A foundation model with multi-variate parallel attention to generate neuronal activity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning from multi-variate time-series with heterogeneous channel configurations remains a fundamental challenge for deep neural networks, particularly in clinical domains such as intracranial electroencephalography (iEEG), where channel setups vary widely across subjects. In this work, we introduce multi-variate parallel attention (MVPA), a novel self-attention mechanism that disentangles content, temporal, and spatial attention, enabling flexible, generalizable, and efficient modeling of time-series data with varying channel counts and configurations. We use MVPA to build MVPFormer, a generative foundation model for human electrophysiology, trained to predict the evolution of iEEG signals across diverse subjects. To support this and future efforts by the community, we release the Long-term iEEG dataset, the largest publicly available iEEG dataset to date, comprising nearly 10,000 hours of recordings from heterogeneous clinical sources. MVPFormer leverages MVPA to achieve strong generalization across subjects, demonstrating expert-level performance in several iEEG tasks. MVPFormer surpasses state-of-the-art (SOTA) Transformer baselines in seizure detection across the Long-term, the MAYO, and the FNUSA dataset, while also achieving SOTA performance on four Brain Tree-Bank iEEG decoding tasks (volume, pitch, onset, and speech). Together, our contributions establish MVPFormer as the first open-source, open-weights, and open-data iEEG foundation model with SOTA clinical performance.

1 Introduction

The increasing availability of biosignals from the human brain and body has driven demand for neural architectures capable of learning from such data [1–3]. A fundamental challenge in this setting is channel heterogeneity: different sensors (or channels) often carry information that is both structurally and semantically non-uniform, while the number and the location of channels may vary across instances, especially in intracranial electroencephalography (iEEG) [4]. iEEG models [5–9] often require subject-specific adaptation to account for new setups, yet they still struggle to generalize. Consequently, effective learning from multi-variate time-series demands models that are flexible and channel-agnostic, without sacrificing locality or the ability to generalize.

In this work, we introduce multi-variate parallel attention (MVPA), a novel self-attention mechanism addressing the structural challenges of channel heterogeneity. MVPA decomposes attention into three components: content-, time-, and channel-based components. Thus, it allows the model to separately learn the semantics of the signal, its temporal dynamics, and spatial (inter-channel) structure. MVPA enables flexible and efficient processing of multi-variate time-series data, without relying on fixed channel positions or global positional encodings.

We use MVPA to build MVPFormer, a foundation model for human electrophysiology trained via generative pre-training to predict the evolution of brain signals. MVPFormer is trained on the

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Long-term iEEG dataset, the largest available iEEG corpus to date with nearly 10,000 hours of multi-channel recordings (or 540,000 channel-hours), collected over a decade in clinical settings and made publicly available as part of this work. We show that MVPFormer not only models neuronal activity during both normal and ictal states, but also enables diverse downstream applications, including seizure detection on multiple institutional datasets [10] and four iEEG decoding tasks from the Brain TreeBank dataset [11]. Remarkably, MVPFormer surpasses an equivalent purely discriminative version which has not undergone generative pre-training, strengthening the validity of foundation models in iEEG.

Our contributions are: 1) **Multi-variate parallel attention (MVPA)**, a novel self-attention mechanism that enables strong generalization across multi-variate time-series with heterogeneous channels; 2) **MVPFormer**, a foundation model for human electrophysiology; 3) The release of the **Long-term iEEG dataset**, the largest iEEG dataset publicly available to date, with almost 10,000 hours of highly curated and labeled iEEG recordings.

Moreover, we make all our contributions open source, realizing the first open-data, open-code, and open-weights iEEG foundation model.

2 Multi-variate parallel attention (MVPA)

Multi-variate parallel attention (MVPA) is based on vanilla attention [12], but unlike the latter it operates on 2D sequences of embeddings $(\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{C,T})$ with a channel dimension C and a time dimension T. Each embedding is of dimension d, i.e., $\boldsymbol{x}_{c,t} \in \mathbb{R}^d$.

We introduce two separate learnable positional codebooks, representing space ($\mathcal{C}$) and time ($\mathcal{T}$). By equipping self-attention with this dual encoding, we can treat the two dimensions individually, which is fundamental in recovering their interplay and would not be possible with vanilla attention:

$$\boldsymbol{a}^{\text{dual}}_{c,t,c',t'} = (\boldsymbol{x}_{c,t} + \mathcal{T}_t + \mathcal{C}_c)^T W_q^T W_k (\boldsymbol{x}_{c',t'} + \mathcal{T}_{t'} + \mathcal{C}_{c'}).$$

Dual attention allows us to exploit the relationship between time and space at the attention level, the most basic computational unit of a Transformer. However, the dual attention mechanism is computationally expensive, as it computes second-order correlations between time and space. Inspired by Transformer-XL [13], we encode the relative distance in the two dimensions between the segments separately and introduce new learnable bias terms (u, v, w). Finally, we remove the correlation terms and rearrange the expanded equation into three related groups:

$$\begin{aligned}
\boldsymbol{a}^{\text{MVPA}}_{c,t,c',t'} &= \boldsymbol{x}^T_{c,t} W_q^T W_{k_e} \boldsymbol{x}_{c',t'} + u^T W_{k_e} \boldsymbol{x}_{c',t'} && \text{Content-based attention} && (1)\\
&+ \boldsymbol{x}^T_{c,t} W_q^T W_{k_t} \mathcal{T}_{t-t'} + v^T W_{k_t} \mathcal{T}_{t-t'} && \text{Time-based attention} && (2)\\
&+ \boldsymbol{x}^T_{c,t} W_q^T W_{k_c} \mathcal{C}_{c-c'} + w^T W_{k_c} \mathcal{C}_{c-c'} && \text{Channel-based attention} && (3)
\end{aligned}$$

The three terms above are the attentional components of MVPA (see Figure S1). Content-based attention only attends to the content of query and key, without any positional encoding. To further reduce the cost, content-based attention makes use of a local attention window [14] that focuses on the most recent L (in our case 10 segments, or 50 seconds) time points. Time-based attention only attends to the query and the distance in time with the key. Finally, channel-based attention only attends to the query and the distance in space with the key. The relative encoding scheme allows the channel component to uncover the hidden connection map between the spatial locations from its initial random initialization, as shown in Appendix G.11.

3 MVPFormer

MVPFormer is our novel Transformer-based predictive foundational model equipped with MVPA, that processes heterogeneous multi-variate iEEG data (see Figure 1). MVPFormer predicts the development of neuronal activity in a *continuous* embedding space governed by a wavelet encoder. We build MVPFormer following the foundational paradigm, with a pre-training dedicated to predicting the future iEEG embedding using a contrastive loss function. Moreover, we show that a successive fine-tuning using LoRA [15] and a simple classification head allows MVPFormer to perform downstream classification tasks. In particular, our results indicate that a model trained in this fashion surpasses an equivalent purely discriminative model (i.e., without generative pre-training), strengthening the validity of foundation models in the iEEG domain as well.

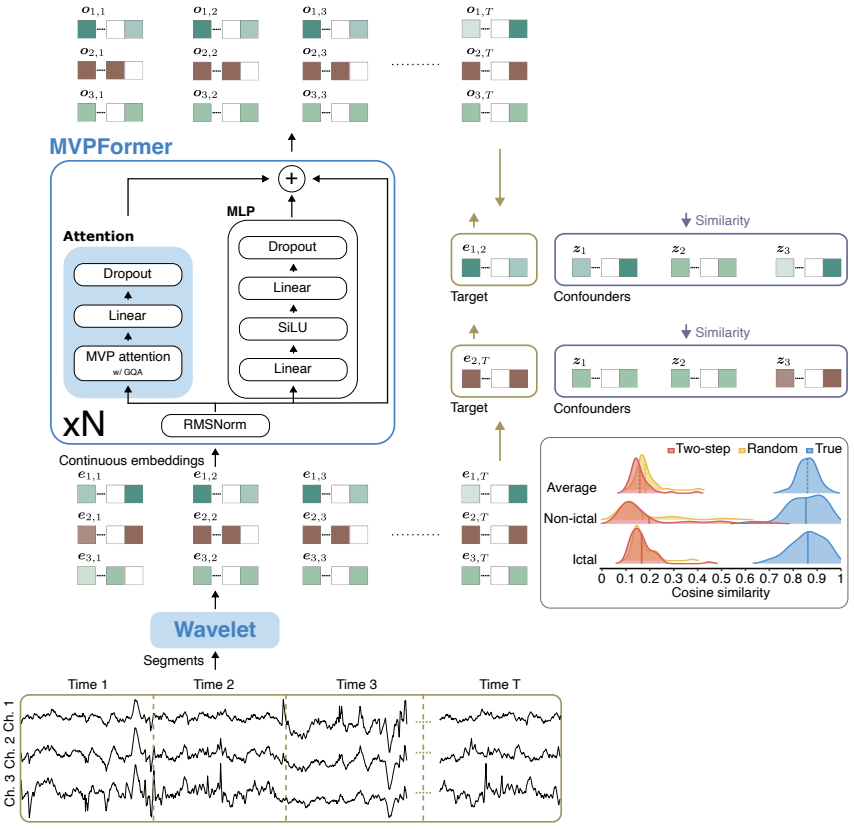

Figure 1: **MVPFormer architecture and forward pass.** iEEG signals are segmented in time and space, encoded via a wavelet-based encoder, and arranged into a 2D embedding grid. These continuous embeddings are processed by MVPA to model temporal, spatial, and content-based dependencies. MVPFormer predicts the next-in-time embedding while reducing similarity to confounders from the same or other subjects. Notched in the bottom right is the resulting cosine similarity with the true target and the confounders after training.

3.1 Training

Generative pre-training MVPFormer is pre-trained to generate neuronal activity by predicting successive input segments in time. During pre-training, random input segments from batched windows are used as confounding targets, which are plausible but different from the true target. See Appendix C.1 for more details on the pre-training.

LoRA fine-tuning for downstream tasks For downstream tasks, we use a linear classification head. We further adopt LoRA [15] to perform parameter-efficient fine-tuning. This leads to a number of trainable parameters during fine-tuning of approximately 0.1% of the base model.

4 Long-term iEEG dataset

The lack of publicly available large-scale iEEG datasets has been a significant obstacle to the development of foundation models for this modality. In an effort towards addressing this issue, together with this work, we open-source the Long-term iEEG dataset, a large-scale iEEG dataset consisting of a total of 68 subjects, 9328 hours of recording, and 704 ictal events. To our knowledge, the Long-term iEEG dataset is the largest publicly available iEEG dataset, fully curated and labeled by experienced clinicians. Appendix D reports more details and illustrates two example recordings.

The iEEG signals were recorded intracranially for clinical purposes with a sampling rate of either 512 Hz or 1024 Hz. The signals were median-referenced and digitally band-pass filtered between 0.5

Table 1: **Results on the iEEG tasks.** We compare MVPFormer with multiple baselines across 4 iEEG datasets and 5 tasks. The models requiring the electrodes' position are indicated by †. The best results without position are bolded, while the results where the position is beneficial are underlined.

		SWEC		MAYO	FNUSA	Brain TreeBank			
		Seizure		Seizure	Seizure	Pitch	Volume	Onset	Speech
Model	Attention	Kappa	f1	f1	f1	acc	acc	acc	acc
MVPFormer	MVPA	**0.57**	**0.56**	**0.36**	**0.46**	/	/	/	/
MVPFormer-S	MVPA	0.54	0.51	0.35	0.46	**0.83**	**0.88**	**0.87**	**0.90**
MV-Llama	Vanilla	0.05	0.02	/	/	0.63	0.77	0.80	0.81
Brant	Vanilla	/	/	/	/	0.61	0.74	0.80	0.80
Brant-2	Vanilla	0.08	0.00	0.19	0.46	/	/	/	/
BrainBERT	Vanilla	0.00	0.00	/	/	0.59	0.66	0.70	0.71
PopT †	Vanilla	/	/	/	/	0.74	0.87	<u>0.90</u>	<u>0.93</u>
PopT	Vanilla	/	/	/	/	0.62	0.76	0.81	0.83

and 120 Hz. All the recordings were inspected by an expert for identification of seizure onsets and offsets, and to remove channels corrupted by artifacts.

5 Results

5.1 Seizure detection task

We first consider a clinically realistic setup that compares model predictions to a board-certified neurologist's annotations using Cohen's Kappa (see Appendix E). Expert-level performance in the seizure classification task varies considerably, from 0.58 [16] to 0.53 [17] in EEG, to 0.57 [18] in iEEG. We consider Kappa values above 0.53 to be expert-level. MVPFormer achieves an average Kappa of 0.57 across 50 unseen subjects from the Long-term iEEG dataset, matching human expert performance (see Figure 2). Importantly, agreement varies by subject, reflecting the clinical reality that seizure presentation complexity strongly affects classification (see Appendix G.10). Overall, MVPFormer demonstrates expert-level seizure classification across a large, heterogeneous cohort. This performance, combined with its low false positive rate (0.17 fp/h), positions MVPFormer as a promising clinical assistant for real-world iEEG analysis.

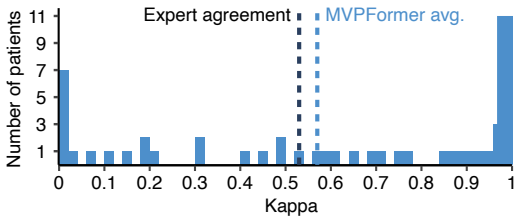

Figure 2: **MVPFormer's inter-rater agreement on seizure detection.** The average Kappa score between MVPFormer-M and the human expert is 0.57, competitive with the values obtained between human experts.

Second, we consider a conventional evaluation (see Appendix G.3) based on F1-score, sensitivity, specificity, and number of false positives per hour (fp/h). We compare MVPFormer against three strong baselines: (1) Brant-2 [19], a SOTA Transformer for iEEG, is fine-tuned with published weights. (2) BrainBERT [8], another SOTA iEEG model with public weights. (3) MV-Llama, an ablation of MVPFormer-S that replaces MVPA with vanilla attention, is trained identically to MVPFormer. We test the considered models zero-shot on our Long-term iEEG dataset and also apply them on the MAYO and FNUSA datasets [10] (see Appendix G.12). The full set of results can be found in Appendix G.1.

As shown in Table 1, both baselines fail to generalize on our Long-term iEEG dataset, achieving Kappa scores of just 0.08 and 0.05, while MVPFormer achieves 0.57 and 0.54 in medium and small configurations, respectively. Moreover, MVPFormer outperforms the baselines on MAYO (highest f1-score of 0.36) and is on par on FNUSA. As a further baseline — to validate our choice of pre-training — we also compare MVPFormer-S with an equivalent model built without the generative

base task (see Appendix G.13). With this setup, the purely discriminative model only reaches a Kappa score of 0.52, inferior to the equivalent MVPFormer-S which reached 0.54.

5.2 Brain TreeBank decoding tasks

We validate the generalization of MVPFormer by testing it on the four tasks of the Brain TreeBank dataset [11], as described in [8, 20]. The four tasks are: 1) discrimination of volume level (volume), 2) discrimination of pitch (pitch), 3) classification of sentence onset (onset), and 4) classification of speech (speech). All four tasks involve the discrimination of high-level cognitive behaviors from iEEG recordings. As such, they represent a significant testbed for MVPFormer outside of its design environment of seizure detection. Table 1 shows the results of MVPFormer against the SOTA baselines represented by PopT [20], BrainBERT [8], and Brant [21], as reported by PopT [20]. The full results can be found in Appendix G.2.

MVPFormer surpasses all baselines on the pitch and volume tasks, providing further evidence that MVPA is well-suited to iEEG tasks and generalizing the model's ability beyond its pre-training dataset. On the remaining two tasks, MVPFormer places second behind PopT, but is still superior to PopT without the electrodes' location. These results indicate that, while the electrodes' position might be beneficial in some instances, MVPA's implicit channel map produces superior results overall by being more flexible and adaptable to a wider variety of existing datasets. We also evaluate the robustness of MVPFormer with respect to the channel selection in Appendix G.14.

6 Related Works

For multi-variate time-series, such as EEG, Transformers face challenges due to the need to preserve both time and channel information [22]. Channel-independent approaches [3] reuse vanilla self-attention and discard all information content in the time dimension, while channel-mixing promises to preserve it by either fusing the channels [23] or processing them sequentially [24]. Specifically for iEEG and EEG, there exist few Transformer-based solutions [21, 19]. Since electrode placements vary widely across subjects, these models struggle with the heterogeneous nature of the data. The complex interplay between time and space, where distant brain regions may be more strongly connected than nearby ones, makes it difficult for conventional attention mechanisms to effectively process iEEG signals.

7 Conclusion

We introduce MVPA, a novel attention mechanism designed to effectively process multi-variate time-series data, and apply it to iEEG signal analysis. MVPA enables MVPFormer, a foundation model trained on our novel Long-term iEEG dataset, to capture complex interactions between time and spatial dimensions in multi-variate time-series. We also contribute the Long-term iEEG dataset itself, as the largest iEEG dataset currently publicly available. MVPFormer is trained following the foundational paradigm to predict the next brain states, and then further fine-tuned on multiple iEEG tasks. MVPFormer achieves SOTA results in several tasks (seizure detection, four decoding tasks) on multiple datasets (our Long-term iEEG dataset, MAYO, FNUSA, and Brain TreeBank).

Limitations While our Long-term iEEG dataset represents a notable step towards increasing the size of iEEG datasets, large language models have clearly shown [25] that ever larger amounts of data are necessary to further develop a foundation model. Moreover, recent works have highlighted the relevance of high-quality data to the performance of the final model [26], also in the (i)EEG domain [27].

8 Broader impact

Effective tools for iEEG analysis provide many practical use-cases [28], including supporting neurologists by processing large amounts of data collected through long-term implantations [29] and wearable devices [30]. Nonetheless, any tool developed for the clinical domain poses risks. For example, over-reliance on automated decision-making could lead to reduced oversight by human experts, diminishing the effectiveness of care.

References

[1] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *Iternational Conference on Learning Representations (ICLR)*, 2024.

[2] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. ZHOU, "Timemixer: Decomposable multiscale mixing for time series forecasting," in *International Conference on Learning Representations (ICLR)*, 2024.

[3] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations (ICLR)*, 2023.

[4] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain*. Oxford University Press, Jan. 2006.

[5] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction — ready for a new era," *Nature Reviews Neurology*, vol. 14, no. 10, 2018.

[6] K.-O. Cho and H.-J. Jang, "Comparison of different input modalities and network structures for deep learning-based seizure detection," *Scientific Reports*, 2020.

[7] P. Thuwajit, P. Rangpong, P. Sawangjai, P. Autthasan, R. Chaisaen, N. Banluesombatkul, P. Boonchit, N. Tatsaringkansakul, T. Sudhawiyangkul, and T. Wilaiprasitporn, "EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, 2022.

[8] C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "BrainBERT: Self-supervised representation learning for intracranial recordings," in *International Conference on Learning Representations (ICLR)*, 2023.

[9] K. Saab, S. Tang, M. Taha, C. Lee-Messer, C. Ré, and D. L. Rubin, "Towards trustworthy seizure onset detection using workflow notes," *npj Digital Medicine*, vol. 7, no. 1, 2024.

[10] P. Nejedly, V. Kremen, V. Sladky, J. Cimbalnik, P. Klimes, F. Plesinger, F. Mivalt, V. Travnicek, I. Viscor, M. Pail, J. Halamek, B. H. Brinkmann, M. Brazdil, P. Jurak, and G. Worrell, "Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals," *Scientific Data*, vol. 7, no. 1, 2020.

[11] C. Wang, A. U. Yaari, A. K. Singh, V. Subramaniam, D. Rosenfarb, J. DeWitt, P. Misra, J. R. Madsen, S. Stone, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "Brain Treebank: Large-scale intracranial recordings from naturalistic language stimuli," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[13] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[14] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[15] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[16] J. Halford, D. Shiau, J. Desrochers, B. Kolls, B. Dean, C. Waters, N. Azar, K. Haas, E. Kutluay, G. Martz, S. Sinha, R. Kern, K. Kelly, J. Sackellares, and S. LaRoche, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in icu eeg recordings," *Clinical Neurophysiology*, vol. 126, no. 9, 2015.

[17] A. C. Grant, S. G. Abdel-Baki, J. Weedon, V. Arnedo, G. Chari, E. Koziorynska, C. Lushbough, D. Maus, T. McSween, K. A. Mortati, A. Reznikov, and A. Omurtag, "EEG interpretation reliability and interpreter confidence: A large single-center study," *Epilepsy & Behavior*, vol. 32, 2014.

[18] M. Quigg, F. Sun, N. B. Fountain, B. C. Jobst, V. S. S. Wong, E. Mirro, S. Brown, and D. C. Spencer, "Interrater reliability in interpretation of electrocorticographic seizure detections of the responsive neurostimulator," *Epilepsia*, vol. 56, no. 6, 2015.

