Towards a generalizable, unified framework for multimodal neural decoding

Nanda H Krishna[1,2,✉] **Mathys Loiselle**[1,3] **Avery Hee-Woon Ryoo**[1,2,✉]
Matthew G Perich[1,2,†] **Guillaume Lajoie**[1,2,†]
[1]Mila [2]Université de Montréal [3]Concordia University [†]Co-senior authors
[✉]{nanda.harishankar-krishna,hee-woon.ryoo}@mila.quebec

Abstract

Recent advances in neural decoding have led to the development of large-scale deep learning-based neural decoders that can generalize across sessions and subjects. However, existing approaches predominantly focus on single modalities of neural activity, limiting their applicability to specific modalities and tasks. In this work, we present a unified, multimodal framework that jointly processes neuronal spikes and local field potentials (LFPs) for behavioural decoding. Our approach employs flexible tokenization schemes for both spikes and LFPs, enabling efficient processing of heterogeneous neural populations. Through experiments on data from nonhuman primates performing motor tasks, we demonstrate that multimodal pretraining yields superior decoding performance compared to unimodal baselines. We also show evidence of cross-modal transfer: models pretrained on both modalities outperform LFP-only models when finetuned solely on LFPs, suggesting a path toward more cost-effective brain-computer interfaces that can use performant LFP-based decoders. Our models also exhibit robustness to missing modalities during inference when trained with modality masking, and scale effectively with both model size and pretraining data. Overall, this work represents an important first step towards unified, general-purpose neural decoders capable of leveraging diverse neural signals for a variety of brain-computer interface applications.

1 Introduction

Brain-computer interfaces (BCIs) rely on neural decoders to translate brain activity into control signals. Recent large-scale deep learning models for neural decoding, such as the NDT [1, 2] and POYO [3–5] families, and others [6, 7] have shown that pretraining on large-scale datasets enables robust generalization to new sessions and datasets with minimal adaptation. However, these models process only single modalities of neural signals, missing opportunities to leverage the complementary information present across different recording types.

Neural activity spans multiple scales, from precise action potentials (spikes) to population-level local field potentials (LFPs). Spikes provide single-neuron resolution, while LFPs capture aggregate dynamics and motor-related oscillations. Despite this complementarity, existing decoders typically use or are capable of processing only one recording modality. This limitation has practical consequences, as spike-based BCIs require expensive, high-bandwidth systems, while LFP-based alternatives offer simpler hardware but historically lower accuracy [8–10].

To alleviate this limitation, we present a multimodal extension of the POYO framework [3] that is capable of jointly processing both spikes and LFPs for behavioural decoding. Our approach introduces a flexible LFP tokenization scheme that is close to POYO's individual spike tokenization scheme, which preserves each modality's characteristics while enabling unified processing.

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

Our key contributions include:

- A unified tokenization scheme for joint processing of spikes and LFPs, enabling strong multimodal decoding performance.
- A demonstration of cross-modal transfer that particularly enhances LFP-only decoding.
- Improving robustness to missing modalities, during both finetuning and inference, via modality masking.

2 Related Work

Neural decoding. Traditionally, neural decoding for continuous tasks was accomplished with statistical models such as the Kalman filter [11–14]. While performant, these methods required careful calibration [15, 16] to new sessions and a considerable amount of training data for a new user. With the availability of large-scale public neural datasets, several recent works have been proposed with the goal of building "foundation model"-like decoders for neural activity [1–5, 7, 17, 18], leveraging advances in large-scale deep learning. Of particular relevance to this work are two classes or families of neural decoding models: NDT-style models [1, 2] and POYO-style models [3–5], which were originally developed for neuronal spiking data and motor decoding tasks similar to those we will consider [19]. The former use binning or patch-based tokenization schemes coupled with generic ViT-like [20] or causal transformer [21] backbones, while the latter use individual spike tokenization coupled with PerceiverIO-style [22] input-output cross-attention and a backbone comprising either Transformer blocks [21] or recurrent networks [5, 23–25]. While these pretraining-based approaches to neural decoding are highly performant, they have not yet been adapted to process multiple neural recording modalities (models like NDT-3 [2] process behaviour as an additional input modality). In this work, we develop upon POYO-style models and aim to adapt them to process and decode behaviour from multiple neural recording modalities.

