Unified Pretraining on Mixed Optophysiology and Electrophysiology Data Across Brain Regions

Ian J. Knight
University of Pennsylvania
Philadelphia, PA, USA
ijknight@upenn.edu

Vinam Arora
University of Pennsylvania
Philadelphia, PA, USA
vinam@upenn.edu

Mehdi Azabou
Columbia University
New York, NY, USA
ma4766@columbia.edu

Zihao Chen
University of Pennsylvania
Philadelphia, PA, USA
zchen95986@upenn.edu

Eva L. Dyer
University of Pennsylvania
Philadelphia, PA, USA
eva.dyer@upenn.edu

Abstract

Building models that unify diverse neural recordings is a crucial step toward scalable foundation models for neuroscience. However, most large-scale models remain tied to a single modality, which limits our ability to integrate information across different spatiotemporal scales. We introduce a POYO-based universal encoder that learns a shared latent representation of electrophysiology (irregular spike times) and optophysiology (regular calcium fluorescence timeseries) without requiring simultaneous recordings. Across large datasets from the Allen Institute spanning both calcium imaging and Neuropixels, we show that joint pretraining outperforms uni-modal baselines and strengthens cross-region transfer. These results show that our mixed-modality pretraining framework can integrate independently collected recordings into a common representational space, advancing the path toward foundation models for diverse multi-modal neural data.

1 Introduction

Neural decoding is entering a new era. Advances in large-scale recording technologies have made it possible to measure brain activity across thousands of neurons, using diverse modalities such as high-density electrophysiology (e.g., Neuropixels (1)) and optophysiology (e.g., calcium imaging (2; 3)). At the same time, machine learning has shown that large pretrained models can achieve strong generalization across tasks, experimental conditions, and neural populations (4; 5; 6; 7; 8; 9; 10; 11). Together, these developments are opening the door to foundation models that can integrate knowledge across datasets.

Multi-modal pretraining has proven highly effective in domains such as vision-language learning, audio-visual representation, and video understanding (12; 13; 14), yet it remains largely unexplored in neuroscience. The diversity of neural datasets arises from the need to obtain complementary views into brain activity with each modality offering distinct strengths. Electrophysiology (EPhys) provides high temporal resolution and precise spike timing, making it possible to study fast neural dynamics. Optophysiology (OPhys), by contrast, offers single-cell resolution across large populations and can leverage genetic tools to target specific neuronal types and classes. These complementary perspectives are typically collected in separate experiments and rarely recorded simultaneously (15; 16). This creates both a challenge and an opportunity: how can we design a universal encoder that learns from independently acquired datasets and builds a common representation across modalities? An

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

ideal solution would allow us to build joint models of neural activity that take advantage of the complementary qualities that distinguish these modalities. In this work, we take a step in this direction by pre-training models on different mixtures of EPhys and OPhys data and evaluating how multi-modal training impacts decoding performance, cross-region transfer, and generalization to data-scarce settings. We introduce a POYO-based (4) universal encoder that unifies EPhys and OPhys in a shared latent space. Our results show that joint pretraining across modalities can provide substantial improvements over modality-specific training. We demonstrate that multi-modal models not only outperform single-modality baselines on held-out sessions, but also provide stronger generalization in cross-region transfer tasks. The main contributions of this work are:

- **Universal encoder for mixed-modality training:** We design a model that ingests both spikes and calcium activity into a unified latent representation without requiring simultaneous recordings.
- **Systematic study of multi-modal pretraining:** We evaluate how different mixtures of modalities affect scaling trends, showing that multi-modal data improves decoding and generalization compared to single-modality pretraining.
- **Cross-region and region-by-region analyses:** We demonstrate that exposure to regions through EPhys boosts transfer to unseen OPhys regions, and we dissect where EPhys-derived improvements arise across the visual cortex.

2 Methods

2.1 Tokenization

We develop a unified tokenization scheme for handling both electrophysiology data (irregular spikes) and optophysiology data (calcium traces).

Spikes. Following POYO (4), each spike is treated as an individual event in our model. For a neuron u that emits spikes at times $t_{u,i}$, we create a token $(\mathbf{x}_u, t_{u,i})$ for each event. Here, $\mathbf{x}_u \in \mathbb{R}^D$ is a learned embedding for the neuron u.

Calcium Traces. Following POYO+ (6), for each neuron u at time-step i, we create a token $(\mathbf{x}_u, \mathbf{f}_{u,i}, t_{u,i})$, where $\mathbf{x}_u$ is the neuron's learned embedding and $\mathbf{f}_i$ is the fluorescence value at the time-step. We embed this token into the latent space as the tuple $(\mathbf{x}_{u,i}, t_{u,i})$, where $\mathbf{x}_{u,i} = [\mathbf{x}_u, \mathbf{W}_f \mathbf{f}_{u,i}]$.