[19] Z. Yuan, F. Shen, M. Li, Y. Yu, C. Tan, and Y. Yang, "Brant-2: Foundation model for brain signals," *arXiv preprint arXiv:2402.1025*, 2024.

[20] G. Chau, C. Wang, S. J. Talukder, V. Subramaniam, S. Soedarmadji, Y. Yue, B. Katz, and A. Barbu, "Population Transformer: Learning population-level representations of neural activity," in *International Conference on Learning Representations (ICLR)*, 2025.
[21] D. Zhang, Z. Yuan, Y. Yang, J. Chen, J. Wang, and Y. Li, "Brant: Foundation model for intracranial neural signal," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
[22] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," *arXiv preprint arXiv:2202.07125*, 2022.
[23] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022.
[24] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *International Conference on Learning Representations (ICLR)*, 2023.
[25] N. Muennighoff, A. M. Rush, B. Barak, T. L. Scao, N. Tazi, A. Piktus, S. Pyysalo, T. Wolf, and C. Raffel, "Scaling data-constrained language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[26] G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. V. Werra, and T. Wolf, "The fineweb datasets: Decanting the web for the finest text data at scale," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.
[27] F. S. Carzaniga, G. T. Hoppeler, M. Hersche, K. Schindler, and A. Rahimi, "The case for cleaner biosignals: High-fidelity neural compressor enables transfer from cleaner iEEG to noisier EEG," in *International Conference on Learning Representations (ICLR)*, 2025.
[28] U. K. Patel, A. Anwar, S. Saleem, P. Malik, B. Rasul, K. Patel, R. Yao, A. Seshadri, M. Yousufuddin, and K. Arumaithurai, "Artificial intelligence as an emerging technology in the current care of neurological disorders," *Journal of Neurology*, vol. 268, no. 5, 2019.
[29] G. K. Bergey, M. J. Morrell, E. M. Mizrahi, A. Goldman, D. King-Stephens, D. Nair, S. Srinivasan, B. Jobst, R. E. Gross, D. C. Shields, G. Barkley, V. Salanova, P. Olejniczak, A. Cole, S. S. Cash, K. Noe, R. Wharen, G. Worrell, A. M. Murro, J. Edwards, M. Duchowny, D. Spencer, M. Smith, E. Geller, R. Gwinn, C. Skidmore, S. Eisenschenk, M. Berg, C. Heck, P. Van Ness, N. Fountain, P. Rutecki, A. Massey, C. O'Donovan, D. Labar, R. B. Duckrow, L. J. Hirsch, T. Courtney, F. T. Sun, and C. G. Seale, "Long-term treatment with responsive brain stimulation in adults with refractory partial seizures," *Neurology*, 2015.
[30] E. Donner, O. Devinsky, and D. Friedman, "Wearable digital health technology for epilepsy," *New England Journal of Medicine*, vol. 390, no. 8, 2024.
[31] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, 2019.
[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.
[34] X. Si, Z. Yang, X. Zhang, Y. Sun, W. Jin, L. Wang, S. Yin, and D. Ming, "Patient-independent seizure detection based on long-term iEEG and a novel lightweight CNN," *Journal of Neural Engineering*, vol. 20, no. 1, 2023.
[35] W. Cui, W. Jeong, P. Thölke, T. Medani, K. Jerbi, A. A. Joshi, and R. M. Leahy, "Neuro-GPT: Developing a foundation model for EEG," *arXiv preprint arXiv:2311.03764*, 2023.
[36] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations (ICLR)*, 2021.
[37] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.
[38] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
[39] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *International Conference on Learning Representations (ICLR)*, 2020.

[40] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, and S. Sanghai, "GQA: Training generalized multi-query transformer models from multi-head checkpoints," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
[41] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
[42] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2024.
[43] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform," *Journal of neuroscience methods*, vol. 123, no. 1, pp. 69–87, 2003.
[44] M. Shen, P. Wen, B. Song, and Y. Li, "An EEG based real-time epilepsy seizure detection approach using discrete wavelet transform and machine learning methods," *Biomedical Signal Processing and Control*, vol. 77, 2022.
[45] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
[46] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.
[47] B. Wang, "Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX," 2021. [Online]. Available: https://github.com/kingoflolz/mesh-transformer-jax
[48] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[49] M. P. Beenhakker and J. R. Huguenard, "Neurons that fire together also conspire together: Is normal sleep circuitry hijacked to generate epilepsy?" *Neuron*, 2009.
[50] H. P. Zaveri, B. Schelter, C. A. Schevon, P. Jiruska, J. G. Jefferys, G. Worrell, A. Schulze-Bonhage, R. B. Joshi, V. Jirsa, M. Goodfellow, C. Meisel, and K. Lehnertz, "Controversies on the network theory of epilepsy: Debates held during the ictals 2019 conference," *Seizure*, vol. 78, 2020.
[51] M. L. Martini, A. A. Valliani, C. Sun, A. B. Costa, S. Zhao, F. Panov, S. Ghatan, K. Rajan, and E. K. Oermann, "Deep anomaly detection of seizures with paired stereoelectroencephalography and video recordings," *Scientific Reports*, vol. 11, no. 1, 2021.
[52] S. Ziyabari, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, "Objective evaluation metrics for automatic classification of EEG events," *arXiv preprint arXiv:1712.10107*, 2017.
[53] V. Shah, I. Obeid, J. Picone, G. Ekladious, R. Iskander, and Y. Roy, "Validation of temporal scoring metrics for automatic seizure detection," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020.
[54] A. Van de Vel, K. Verhaert, and B. Ceulemans, "Critical evaluation of four different seizure detection systems tested on one patient with focal and generalized tonic and clonic seizures," *Epilepsy & Behavior*, vol. 37, 2014.
[55] E. Bruno, P. F. Viana, M. R. Sperling, and M. P. Richardson, "Seizure detection at home: Do devices on the market match the needs of people living with epilepsy and their caregivers?" *Epilepsia*, 2020.
[56] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2022.
[57] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
[58] L. A. Shokooh, D. H. Toffa, P. Pouliot, F. Lesage, and D. K. Nguyen, "Intracranial eeg seizure onset and termination patterns and their association," *Epilepsy Research*, vol. 176, 2021.
[59] J. Gotman, "A few thoughts on "what is a seizure?"," *Epilepsy & Behavior*, vol. 22, pp. S2–S3, Dec. 2011.
[60] R. F. Betzel, J. D. Medaglia, A. E. Kahn, J. Soffer, D. R. Schonhaut, and D. S. Bassett, "Structural, geometric and genetic factors predict interregional brain connectivity patterns probed by electrocorticography," *Nature Biomedical Engineering*, 2019.

[61] J. C. Pang, K. M. Aquino, M. Oldehinkel, P. A. Robinson, B. D. Fulcher, M. Breakspear, and A. Fornito, "Geometric constraints on human brain function," *Nature*, vol. 618, no. 7965, 2023.
[62] H. H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography: 1957," *Electroencephalography and Clinical Neurophysiology*, vol. 10, no. 2, pp. 370–375, 1958.

A Details on multi-variate parallel attention (MVPA)

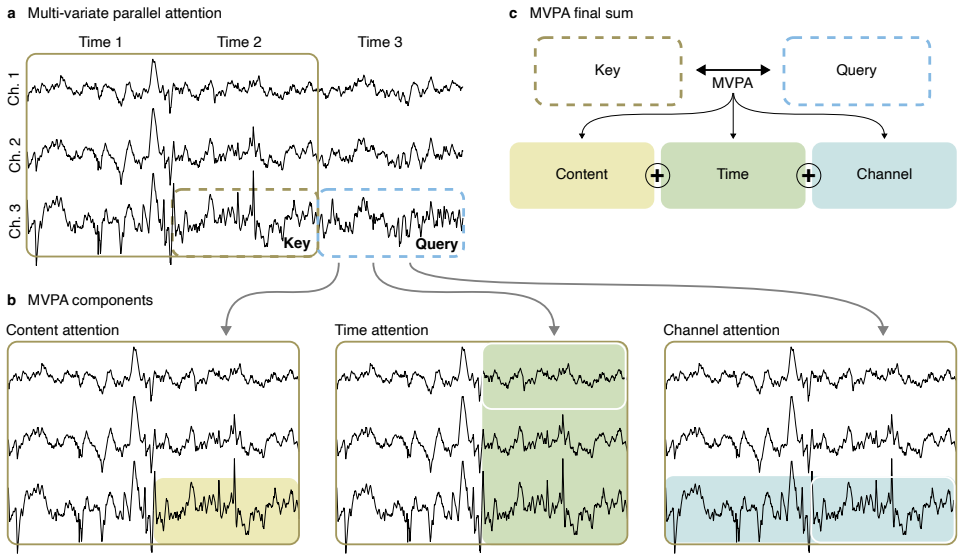

Figure S1: **Multi-variate parallel attention (MVPA). (a)** The input signal is divided into temporal and spatial segments. Each query-key interaction is computed for all keys within a local window. **(b)** MVPA decomposes attention into three components: content-based, computed per segment without positional encoding; time-based, shared across channels and dependent only on temporal distance; and channel-based, shared across time steps and dependent only on spatial distance. **(c)** The final attention is the sum of these three components, each capturing a distinct aspect of the data.

Algorithm 1 illustrates the multi-variate parallel attention algorithm.

A.1 Further motivation of MVPA

Single-channel data can be treated equivalently to sentences, by dividing the signal into 1D patches, which form the tokens. This modality has attracted considerable interest [31, 32], frequently for speech recognition tasks that are once again related to the natural language domain.

There is no straightforward extension of vanilla attention to the 2D case. The Vision Transformer [33] processes images by extending the notion of the patches to the 2D case. It carves images into a collection of patches, which it then flattens into a 1D sequence. Each patch has 2D coordinates (i, j) which get flattened by an arbitrary function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ into a 1D index (l). This is a simple way to recover the 1D case, but it has several drawbacks. First, by flattening the patches we lose any notion of spatial structure, as nearby patches in space are no longer necessarily close in the sequence. Any information about the structure of the patches is lost. However, if the size of the images, the number of patches, and the flattening direction are kept constant, then the Transformer might autonomously learn it. If it learns the structure, then it cannot be exposed to different images as it would completely misinterpret them; if it does not learn the structure, then it is missing critical information. This leads to an inflexible model which cannot easily generalize to different images. One possible solution is to choose a bijective f, such as the Cantor pairing function, to have a one-to-one correspondence between the position of the patch in the image and in the sequence. This solution is, however, quite unintuitive. Second, the Vision Transformer does not distinguish between the two dimensions of height and width, i.e., it does not distinguish between up, down, left, and right. For images this limitation is not too impactful, as most of the information is conveyed in the closeness of two patches and not their relative position in any dimension.

The patching schema of the Vision Transformer is unsuitable to multi-variate time-series, as the two dimensions of time and channels require delicate handling. Transformers for time-series are a well-known problem in the field [22]. Channel-independent approaches [3] reuse vanilla self-attention and discard all information content in the time dimension, while channel-mixing promises to preserve it by either fusing the channels [23] or processing them sequentially [24]. The second

Algorithm 1: Computation of MVPA

Input: $\boldsymbol{x}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ output token of Encoder; $n_{\text{embed}} = 768$
Output: $\boldsymbol{o}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ output attention
Data: $\boldsymbol{t}$ time encoding; $\boldsymbol{c}$ channel encoding; $\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{w}$ biases; $h \in [1, \ldots, n_{\text{heads}}]$; $\boldsymbol{h}_{k,v} \in [1, \ldots, n_{\text{gqa}}]$

def `MVPAttention(`$\boldsymbol{x}_{c,t}$`):`

`# Compute query separately from key and value due to GQA`

$\boldsymbol{q}^h_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{x}_{c,t})$

$\boldsymbol{k}^{h_{k,v}}_{c,t}, \boldsymbol{v}^{h_{k,v}}_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{x}_{c,t})$

`# Compute the three components of MVPA`

$\boldsymbol{g}^h_{(c,t),(c',t')} \leftarrow (\boldsymbol{q}^h_{c,t} + \boldsymbol{u}^{h_{k,v}})^T \boldsymbol{q}^h_{c',t'}$

`# Time and channel components are independent of the key content, so they do not need to be recomputed`

$\boldsymbol{s}^h_t \leftarrow (\boldsymbol{q}^h_{c,t} + \boldsymbol{y}^{h_{k,v}})^T \boldsymbol{t}^{h_{k,v}}$

$\boldsymbol{l}^h_c \leftarrow (\boldsymbol{q}^h_{c,t} + \boldsymbol{w}^{h_{k,v}})^T \boldsymbol{c}^{h_{k,v}}$

`# Shift the time and channel components to avoid recomputation, from Transformer-XL`

$\boldsymbol{s}^h_{t,t'} \leftarrow \text{SHIFTTIME}_{t'}(\boldsymbol{s}^h_t)$

$\boldsymbol{l}^h_{c,c'} \leftarrow \text{SHIFTCHANNEL}_{c'}(\boldsymbol{l}^h_c)$

`# Apply window and causal mask`

$\boldsymbol{m}^h_{(c,t),(c',t')} \leftarrow \text{CAUSALMASK}(\boldsymbol{g}^h_{(c,t),(c',t')} + \boldsymbol{s}^h_{t,t'} + \boldsymbol{l}^h_{c,c'})$

$\boldsymbol{n}^h_{(c,t),(c',t')} \leftarrow \text{WINDOWMASK}_{10}(\boldsymbol{m}^h_{(c,t),(c',t')})$

`# Apply structured dropout`

$\boldsymbol{d}^h_{(c,t),(c',t')} \leftarrow \text{STRUCTUREDDROPOUT}_{0.1}(\boldsymbol{n}^h_{(c,t),(c',t')})$

`# Compute final attention value`

$\boldsymbol{a}^h_{(c,t),(c',t')} \leftarrow \text{SIGMOID}(\frac{1}{\sqrt{n_{\text{embed}}}} \boldsymbol{d}^h_{(c,t),(c',t')})$

$\boldsymbol{o}^h_{c,t} \leftarrow \sum_{c',t'} \boldsymbol{a}^h_{(c,t),(c',t')} \cdot \boldsymbol{v}^{h_{k,v}}_{(c,t),(c',t')}$

return $\boldsymbol{o}^h_{c,t}$

family of solutions is more promising in addressing the issue but is still limited either with respect to computational expense or expressiveness.

EEG signals are multi-variate recordings of the brain. Transformer-based approaches to EEG are sparse [34, 35], due to the often unmanageable complexity of the data. In iEEG recordings, the subjects are implanted with electrodes directly in multiple areas of the brain for the purpose of clinical diagnosis. There is no standardized location, or even number of electrodes, for intracranial implants. This makes iEEG an extremely heterogeneous data modality, intractable for conventional attention approaches. The channels present a fundamental source of information, as electric fields spread in different areas of the brain on different time-scales and with different intensities depending on the strength of the connection between the areas. Moreover, the relationship between brain regions is not always proportional to their spatial closeness, as distant areas might be more strongly connected than close ones. There is a tremendously intricate interplay between space and time, which the Transformer must exploit.

A.2 Comparison with alternative attention mechanisms

We further compare MVPA with other existing alternatives to better characterize the features of MVPA. In particular, we draw our main inspiration for the disentanglement and relative positional encoding from Transformer-XL [13] and DeBERTa [36], which were the first to introduce this concept. We now compare MVPA against a selection of relevant alternative attention mechanisms.

Axial attention [37] consists of two separate attention mechanisms, *RowAttention* and *ColumnAttention*, each of which attends to one row (one channel) or one column (one timepoint) only. The layers

Table S1: **Summary of the differences between MVPA and existing attention mechanisms.**

	Domain	Complexity	Disentangled	Relative position	Simultaneous time and space	Receptive field
Vanilla	1D	Quadratic	No	No	No	Global
Transformer-XL [13]	1D	Quadratic	Yes	Yes	No	Global
DeBERTa [36]	1D	Quadratic	Yes	No	No	Global
Axial [37]	2D	Subquadratic	No	No	No	Local
Criss-cross [38]	2D	Subquadratic	No	No	Yes	Local
Localized [14]	2D	Quadratic	No	No	Yes	Local
Ours	**2D**	**Subquadratic**	**Yes**	**Yes**	**Yes**	**Global**

are then stacked sequentially to recover the full receptive field. MVPA, in contrast, attends to both time and space simultaneously, and has a global receptive field built-in at every leayer.

Criss-cross attention [38] computes the attention between each point and all the other points in its row or column via the *affinity* operation. Once again, the layers are applied recursively to obtain the full receptive field. One of the most signficant differences between MVPA and criss-cross attention is in the encoding of the position. In fact, MVPA treats rows and columns differently through two independent positional codebooks, while in criss-cross attention distance in rows and heights is equivalent. This is a natural consequence of the design choices, as criss-cross attention is designed for images, where the two dimensions are indeed equivalent. Moreover, MVPA again has a global receptive field.

Finally, localized sparse transformers [14] use separate heads with separate connectivity patterns to improve on the computational requirements of the full attention. As before, the full receptive field is only recovered with multiple applications. MVPA, on the other hand, computes the full 2D attention over the entire input in every head. Moreover, the separate positional codebooks allow MVPA to treat the dimensions differently, which localized sparse transformers cannot do.

Table S1 summarizes the main features of MVPA with respect to the considered alternatives.

A.3 Efficient computation of MVPA

Vanilla attention is quadratic in the number of input elements, and this often represents a significant computational roadblock [39]. The input becomes intractable as the number of channels increases, especially for multi-variate time-series. At the same time, more channels imply more sources of information, and we cannot simply discard them.

MVPA is also quadratic, but we employ a number of techniques to significantly reduce the computational complexity and make the processing of very large signals feasible. Letting T be the number of time segments and C be the number of channels, the context length of the Transformer becomes $T \times C$ and number of terms necessary to compute for vanilla attention $O(T^2 \times C^2)$. Given a reasonable estimation of 100 segments and 50 channels the context length would be 5,000, until recently intractable even for language models.

By dividing MVPA into three components we gain considerable advantages (see Table S2 for the complexity of each term). Computing the time- and channel-based terms efficiently requires two main techniques. First, we recognize it is unnecessary to compute the full attention matrix, which is quadratic in the context length (i.e., both time and space). By design, all elements of the time-based attention are the same for each channel (see Figure S1b, the green components are all equal), and all elements of the channel-based attention are the same for each time point (see Figure S1c, the blue components are all equal). Hence, complexity is quadratic in one dimension and constant in the other. We then simply repeat the elements along the right dimension at no additional cost. Second, we employ the shifting operation described in Supplementary Section B of Transformer-XL [13] to compute all relative embeddings in one pass.

Content attention, though stripped of positional encoding, remains the most expensive component. To further reduce computational cost, with little impact to performance, we make use of a local attention window [14] which focuses on the most recent L time points discarding ones that have little information content. Since time-based attention is not limited, the lookup window still spans the entire context (though it is affected, see Figure S18b). Thus, for $L \ll T$, the total complexity of

	Time-based	Channel-based	Content-based (w/ window)	Content-based (vanilla)
Complexity	$O(T^2 \times C)$	$O(T \times C^2)$	$O(L^2 \times C^2)$	$O(T^2 \times C^2)$

Table S2: **Complexity of each component of MVPA.** T is the number of time segments in the signal, C is the number of channels, and L is the size of the local window. Content-based attention without window has the same complexity as vanilla attention.

MVPA is $O(T^2 \times C + T \times C^2)$, quadratic in each dimension but subquadratic in the context length. Combining all techniques, MVPA pushes the effective total context length to over 10,000.

Given the three components are independent of each other, it is possible to exclude any one and reduce computations even more. As an additional cost-saving measure, we use grouped query attention [40] to reduce the number of heads without loss of performance. In summary, MVPA correctly treats time and space as unrelated dimensions, forcing the model to consider them separately, all with little computational overhead.

A.4 Triton implementation of MVPA

MVPFormer's training effectiveness is heavily affected by batch size, as its training routine draws the negative samples from the batch. The bigger the batch size, the more variety in the negative samples and the better the model generalizes. Given the large context size of MVPFormer, up to 10k, a pure Python implementation of scaled dot product attention would consume too much VRAM to be useful. FlashAttention [41] and FlashAttention-2 [42] provide the blueprint to solve this problem, though they only apply to vanilla attention. Using tiling, FlashAttention makes VRAM consumption linear instead of quadratic in the context length, enabling training on much longer context.

We develop FlashMVPA using the same technique in the OpenAI Triton language, which gives lower-level access to CUDA primitives. While a CUDA implementation could likely deliver better raw performance, the choice of Triton is dictated by the much lower coding time, though Triton is less robust and more prone to unexpected behaviors at this point. The time-based and channel-based components of MVPA are computed using PyTorch's own matrix-multiply, but are then shifted (Transformer-XL trick) and added using Triton, while the content-based component is fully implemented in Triton. This is due to limitations in Triton. FlashMVPA reaches 20 TFlops on an A100.