Multimodal deep learning. Processing multiple input modalities is a key problem in deep learning [26], with several methods being actively developed to jointly process modalities such as vision and language [27], or even video and audio [28]. Particularly relevant to our work are the explorations of multimodal machine learning for healthcare [29]. Ongoing research aims to tackle key problems such as identifying the right pretraining objectives [30, 31], investigating cross-modal transfer [32], and studying modality fusion approaches to develop better architectures for multimodal processing [33]. Most relevant to this work are neural decoding models which attempt to integrate both neuronal spiking activity and spectral features such as spiking-band powers [34] or local field potentials (LFPs) [35] to improve neural decoding performance.

3 Methods

Data preprocessing. In this work, we attempt to minimize the amount of data preprocessing through the use of well-designed tokenization schemes and model architectures. Due to the use of a flexible individual spike tokenization scheme [3] (described below), we avoid having to bin spike counts and are able to process spiking data from heterogeneous neural populations, including single-neuron spikes and multi-unit activity. Meanwhile, LFP powers are extracted from the raw time-domain voltage signal through causal band-pass filtering. The extracted bands include the local motor potential (LMP; 0.1-20 Hz), delta (1-4 Hz), theta (3-10 Hz), alpha (12-23 Hz), beta (27-38 Hz), and gamma (50-300 Hz) bands. The only other preprocessing we do is to remove outliers in the behaviour using a simple acceleration-based threshold.

Spike tokenization. We adopt a recently proposed individual spike tokenization framework [3] for our model. In this framework, each neuronal spike is represented by the identity of the neural "unit" it came from and the timestamp at which the spike occurred. Information about unit identity is provided through a unique, learnable unit embedding for each neuron, while spike-timing information is provided through rotary positional embeddings (RoPE) [36]. This positional embedding scheme allows models to learn the relative timing between tokens rather than the absolute spike-times, thus enabling better generalization. Overall, each individual spike token from a unit with ID i is a D-dimensional vector, given by $\mathbf{x} = (\mathrm{UnitEmb}(i), t)$, where $\mathrm{UnitEmb} : \mathbb{Z} \to \mathbb{R}^D$. Note that we opt to use the term “unit” rather than neuron, since spikes may be assigned at a coarser specificity than at the single-neuron level (e.g., multi-unit activity). The flexibility of this tokenization scheme and the model architecture described below allows us to efficiently process variable-sized neural populations efficiently and facilitates training and efficient finetuning on multiple datasets and recording sessions.

LFP tokenization. We introduce a method to tokenize LFP powers that is similar to the individual spike tokenization scheme and a tokenization scheme for Calcium imaging [4]. Each individual channel is provided a unique, D-dimensional learnable channel embedding, akin to unit embeddings. This embedding is added to a linear projection of the LFP powers to D dimensions to give one LFP token. Finally, just as with the spike tokens, the timestamps associated with the LFP tokens are provided through RoPE. Overall, each LFP token from a channel c with LFP powers $\mathbf{p} \in \mathbb{R}^6$ is given by $\mathbf{x} = (\mathrm{ChannelEmb}(c) + f(\mathbf{p}), t)$, where $\mathrm{ChannelEmb} : \mathbb{Z} \to \mathbb{R}^D$ and $f : \mathbb{R}^6 \to \mathbb{R}^D$. Note that this tokenization scheme for LFPs retains the flexibility of the spike tokenization scheme in supporting a variable number of channels.

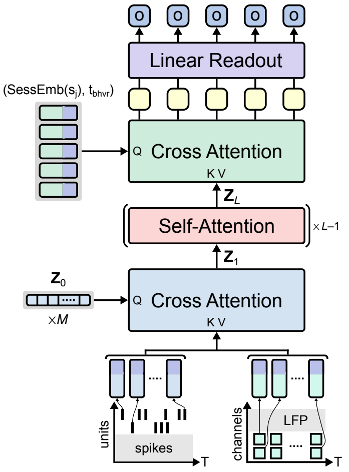

Figure 1: **An architecture for multimodal neural decoding.** We adapt POYO [3] to decode behaviour from both spikes and local field potentials (LFPs).

