensor layouts, limiting their ability to capture population-level interactions across heterogeneous datasets. More recent efforts jointly pretrain spatial and temporal dimensions to handle variable inputs [Zhang et al., 2024, Yang et al., 2024, Jiang et al., 2024, Ye et al., 2024, Cai et al., 2023], but their coupled design increases computational cost and sensitivity to preprocessing. The Population Transformer (PopT) [Chau et al., 2025] addresses this by combining pretrained temporal encoders with a spatial aggregation transformer to enable population-level modeling across electrode configurations, yet it remains constrained by the fixed-duration temporal inputs of its encoders, which can hinder transfer across mismatched time intervals and preprocessing schemes.

General Time-Series Foundation Models and Variable-Length Training. Beyond the neural domain, several time-series approaches highlight the benefits of training with variable-length inputs. TS-Rep [Somaiya et al., 2022] encourages duration-agnostic representations via a triplet-based objective, while Time Warping with a Discriminative Teacher [Iwana and Uchida, 2021] introduces controlled distortions through dynamic time warping. Other methods exploit frequency-domain consistency, such as TF-C [Zhang et al., 2022] and BioFAME [Liu et al., 2023], to promote invariance to temporal scale. These works demonstrate that temporal augmentation and multi-view learning can yield scale-robust features, but they are largely limited to univariate or single-modality settings. Additionally, many popular foundation models for time series [Ansari et al., 2024, Yue et al., 2022, Talukder et al., 2024, Wang et al., 2023] still rely on fixed-length windows, requiring ad hoc strategies to handle context mismatches. This motivates our approach: augmenting PopT's pretraining data with variable-length intervals to extend population-level neural representation learning toward temporal invariance in a multivariate setting.

B Training Details

To run all our experiments (data processing, pretraining, evaluations, interpretability), one only needs 1 NVIDIA RTX A6000 (50GB GPU RAM). Pretraining PopT on a single interval takes approximately 1.5 days on 1 GPU and pretraining TSAP on a single interval takes approximately 3 days on 1 GPU. Our downstream evaluations take a few minutes to run each. For the purposes of data processing and gathering all the results in the paper, we parallelized the experiments on 6 GPUs.

C Sentence Onset Results

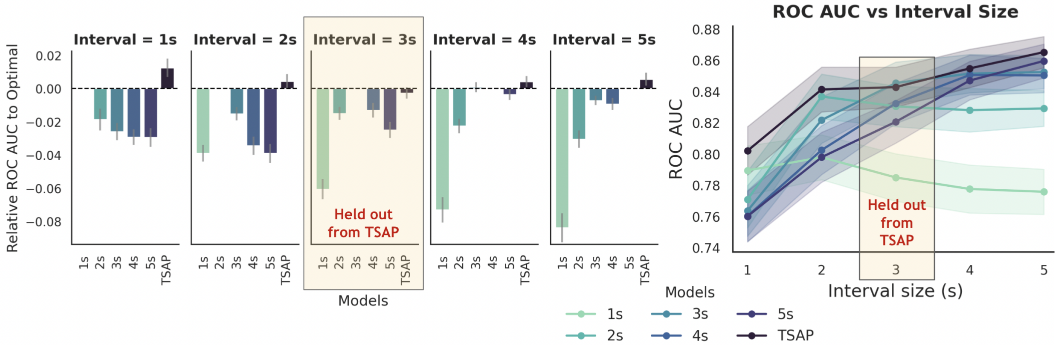

Figure 4: **Performance drop from mismatch in input time-scales is recovered by TSAP.** (a) Compared to the optimal (dotted line), models (x-axis) trained with mismatched time-scales perform much worse (below the line), while TSAP (dark blue) generally improves upon the optimal baseline. Shown are the Sentence Onset relative ROC AUC difference means and standard error across subjects and 5 seeds. (b) We see TSAP (dark blue) closely matches or outperform other models across all input time-scales. Shown are the Sentence Onset ROC AUC mean and standard error across subjects and 5 seeds.

D Cluster Analysis

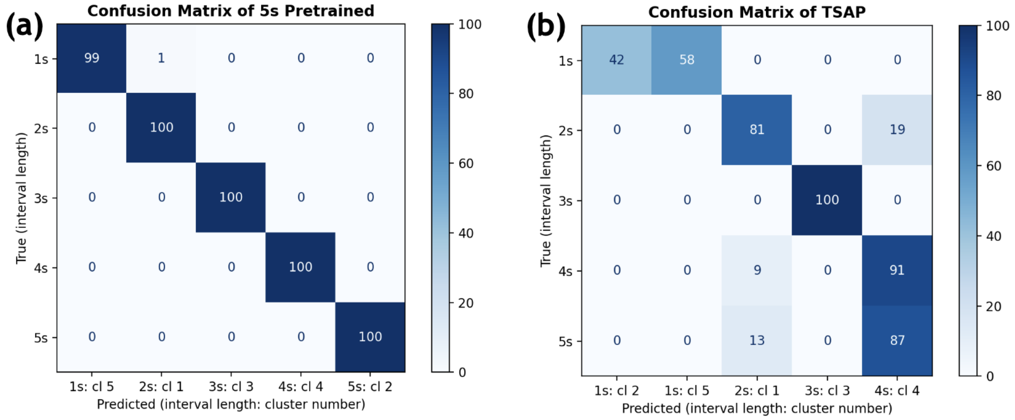

Figure 5: **Confusion matrix following K-Means clustering of CLS tokens.** For (a) 5s Pretrained PopT and (b) TSAP model. We see clean clustering in the 5s, but much more confusion in the TSAP version.

E Statistical Analysis

We chose to conduct a paired t-test due to the high variation between subjects and seeded splits. We see that most of the TSAP models are significantly better than the Optimal Baselines based on the p-values from this paired t-test. Additionally, the 95% confidence interval have very small negative bounds which indicate high confidence that our model will perform at or better than the Optimal Baseline. The held-out 3-second interval has the least significant improvement for our TSAP model, but we still see that it occasionally has improvement upon the optimal baseline (1/2 wins).

Interval	**Mean**	**Std. Err**	t	p	**95% CI**	N
Word Onset						
1s	0.0074	0.0029	2.516	0.01675∗	(0.0014, 0.0134)	35
2s	0.0061	0.0034	1.774	0.08506	(−0.0009, 0.0130)	35
3s	0.0021	0.0027	0.777	0.4423	(−0.0034, 0.0076)	35
4s	0.0083	0.0018	4.647	0.00005∗	(0.0047, 0.0120)	35
5s	0.0062	0.0020	3.079	0.00409∗	(0.0021, 0.0102)	35
Sentence Onset						
1s	0.0126	0.0056	2.258	0.0305∗	(0.0013, 0.0239)	35
2s	0.0045	0.0042	1.062	0.2958	(−0.0041, 0.0131)	35
3s	−0.0026	0.0034	−0.782	0.4396	(−0.0095, 0.0042)	35
4s	0.0042	0.0033	1.297	0.2034	(−0.0024, 0.0109)	35
5s	0.0057	0.0039	1.463	0.1526	(−0.0022, 0.0135)	35

Table 2: **Paired t-test results (TSAP vs Optimal Baseline)** for ROC AUC across subject/seed pairs. Shown are results for Word Onset (top) and Sentence Onset (bottom). Entries are the mean paired difference, standard error, t statistic, p-value, 95% confidence interval, and sample size (N).