Algorithm 2: Computation of FlashMVPA

Input: $\boldsymbol{x}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ output token of Encoder; $n_{\text{embed}} = 768$
Output: $\boldsymbol{o}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ output attention
Data: $\boldsymbol{t}$ time encoding; $\boldsymbol{c}$ channel encoding; $\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{w}$ biases; $h \in [1, \ldots, n_{\text{heads}}]$; $h_{k,v} \in [1, \ldots, n_{\text{gqa}}]$

def `FlashMVPAttention(`$\boldsymbol{x}_{c,t}$`):`
  `# Compute query separately from key and value due to GQA`
  $\boldsymbol{q}^h_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{x}_{c,t})$
  $\boldsymbol{k}^{h_{k,v}}_{c,t}, \boldsymbol{v}^{h_{k,v}}_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{x}_{c,t})$
  `# Need to compute time and channel components outside Triton`
  $\boldsymbol{s}^h_t \leftarrow (\boldsymbol{q}^h_{c,t} + v^{h_{k,v}})^T \boldsymbol{t}^{h_{k,v}}$
  $\boldsymbol{l}^h_c \leftarrow (\boldsymbol{q}^h_{c,t} + w^{h_{k,v}})^T \boldsymbol{c}^{h_{k,v}}$
  `# Triton MVPA combines all computations into one kernel`
  $\boldsymbol{o}^h_{c,t} \leftarrow \text{TRITONMVPA}(\boldsymbol{q}^h_{c,t}, \boldsymbol{s}^h_t, \boldsymbol{l}^h_t, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{w})$
  return $\boldsymbol{o}^h_{c,t}$

To provide a clearer evaluation of the computational performance benefits of FlashMVPA, we compare the inference speed and VRAM usage of the naïve (purely PyTorch-based) implementation of MVPA and FlashMVPA. We test both implementations with a batch size of 64 and a size of 768, to maintain a realistic scenario. We vary both the number of time windows (T) and the number of channels (C) from 1 to 50, and report the runtime and memory consumption for a forward pass. We test the

attention module in isolation to avoid introducing confounding variables, and we follow the best practices in GPU benchmarking. Table S3 shows the full quadratic scaling of MVPA, which runs out of resources at 50 windows and channels.

Table S3: **Computational performance of MVPA**. Runtime and VRAM consumption of the naïve implementation of MVPA.

T \ C	1	10	20	30	40	50
1	1.21 us / 0.02 GB	1.36 us / 0.04 GB	1.38 us / 0.06 GB	1.44 us / 0.09 GB	1.55 us / 0.11 GB	1.73 us / 0.14 GB
10	1.32 us / 0.04 GB	2.95 us / 0.29 GB	6.24 us / 0.80 GB	11.18 us / 1.55 GB	17.30 us / 2.54 GB	23.10 us / 3.76 GB
20	1.31 us / 0.06 GB	6.25 us / 0.81 GB	17.23 us / 2.53 GB	31.28 us / 5.19 GB	52.11 us / 8.82 GB	75.18 us / 13.39 GB
30	1.34 us / 0.09 GB	11.41 us / 1.55 GB	31.26 us / 5.19 GB	64.52 us / 10.97 GB	115.98 us / 18.88 GB	168.96 us / 28.90 GB
40	1.35 us / 0.11 GB	17.31 us / 2.54 GB	52.03 us / 8.82 GB	115.91 us / 18.88 GB	190.35 us / 32.70 GB	283.36 us / 50.31 GB
50	1.49 us / 0.14 GB	22.97 us / 3.76 GB	75.03 us / 13.39 GB	168.77 us / 28.91 GB	283.26 us / 50.31 GB	OOM

On the other hand, the figures in Table S4 indicate a subquadratic scaling for FlashMVPA, which can easily handle the maximum dimensions of this test.

Table S4: **Computational performance of FlashMVPA**. Runtime and VRAM consumption of the Triton implementation of MVPA.

T \ C	1	10	20	30	40	50
1	1.04 us / 0.02 GB	1.09 us / 0.04 GB	1.14 us / 0.06 GB	1.27 us / 0.07 GB	1.45 us / 0.09 GB	1.71 us / 0.11 GB
10	1.07 us / 0.04 GB	2.36 us / 0.17 GB	5.37 us / 0.32 GB	7.68 us / 0.48 GB	11.18 us / 0.65 GB	11.89 us / 0.81 GB
20	1.08 us / 0.06 GB	4.96 us / 0.32 GB	10.40 us / 0.63 GB	15.11 us / 0.94 GB	20.54 us / 1.27 GB	28.53 us / 1.62 GB
30	1.06 us / 0.07 GB	7.42 us / 0.47 GB	14.43 us / 0.94 GB	25.30 us / 1.43 GB	37.05 us / 1.94 GB	47.04 us / 2.45 GB
40	1.17 us / 0.09 GB	10.76 us / 0.64 GB	20.30 us / 1.27 GB	38.18 us / 1.93 GB	49.75 us / 4.59 GB	79.12 us / 3.32 GB
50	1.33 us / 0.11 GB	11.77 us / 0.80 GB	37.65 us / 1.61 GB	46.53 us / 2.45 GB	77.39 us / 3.32 GB	101.46 us / 4.20 GB

A.5 Relative shifting

By design, MVPA requires the computation of relative time and channel encodings, which can notably slow down the overall operation. While this does not affect vanilla attention, other relative attentions provide us with an elegant solution to this problem. In particular, the shifting operation from Transformer-XL provides us with an efficient alternative to recomputing the time- and channel-based attention components. To keep notation simple, let $\boldsymbol{q}_t = \boldsymbol{x}_{:,t}^T W_q^T$, $\boldsymbol{p}_c = \boldsymbol{x}_{c,:}^T W_q^T$, $\boldsymbol{T}_i = W_{k_t} \mathcal{T}_{T-1-i}$, and $\boldsymbol{C}_i = W_{k_C} \mathcal{C}_{C-1-i}$. The shift in time can be performed as in the original implementation

$$\begin{pmatrix} \boldsymbol{q}_0\boldsymbol{T}_0 & \boldsymbol{q}_0\boldsymbol{T}_1 & \dots & \dots & \boldsymbol{q}_0\boldsymbol{T}_{T-1} \\ \boldsymbol{q}_1\boldsymbol{T}_0 & \boldsymbol{q}_1\boldsymbol{T}_1 & \dots & \dots & \boldsymbol{q}_1\boldsymbol{T}_{T-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{q}_{T-1}\boldsymbol{T}_0 & \boldsymbol{q}_{T-1}\boldsymbol{T}_1 & \dots & \dots & \boldsymbol{q}_{T-1}\boldsymbol{T}_{T-1} \end{pmatrix} \xrightarrow{\text{SHIFTTIME}} \begin{pmatrix} \boldsymbol{q}_0\boldsymbol{T}_{T-1} & 0 & \dots & \dots & 0 \\ \boldsymbol{q}_1\boldsymbol{T}_{T-2} & \boldsymbol{q}_1\boldsymbol{T}_{T-1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{q}_{T-1}\boldsymbol{T}_0 & \dots & \dots & \dots & \boldsymbol{q}_T\boldsymbol{T}_{T-1} \end{pmatrix} \tag{4}$$

The right triangular matrix is zeroed out as a requisite of autoregressive training, i.e., we cannot attend to keys in the future. The entire time shifting operation can be performed efficiently and quickly using tensor manipulation in PyTorch.

Thus, the time attention component does not require recomputation for each time position, i.e., each row in the matrix of the time component.

The shift in channels is more involved

$$\begin{pmatrix} \boldsymbol{p}_0\boldsymbol{C}_0 & \boldsymbol{p}_0\boldsymbol{C}_1 & \dots & \dots & \boldsymbol{p}_0\boldsymbol{C}_{C-1} \\ \boldsymbol{p}_1\boldsymbol{C}_0 & \boldsymbol{p}_1\boldsymbol{C}_1 & \dots & \dots & \boldsymbol{p}_1\boldsymbol{C}_{C-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{p}_{C-1}\boldsymbol{C}_0 & \boldsymbol{p}_{C-1}\boldsymbol{C}_1 & \dots & \dots & \boldsymbol{p}_{C-1}\boldsymbol{C}_{C-1} \end{pmatrix} \xrightarrow{\text{SHIFTCHANNEL}} \begin{pmatrix} \boldsymbol{p}_0\boldsymbol{C}_{C-1} & \boldsymbol{p}_0\boldsymbol{C}_{C-2} & \dots & \dots & \boldsymbol{p}_0\boldsymbol{C}_0 \\ \boldsymbol{p}_1\boldsymbol{C}_{C-2} & \boldsymbol{p}_1\boldsymbol{C}_{C-1} & \boldsymbol{p}_1\boldsymbol{C}_{C-2} & \dots & \boldsymbol{p}_1\boldsymbol{C}_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{p}_{C-1}\boldsymbol{C}_0 & \dots & \dots & \dots & \boldsymbol{p}_C\boldsymbol{C}_{C-1} \end{pmatrix} \tag{5}$$

Here, no element is zeroed out, as all channels can attend to all other channels. The channel shifting operation does not (to our knowledge) enjoy an implementation which as efficient as the time shifting one in PyTorch, but requires relatively complex index manipulation which cannot be streamlined.

As before, thanks to this shifting operation the channel attention component does not require recomputation for each channel position.

We provide a Triton implementation for both operations which is much more efficient and must be preferred when training a model.

A.6 Structured attention dropout

Dropout is a common technique to improve the generalization performance of neural networks. In Transformers, it is often applied inside the attention block to randomly zero-out some query-key attentions, to avoid over-reliance of the model on specific connections.

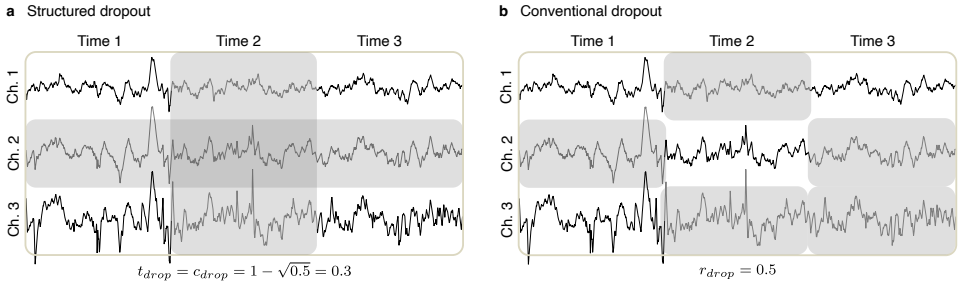

Figure S2: **Structured dropout. (a)** Our structured dropout blanks entire channels and time steps, to reduce the number of correlated segments. The dropout rate is computed to maintain the same number of dropped out segments as conventional dropout. **(b)** Conventional dropout blanks segments randomly. This is less effective with time-series because adjacent segments in time or space contain much of the same information.

Dropout usually applies to all elements with equal probability and creates uniform holes in the attention matrix. This is not efficient in the case of multi-variate time-series, as for each hole the neighboring segments are likely to carry very similar information, reducing dropout's effectiveness. We introduce a structured dropout technique which blanks entire channels and time points instead of individual segments. This technique is in principle much more effective by removing all segments which are more likely to be strongly correlated. We keep the same parameters as in conventional Dropout and compute the channel-specific and time-specific dropout rates as

$$t_{\text{drop}} = c_{\text{drop}} = 1 - \sqrt{1 - r_{\text{drop}}} \tag{6}$$

This ensures that approximately the same overall number of elements are zeroed (see Figure S2).

For the specific dropout rates and the location of the structured dropout layers refer to the description of the architecture in Appendix B.

B MVPFormer architecture

MVPFormer is part of a family of predictive deep learning models with 74 million (MVPFormer-S) to 1.2 billion (MVPFormer-M, or simply MVPFormer for brevity) parameters based on the Transformer [12] architecture, capable of generating iEEG signals. See Tables S5 and S6 for a breakdown of the models sizes and hyperparameters.

Wavelet encoder The first processing step maps the raw iEEG signal to continuous embeddings. We begin by partitioning the raw data into segments of five seconds. Each segment passes independently through a db4 wavelet decomposition, which has been shown to be highly effective for biosignals [43, 44]. Depending on the model's overall size, it is then linearly projected onto a smaller latent space. This projection, or feature vector, is the embedding. Our method is inspired by wav2vec [31], though we use learnable embeddings. We apply the encoding channel-wise, meaning each segment remains one-dimensional.

Decoder MVPFormer is based on the Llama2 architecture [45] with parallel attention and MLP blocks inspired by Megatron-LM [46]. This choice was informed by the selection of a generative model powerful enough to process brain iEEG signals and computationally light enough to enable extensive testing. We provide two models to evaluate the scaling of our foundational model: MVPFormer-S with 75M parameters and MVPFormer-M (or simply MVPFormer) with 1.2B parameters.

Table S5: **Breakdown of the parameters of MVPFormer-S.** The dimensions are indicated for each of the components of MVPFormer-S.

Transformer	Encoder	Signal
$n_{\text{layers}} \leftarrow 12$	$n_{\text{input}} \leftarrow 2560$	$w_{\text{length}} \leftarrow 500\,\text{s}$
$n_{\text{heads}} \leftarrow 12$	$n_{\text{embed}} \leftarrow 768$	$n_{\text{segments}} \leftarrow 100$
$n_{\text{gqa}} \leftarrow 4$		$w_{\text{segment}} \leftarrow 5\text{s}$
$n_{\text{embed}} \leftarrow 768$		$n_{\text{negatives}} \leftarrow 30$
$n_{\text{inner}} \leftarrow 1728$		$n_{\text{local}} \leftarrow 10$
$r_{\text{drop}} \leftarrow 0.1$		

Table S6: **Breakdown of the parameters of MVPFormer-M.** The dimensions are indicated for each of the components of MVPFormer-M.

Transformer	Encoder	Signal
$n_{\text{layers}} \leftarrow 24$	$n_{\text{input}} \leftarrow 2560$	$w_{\text{length}} \leftarrow 500\,\text{s}$
$n_{\text{heads}} \leftarrow 16$	$n_{\text{embed}} \leftarrow 1024$	$n_{\text{segments}} \leftarrow 100$
$n_{\text{gqa}} \leftarrow 8$		$w_{\text{segment}} \leftarrow 5\text{s}$
$n_{\text{embed}} \leftarrow 2048$		$n_{\text{negatives}} \leftarrow 30$
$n_{\text{inner}} \leftarrow 5362$		$n_{\text{local}} \leftarrow 10$
$r_{\text{drop}} \leftarrow 0.1$		

B.1 Inference

The full end-to-end inference procedure is reported in Algorithm 3.

B.2 Encoder

The Encoder block is detailed above. The algorithmic overview is presented in Algorithm 4.

Algorithm 3: Full inference with MVPFormer

Input: $\boldsymbol{x} \in \mathbb{R}^{C \times T}$ raw input; C, T number of channels and length resp.; $n_{\text{layers}} = 12$
Data: $c \in [1, \ldots, C]$; $t \in [1, \ldots, T//n_{\text{segments}} + 1]$
Output: $\boldsymbol{o}_{c,(t-1)} \in \mathbb{R}^{n_{\text{embed}}}$ generated embedding; $n_{\text{embed}} = 768$

def `Inference(`$\boldsymbol{z}_{c,t}$`):`
  $\boldsymbol{x}_{c,t} \leftarrow \text{SEGMENT}(\boldsymbol{x})$
  $\boldsymbol{e}_{c,t} \leftarrow \text{ENCODER}(\boldsymbol{x}_{c,t})$
  for $l \leftarrow 1$ **to** n_{layers} **do**
    $\boldsymbol{e}_{c,t} \leftarrow \text{DECODER}(\boldsymbol{e}_{c,t})$
  $\boldsymbol{o}_{c,(t-1)} \leftarrow \boldsymbol{e}_{c,t}$
  return $\boldsymbol{o}_{c,(t-1)}$

Algorithm 4: Encoder block of MVPFormer

Input: $\boldsymbol{x}_{c,t} \in \mathbb{R}^{n_{\text{input}}}$ raw input segment; $n_{\text{input}} = 2560$; $c \in [1, \ldots, C]$; $t \in [1, \ldots, T]$
Output: $\boldsymbol{o}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ output token; $n_{\text{embed}} = 768$
Data: $l = 8$ maximum decomposition level given n_{input}

def `Encoder(`$\boldsymbol{x}_{c,t}$`):`
  $\boldsymbol{d}_{c,t} \leftarrow \text{DISCRETEWAVELETDECOMPOSITION}_{db4}(\boldsymbol{x}_{c,t}, l)$
  $\boldsymbol{z}_{c,t} \leftarrow \text{RMSNORM}(\boldsymbol{d}_{c,t})$
  $\boldsymbol{o}_{c,t} \leftarrow \text{LINEAR}(\boldsymbol{z}_{c,t})$
  return $\boldsymbol{o}_{c,t}$

B.3 Decoder

The collection of vectors resulting from the Encoder block is flattened into a 1D sequence to provide a unified input interface to the Transformer decoder blocks, consistent with conventional Transformers. All the encoded segments corresponding to a window form the input to the Transformer module, which computes the MVPA among all the segments. The segments are sequentially processed by multiple Transformer layers, composed of attention and MLP blocks in a deep network configuration. The attention blocks are masked to guarantee that MVPFormer only has access to past segments to generate the target. The model produces one output embedding for each input segment. The algorithmic overview is presented in Algorithm 5, while the MLP block in Algorithm 6.

Algorithm 5: Decoder block of MVPFormer

Input: $\boldsymbol{o}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ input tokens; $n_{\text{embed}} = 768$
Output: $\boldsymbol{o}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$

def `Decoder(`$\boldsymbol{o}_{c,t}$`):`
  $\boldsymbol{z}_{c,t} \leftarrow \text{RMSNORM}(\boldsymbol{o}_{c,t})$
  `# Compute attention`
  $\boldsymbol{a}_{c,t} \leftarrow \text{MVPATTENTION}(\boldsymbol{z}_{c,t})$
  $\boldsymbol{d}_{c,t} \leftarrow \text{DROPOUT}(\text{LINEARNOBIAS}((\boldsymbol{a}_{c,t}))$
  `# Compute feedforward residuals in parallel with attention [47]`
  $\boldsymbol{s}_{c,t} \leftarrow \text{MLP}(\boldsymbol{z}_{c,t})$
  `# Sum residuals and attention`
  $\boldsymbol{o}_{c,t} \leftarrow \boldsymbol{o}_{c,t} + \boldsymbol{d}_{c,t} + \boldsymbol{s}_{c,t}$
  return $\boldsymbol{o}_{c,t}$

Algorithm 6: MLP block of MVPFormer

Input: $\boldsymbol{z}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$ normalised Decoder output; $n_{\text{embed}} = 768$
Data: $\boldsymbol{u}_{c,t}, \boldsymbol{g}_{c,t} \in \mathbb{R}^{n_{\text{inner}}}$; $n_{\text{inner}} = 1728$
Output: $\boldsymbol{s}_{c,t} \in \mathbb{R}^{n_{\text{embed}}}$

def `MLP(`$\boldsymbol{z}_{c,t}$`)`**:**
  $\boldsymbol{u}_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{z}_{c,t})$
  $\boldsymbol{g}_{c,t} \leftarrow \text{SILU}(\text{LINEARNOBIAS}(\boldsymbol{z}_{c,t}))$
  $\boldsymbol{s}_{c,t} \leftarrow \text{LINEARNOBIAS}(\boldsymbol{u}_{c,t} + \boldsymbol{g}_{c,t})$
  return $\boldsymbol{s}_{c,t}$

C Details on training

C.1 Generative pre-training

MVPFormer is used to generate neuronal activity while in the base prediction task. During training, the target for each output is the successive input segment in time, not in space. We sample random input segments from the rest of the batched windows to create the confounding targets $Z = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$. These segments still represent actual iEEG signals, so they are plausible, but they are expected to be very different from the true target.

This scheme strikes the correct balance between too much similarity and too little. The objective of MVPFormer is to generate future iEEG signals, so we choose a contrastive loss to increase the cosine similarity of its output with the true target, while decreasing it with the confounding targets. As training progresses, MVPFormer starts to produce outputs that look like encoded segments, i.e., its inputs. MVPFormer becomes more and more capable of choosing the right target and thus is able to predict the future signal.

Extraction of positive and negative examples Out of the entire dataset, B windows are chosen at random to form a batch. Each window $W_{i\in[1..B]}$ has an arbitrary sample rate and C_i channels. First, the sampling rate is normalized to 512 Hz, then the windows are divided into T non-overlapping segments per-channel, leaving us with $C_i \times T$ segments per window. Each segment is passed in parallel through the encoder. For the sake of simplicity, suppose one window W^* (with C^* channels) is selected at random as the positive window, and all the others as the confounding windows. The embeddings of W^* form the input context E with length $C^* \times T$.