Model architecture. We employ the original POYO architecture [3] as-is for decoding behaviour from multimodal neural activity, i.e., the spike and LFP tokens. The POYO architecture, based on the PerceiverIO framework [22], can be decomposed into three blocks: the input cross-attention module, the self-attention backbone, and the output cross-attention and readout module. First, the input cross-attention module is used to compress a variable length sequence of input spike and LFP tokens from a context window $[0, T]$ into a fixed-length sequence of latent representations. Consider a sequence of N input tokens $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N\times K}$ and a sequence of M learnable latents $\mathbf{Z}_0 = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M] \in \mathbb{R}^{M\times D}$. The queries for the input cross-attention are given by the projection $\mathbf{Q} = \mathbf{Z}_0\mathbf{W}_q$, while the keys and values are given by $\mathbf{K} = \mathbf{X}\mathbf{W}_k$ and $\mathbf{V} = \mathbf{X}\mathbf{W}_v$ respectively. Overall, the result of the cross-attention is given by:

$$\mathbf{Z}_1 = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{1}$$

We see that $\mathbf{Z}_1$ is a fixed-length latent sequence whose dimensions depend only on the number of queries M and the dimensionality of the values in the attention module, which is a hyperparameter. $\mathbf{Z}_1$ is processed by a series of $(L-1)$ self-attention layers to give a latent sequence $\mathbf{Z}_L$. Finally, projections of these latents are used as the keys and values of an output cross-attention module, where a variable number of output queries is used to decode behaviour at P timestamps in the context window $[0, T]$. The output queries are made up of learnable session embeddings which are unique to each recording session, while timestamp information is provided through RoPE to the the cross-attention module.

Overall, this architecture provides several key advantages over typical Transformer-based approaches: (1) the computational complexity scales linearly with the number of input tokens due to self-attention being applied in a fixed-dimensional latent space, (2) the input module allows for processing a highly variable number of input tokens in each context window, thus allowing us to forego binning and preventing the loss of spike-timing information, and (3) the output module allows for the prediction of multiple behavioural outputs per context window and alleviates the need for trial-alignment.

Finetuning to new sessions. Given a model that has been pretrained on multiple sessions or datasets, we outline two strategies adapted from POYO [3] for finetuning the model on downstream sessions with new sets of units and channels. The first, *unit identification (UI)*, is a highly efficient scheme in which all model weights are frozen and only the embeddings (i.e., unit, channel, and session embeddings) are reinitialized and trained on data from the new session. This scheme allows the model to preserve the task-relevant neural dynamics learned during pretraining and typically updates fewer than 1% of the total number of model parameters. The second, *full finetuning (FT)*, involves first performing UI for a few epochs, followed by unfreezing and training all model parameters end-to-end. While this scheme is more compute-intensive, it yields maximum performance in our evaluations.

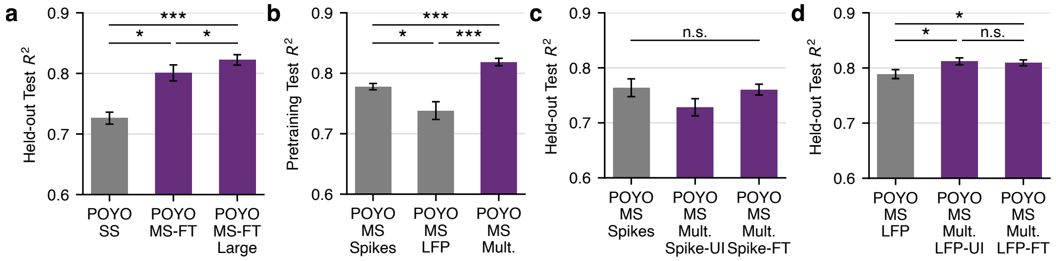

Figure 2: **Multimodal pretraining improves decoding performance.** **(a)** Pretrained multimodal models outperform single-session multimodal models; performance improves further with larger model size (12M vs 2M). **(b)** Multimodal models outperform unimodal models. **(c)** Multimodal models finetuned on just spikes match (FT) or are close to (UI) fully-finetuned unimodal spike-based models. **(d)** Multimodal models finetuned on just LFPs outperform fully-finetuned unimodal LFP-based models. **Key:** Error bars: ± SEM over 5 (a,c,d) or 25 (b) sessions; *,***: $p \leq 0.05, 0.001$ on a paired t-test across sessions; SS: Single-session; MS: Multi-session.

4 Experiments

In this section, we show how multimodality and multi-session pretraining can help improve neural decoding performance over unimodal and single-session approaches. We also discuss the robustness of our model to the absence of either spikes or LFPs and demonstrate its ability to effectively learn from each modality.

Table 1: **Behavioural decoding results on heldout sessions.** We report the mean R^2 (↑) ± SD across 5 sessions. SS: Single-session, MS: Multi-session.