Both irregular spike trains and regularly-sampled calcium timeseries are mapped to a common sequence of tokens $(\mathbf{x}_{u,i}, t_{u,i})$ that serve as inputs to our encoder.

2.2 Encoder

We employ a PerceiverIO-style encoder (17) adapted for multi-modal inputs. The encoder starts with two separate cross-attention blocks, one for each modality. The EPhys-specific cross-attention block compresses the spike tokens, while the OPhys-specific cross-attention block compresses the calcium tokens. We use modality-specific query tokens for each cross-attention blocks, these query tokens share the same number and shape to allow identical processing by subsequent layers. In our ablation study, we find that a two cross-attention head solution is better than using separate encoders (see Appendix Table 1 in Section E).

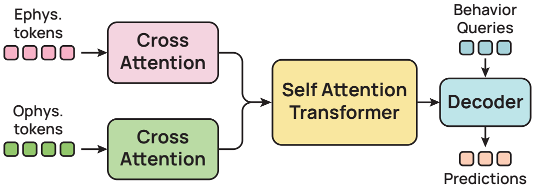

Figure 1: **Model Architecture.** Tokens from each input modality are projected into a common latent space using cross-attention blocks. Latents are further processed using a transformer, and finally behavior is queried through a decoder cross-attention block.

Formally, we initialize $M \times N$ latent token tuples for each modality, $(\mathbf{z}_{mn}^{(0,\text{EPhys})}, \tau_{mn})$ and $(\mathbf{z}_{mn}^{(0,\text{OPhys})}, \tau_{mn})$, where $\mathbf{z}_{mn}^{(0,\text{EPhys})}, \mathbf{z}_{mn}^{(0,\text{OPhys})} \in \mathbb{R}^D$ are learned vectors and τ_{mn} denote virtual timesteps which are uniformly spaced across the context window. Each cross-attention maps the input tokens into latents:

$$\mathbf{z}_{mn}^{(1)} = \mathbf{z}_{mn}^{(0)} + \sum_{u=1}^{U} \sum_{i=1}^{T_u} \text{softmax}\Big(\big(\mathbf{R}(\tau_{mn})\mathbf{q}_{mn}\big)^\top \big(\mathbf{R}(t_{u,i})\mathbf{k}_{u,i}\big)\Big)\, \mathbf{v}_{u,i}, \tag{1}$$

where values $\mathbf{v}_{u,i} = \mathbf{W}_V \mathbf{x}_{u,j}$ and keys $\mathbf{k}_{u,i} = \mathbf{W}_K \mathbf{x}_{u,i}$ are derived from the input tokens, queries $\mathbf{q}_{mn} = \mathbf{W}_Q \mathbf{z}^{(0)}_{mn}$ from latent tokens, and $\mathbf{R}(t)$ are rotary position embedding matrices (18; 4). Note that $\mathbf{z}^{(0)}, \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V$ are different for the two modalities.

After the cross-attention based compression, we treat the latent tokens from both modalities identically. This means, during training with mini-batches, we can simply stack the EPhys and OPhys. latent sequences, $\mathbf{z}^{(1)}$, into a batch and pass them through the subsequent layers. All latent tokens go through the same set of self-attention transformer blocks for further refinement (Figure 1) :

$$\mathbf{z}^{(l+1)}_{mn} = \mathbf{z}^{(l)}_{mn} + \sum_{m',n'} \mathrm{softmax}\Big(\big(\mathbf{R}(\tau_{mn}) \mathbf{W}_Q \mathbf{z}^{(l)}_{mn}\big)^\top \big(\mathbf{R}(\tau_{m'n'}) \mathbf{W}_K \mathbf{z}^{(l)}_{m'n'}\big) \Big)\, \mathbf{W}_V \mathbf{z}^{(l)}_{m'n'}. \tag{2}$$

2.3 Decoder

To train the model, we employ a single layer cross-attention decoder that maps the shared latent representation to task-specific outputs. Given the final latent tokens $\mathbf{z}^{(L)}_{mn}$, a *learned* query token $\mathbf{o}$ is used to attend over the latent sequence. The query token also has an associated timestamp t_i to facilitate sequence-to-sequence tasks:

$$\mathbf{y}_i = \mathbf{o} + \sum_{m,n} \mathrm{softmax}\Big(\big(\mathbf{R}(t_i) \mathbf{W}_Q \mathbf{o}\big)^\top \big(\mathbf{R}(\tau_{mn}) \mathbf{W}_K \mathbf{z}^{(L)}_{mn}\big) \Big)\, \mathbf{W}_V \mathbf{z}^{(L)}_{mn}. \tag{3}$$

The resulting outputs $\mathbf{y}_i$ are passed through a linear projection $\mathbf{W}_{\text{task}} : \mathbb{R}^D \rightarrow \mathbb{R}^{D_{\text{task}}}$, where D_{task} is the dimensionality of the target variable. For regression tasks (e.g., continuous stimulus decoding), we use mean-squared error loss, whereas for classification tasks we use cross-entropy loss.