For each segment, n embeddings are selected at random from the confounding windows to form the negative samples Z. Each $Z_{c,t}$ has n elements, thus Z has size $C^* \times T \times n$. Z is excluded from backpropagation.

MVPFormer processes the entire E at once and produces an output O also of size $C^* \times T$. We then compute the losses and iteratively optimize to train the model.

Contrastive loss We train MVPFormer using a contrastive loss [48] and an auxiliary loss. To compute the contrastive loss, we rely on having other windows in the batch, so a larger batch size leads in general to a more stable training and better generalization performance. Let $\boldsymbol{e}^i,\ i \in [1, \ldots, B]$ be the outputs of the signal Encoder and $\boldsymbol{o}^i,\ i \in [1, \ldots, B]$ the outputs of the Decoder stack, for B the batch size. For each i^*, we select at random $n_{\text{negatives}}$ elements from $\boldsymbol{e}^i, i \neq i^*$ to act as our negative samples $\boldsymbol{n}^{i*}$. Clearly, the bigger the batch the greater the entropy. We compute the contrastive loss for each i^* as follows:

$$\mathcal{L}^i_{c,t} = -\log \frac{\exp(\text{sim}(\boldsymbol{o}_{c,t}, \boldsymbol{e}_{c(t+1)})/\tau)}{\sum_{\boldsymbol{z}_k \in Z} \exp(\text{sim}(\boldsymbol{o}_{c,t}, \boldsymbol{z}_k)/\tau)} \tag{7}$$

Finally summing over every i, c, t gives us the optimization target for the generative task.

The loss is invariant to the channel c, which encourages all the outputs to be the same regardless of channel. The temperature τ is 0.1.

C.2 Generation of neuronal activity

The generation of brain signals during inference proceeds analogously as during training. However, we do not have access to the same source of entropy as in training since the batches are limited to one subject at a time. This limitation implies that the evaluation scores of MVPFormer must be more punishing than the training objective, since we cannot reliably estimate the accuracy with which MVPFormer chooses the right target. For this reason, we measure the cosine similarity directly in a three-way reference scheme. First, we consider the cosine similarity of the output with the true target. Second, we consider the similarity with the maximally correlated target. Third, we measure the cosine similarity with the highest form of entropy available, random segments in the batch that are still close by in time. This measurement ensures that the difference in similarity between the true and confounding targets remains significant.

D Long-term iEEG dataset

The Long-term iEEG dataset is presented in Section 4. Due to institutional data privacy concerns, the dataset does not contain information about the location of the channels in the brain. All the subjects gave written informed consent that their iEEG data might be used for research and teaching purposes. The decision on the necessity for iEEG recordings, the electrode implantation scheme, and the decision about surgical therapy were made entirely on clinical grounds. These decisions were taken prior to and completely independently from the compilation of this dataset. This dataset may only be used for research. For other applications any liability is denied. In particular, the dataset must not be used for diagnostic purposes.

Here, Table S7 shows the full details of the dataset in a subject-by-subject breakdown. Finally, Figure S3 shows two annotated seizures in the dataset.

Table S7: **Per-subject details of our Long-term iEEG dataset.** Ch. is the number of electrodes, f_s is the sampling frequency in Hz, Rec. [h] is the length of the recording in hours, and Ev. is the number of seizures. The entire dataset contains 68 subjects, 9328 hours of recording and 704 ictal events.

Subject	Ch.	f_s [Hz]	Rec [h]	Ev.	Subject	Ch.	f_s [Hz]	Rec [h]	Ev.	Subject	Ch.	f_s [Hz]	Rec [h]	Ev.
ID01	88	512	293.4	2	ID24	32	1024	40.7	14	ID47	32	1024	330.4	3
ID02	66	512	235.2	2	ID25	128	512	109.4	4	ID48	57	1024	28.4	6
ID03	64	512	158.4	4	ID26	34	1024	87.6	1	ID49	60	512	140.4	6
ID04	32	1024	40.7	14	ID27	32	1024	146	8	ID50	64	1024	177.2	2
ID05	128	512	109.4	4	ID28	75	512	69	4	ID51	89	512	161.5	1
ID06	32	1024	146	8	ID29	61	1024	143.8	70	ID52	69	512	112.6	2
ID07	75	512	69	4	ID30	48	1024	40.9	27	ID53	22	1024	134.9	1
ID08	61	1024	143.8	70	ID31	32	1024	42.4	17	ID54	54	1024	202	3
ID09	48	1024	40.9	27	ID32	32	1024	212.2	2	ID55	24	1024	152.1	2
ID10	32	1024	42.4	17	ID33	104	512	53.6	1	ID56	62	1024	130.5	3
ID11	32	1024	212.2	2	ID34	56	1024	191.4	9	ID57	40	1024	90.7	12
ID12	56	1024	191.4	9	ID35	64	1024	104	7	ID58	92	512	138.2	7
ID13	64	1024	104	7	ID36	24	1024	161.4	60	ID59	54	1024	107.3	15
ID14	24	1024	161.4	60	ID37	98	512	195.9	2	ID60	74	512	50.7	8
ID15	98	512	195.9	2	ID38	34	1024	177.1	5	ID61	76	512	89.6	6
ID16	34	1024	177.1	5	ID39	60	1024	129.6	2	ID62	60	1024	235.1	7
ID17	60	1024	129.6	2	ID40	42	1024	205.1	5	ID63	64	512	179.8	4
ID18	42	1024	205.1	5	ID41	33	1024	82.7	3	ID64	56	1024	36.3	20
ID19	29	1024	21.7	25	ID42	63	1024	87.8	2	ID65	49	1024	139.7	8
ID20	88	512	293.4	2	ID43	126	512	63.2	2	ID66	39	1024	212.3	2
ID21	66	512	235.2	2	ID44	60	1024	150.3	2	ID67	63	512	111.7	4
ID22	64	512	158.4	4	ID45	47	1024	157.3	1	ID68	32	1024	167.8	3
ID23	32	1024	42.4	33	ID46	86	512	140.5	21					

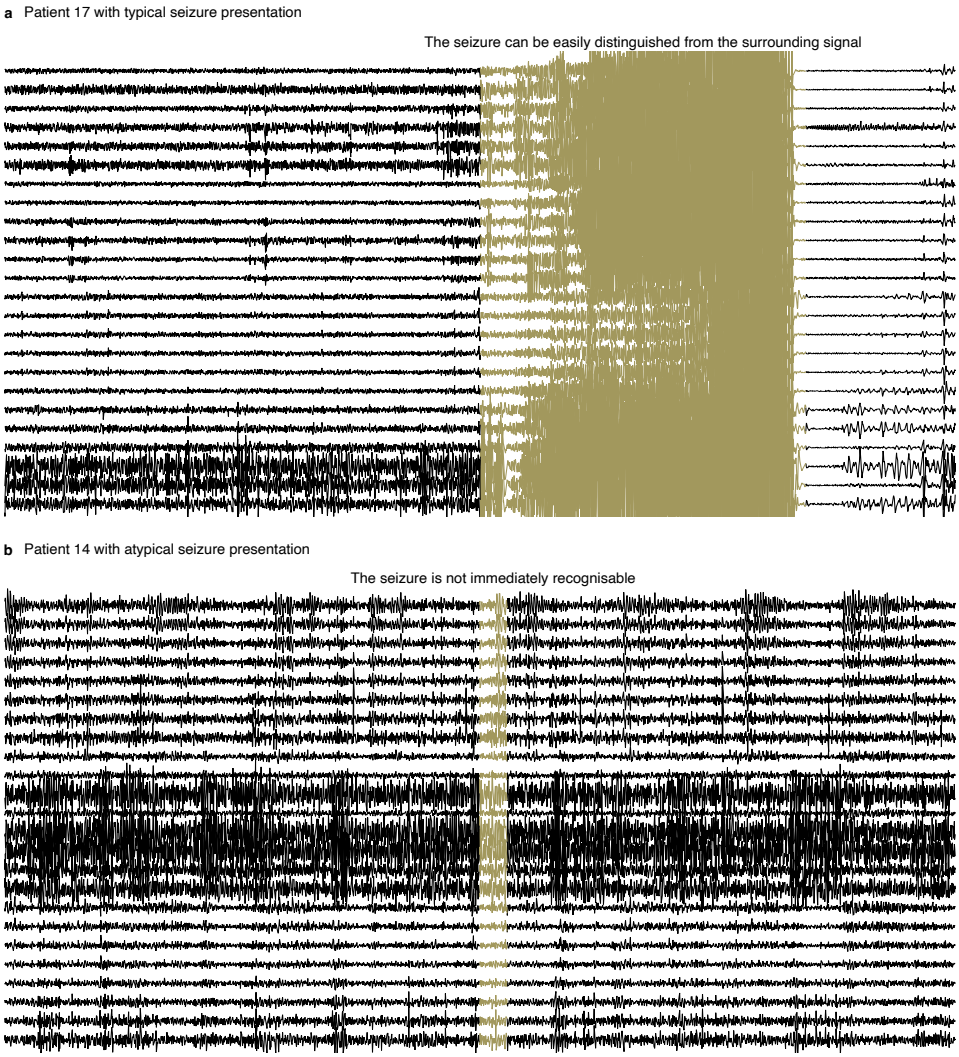

Figure S3: **iEEG activity of two patients with different ictal patterns. (a)** Patient 17 of the Long-term iEEG dataset presents typical ictal events. The seizure can be clearly distinguished even by a non-expert, and MVPFormer performs very well on this patient. The number of channels is reduced from the original recording to facilitate comparison with the more difficult presentation. **(b)** Patient 14 of the Long-term iEEG dataset does not have typical events. The neuronal activity during seizures for this patient cannot be clearly distinguished, and assessment by experts would diverge considerably. As expected, MVPFormer has a high level of disagreement on this patient. All the channels of the original recording are presented to exclude the chance of some channels carrying additional information.

E Generation of results

E.1 Pre-training

We divide each recording into windows of 500 seconds each, with a stride of five seconds. Then, each window is divided into 100 segments (each five seconds long), yielding a total of 39B total training segments of which 390M are unique. We pre-train MVPFormer on 18 subjects, leaving the remaining 50 subjects for testing. MVPFormerM is pre-trained for 1.2M steps on a single node with 8 NVIDIA A100-80GB GPUs for two weeks. The chosen optimizer is FusedAdam with 0.1 weight decay, from the Deepspeed library compiled on the specific machine. The training used bf16-mixed DeepSpeed stage 2 without activation checkpointing. The learning rate is fixed to 10^{-4}. The training environment includes PyTorch 2.0, PyTorch Lightning 2.0, and Triton 2.1.0.

E.2 Fine-tuning

After pre-training, we further fine-tune the MVPFormer for each task. For the seizure detection task, we fine-tune on the same 18 subjects of the Long-term iEEG dataset, and then test in a zero-shot manner on nearly 7,000 hours of iEEG data from 50 unseen subjects, all of them suffering from epilepsy. This allows us to assemble a rigorous set of results which are relevant to a real-world application. Moreover, to keep computational cost moderate, we use a subset of the channels of each subject for testing. We select the channels based on a visual inspection by a non-expert to include only the least noisy ones. In a real-world clinical scenario selection and validation would comprise a minimal additional burden for the expert. The number of the chosen channels and their position varies for each subject, from 30 up to 50 channels (see Appendix G.9).

For the four tasks of the Brain TreeBank dataset, we follow the same procedure as BrainBERT [8] and PopT [20] by dividing the data into a training and testing split. As before, we also evaluate the robustness of MVPFormer with respect to the channel selection in Appendix G.14.

E.3 Episodic seizure post-processing

For episodic evaluation we apply three post-processing steps to the model output:

- Merge events happening within 5 minutes of each other
- Remove events shorter than 20 seconds in length
- Remove events with less than 5 positive responses

Moreover, when the subject has multiple seizures in one minute we merge them into one.

E.4 Online seizure thresholding

In the clinical evaluation setup we apply a simple thresholding to decide whether to report a seizure or not. We set 3 positive seconds out of 10 to be the lower limit for detecting a seizure, to deter false positives; events shorter than 3 seconds are thus not reported, and an additional latency of 3 seconds is to be considered. We find this trade-off has limited drawbacks in practice, as there is often large disagreement even among neurologists about very short events.

E.5 Kappa score estimation

To estimate the Kappa score, we choose 300 random segments per subject to compare their classification from MVPFormer and the labels. We perform multiple iterations to ensure no bias in this computation. Figure S4 indicates that our choice of 250 iterations is sufficient for stable results.

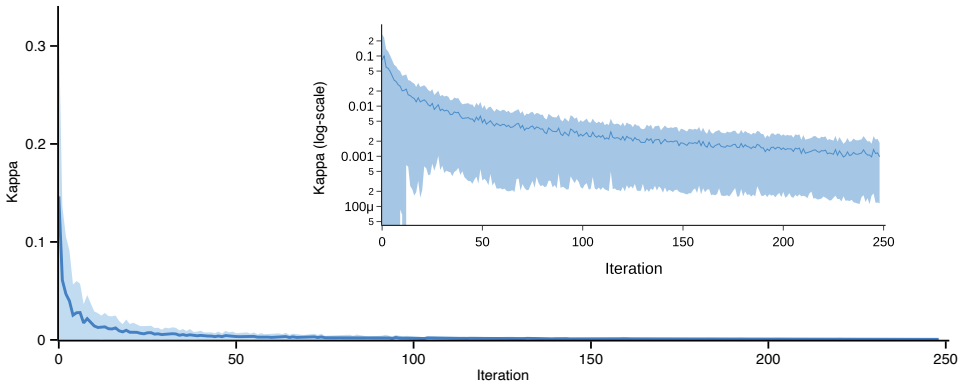

Figure S4: **Mean absolute error in Cohen's kappa estimation.** Our estimation scheme for Cohen's kappa converges after very few iterations. The error is computed per-subject as the absolute difference between the running averages at each two consecutive iterations; the running average is the average of all preceding steps. The average and standard deviation across all subjects is reported here. We compute up to 250 random iterations to ensure precise reporting.

F Prediction of iEEG signals

We evaluate the effects of the size of the model and the attention mechanism on the iEEG prediction task. Figure S5 shows that both MVPA and the vanilla attention are effective at predicting the next brain states. Scaling up the model size from MVPFormer-S to MVPFormer-M has the effect of shortening the tail of the true distribution, effectively increasing the concentration of the cosine similarity towards the maximum.

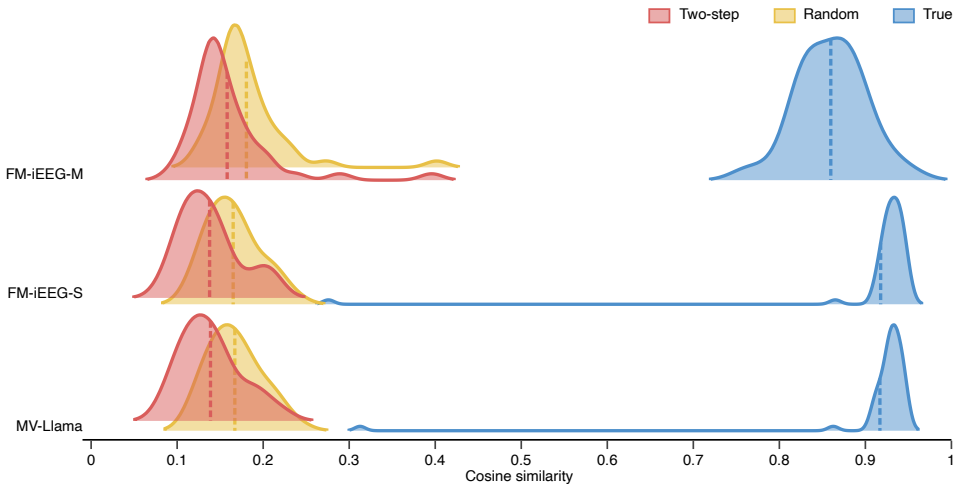

Figure S5: **Performance of MVPFormer-M, MVPFormer-S, and MV-Llama on the iEEG prediction task.** We report the average two-step, random, and true target cosine similarities for the three different models. All three are effective at predicting iEEG activity, while the larger model takes advantage of the increased embedding size by increasing the concentration of cosine similarities towards 1.

Complete details are available for the iEEG prediction performance of MVPFormer-M (see Appendix F.1, MVPFormer-S (see Appendix F.2), and MV-Llama (see Appendix F.3).

F.1 Prediction of iEEG and ictal activity

MVPFormer is primarily a neuronal prediction model, trained to generate neuronal activity regardless of whether such activity is pathological or physiological. To understand the behavior of MVPFormer with anomalous brain states, we evaluate its performance in generating ictal neuronal activity. The precise relationship between ictal and interictal states is a point of ongoing discussion [49, 50], but many consider an approach to seizures as anomalies [51] the most appropriate. The Long-term iEEG dataset contains many ictal events, so we are able to evaluate the performance of MVPFormer in generating anomalous activity. In particular, in this dataset the ratio between non-ictal and ictal states is approximately 500:1.

Figure S6 shows that ictal states are not anomalous for MVPFormer. In particular, the prediction similarity of MVPFormer does not degrade when generating ictal activity. Moreover, the prediction similarity in the ictal state is neither significantly different from the average similarity nor from the non-ictal similarity. This indicates that MVPFormer's understanding of the mechanisms of generation of neuronal activity encompasses the pathological ictal state as well. Therefore, MVPFormer must model patterns found both in physiological and pathological brain states. Finally, MVPFormer incorporates a model of seizure generation as a by-product of its predictive task, which is particularly noteworthy.

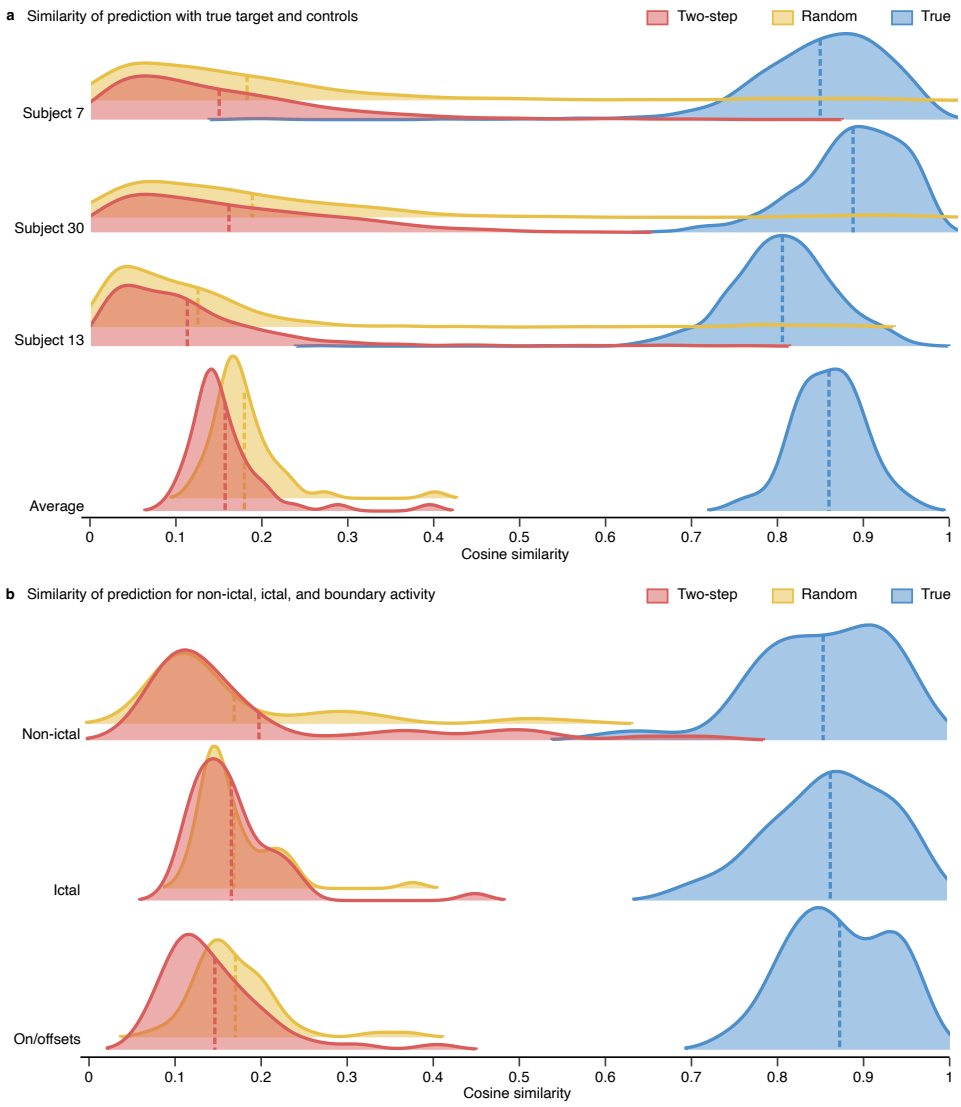

Figure S6: **Performance of MVPFormer on the prediction task. (a)** A three-reference evaluation scheme is used to assess MVPFormer's performance. The true target is the immediate future, i.e., the next five seconds of iEEG signal. The two-step target is the signal twice removed in the future, i.e., the five seconds of iEEG signal coming after the true target. Finally, the random target is chosen from iEEG signals which are close by in time with the true target. The distribution of the average similarity across the entire recording is shown together with the similarity within three representative subjects (with maximum, median, and minimum average similarity). **(b)** The prediction similarity is computed again for all three targets, distinguishing between targets which lie within an ictal event, without, or at the boundary. There is no significant difference in the performance of MVPFormer in predicting ictal or non-ictal activity, indicating that MVPFormer can encompass anomalous brain states as well, together with the transitions between physiological and anomalous.