Model	Modality	Evaluation R^2
GRU SS	Spikes	0.7122 ± 0.0355
	LFP	0.3381 ± 0.0437
	Multimodal	0.6225 ± 0.0450
POYO SS	Spikes	0.7183 ± 0.0348
	LFP	0.1671 ± 0.1086
	Multimodal	**0.7262 ± 0.0197**
GRU MS FT	Spikes	0.7168 ± 0.0307
	LFP	0.6059 ± 0.0852
	Multimodal	0.7136 ± 0.0433
POYO MS UI	Spikes	0.7457 ± 0.0433
	LFP	0.7833 ± 0.0169
	Multimodal	0.7889 ± 0.0287
POYO MS FT	Spikes	0.7638 ± 0.0362
	LFP	0.7888 ± 0.0182
	Multimodal	**0.8008 ± 0.0263**

4.1 Data and experimental setup

Our models are trained on data recorded from nonhuman primates performing a touchscreen-based random target task [19]. For all our experiments, we consider 30 sessions from Monkey I, which contain both spikes and LFPs. Of these 30 sessions, 25 are used for pretraining our multi-session models, while 5 are heldout for evaluation. We train our POYO models in a supervised manner using the mean squared error loss to accurately predict two-dimensional finger/cursor velocities. The inputs to the models are context windows of neural activity which are one second long. The models are trained to observe these context windows of activity and predict all behaviour from within each window. Note that we do not segment our data into trials during training and evaluation, thanks to the flexibility of the POYO framework.

4.2 Pretraining and generalization performance

Firstly, we found that multimodal models pretrained on multiple sessions' worth of data outperformed single-session multimodal models on the 5 heldout sessions from our dataset. As shown in Figure 2a, we also see further improvements in performance when scaling the size of the models in addition to increasing the amount of pretraining data.

A key observation during pretraining is the considerable improvement in decoding performance with LFPs when training on multiple sessions, as opposed to a single session. We also observe that the multimodal model outperforms unimodal models in both pretraining and finetuning performance (Figure 2b and Table 1). This is especially the case when incorporating regularization techniques such as unit dropout and modality masking. Overall, our results indicate that decoding from LFPs benefits greatly from scale and serves to enhance multimodal performance by a significant margin.

Turning to finetuning performance, we observe that the multimodally pretrained POYO model finetuned on LFPs outperforms the LFP-pretrained POYO model on decoding from just LFPs (Figure 2d). This is a positive indication of cross-modal transfer, and suggests that a multimodally pretrained model may be finetuned to yield performant LFP-based BCI decoders, which could be a path towards more cost-effective BCI devices compared to those that process spikes [8, 10].

4.3 Robustness to the absence of a modality

As briefly described above and in addition to the same-modality finetuning experiments shown in Table 1, we investigated the performance of a multimodally pretrained model finetuned on a single modality, in order to test the multimodal model's robustness to the absence of a modality during finetuning. When finetuning using either UI or FT on either spikes or LFPs, we found the models' performance to be comparable to (spikes; Figure 2c) or better than (LFPs; Figure 2d) unimodal models finetuned on the same modality – with the most noticeable improvements being on LFPs as discussed above. These results underscore the potential of cross-modal transfer in building performant models for modalities such as LFPs, for which it is hard to do so in isolation and at smaller scales [9].

We also tested the robustness of our models to missing modalities during inference. When trained with modality masking, our multimodal models performed fairly well when provided only a single modality as input, although they slightly underperformed unimodal models (R^2 = 0.7537 ± 0.0361 and 0.6920 ± 0.0357 for LFPs and spikes respectively). However, models trained without any modality masking were far more brittle and failed to perform well when one modality was missing during inference (R^2 = 0.3967 ± 0.0693 and 0.5073 ± 0.0548 for LFPs and spikes respectively). While our experiments with modality masking were restricted to one configuration (masking probability = 0.2), we believe that with some tuning we can obtain more resilient models.

5 Discussion

Summary. In this work, we extended POYO [4], a spike-based neural decoding model, to handle multimodal neural data consisting of both spikes and LFPs. Quite like POYO, we demonstrated the benefits of multi-session pretraining and large-scale models for downstream generalization performance. Importantly, we obtained stronger performance with multimodal models compared to unimodal models, and also showed the robustness of our multimodal models to the absence of some modalities during finetuning and inference. In fact, we found that a multimodally pretrained model consistently outperformed unimodal LFP-based models when generalizing to new sessions of data with just the LFPs, thus showing evidence of cross-modal transfer. This could help enable low-power, efficient BCIs with performant LFP-based decoders [8, 10].