2.4 Universal Multi-modal Training

A key property of our approach is that it does not require simultaneously recorded multi-modal data. Instead, we treat EPhys and OPhys recordings as complementary but independently collected views of neural activity. During training, datasets from different modalities, regions, and cre-lines are combined into a single supervised learning framework. Each batch may contain trials from either modality, and the shared encoder learns representations that generalize across them. This strategy enables the model to function as a universal encoder: on shared tasks, it can integrate information from independently acquired datasets and improve generalization in new experimental contexts where only one modality is available.

3 Results

Setup. We train our models on two large-scale datasets from the Allen Institute: the Allen Brain Observatory (OPhys) (2) and the Allen Neuropixels survey (EPhys) (1). These datasets share an experimental protocol during which neural signals of mice are recorded in response to visual stimuli (see Appendix A for more details). From these large datasets, we randomly select 100 OPhys sessions containing the 5 most prevalent regions (VISp, VISpm, VISam, VISal, VISl) in equal proportion and 100 EPhys sessions in the same fashion. Our model is trained to decode three of these stimuli: static gratings orientation (6 classes), drifting gratings orientation (6 classes), and natural scenes (119 classes). We holdout two sessions from each region in each modality for evaluation purposes; the remaining sessions are used for pre-training. See Appendices B and C for more details on our training and finetuning setup.

3.1 Scaling trends in multi-modal pre-training as a function of data mixtures

In this first set of experiments, we ask whether mixing in EPhys data during pre-training can help improve downstream performance on OPhys data. In particular, we devise different mixtures of OPhys and EPhys data with different modality ratios (see Appendix E). We pre-train models on each mixture and then finetune each on held-out OPhys session to evaluate their generalization. For each of the tasks, we report model performance as a function of the pre-training set size in Figure 2.

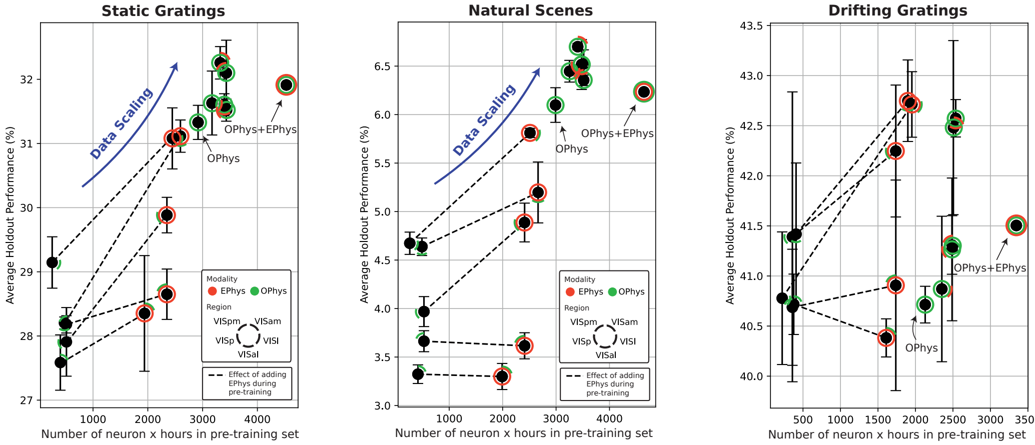

Figure 2: *Scaling trends in multi-modal pre-training as a function of data mixtures. Reported accuracy is the average of 10 evaluation sessions across 3 seeds. Error bars denote standard error of mean.*

Observation 1: For both static gratings and natural scenes, we observe a positive scaling trend in the decoding of these stimuli as we increase the amount of pre-training data. This remains true for various mixtures of EPhys and OPhys.

Observation 2: A model pre-trained only on multi-region OPhys data is never the top model for any of the tasks. This is true even when compared to models that are pre-trained on similarly-sized datasets. In particular, we find that for the drifting gratings task, the OPhys only model is outperformed by models that were pretrained on primarily EPhys data. This result suggests that given the same pre-training data budget, multi-modal pretraining brings diversity that boosts generalization.

Observation 3: In the common scenario where OPhys data is only collected in one region, we can supplement the pre-training set using pre-existing multi-region EPhys data. This trend is highlighted by the dotted lines in Figure 2. We find that this leads to significantly improved performance across the board. This demonstrates the enormous potential of multi-modal training in regimes where access to data is easier for one modality (EPhys) compared to the other (OPhys).