F.2 Effects of the scale of the model

Figure S7 shows the full details on the performance of MVPFormer-S on the Long-term iEEG dataset in the iEEG prediction task.

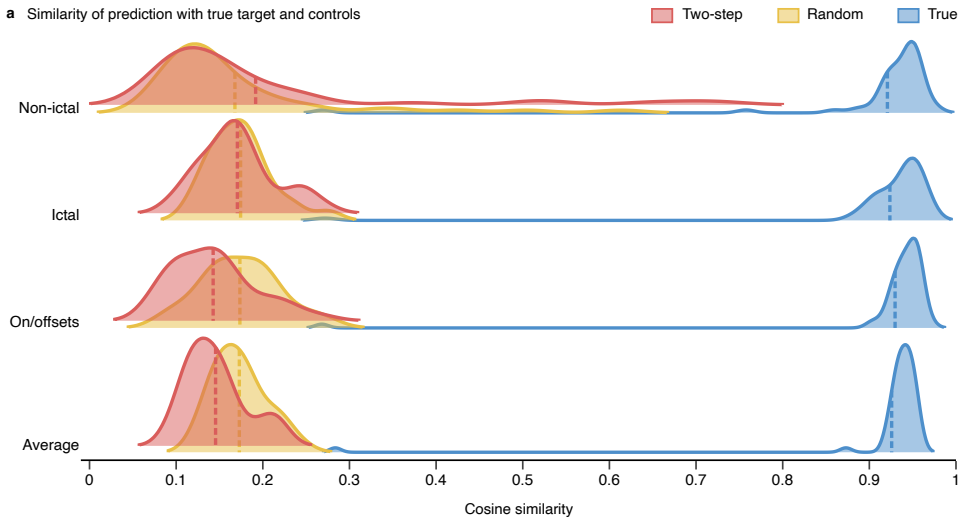

Figure S7: **Performance of MVPFormer-S on the prediction task.** Performance of MVPFormer-S in generating neuronal activity of unseen test subjects. The prediction similarity is computed for all three targets, distinguishing between targets which lie within an ictal event or without. There is no significant difference in the performance of MVPFormer-S in predicting ictal or non-ictal activity.

F.3 Effects of the attention mechanism

Figure S8 shows that vanilla attention is also effective in predicting the development of iEEG signal.

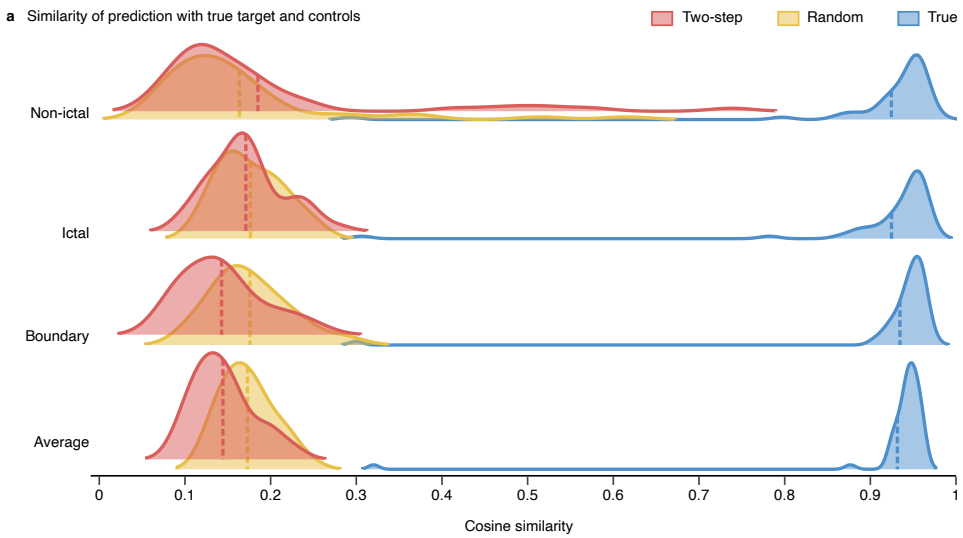

Figure S8: **Performance of MV-Llama on the prediction task.** The prediction similarity is computed for all three targets, distinguishing between targets which lie within an ictal event or without. There is no significant difference in the performance of MV-Llama in predicting ictal or non-ictal activity. This indicates that vanilla attention with a proper positional encoding scheme can effectively generate neuronal activity. However, this does not translate to improved performance in the seizure classification task (see Table S12 vs. Table S11)

F.4 Per-subject cosine similarity

We provide a detailed per-subject breakdown of the maximum cosine similarity measure for MVP-Former. Figure S9 shows the per-patient global similarity. Figure S10 shows the per-patient similarity within an anomaly. Figure S11 shows the per-patient similarity at the boundary of an anomaly.

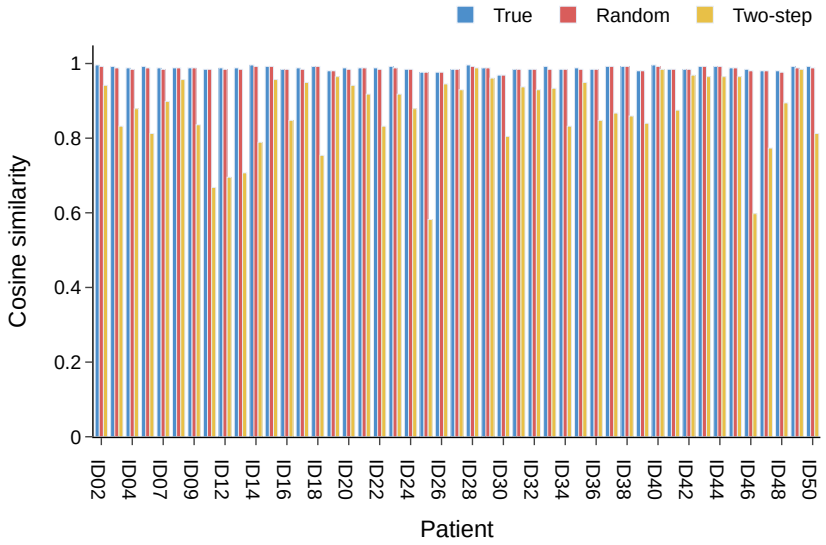

Figure S9: **Breakdown of total cosine similarity per-patient.** Maximum cosine similarity of MVPFormer's output with the true, random, and two-step targets over the entire Long-term iEEG dataset. The data is shown patient-by-patient.

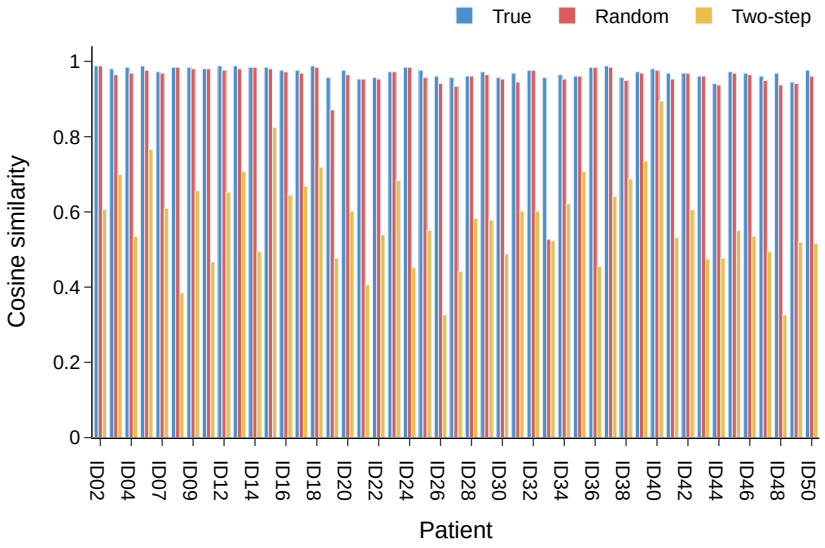

Figure S10: **Breakdown of anomaly cosine similarity per-patient.** Maximum cosine similarity of MVPFormer's output with the true, random, and two-step targets while within an anomaly (seizure) in the Long-term iEEG dataset. The data is shown patient-by-patient.

Figure S11: **Breakdown of boundary cosine similarity per-patient.** Maximum cosine similarity of MVPFormer's output with the true, random, and two-step targets while in the onset and offset zones of seizures in the Long-term iEEG dataset. The data is shown patient-by-patient.

G Additional results

G.1 Seizure detection

Table S8 reports a summary of the seizure detection results across all datasets and architectures.

Table S8: **Results on seizure detection.** We compare MVPFormer with Brant-2, the current SOTA Transformer model for iEEG, and MV-Llama, our vanilla attention-based baseline.

		Long-term					MAYO			FNUSA		
		Episodic				Raw	Raw			Raw		
Model	Attention	**Kappa**	f1	sens	fp/h	f1	f1	sens	spec	f1	sens	spec
MVPFormer	MVPA	**0.57**	**0.56**	0.73	0.17	**0.50**	**0.36**	0.38	0.91	**0.46**	0.94	0.10
MVPFormer-S	MVPA	0.54	0.51	0.69	0.11	0.49	0.35	0.41	0.88	**0.46**	0.99	0.03
Brant-2 [19]	Vanilla	0.08	0.00	0.00	0.06	0.00	0.19	1.00	0.18	**0.46**	0.99	0.02
MV-Llama	Vanilla	0.05	0.02	0.01	0.00	0.01	/	/	/	/	/	/

G.2 Brain TreeBank decoding tasks

Table S9 reports a summary of the decoding tasks of the Brain TreeBank dataset.

Table S9: **Results on Brain TreeBank iEEG tasks.** We compare MVPFormer with multiple Transformer-based architectures on the four tasks of the Brain TreeBank dataset [11]. The best results without electrode's position are bolded, while the results where the electrodess position is beneficial are underlined.

Model	Attention	Electrode location	Pitch	Volume	Onset	Speech
MVPFormer	MVPA	No	**0.83** (0.02)	**0.88** (0.01)	**0.87** (0.02)	**0.90** (0.02)
MV-Llama	Vanilla	No	0.62 (0.03)	0.77 (0.02)	0.80 (0.03)	0.81 (0.02)
Brant [21]	Vanilla	No	0.61 (0.03)	0.74 (0.03)	0.80 (0.04)	0.80 (0.03)
PopT w/o encoding [20]	Vanilla	No	0.62 (0.07)	0.76 (0.07)	0.81 (0.09)	0.83 (0.10)
PopT (BrainBERT) [20]	Vanilla	Yes	0.74 (0.03)	0.87 (0.03)	0.90 (0.01)	0.93 (0.02)
PopT (TOTEM) [20]	Vanilla	Yes	0.64 (0.03)	0.79 (0.02)	0.90 (0.02)	0.88 (0.05)

G.3 Conventional evaluation

In addition to our clinically motivated evaluation (see Section 5.1), we assess all our models using conventional machine learning metrics for seizure detection: F1-score, sensitivity, and false positive rate. These metrics are commonly used in benchmarking seizure detection models [52, 53], and allow comparison with prior work. The full seizure detection results of MVPFormer are shown in Figure S12.

We evaluate against two baselines: Brant-2 [19], a SOTA Transformer model for iEEG, and MV-Llama, an ablation of MVPFormer-S that uses standard attention instead of MVPA (see Appendix G.5). Brant-2 is fine-tuned with its published pre-trained weights and protocol. MV-Llama is trained identically to MVPFormer-S.

We report both raw and episodic metrics. Episodic metrics reflect clinically meaningful detections by grouping predictions into events [52]. The detailed results are provided in Tables S11 and S10.

The similarity between raw and episodic F1-scores suggests that MVPFormer naturally learns to detect seizure episodes of realistic length and frequency. On the 50-subject Long-term test set, the false positive rate is 0.17 fp/h (0.11 for MVPFormer-S), comparable to commercial EEG devices used in clinical practice [54, 55]. As expected in a real-world dataset, false positive rates vary considerably across subjects, with 78% having fewer than 0.05 fp/h.

These results confirm that MVPFormer performs competitively on conventional seizure detection benchmarks, while also offering robust generalization to clinically realistic evaluation settings.

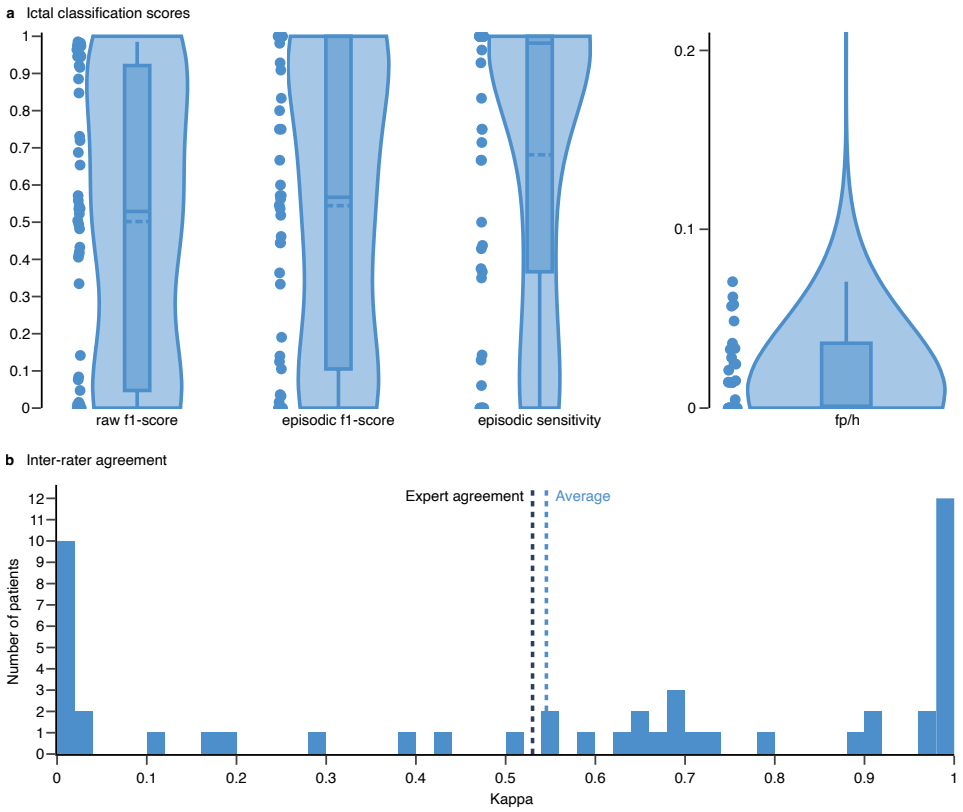

Figure S12: **Performance of MVPFormer on the classification task. (a)** The F1-score, sensitivity, and fp/h are reported. Raw results are computed without any post-processing of MVPFormer's output, while episodic results follow a common post-processing procedure which merge close ictal classifications. **(b)** Cohen's Kappa is used to measure the agreement between the artificial assistant and the human expert. The average kappa is 0.57, competitive with the values obtained between human experts. The distribution of kappa values clearly indicates that a minority of subjects are the source of most disagreement, consistent with the variability of inter-rater agreement among human experts.

Table S10: **Details of seizure detection results of MVPFormer with 18 subject pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task. The similarity reports the breakdown of the cosine similarity in each of the considered scenarios.

		Classification metrics				**Similarity**								
		Raw	Episodic			True			Random			Two-step		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h	average	ictal	non-ictal	average	ictal	non-ictal	average	ictal	non-ictal
ID19	0.01	0.01	0.14	0.44	3.55	0.30	0.27	0.25	0.14	0.15	0.10	0.16	0.14	0.10
ID20	1	0.95	1	1	0	0.89	0.95	0.91	0.17	0.19	0.12	0.14	0.17	0.12
ID21	0.99	0.96	1	1	0	0.92	0.92	0.93	0.16	0.15	0.16	0.13	0.22	0.18
ID22	0.99	0.95	1	1	0	0.89	0.92	0.90	0.15	0.15	0.10	0.12	0.15	0.08
ID23	0.1	0.14	0.11	0.06	0.07	0.88	0.95	0.94	0.22	0.22	0.20	0.20	0.21	0.20
ID24	0.92	0.92	0.93	0.93	0.02	0.80	0.93	0.89	0.13	0.13	0.09	0.14	0.11	0.14
ID25	0	0	0	0	0	0.84	0.90	0.90	0.17	0.24	0.14	0.15	0.20	0.13
ID26	0.01	0	0	0	0.59	0.88	0.96	0.94	0.16	0.21	0.09	0.13	0.16	0.10
ID27	1	0.95	1	1	0	0.88	0.95	0.95	0.20	0.23	0.20	0.17	0.23	0.21
ID28	0.75	0.85	0.75	0.75	0.01	0.85	0.89	0.61	0.26	0.14	0.07	0.16	0.15	0.10
ID29	0.29	0.33	0.19	0.13	0.05	0.86	0.92	0.92	0.14	0.14	0.11	0.11	0.12	0.13
ID30	0.96	0.92	0.98	0.96	0	0.89	0.95	0.94	0.17	0.17	0.13	0.14	0.16	0.14
ID31	0.98	0.96	1	1	0	0.88	0.95	0.95	0.18	0.21	0.13	0.15	0.22	0.17
ID32	1	0.97	1	1	0	0.93	0.93	0.94	0.17	0.15	0.12	0.13	0.16	0.11
ID33	0	0	0	0	0.37	0.95	0.96	0.95	0.39	0.38	0.31	0.39	0.45	0.46
ID34	0.98	0.97	1	1	0	0.89	0.94	0.92	0.14	0.14	0.09	0.11	0.14	0.12
ID35	0.98	0.94	1	1	0	0.86	0.93	0.94	0.18	0.15	0.25	0.17	0.17	0.35
ID36	0.58	0.73	0.54	0.37	0	0.91	0.95	0.94	0.19	0.20	0.13	0.16	0.18	0.14
ID37	1	0.97	1	1	0	0.77	0.74	0.66	0.16	0.12	0.29	0.14	0.15	0.49
ID38	0.98	0.98	1	1	0	0.82	0.85	0.77	0.15	0.13	0.13	0.13	0.14	0.09
ID39	1	0.98	1	1	0	0.83	0.70	0.83	0.16	0.12	0.10	0.14	0.12	0.10
ID40	1	0.99	1	1	0	0.87	0.88	0.90	0.15	0.13	0.09	0.13	0.13	0.11
ID41	0.68	0.72	0.67	1	0.04	0.86	0.79	0.86	0.16	0.15	0.10	0.13	0.15	0.10
ID42	0.99	0.89	1	1	0	0.82	0.80	0.79	0.15	0.14	0.07	0.13	0.15	0.07
ID43	0.92	0.69	1	1	0	0.80	0.77	0.84	0.13	0.14	0.12	0.10	0.15	0.11
ID44	0	0	0	0	0	0.82	0.89	0.89	0.11	0.14	0.49	0.08	0.09	0.65
ID45	0.02	0.02	0.04	1	0.34	0.82	0.87	0.77	0.14	0.16	0.09	0.12	0.14	0.16
ID46	0.5	0.5	0.56	0.43	0.01	0.85	0.87	0.61	0.18	0.14	0.28	0.16	0.11	0.51
ID47	0.72	0.54	0.46	1	0.02	0.84	0.81	0.83	0.16	0.15	0.12	0.14	0.16	0.09
ID48	0.78	0.65	0.91	0.83	0	0.80	0.83	0.79	0.12	0.12	0.08	0.10	0.11	0.09
ID49	0.52	0.54	0.6	1	0.06	0.89	0.79	0.86	0.23	0.14	0.35	0.23	0.17	0.53
ID50	0.67	0.57	0.44	1	0.03	0.86	0.90	0.84	0.16	0.24	0.11	0.14	0.21	0.10
ID51	0.74	0.41	0.33	1	0.02	0.87	0.78	0.89	0.26	0.14	0.28	0.28	0.24	0.39
ID52	0.44	0.48	0.36	1	0.06	0.82	0.84	0.79	0.21	0.18	0.11	0.19	0.16	0.14
ID53	0	0.01	0	1	3.56	0.82	0.77	0.79	0.17	0.17	0.09	0.14	0.15	0.11
ID54	0.01	0.01	0.02	0.67	1.21	0.82	0.71	0.83	0.16	0.14	0.12	0.13	0.14	0.09
ID55	0.63	0.49	0.44	1	0.03	0.84	0.87	0.80	0.18	0.18	0.11	0.16	0.19	0.14
ID56	0.62	0.52	0.57	0.67	0.02	0.87	0.81	0.79	0.19	0.18	0.12	0.17	0.20	0.18
ID57	0	0	0	0	0	0.87	0.87	0.81	0.16	0.21	0.14	0.13	0.17	0.17
ID58	0.18	0.08	0.12	0.14	0.06	0.86	0.84	0.77	0.21	0.22	0.33	0.20	0.24	0.30
ID59	0	0	0	0	0	0.82	0.85	0.80	0.17	0.14	0.09	0.16	0.12	0.11
ID60	0.16	0.08	0.55	0.38	0	0.87	0.88	0.93	0.15	0.14	0.56	0.12	0.11	0.37
ID61	0	0	0	0	0	0.89	0.87	0.89	0.18	0.16	0.11	0.15	0.13	0.11
ID62	0.7	0.41	0.83	0.71	0	0.83	0.86	0.92	0.14	0.13	0.52	0.13	0.12	0.72
ID63	0.41	0.43	0.57	1	0.03	0.82	0.86	0.86	0.15	0.16	0.09	0.13	0.15	0.06
ID64	0.52	0.56	0.52	0.35	0	0.86	0.87	0.85	0.15	0.15	0.16	0.12	0.14	0.10
ID65	0.72	0.54	0.75	0.75	0.01	0.83	0.87	0.85	0.13	0.15	0.09	0.10	0.13	0.08
ID66	0.85	0.42	0.8	1	0	0.75	0.80	0.76	0.12	0.13	0.06	0.10	0.11	0.08
ID67	0.04	0.05	0.03	0.5	1.06	0.83	0.82	0.81	0.20	0.15	0.07	0.19	0.12	0.10
ID68	0.66	0.00	0.00	0.00	0.01	0.84	0.80	0.82	0.17	0.22	0.15	0.14	0.19	0.16

G.4 Effects of the scale of the model

The performance improvements of LLMs as a function of their model sizes have also been widely reported [56, 57]. According to Chinchilla's scaling law the training dataset is already not large enough to fully train MVPFormer-S (75M parameters), so we investigate whether a larger model (MVPFormer-M, 1.2B parameters) can provide any improvement in performance.