Limitations and Future Work. While we presented a preliminary investigation on large-scale multimodal neural decoders in this work, there are several outstanding directions which we are keen to explore. For example, we have not yet fully explored the design space of encoder architectures. While we used a single encoder for all modalities in this work (unlike some previous work [32]), it remains to be seen what approach would scale best as the number of modalities increases. We are also actively looking at other pretraining objectives such as masked autoencoding, incorporating cross-modal masking and reconstruction during training [30–32]. Finally, we wish to train our models on many more datasets, modalities, and tasks, which will move us closer towards our goal of building a general-purpose neuro-foundation model.

Acknowledgments and Disclosure of Funding

MGP acknowledges support of a Future Leaders award from the Brain Canada Foundation and a J1 Chercheurs-boursiers en intelligence artificielle from the Fonds de recherche du Québec – Santé. GL acknowledges support from NSERC Discovery Grant RGPIN-2018-04821, the Canada Research Chair in Neural Computations and Interfacing, a Canada-CIFAR AI Chair, IVADO, and the Canada First Research Excellence Fund. The authors also acknowledge the support of computational resources provided by Mila, the Digital Research Alliance of Canada, and NVIDIA that enabled this research.

References

[1] J. Ye, J. Collinger, L. Wehbe, and R. Gaunt. "Neural Data Transformer 2: Multi-context Pretraining for Neural Spiking Activity". *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 80352–80374.

[2] J. Ye, F. Rizzoglio, X. Ma, A. Smoulder, H. Mao, G. H. Blumenthal, W. Hockeimer, N. G. Kunigk, D. D. Moore, P. J. Marino, et al. "A Generalist Intracortical Motor Decoder". *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025.

[3] M. Azabou, V. Arora, V. Ganesh, X. Mao, S. Nachimuthu, M. Mendelson, B. Richards, M. Perich, G. Lajoie, and E. Dyer. "A Unified, Scalable Framework for Neural Population Decoding". *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 44937–44956.

[4] M. Azabou, K. X. Pan, V. Arora, I. J. Knight, E. L. Dyer, and B. A. Richards. "Multi-session, multi-task neural decoding from distinct cell-types and brain regions". *The Thirteenth International Conference on Learning Representations*. 2025.

[5] A. H.-W. Ryoo, N. H. Krishna, X. Mao, M. Azabou, E. L. Dyer, M. G. Perich, and G. Lajoie. "Generalizable, real-time neural decoding with hybrid state-space models". *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025.

[6] W. Jiang, L. Zhao, and B.-l. Lu. "Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI". *The Twelfth International Conference on Learning Representations*. 2024.

[7] 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". *The Thirteenth International Conference on Learning Representations*. 2025.

[8] B. M. Karpowicz, B. Bhaduri, S. R. Nason-Tomaszewski, B. G. Jacques, Y. H. Ali, R. D. Flint, P. H. Bechefsky, L. R. Hochberg, N. AuYong, M. W. Slutzky, et al. "Reducing power requirements for high-accuracy decoding in iBCIs". *Journal of Neural Engineering* 21.6 (2024), p. 066001.

[9] N. Ahmadi, T. G. Constandinou, and C.-S. Bouganis. "Decoding Hand Kinematics from Local Field Potentials Using Long Short-Term Memory (LSTM) Network". *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. 2019, pp. 415–419.

[10] N. Even-Chen, D. G. Muratore, S. D. Stavisky, L. R. Hochberg, J. M. Henderson, B. Murmann, and K. V. Shenoy. "Power-saving design opportunities for wireless intracortical brain–computer interfaces". *Nature Biomedical Engineering* 4.10 (2020), pp. 984–996.

[11] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". *Journal of Basic Engineering* 82.1 (1960), pp. 35–45.

[12] W. Wu, M. Black, Y. Gao, M. Serruya, A. Shaikhouni, J. Donoghue, and E. Bienenstock. "Neural Decoding of Cursor Motion Using a Kalman Filter". *Advances in Neural Information Processing Systems*. Vol. 15. 2002.

[13] S. Koyama, S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. "Comparison of brain–computer interface decoding algorithms in open-loop and closed-loop control". *Journal of Computational Neuroscience* 29.1–2 (2010), pp. 73–87.

[14] F. R. Willett, D. R. Young, B. A. Murphy, W. D. Memberg, C. H. Blabe, C. Pandarinath, S. D. Stavisky, P. Rezaii, J. Saab, B. L. Walter, et al. "Principled BCI Decoder Design and Parameter Selection Using a Feedback Control Model". *Scientific Reports* 9.1 (2019).