Figure S13 shows the seizure detection performance of MVPFormer-S on the Long-term iEEG dataset (see Table S11). As noted in the main results, MVPFormer-M marginally improves seizure detection results over MVPFormer-S. In particular, it reaches higher F1-score but higher fp/h rate as well, with small net improvement. Therefore, we have shown that the amount of iEEG data currently available is not sufficient to fully take advantage of the increase in model size of Transformers. We hope that making the Long-term iEEG dataset publicly available will increase overall availability and unlock further model scaling potential.

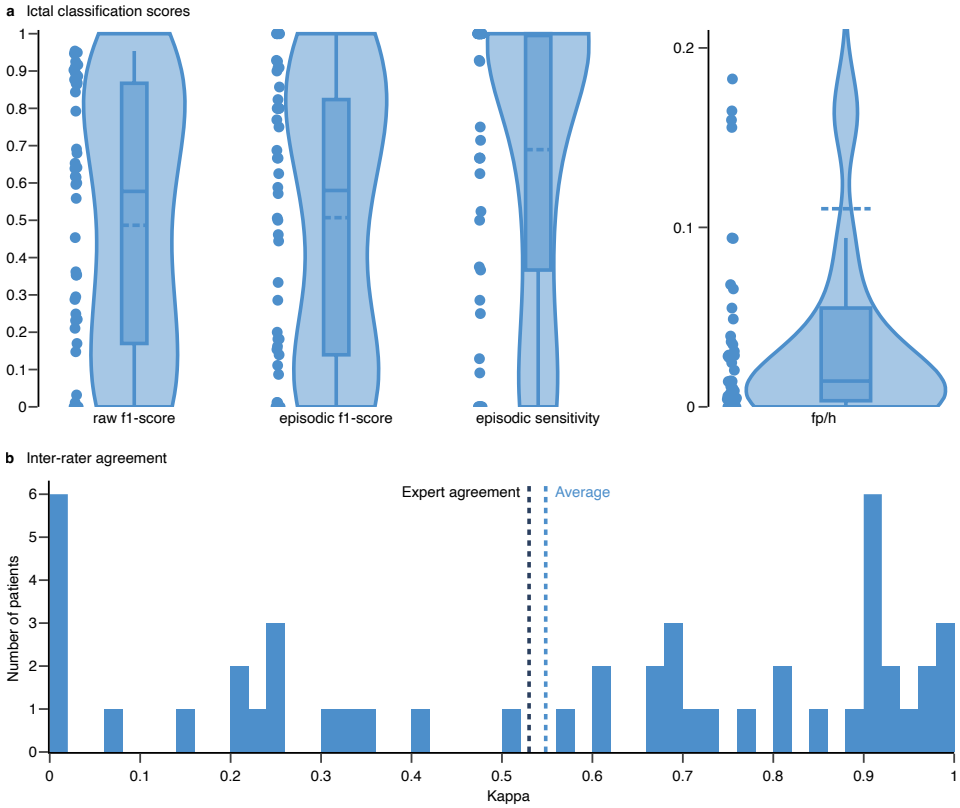

Figure S13: **Performance of MVPFormer-S on the classification task. (a)** Seizure detection results of MVPFormer-S on unseen subjects: the F1-score, sensitivity, and fp/h are reported. Raw results are computed without any post-processing of MVPFormer's output, while episodic results follow a common post-processing procedure which merge close ictal classifications. **(b)** Cohen's Kappa is used to measure the agreement between MVPFormer-S and the human expert. The average kappa is 0.54, competitive with the values obtained between human experts. The distribution of kappa values clearly indicates that a minority of subjects are the source of most disagreement, consistent with the variability of inter-rater agreement among human experts.

Table S11: **Details of seizure detection results of MVPFormer-S with 18 subject pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task. The similarity reports the breakdown of the cosine similarity in each of the considered scenarios.

		Classification metrics				**Similarity**								
		Raw	Episodic			True			Random			Two-step		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h	average	ictal	non-ictal	average	ictal	non-ictal	average	ictal	non-ictal
ID19	-0.05	0	0	0	0.97	0.28	0.27	0.27	0.21	0.2	0.25	0.21	0.2	0.23
ID20	0.98	0.87	0.8	1	0	0.95	0.96	0.95	0.18	0.21	0.12	0.15	0.17	0.12
ID21	0.92	0.86	0.8	1	0	0.94	0.93	0.95	0.16	0.18	0.18	0.13	0.24	0.2
ID22	0.98	0.89	1	1	0	0.94	0.89	0.95	0.15	0.16	0.09	0.12	0.17	0.07
ID23	0.26	0.17	0.14	0.09	0.16	0.95	0.95	0.95	0.25	0.28	0.23	0.22	0.24	0.24
ID24	0.91	0.9	0.93	0.93	0.02	0.93	0.94	0.93	0.19	0.19	0.12	0.16	0.17	0.17
ID25	0.35	0	0	0	0.03	0.94	0.96	0.96	0.17	0.19	0.15	0.14	0.18	0.15
ID26	0.08	0.03	0.11	1	0.18	0.95	0.97	0.96	0.17	0.22	0.12	0.14	0.18	0.11
ID27	0.99	0.93	1	1	0	0.95	0.96	0.96	0.22	0.24	0.21	0.19	0.25	0.24
ID28	0.71	0.68	0.75	0.75	0.01	0.95	0.96	0.95	0.17	0.19	0.09	0.13	0.17	0.12
ID29	0.25	0.29	0.2	0.13	0.03	0.95	0.95	0.95	0.16	0.15	0.14	0.12	0.13	0.13
ID30	0.92	0.88	0.93	0.93	0.05	0.95	0.96	0.96	0.19	0.19	0.15	0.16	0.18	0.16
ID31	0.98	0.95	1	1	0	0.95	0.96	0.96	0.19	0.24	0.11	0.16	0.24	0.18
ID32	1	0.92	1	1	0	0.95	0.91	0.96	0.18	0.13	0.16	0.15	0.18	0.12
ID33	0	0	0	0	0.26	0.93	0.95	0.93	0.23	0.28	0.19	0.21	0.26	0.21
ID34	0.89	0.91	0.9	1	0.01	0.95	0.95	0.95	0.15	0.15	0.1	0.11	0.15	0.13
ID35	0.85	0.87	0.82	1	0.03	0.92	0.94	0.92	0.17	0.19	0.19	0.15	0.17	0.22
ID36	0.5	0.62	0.51	0.37	0.03	0.96	0.96	0.95	0.22	0.23	0.14	0.19	0.21	0.17
ID37	0.74	0.69	0.5	1	0.02	0.92	0.91	0.91	0.16	0.15	0.33	0.14	0.17	0.69
ID38	0.94	0.95	0.91	1	0.01	0.93	0.95	0.91	0.14	0.15	0.11	0.11	0.16	0.08
ID39	0.92	0.89	0.8	1	0.01	0.93	0.9	0.92	0.16	0.13	0.11	0.14	0.13	0.11
ID40	0.95	0.95	0.91	1	0	0.94	0.95	0.95	0.15	0.16	0.1	0.13	0.14	0.12
ID41	0.61	0.65	0.67	1	0.04	0.95	0.91	0.94	0.16	0.16	0.14	0.13	0.16	0.11
ID42	0.86	0.84	0.8	1	0.01	0.93	0.91	0.92	0.15	0.17	0.08	0.12	0.16	0.08
ID43	0.9	0.64	1	1	0	0.93	0.9	0.95	0.14	0.14	0.1	0.11	0.16	0.12
ID44	0	0	0	0	0	0.93	0.95	0.92	0.12	0.12	0.51	0.09	0.1	0.64
ID45	0.01	0	0.01	1	1.06	0.94	0.95	0.93	0.15	0.18	0.13	0.12	0.14	0.14
ID46	0.6	0.56	0.69	0.52	0	0.94	0.95	0.76	0.18	0.15	0.36	0.16	0.12	0.53
ID47	0.26	0.29	0.16	1	0.09	0.94	0.93	0.94	0.18	0.2	0.12	0.15	0.18	0.12
ID48	0.68	0.62	0.8	0.67	0	0.93	0.93	0.92	0.12	0.11	0.06	0.1	0.12	0.09
ID49	0.85	0.79	0.86	1	0.01	0.95	0.92	0.93	0.22	0.14	0.26	0.21	0.19	0.51
ID50	0.69	0.6	0.44	1	0.03	0.93	0.95	0.94	0.17	0.19	0.12	0.14	0.23	0.11
ID51	0.35	0.23	0.15	1	0.07	0.93	0.91	0.93	0.23	0.15	0.2	0.23	0.28	0.26
ID52	0.23	0.35	0.18	1	0.16	0.93	0.94	0.95	0.16	0.18	0.1	0.13	0.16	0.1
ID53	0.12	0.21	0.09	1	0.16	0.93	0.9	0.93	0.19	0.17	0.13	0.15	0.16	0.12
ID54	0.01	0.01	0.01	0.67	1.72	0.94	0.88	0.94	0.18	0.19	0.14	0.14	0.16	0.1
ID55	0.37	0.35	0.29	1	0.07	0.94	0.94	0.94	0.2	0.18	0.14	0.17	0.2	0.17
ID56	0.72	0.6	0.67	0.67	0.01	0.94	0.91	0.88	0.19	0.18	0.13	0.17	0.19	0.16
ID57	0	0	0	0	0	0.94	0.94	0.89	0.14	0.13	0.14	0.11	0.09	0.16
ID58	0.22	0.25	0.18	0.29	0.09	0.94	0.94	0.86	0.21	0.22	0.18	0.2	0.24	0.15
ID59	0	0	0	0	0	0.92	0.94	0.91	0.14	0.14	0.11	0.12	0.12	0.09
ID60	0.33	0.23	0.46	0.38	0.04	0.95	0.96	0.98	0.14	0.14	0.61	0.11	0.12	0.37
ID61	0	0	0	0	0	0.96	0.96	0.96	0.19	0.17	0.16	0.16	0.16	0.14
ID62	0.69	0.45	0.77	0.71	0	0.92	0.93	0.97	0.14	0.14	0.43	0.11	0.11	0.73
ID63	0.94	0.9	1	1	0	0.93	0.94	0.95	0.15	0.17	0.09	0.12	0.16	0.07
ID64	0.6	0.64	0.62	0.5	0.06	0.95	0.95	0.95	0.16	0.17	0.13	0.13	0.14	0.11
ID65	0.5	0.35	0.59	0.62	0.03	0.94	0.95	0.94	0.14	0.17	0.08	0.11	0.14	0.09
ID66	0.7	0.36	0.57	1	0.01	0.87	0.92	0.92	0.13	0.13	0.08	0.1	0.13	0.08
ID67	0.26	0.15	0.33	0.25	0.01	0.94	0.94	0.94	0.22	0.17	0.11	0.21	0.13	0.1
ID68	0.65	0.00	0.00	0.00	0.01	0.94	0.92	0.95	0.15	0.13	0.31	0.10	0.10	0.10

G.5 Effects of the attention mechanism

To assess the validity of our MVPA scheme, we train MV-Llama, a model almost equivalent to MVPFormer-S that uses vanilla attention instead of MVPA. While MV-Llama uses vanilla attention, it is still based on the SOTA Llama2 architecture. We also re-use the vanilla positional encoding, with a simple adjustment to recover a one-to-one correspondence between the positional encoding and the position of the patch in the time-series (the Cantor pairing function, see App. A.1).

Table S8 in the main text indicates that vanilla attention does not perform seizure detection at a level comparable to MVPA. In particular, the performance of MV-Llama is poor, indicating that it cannot generalize to this task. We argue this is due to higher flexibility of the internal representations generated by MVPA, which better lend themselves to further tasks, such as seizure classification.

Table S12: **Details of seizure detection results of MV-Llama with 18 subject pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task. The similarity reports the breakdown of the cosine similarity in each of the considered scenarios.

		Classification metrics				**Similarity**								
		Raw	Episodic			True			Random			Two-step		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h	average	ictal	non-ictal	average	ictal	non-ictal	average	ictal	non-ictal
ID19	0.00	0.01	0.12	0.06	0.00	0.32	0.31	0.29	0.23	0.23	0.23	0.23	0.22	0.22
ID20	0.00	0.00	0.00	0.00	0.00	0.96	0.97	0.95	0.18	0.21	0.14	0.15	0.18	0.13
ID21	0.00	0.00	0.00	0.00	0.00	0.95	0.90	0.96	0.16	0.15	0.15	0.13	0.23	0.20
ID22	0.00	0.00	0.00	0.00	0.00	0.95	0.78	0.95	0.15	0.13	0.08	0.12	0.16	0.07
ID23	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.25	0.26	0.20	0.22	0.24	0.24
ID24	0.00	0.00	0.00	0.00	0.00	0.94	0.94	0.93	0.19	0.20	0.19	0.16	0.18	0.17
ID25	0.00	0.00	0.00	0.00	0.00	0.95	0.96	0.96	0.17	0.19	0.16	0.14	0.18	0.15
ID26	0.00	0.00	0.00	0.00	0.00	0.96	0.97	0.96	0.18	0.22	0.10	0.14	0.18	0.12
ID27	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.22	0.25	0.19	0.19	0.25	0.24
ID28	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.17	0.19	0.08	0.13	0.17	0.13
ID29	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.96	0.16	0.15	0.09	0.12	0.13	0.13
ID30	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.19	0.20	0.15	0.16	0.19	0.17
ID31	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.20	0.24	0.16	0.16	0.24	0.18
ID32	0.00	0.00	0.00	0.00	0.00	0.95	0.88	0.96	0.18	0.16	0.16	0.15	0.18	0.12
ID33	0.00	0.00	0.00	0.00	0.00	0.93	0.95	0.92	0.23	0.24	0.19	0.21	0.25	0.20
ID34	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.96	0.15	0.16	0.12	0.11	0.15	0.12
ID35	0.00	0.00	0.00	0.00	0.00	0.93	0.94	0.92	0.17	0.18	0.16	0.14	0.17	0.18
ID36	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.96	0.22	0.24	0.16	0.19	0.21	0.17
ID37	0.00	0.00	0.00	0.00	0.00	0.92	0.88	0.87	0.16	0.15	0.36	0.14	0.16	0.57
ID38	0.30	0.00	0.00	0.00	0.01	0.93	0.95	0.91	0.14	0.16	0.10	0.11	0.16	0.07
ID39	0.00	0.00	0.00	0.00	0.00	0.93	0.87	0.94	0.16	0.14	0.08	0.13	0.13	0.11
ID40	0.00	0.00	0.00	0.00	0.00	0.94	0.95	0.95	0.15	0.17	0.10	0.12	0.15	0.12
ID41	0.00	0.00	0.00	0.00	0.00	0.96	0.92	0.95	0.16	0.15	0.13	0.13	0.15	0.11
ID42	0.00	0.00	0.00	0.00	0.00	0.94	0.92	0.94	0.15	0.18	0.09	0.12	0.17	0.08
ID43	0.00	0.00	0.00	0.00	0.00	0.94	0.91	0.96	0.15	0.11	0.12	0.11	0.17	0.12
ID44	0.00	0.00	0.00	0.00	0.00	0.94	0.96	0.86	0.12	0.12	0.37	0.09	0.10	0.50
ID45	0.00	0.00	0.00	0.00	0.00	0.94	0.95	0.93	0.15	0.15	0.12	0.12	0.14	0.15
ID46	0.00	0.00	0.00	0.00	0.00	0.94	0.96	0.80	0.18	0.14	0.28	0.16	0.12	0.53
ID47	0.74	0.00	0.00	0.00	0.00	0.95	0.94	0.95	0.18	0.20	0.14	0.15	0.18	0.13
ID48	0.00	0.00	0.00	0.00	0.00	0.93	0.94	0.93	0.12	0.12	0.07	0.09	0.12	0.09
ID49	0.00	0.00	0.00	0.00	0.00	0.95	0.93	0.92	0.21	0.14	0.29	0.20	0.19	0.46
ID50	0.00	0.00	0.00	0.00	0.00	0.94	0.96	0.94	0.17	0.21	0.12	0.14	0.23	0.11
ID51	0.00	0.00	0.00	0.00	0.00	0.93	0.90	0.92	0.22	0.15	0.18	0.21	0.28	0.21
ID52	0.60	0.09	0.67	0.50	0.00	0.94	0.95	0.95	0.16	0.17	0.12	0.13	0.17	0.10
ID53	0.00	0.00	0.00	0.00	0.00	0.95	0.93	0.94	0.19	0.20	0.11	0.16	0.17	0.12
ID54	0.00	0.00	0.00	0.00	0.00	0.94	0.89	0.95	0.18	0.16	0.12	0.15	0.16	0.10
ID55	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.95	0.20	0.21	0.13	0.17	0.20	0.17
ID56	0.00	0.00	0.00	0.00	0.00	0.94	0.91	0.89	0.19	0.18	0.15	0.16	0.18	0.15
ID57	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.90	0.14	0.13	0.10	0.10	0.09	0.14
ID58	0.00	0.00	0.00	0.00	0.00	0.94	0.94	0.88	0.21	0.22	0.15	0.19	0.23	0.17
ID59	0.00	0.00	0.00	0.00	0.00	0.93	0.94	0.93	0.14	0.15	0.06	0.11	0.12	0.10
ID60	0.00	0.00	0.00	0.00	0.00	0.95	0.96	0.98	0.14	0.15	0.62	0.11	0.12	0.41
ID61	0.00	0.00	0.00	0.00	0.00	0.96	0.96	0.97	0.19	0.20	0.18	0.16	0.16	0.15
ID62	0.00	0.00	0.00	0.00	0.00	0.92	0.93	0.97	0.13	0.13	0.52	0.10	0.10	0.74
ID63	0.00	0.00	0.00	0.00	0.00	0.94	0.95	0.95	0.16	0.18	0.11	0.12	0.16	0.07
ID64	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.96	0.16	0.18	0.12	0.13	0.15	0.11
ID65	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.95	0.14	0.15	0.11	0.11	0.14	0.09
ID66	0.74	0.00	0.00	0.00	0.00	0.88	0.93	0.92	0.12	0.14	0.08	0.09	0.13	0.08
ID67	0.00	0.00	0.00	0.00	0.00	0.95	0.95	0.94	0.22	0.17	0.09	0.20	0.14	0.10
ID68	0.00	0.00	0.00	0.00	0.00	0.86	0.91	0.89	0.16	0.17	0.10	0.20	0.16	0.08

G.6 Seizure detection with Brant-2

Figure S14 shows the full distribution of the results of Brant-2 on the seizure detection task.