[15] B. Jarosiewicz, N. Y. Masse, D. Bacher, S. S. Cash, E. Eskandar, G. Friehs, J. P. Donoghue, and L. R. Hochberg. "Advantages of closed-loop calibration in intracortical brain–computer interfaces for people with tetraplegia". *Journal of Neural Engineering* 10.4 (2013), p. 046012.

[16] V. Gilja, C. Pandarinath, C. H. Blabe, P. Nuyujukian, J. D. Simeral, A. A. Sarma, B. L. Sorice, J. A. Perge, B. Jarosiewicz, L. R. Hochberg, et al. "Clinical translation of a high-performance neural prosthesis". *Nature Medicine* 21.10 (2015), pp. 1142–1145.

[17] Y. Zhang, Y. Wang, D. M. Jiménez-Benetó, Z. Wang, M. Azabou, B. Richards, R. Tung, O. Winter, T. I. B. Laboratory, E. Dyer, et al. "Towards a "Universal Translator" for Neural Dynamics at Single-Cell, Single-Spike Resolution". *Advances in Neural Information Processing Systems*. Vol. 37. 2024, pp. 80495–80521.

[18] Y. Zhang, Y. Wang, M. Azabou, A. Andre, Z. Wang, H. Lyu, I. B. Laboratory, E. L. Dyer, L. Paninski, and C. L. Hurwitz. "Neural Encoding and Decoding at Scale". *Proceedings of the 42nd International Conference on Machine Learning*. Ed. by A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp,

T. Maharaj, K. Wagstaff, and J. Zhu. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 76175–76192.

[19] J. E. O'Doherty, M. M. B. Cardoso, J. G. Makin, and P. N. Sabes. "Nonhuman Primate Reaching with Multichannel Sensorimotor Cortex Electrophysiology". Zenodo: `10.5281/zenodo.3854034`. 2020.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". *International Conference on Learning Representations*. 2021.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[22] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al. "Perceiver IO: A General Architecture for Structured Inputs & Outputs". *International Conference on Learning Representations*. 2022.

[23] A. Gu and T. Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". 2024. arXiv: `2312.00752`.

[24] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, 2014, pp. 1724–1734.

[25] A. Gu, K. Goel, and C. Re. "Efficiently Modeling Long Sequences with Structured State Spaces". *International Conference on Learning Representations*. 2022.

[26] C. Akkus, L. Chu, V. Djakovic, S. Jauch-Walser, P. Koch, G. Loss, C. Marquardt, M. Moldovan, N. Sauter, M. Schneider, et al. "Multimodal Deep Learning". 2023. arXiv: `2301.04856`.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning Transferable Visual Models From Natural Language Supervision". *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.

[28] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu. "Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration". 2023. arXiv: `2306.09093`.

[29] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi. "Review of multimodal machine learning approaches in healthcare". *Information Fusion* 114 (2025), p. 102690.

[30] A. Hussen Abdelaziz, B.-J. Theobald, P. Dixon, R. Knothe, N. Apostoloff, and S. Kajareker. "Modality Dropout for Improved Performance-driven Talking Faces". *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 378–386.

[31] M. Kleinman, A. Achille, and S. Soatto. "Critical Learning Periods for Multisensory Integration in Deep Networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 24296–24305.

[32] C. Fang, C. Sandino, B. Mahasseni, J. Minxha, H. Pouransari, E. Azemi, A. Moin, and E. Zippi. "Promoting cross-modal representations to improve multimodal foundation models for physiological signals". 2024. arXiv: `2410.16424`.

[33] Y. Zhang, P. E. Latham, and A. M. Saxe. "Understanding Unimodal Bias in Multimodal Deep Linear Networks". *Proceedings of the 41st International Conference on Machine Learning*. Ed. by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 59100–59125.

[34] F. R. Willett, J. Li, T. Le, C. Fan, M. Chen, E. Shlizerman, Y. Chen, X. Zheng, T. S. Okubo, T. Benster, et al. "Brain-to-Text Benchmark '24: Lessons Learned". 2024. arXiv: `2412.17227`.

[35] R. D. Flint, E. W. Lindberg, L. R. Jordan, L. E. Miller, and M. W. Slutzky. "Accurate decoding of reaching movements from field potentials in the absence of spikes". *Journal of Neural Engineering* 9.4 (2012), p. 046006.

[36] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. "RoFormer: Enhanced transformer with Rotary Position Embedding". *Neurocomputing* 568 (2024), p. 127063.