Table S13 presents the detailed subject-by-subject breakdown of the performance of Brant-2.

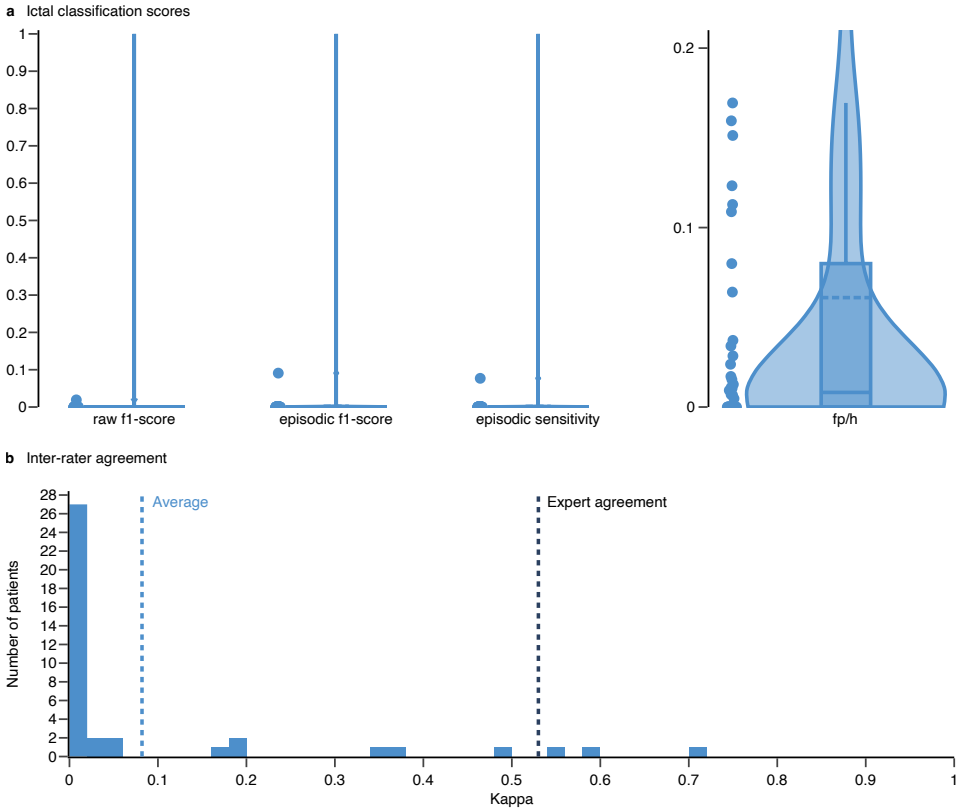

Figure S14: **Performance of Brant-2 on the classification task. (a)** The F1-score, sensitivity, and fp/h are reported. The results are computed using the same schema as MVPFormer. Brant-2, however, is not capable of generalizing to this dataset. **(b)** Cohen's kappa is used to measure the agreement between the artificial assistant and the human expert. The average kappa is 0.08, not competitive with MVPFormer.

Table S13: **Details of seizure detection results of Brant-2 with 18 subject pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID19	N.A.	N.A.	N.A.	N.A.	N.A.
ID20	0.47	0	0	0	0.02
ID21	0	0	0	0	0
ID22	0.4	0	0	0	0.01
ID23	0	0	0	0	0
ID24	-0.02	0	0	0	0.32
ID25	0	0	0	0	0
ID26	0	0	0	0	0
ID27	0	0	0	0	0.12
ID28	-0.01	0	0	0	0.49
ID29	0.01	0	0	0	0.01
ID30	0	0	0	0	0
ID31	0	0	0	0	0
ID32	0	0	0	0	0
ID33	0	0	0	0	0
ID34	0.64	0	0	0	0.01
ID35	0	0	0	0	0
ID36	N.A.	N.A.	N.A.	N.A.	N.A.
ID37	0	0	0	0	0
ID38	0.01	0	0	0	0.17
ID39	0	0	0	0	0
ID40	0	0	0	0	0.29
ID41	0	0	0	0	0.11
ID42	0.07	0	0	0	0.16
ID43	0	0	0	0	0
ID44	0.74	0	0	0	0.01
ID45	0	0	0	0	0
ID46	0.01	0	0	0	0.06
ID47	0.02	0	0	0	0.15
ID48	0	0	0	0	0
ID49	0.14	0	0	0	0.03
ID50	0	0	0	0	0.11
ID51	0	0	0	0	0.04
ID52	0	0	0	0	0.08
ID53	N.A.	N.A.	N.A.	N.A.	N.A.
ID54	0	0	0	0	0
ID55	N.A.	N.A.	N.A.	N.A.	N.A.
ID56	0.56	0	0	0	0.02
ID57	0	0	0	0	0.01
ID58	0	0	0	0	0
ID59	0.14	0	0	0	0.01
ID60	0	0	0	0	0
ID61	0	0	0	0	0
ID62	0.16	0	0	0	0.03
ID63	0	0	0	0	0
ID64	0.07	0.02	0.09	0.08	0.22
ID65	0	0	0	0	0
ID66	0	0	0	0	0.29
ID67	0	0	0	0	0
ID68	0	0	0	0	0

G.7 Seizure detection with BrainBERT

Table S14 presents the detailed subject-by-subject breakdown of the performance of BrainBERT.

Table S14: **Details of seizure detection results of BrainBERT with 18 subject pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID19	0.00	0.00	0.00	0.00	0.00
ID20	0.00	0.00	0.00	0.00	0.00
ID21	0.00	0.00	0.00	0.00	0.00
ID22	0.00	0.00	0.00	0.00	0.00
ID23	0.00	0.00	0.00	0.00	0.00
ID24	0.00	0.00	0.00	0.00	0.00
ID25	0.00	0.00	0.00	0.00	0.00
ID26	0.00	0.00	0.00	0.00	0.00
ID27	0.00	0.00	0.00	0.00	0.00
ID28	0.00	0.00	0.00	0.00	0.00
ID29	0.00	0.00	0.00	0.00	0.00
ID30	0.00	0.00	0.00	0.00	0.00
ID31	0.00	0.00	0.00	0.00	0.00
ID32	0.00	0.00	0.00	0.00	0.00
ID33	0.00	0.00	0.00	0.00	0.00
ID34	0.00	0.00	0.00	0.00	0.00
ID35	0.00	0.00	0.00	0.00	0.00
ID36	0.00	0.00	0.00	0.00	0.00
ID37	0.00	0.00	0.00	0.00	0.00
ID38	0.00	0.00	0.00	0.00	0.00
ID39	0.00	0.00	0.00	0.00	0.00
ID40	0.00	0.00	0.00	0.00	0.00
ID41	0.00	0.00	0.00	0.00	0.00
ID42	0.00	0.00	0.00	0.00	0.00
ID43	0.00	0.00	0.00	0.00	0.00
ID44	0.00	0.00	0.00	0.00	0.00
ID45	0.00	0.00	0.00	0.00	0.00
ID46	0.00	0.00	0.00	0.00	0.00
ID47	0.00	0.00	0.00	0.00	0.00
ID48	0.00	0.00	0.00	0.00	0.00
ID49	0.00	0.00	0.00	0.00	0.00
ID50	0.00	0.00	0.00	0.00	0.00
ID51	0.00	0.00	0.00	0.00	0.00
ID52	0.00	0.00	0.00	0.00	0.00
ID53	0.00	0.00	0.00	0.00	0.00
ID54	0.00	0.00	0.00	0.00	0.00
ID55	0.00	0.00	0.00	0.00	0.00
ID56	0.00	0.00	0.00	0.00	0.00
ID57	0.00	0.00	0.00	0.00	0.00
ID58	0.00	0.00	0.00	0.00	0.00
ID59	0.00	0.00	0.00	0.00	0.00
ID60	0.00	0.00	0.00	0.00	0.00
ID61	0.00	0.00	0.00	0.00	0.00
ID62	0.00	0.00	0.00	0.00	0.00
ID63	0.00	0.00	0.00	0.00	0.00
ID64	0.00	0.00	0.00	0.00	0.00
ID65	0.00	0.00	0.00	0.00	0.00
ID66	0.00	0.00	0.00	0.00	0.00
ID67	0.00	0.00	0.00	0.00	0.00
ID68	0.00	0.00	0.00	0.00	0.00

G.8 Effects of the scale of the pre-training dataset

The performance of LLMs as the size of their training dataset increases has been investigated quite thoroughly [56, 57], giving rise to a variety of scaling laws. Following Chinchilla's scaling law, a model with 75 million parameters like MVPFormer-S should be trained with around 2 billion tokens, while we only have 400 million at our disposal.

The architecture of the model and the nature of the training data, however, make it unclear whether such laws can be adopted for MVPFormer as well. We investigate this behavior by continuing the training of MVPFormer-S on 40 more subjects, to bring the total to 58 pre-training subjects for almost 7,000 hours of iEEG recordings. In particular, MVPFormer is initially trained on 304 ictal events, and then further on 323 more. Therefore, we are left with 10 unseen subject to test the downstream seizure detection task.

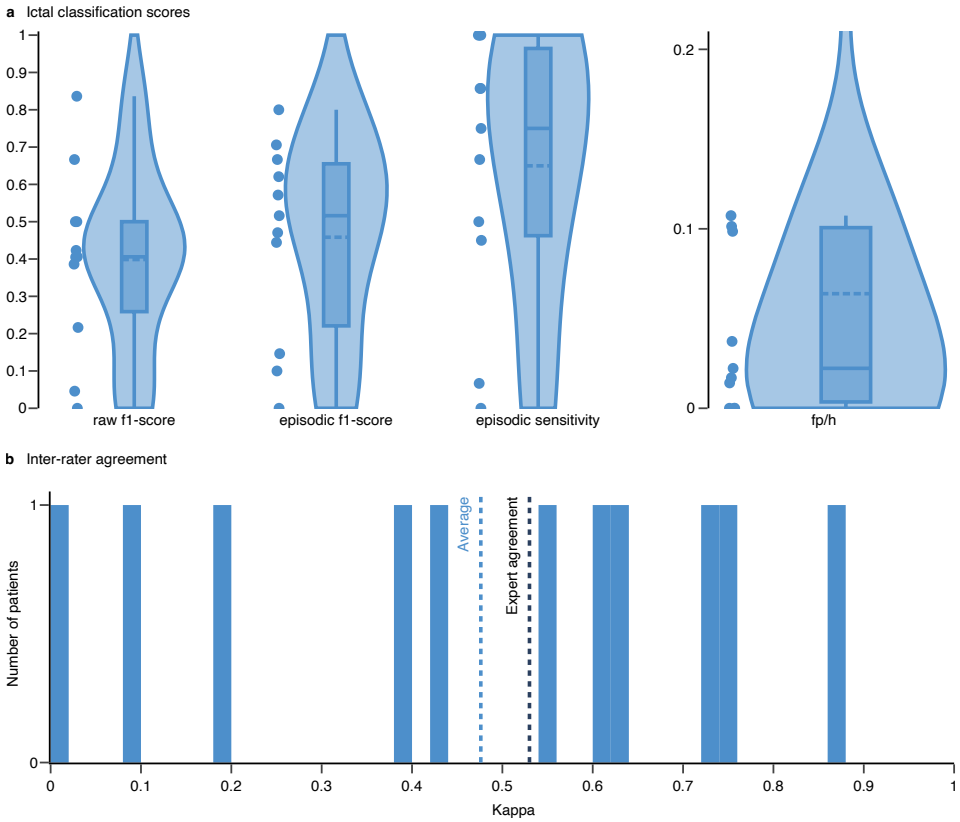

Figure S15: **Seizure detection with 58 patient pre-training.** **(a)** Seizure detection results of MVPFormer-S on the 10 remaining unseen patients: the F1-score, sensitivity, and fp/h are reported. All performance metrics are improved with respect to the original MVPFormer model. The raw and episodic F1-scores are significantly different here, indicating that MVPFormer benefits of the episode merging effect of post-processing on these patients. The false positive rate has decreased further with the scale of the pre-training dataset. **(b)** Inter-rater agreement of MVPFormer with the human expert: Cohen's kappa is used to measure the agreement between the artificial assistant and the human expert. The average kappa is increased to 0.48. The distribution of kappa values again indicates that there is considerable variability in the agreement.

Figure S15 shows the performance of the 58-subject MVPFormer-S on the 10 unseen subjects (for a detailed breakdown see Table S15). On the other hand, Figure S16 shows the results of the original MVPFormer-S model on those same 10 subjects (for a detailed breakdown see Table S16). All performance metrics improve with a growing pre-training dataset size, although on a small test cohort, indicating that increasing the number of subjects in the pre-training dataset has a net positive effect on the downstream classification task.

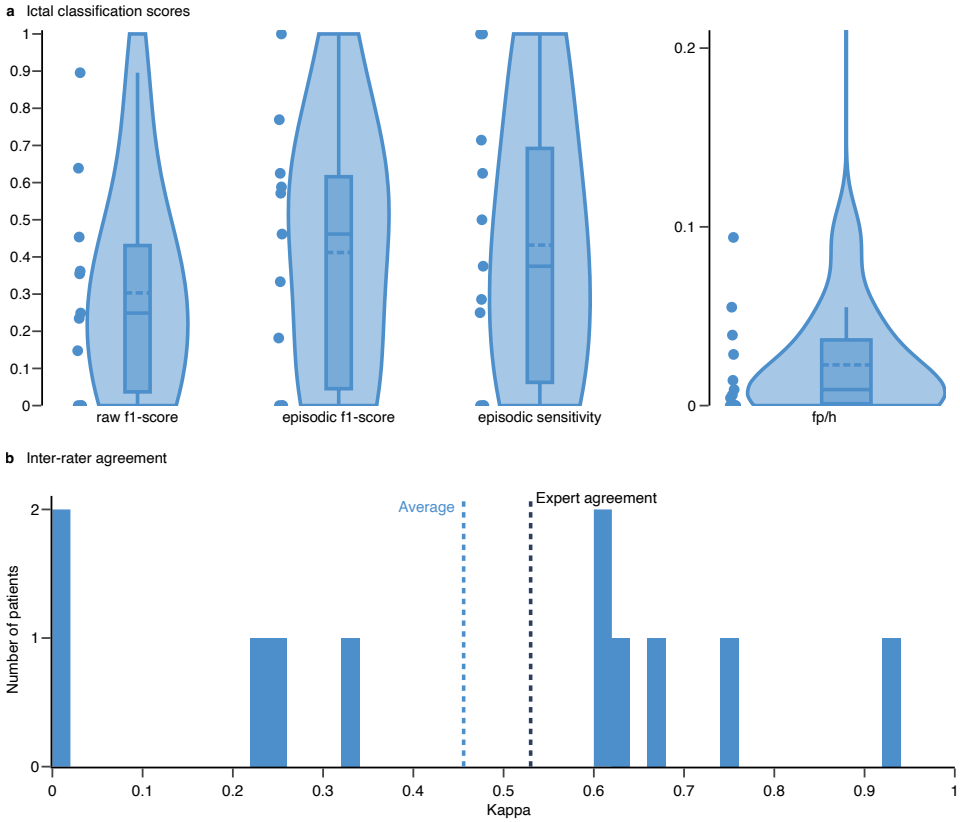

Figure S16: **Seizure detection with 18 subjects pre-training on a selection of 10 unseen subjects. (a)** Seizure detection results of MVPFormer-S on the 10 subjects excluded from the 58-subjects model: the F1-score, sensitivity, and fp/h are reported. The raw and episodic F1-scores are significantly different here, indicating that MVPFormer benefits of the episode merging effect of post-processing on these patients. These results are a subset of those presented in the Results section. **(b)** Inter-rater agreement of MVPFormer with the human expert: Cohen's kappa is used to measure the agreement between the artificial assistant and the human expert. The average kappa is 0.46, reduced from the overall results indicating that these subjects are more difficult than average. The distribution of kappa values again indicates that there is considerable variability in the agreement.

Table S15: **Details of seizure detection results of MVPFormer-S with 58 patient pre-training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID59	0.40	0.41	0.44	0.86	0.86
ID60	0.07	0.05	0.10	0.07	0.07
ID61	0.54	0.39	0.47	0.50	0.50
ID62	0.00	0.00	0.00	0.00	0.00
ID63	0.65	0.50	0.71	0.86	0.86
ID64	0.76	0.84	0.67	1.00	1.00
ID65	0.60	0.67	0.62	0.45	0.45
ID66	0.50	0.41	0.52	1.00	1.00
ID67	0.79	0.50	0.57	1.00	1.00
ID68	0.22	0.22	0.15	0.75	0.75

Table S16: **Details of seizure detection results of MVPFormer-S with 18 patient pre-training on a selection of 10 subjects.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID59	0.22	0.25	0.18	0.29	0.09
ID60	0.00	0.00	0.00	0.00	0.00
ID61	0.33	0.23	0.46	0.38	0.04
ID62	0.00	0.00	0.00	0.00	0.00
ID63	0.69	0.45	0.77	0.71	0.00
ID64	0.94	0.90	1.00	1.00	0.00
ID65	0.60	0.64	0.62	0.50	0.06
ID66	0.50	0.35	0.59	0.62	0.03
ID67	0.70	0.36	0.57	1.00	0.01
ID68	0.26	0.15	0.33	0.25	0.01

G.9 Effects of the selection of channels

As discussed in previous sections, the number of channels can vary considerably across subjects. To speed up the seizure detection task, we quickly select a subset of the channels (up to 50) which visually appear least noisy and most relevant. We also test the effect of including all channels, expecting it to decrease both the speed and performance due to the decrease of the overall signal-to-noise ratio.

Figure S17 shows that the performance decreases when we use all channels (for a detailed breakdown see Table S17). This is expected, as the noise contained in the entire recording increases together with the information content. MVPFormer's ability to generalize is not affected by the number of channels, but the noise affects the performance. Therefore, the optimal real-world operation of MVPFormer is obtained by selecting a subset of channels for detection.

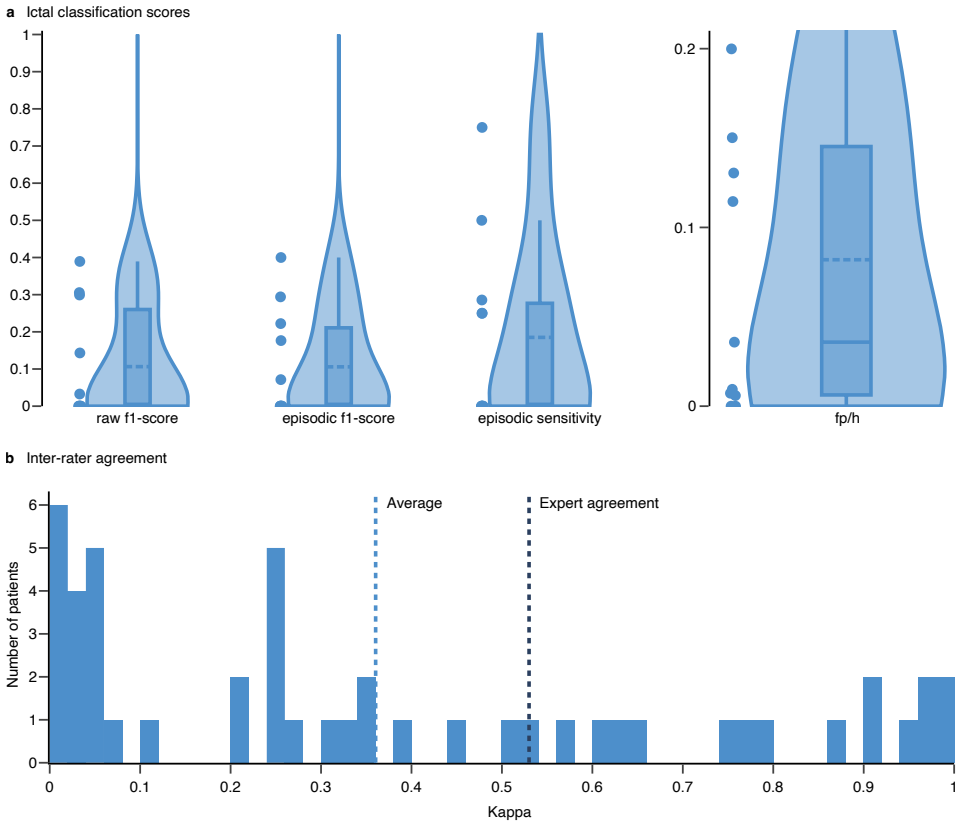

Figure S17: **Seizure detection with 18 patient pre-training and evaluation on all channels. (a)** Seizure detection results of MVPFormer-S on the 10 remaining unseen patients evaluated on all channels: the F1-score, sensitivity, and fp/h are reported. The performance metrics are reduced with respect to the results obtained when selecting a subset of the channels. MVPFormer is not affected by the number of channels, but the increase of noise emerging from all the channels contributes to a reduction in performance. **(b)** The average kappa is 0.36, reduced from the evaluation on a subset of channels.

Table S17: **Details of seizure detection results of MVPFormer-S with evaluation on all channels.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID19	-0.05	0	0	0	0
ID20	0.84	0.47	0.5	0.5	0.5
ID21	0.86	0.4	0.8	1	1
ID22	0.91	0.81	0.89	1	1
ID23	0.25	0.17	0.14	0.09	0.09
ID24	0.91	0.9	0.93	0.93	0.93
ID25	0.45	0.17	0.25	0.25	0.25
ID26	0.25	0.04	0.22	1	1
ID27	0.98	0.93	1	1	1
ID28	0.64	0.42	0.67	0.5	0.5
ID29	0.05	0.04	0.07	0.29	0.29
ID30	0.2	0.31	0.19	0.11	0.11
ID31	0.98	0.95	1	1	1
ID32	0.98	0.92	1	1	1
ID33	0.01	0	0	0	0
ID34	0.04	0.06	0.04	1	1
ID35	0.61	0.63	0.55	0.43	0.43
ID36	0.5	0.62	0.51	0.37	0.37
ID37	0.03	0.02	0.04	1	1
ID38	0.97	0.97	1	1	1
ID39	0.77	0.52	0.67	1	1
ID40	0.37	0.5	0.31	0.8	0.8
ID41	0.52	0.47	0.55	1	1
ID42	0.63	0.54	0.57	1	1
ID43	0.94	0.71	1	1	1
ID44	0	0	0	0	0
ID45	0	0	0	0	0
ID46	0.06	0.03	0.06	0.19	0.19
ID47	0.23	0.29	0.16	1	1
ID48	0.25	0.18	0.31	0.33	0.33
ID49	0.4	0.47	0.4	1	1
ID50	0.05	0.05	0.06	1	1
ID51	0.06	0.05	0.07	1	1
ID52	0	0.01	0.01	1	1
ID53	0.13	0.21	0.09	1	1
ID54	0.01	0.01	0.01	0.67	0.67
ID55	0.37	0.35	0.29	1	1
ID56	0.33	0.34	0.31	0.67	0.67
ID57	0	0	0	0	0
ID58	0.29	0	0	0	0
ID59	0.02	0	0	0	0
ID60	0	0	0	0	0
ID61	0	0	0	0	0
ID62	0.04	0.03	0.07	0.29	0.29
ID63	0.21	0.39	0.18	0.75	0.75
ID64	0.25	0.3	0.29	0.25	0.25
ID65	0.04	0	0	0	0
ID66	0.76	0.31	0.4	0.5	0.5
ID67	0.17	0.14	0.22	0.25	0.25
ID68	0.61	0.00	0.00	0.00	0.01

G.10 Patient classification difficulty

In the medical practice each patient has unique seizure presentations, though they might be broadly grouped into different categories [58]. As an effect, some patients have seizures which can be considered more typical (Figure S3a), and hence easier to detect, while others might have very atypical events (Figure S3b). There might be broad disagreement among neurologists over these atypical seizures, and at the same time no disagreement at all over the typical patients [59].

This phenomenon intuitively creates a difficulty scale among the patients, which also affects MVP-Former and contributes to the spread of performance between the model and the human expert. To better assess the impact of this latent patient classification difficulty we performed a multiple correlation analysis using the total recording length, the number of seizures, and the frequency of seizures to predict the kappa score, yielding an R^2 of 0.054. The model performance is thus independent of the three variables, and we believe the difficulty might help explain most of the variance. The literature supports this hypothesis, as the subjects themselves can account for up to 65% of the variance [17] while the clinical setup itself has no impact.

G.11 Channel connectivity map

Generating future iEEG signal embedding implicitly places the greatest emphasis on the dimension of time, but to do so it is necessary to consider the interactions between channels as well. MVPFormer thus takes into consideration all electrodes concurrently, as electric potentials flow across different areas and circuits in the brain following their intrinsic connections and constraints [60, 61]. The number of channels depends on the number of electrodes decided for a specific patient and the clinical setup. As iEEG implantations are decided on a case-by-case basis by a physician, there is no uniform standard on where to place the electrodes, in contrast with the 10-20 system [62] for EEG. Therefore, we cannot give MVPFormer any a priori knowledge of how the channels will interact in space, and the model has to learn it on its own. MVPA enables our model to dynamically learn these connections to build an internal map of the flow, becoming independent from a specific electrode configuration.

It is well-known that two neighboring brain regions might not be as strongly connected as two faraway regions. The relationship between the electrodes (and hence the channels) mirrors this behavior. To truly understand the link between two channels the model must build a map of the connection strength between different brain areas and how these connections impact the diffusion of electric fields across channels. In MVPA, this understanding is the underpinning of the channel-based component. In particular, the complex interplay between the query, the channel codebook, and the channel attention, acts as the first level of processing. Further, the deep structure with multiple layers provides more representational power, as is typical of deep models.

Figure S18 shows that the channel-based MVPA component encodes a form of the brain connectivity map. Initially, the map is random as MVPA is randomly initialized. As training progresses, the attention magnitude among the channels starts to differentiate, building a map of the connection strength. The map is dominated by the diagonal component, which indicates that in general neighboring channels are more related than distant channels. However, it is possible to clearly distinguish clusters of strongly connected electrodes and also skipped connections, which possibly refer to strong connections between distant regions. Since the channel distance is relative, it can apply to arbitrary clinical setups and is not limited to already seen channels. Moreover, as the channel attention is a function of both the query content and the channel distance, the combination of the two can effectively modulate the attention even on unseen subjects.

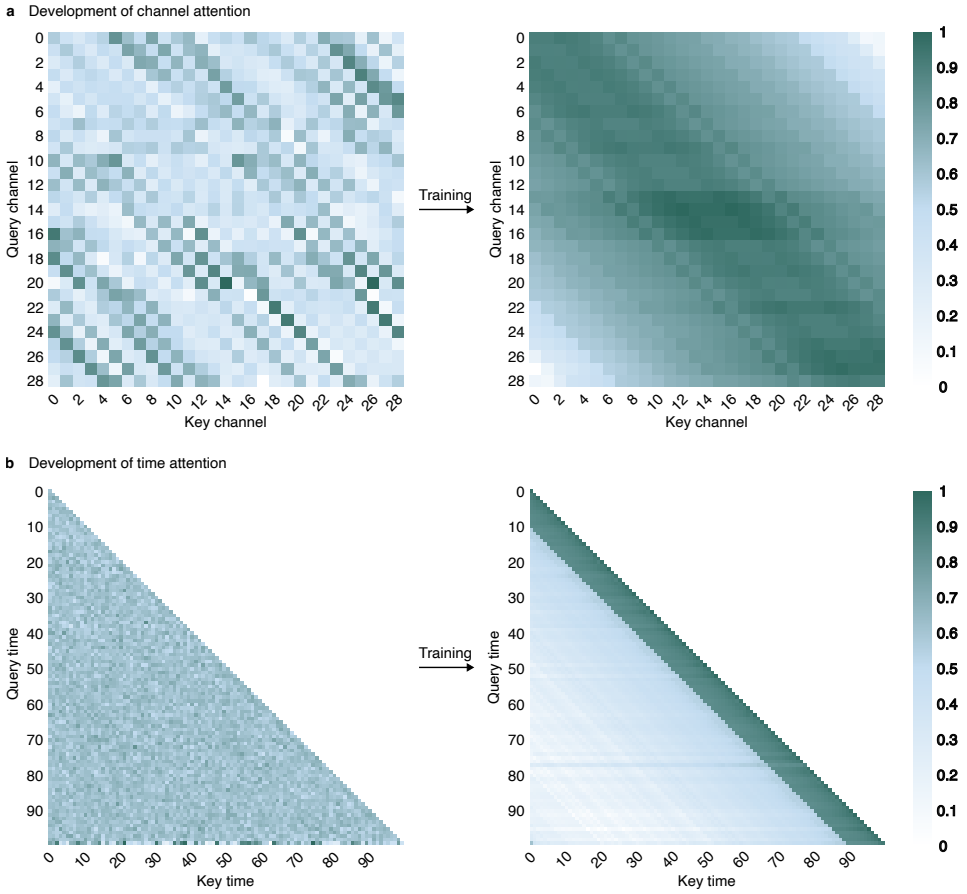

Figure S18: **Attention components before and after training.** **(a)** The channel-based attention component is randomly initialized. At the end of training, it shows the diagonal structure which indicates that the relationship between the channels is mainly one of proximity. This is expected, as nearby channels are expected to be more closely related, and showcases MVPA's learning outcome. **(b)** The time-based attention component is also randomly initialized. At the end of training, it shows that segments which are close in time are more related. Particularly, few closest segments are attended to more strongly, as the content-based attention's lookback windows is limited to a few segments.

G.12 Results on the MAYO and FNUSA datasets

We test MVPFormer additionally against Brant-2 on the iEEG MAYO and FNUSA datasets. Both datasets are single-channel iEEG datasets containing both physiological and pathological activity. We use these datasets to evaluate MVPFormer's performance on extremely noisy data, and to compare its resilience with the SOTA.

The MAYO dataset contains 24 patients for a total of 130 hours of data. In particular, 36% of the data is non-ictal, 9% is ictal, and 53% is noise. Moreover, 18 subjects contain no ictal activity and 9 patients are fully noise, with 13 being majority noise. The FNUSA dataset contains 14 patients for a total of 160 hours. 48% of the dataset is non-ictal, 27% is ictal, and 23% is noise. Moreover, the data of 2 subjects is completely noise, and the data of 3 other subjects is fully ictal. Both datasets are much smaller in scale than the Long-term iEEG dataset, and are heavily dominated by noise, both artifact and powerline. In particular, the Long-term iEEG dataset is almost 2,000 times larger and is also carefully evaluated by an expert neurologist to remove channels which contain too much noise or artifacts.

As we wish to assess MVPFormer in a realistic, real-world scenario, we do not remove any noise from the dataset but test them as-is. In particular, we consider noise and physiological activity as one category, and pathological activity as another. However, kappa scores are not meaningful with such small datasets, therefore we provide the aggregate F1-score, and the average sensitivity and specificity. Specifically, given the fact that many patients do not contain ictal activity, we do not compute the average F1-score across subjects, but pool together all subject's results and compute the aggregate F1-score.

We use the same MVPFormer models pre-trained on our Long-term iEEG dataset, and train a specific classification head for either MAYO or FNUSA by fine-tuning on the first four patients. Then, we test on the remaining patients. We also use the Brant-2 model whose pre-trained weights are publicly available, and fine-tune in the same manner as MVPFormer using the fine-tuning code provided by the authors.

The results can be found in Tables S18 and S19. Given the very low signal-to-noise ratio of both datasets, overall performance is affected. On the FNUSA dataset, where the amount of noise is more moderate, all models perform similarly, with MVPFormer-M showing a higher specificity. However, MVPFormer has a clear advantage on the MAYO dataset, with almost double the F1-score with respect to Brant-2. The difference between MVPFormer-S and MVPFormer-M is minimal, as the sizes of the datasets involved are too small to fully train a very large model such as MVPFormer-M (see Appendix G.4 for more information).

Table S18: **Summary of seizure detection results of all models on the MAYO iEEG dataset.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

Model	F1-score	Sensitivity	Specificity
MVPFormer-M	**0.36**	0.38	**0.91**
MVPFormer-S	0.35	0.41	0.88
Brant-2	0.19	1.00	0.18

Table S19: **Summary of seizure detection results of all models on the FNUSA iEEG dataset.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

Model	F1-score	Sensitivity	Specificity
MVPFormer-M	**0.46**	0.94	**0.10**
MVPFormer-S	0.46	0.99	0.03
Brant-2	0.46	0.99	0.02

G.13 Ablation of the prediction task

We design MVPFormer with a two-phase training regime. First, during the generative pre-training task MVPFormer learns to predict the neuronal activity. Second, during the classification task it needs to correctly classify ictal periods. To determine the significance of the generative task on the classification task, we train MVPFormer only on the classification task and compare its performance with the full architecture. Figure S19 and Table S20 clearly indicate that the generative task is of fundamental importance to the overall architecture, with a Kappa score decrease to 0.52. This is below the original result of 0.54 and below the human agreement threshold. Moreover, the distribution of agreement has flattened, with an overall decrease of performance across the board and an increase of subjects with no agreement. This suggests that without pre-training the generalization capability of MVPFormer suffers. Therefore, the generative task is necessary and is a significant contributor to learning.

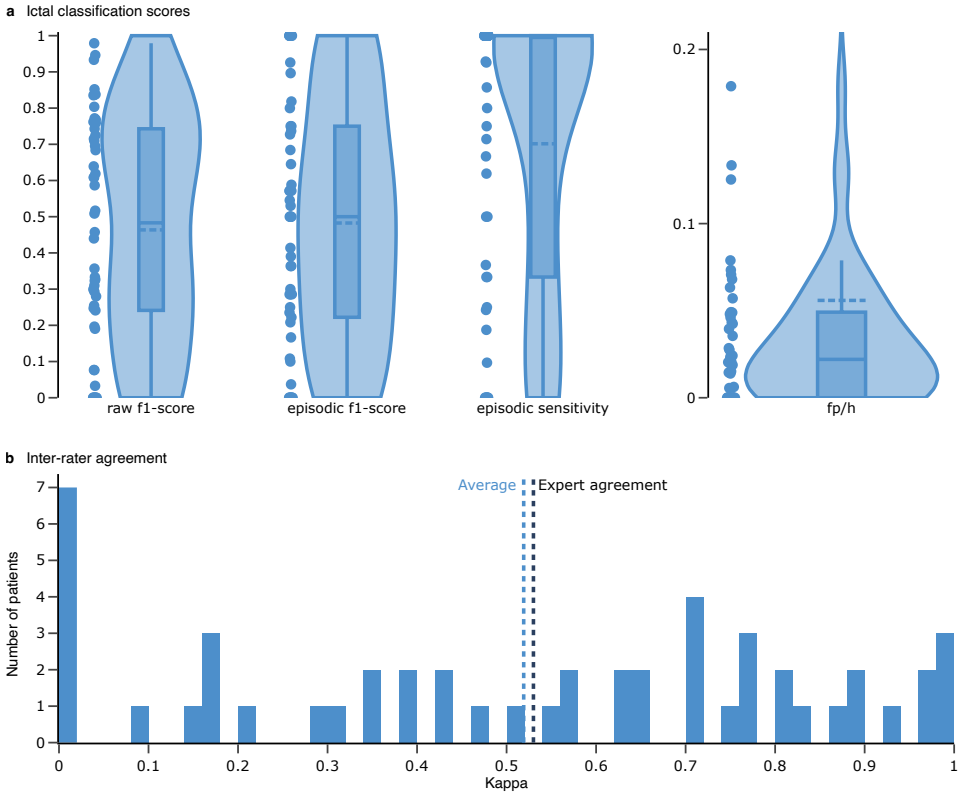

Figure S19: **Seizure detection with no generative pre-training and 18 subjects classification training. (a)** Seizure detection results of MVPFormer-S on 40 unseen subjects which are part of the training set for the 58-subjects model: the F1-score, sensitivity, and fp/h are reported. The raw and episodic F1-scores are notably lower here with respect to the 58-subjects model. This is expected given the 58-subject model is pre-trained on these subjects. These results are a subset of those presented in the Results section. **(b)** Cohen's kappa is used to measure the agreement between the artificial assistant and the human expert. The average kappa is 0.56, competitive with expert agreement but, as expected, reduced from the 58-subjects pre-trained model.

Table S20: **Seizure detection with no generative pre-training and 18 subjects classification training.** Kappa is the inter-rater agreement. The classification metrics report the raw and episodic metrics relevant for the seizure classification task.

		Classification metrics			
		Raw	Episodic		
Subject	**Kappa**	f1-score	f1-score	sensitivity	fp/h
ID19	0.37	0.46	0.3	0.19	0.05
ID20	0.99	0.98	1	1	0
ID21	0.47	0.3	0.29	1	0.04
ID22	0.74	0.7	0.73	1	0.02
ID23	0.37	0.51	0.36	0.24	0.07
ID24	0.88	0.84	0.9	0.93	0.05
ID25	0	0	0	0	0
ID26	0.43	0.19	0.29	1	0.06
ID27	0.95	0.84	1	1	0
ID28	0.73	0.8	0.75	0.75	0.01
ID29	0.17	0.28	0.17	0.1	0.01
ID30	0.9	0.85	0.93	0.93	0.05
ID31	0.98	0.95	1	1	0
ID32	0.98	0.93	1	1	0
ID33	0	0	0	0	0
ID34	0.78	0.84	0.82	1	0.02
ID35	0.71	0.77	0.74	1	0.05
ID36	0.59	0.68	0.53	0.37	0.01
ID37	0.66	0.61	0.5	1	0.02
ID38	0.59	0.64	0.59	1	0.04
ID39	0.76	0.71	0.57	1	0.02
ID40	0.67	0.77	0.57	0.8	0.02
ID41	0.75	0.72	0.75	1	0.02
ID42	0.97	0.74	1	1	0
ID43	0.15	0	0	0	0.06
ID44	0.98	0.76	1	1	0
ID45	0.03	0.03	0.04	1	0.33
ID46	0.6	0.62	0.68	0.62	0.03
ID47	0.79	0.72	0.55	1	0.02
ID48	0.88	0.76	1	1	0
ID49	0.08	0.08	0.11	1	0.71
ID50	0.36	0.36	0.24	1	0.07
ID51	0.52	0.32	0.22	1	0.04
ID52	0.5	0.29	0.5	1	0.04
ID53	0.2	0.33	0.1	1	0.13
ID54	0.61	0.31	0.29	0.33	0.01
ID55	0.35	0.25	0.25	1	0.08
ID56	0.82	0.61	0.8	0.67	0
ID57	0	0	0	0	0
ID58	0.17	0.2	0.21	0.71	0.26
ID59	0	0	0	0	0
ID60	0.16	0.08	0.36	0.25	0.02
ID61	0	0	0	0	0
ID62	0.42	0.44	0.41	0.86	0.07
ID63	0.54	0.25	0.57	0.5	0.01
ID64	0.62	0.59	0.65	0.5	0.03
ID65	0.31	0.24	0.39	1	0.18
ID66	0.69	0.52	0.5	1	0.02
ID67	0	0	0	0	0.13
ID68	0.78	0.33	0.5	0.33	0

G.14 Effects of the number of channels on the Brain TreeBank dataset

To better evaluate the robustness of MVPFormer to an increasing number of channels, we evaluate the four tasks of the BrainTreeBank with a range of 10 to 50 channels. The performance of MVPFormer moderately increases with no reduction with the channel number, as reported in PopT [20] as well, indicating that our model is robust to the number of channels.

Table S21: **Effects of the number of channels on the four tasks of the Brain TreeBank dataset.** Evaluation of the performance of MVPFormer with respect to number of channels for fine-tuning and testing.

Channels	Pitch	Volume	Onset	Speech
10	0.81 (0.01)	0.85 (0.01)	0.86 (0.02)	0.87 (0.02)
20	0.82 (0.01)	0.87 (0.01)	0.87 (0.02)	0.88 (0.02)
30	0.82 (0.02)	0.87 (0.01)	0.87 (0.02)	0.89 (0.02)
40	0.83 (0.01)	0.87 (0.01)	0.87 (0.02)	0.89 (0.02)
50	0.83 (0.01)	0.88 (0.01)	0.87 (0.02)	0.90 (0.02)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and the introduction are thoroughly described and evaluated in the paper and reflect the conclusions reached in it.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in the dedicated Limitations paragraph and throughout the paper, as appropriate.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes in details all the algorithms and the methods used to generate the results, and provides all the data, code, and weights necessary to reproduce them.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and the code will be open-sourced after the review process, and are made available during review in anonymized form.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the experimental settings and details are described in the paper, both in the main text and more in detail in the Appendices.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All results are presented in concise forms as averages or otherwise, but are always accompanied by the full distribution and full tables. Where applicable, we provide the mean, median, and standard deviations for all metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the relevant details regarding the number of GPUs, their type, and the training time required to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: All possible impacts of this paper are dicussed in the appropriate section of the main text.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not foresee high risks associated with this research, and provide the model under the guidelines of it being used for research purposes only.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all models and datasets used in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All new assets will be accompanied by relevant GitHub repositories and/or other web resources documenting their usage. During the review process, we will make them available in anonymized form.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The dataset collected has been made publicly available by the relevant institution, and is there fully compliant.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The dataset collected has been made publicly available by the relevant institution, and is there fully compliant.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs have not been used in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.