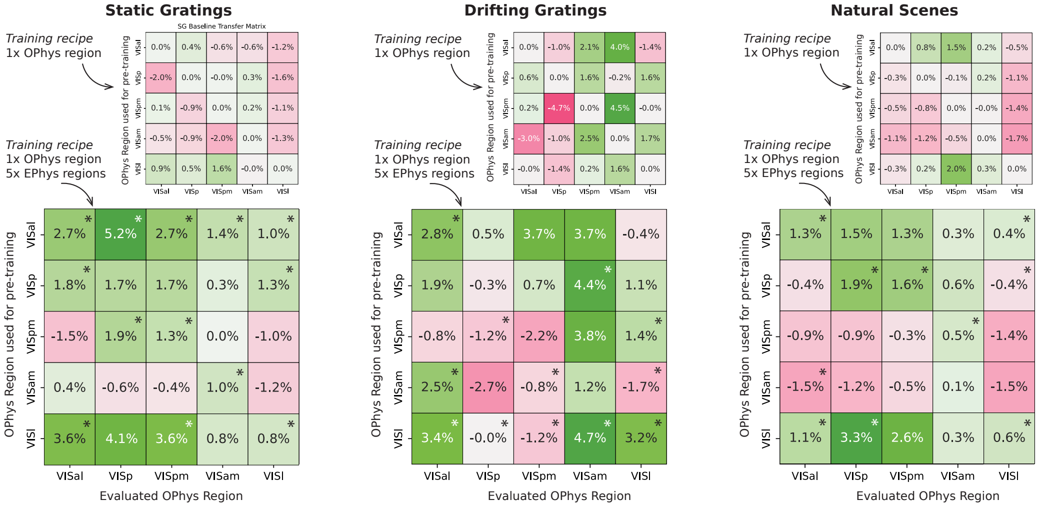

Figure 3: *Comparison of baseline transfer relationship for single-region OPhys models and the change in the transfer relationship induced by adding multi-region EPhys. (*) indicates transfer relationships whose direction of change was consistent across all three seeds (i.e., all increased or decreased in transfer quality).*

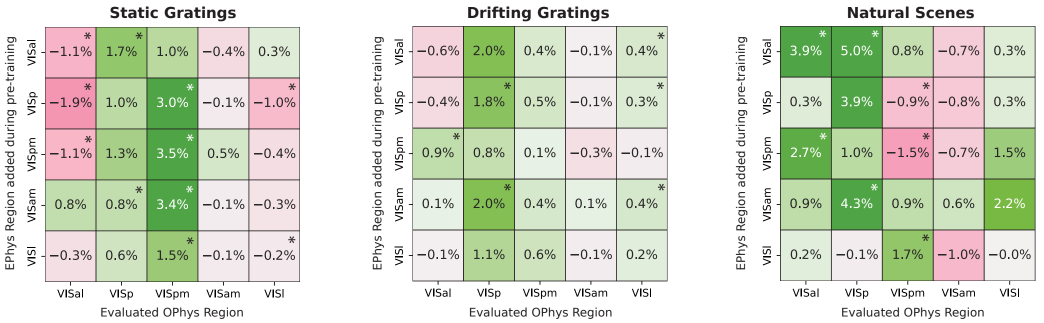

Figure 4: *Change in performance induced by adding region-specific EPhys recording to a multi-region OPhys pre-training dataset, compared to a uni-modal OPhys model. (*) indicates the change was consistent across all three seeds (i.e., all increased or decreased in performance).*

3.2 Cross-region transfer in the context of multi-modal pretraining.

Prior work has investigated the ability of models pretrained on a single region to transfer to an unseen region. In a recent paper introducing the POYO+ model Azabou et al. (6) pre-trained OPhys models on one region, then finetuned them on recordings from other regions for evaluation. Here, we extend those experiments by incorporating multi-region EPhys data into the pre-training. Specifically, we ask whether exposure to a region through EPhys can improve transfer when the model has not seen that region in OPhys. Figure 3 reports the improvements over a baseline where models are pre-trained and evaluated on the same region. We compare (i) a uni-modal OPhys model trained on a single region and (ii) a multi-modal model trained on one OPhys region plus all region in EPhys. When comparing the transfer matrices, we observe an overall improvement in cross-region transfer. This confirms that exposure to a region through EPhys is sufficient to generalize to that region in the OPhys modality.

Importantly, we note that regions that transfer poorly in the uni-modal setting e.g. VISpm-only or VISam in all tasks, demonstrate improved transfer in the multi-modal setting. Other conclusions can be drawn from these matrices. Notably, the multi-modal model trained on VISal generalizes consistently well across all other regions. This suggests that pairing OPhys data from VISal with EPhys data spanning the visual cortex may be an effective recipe for building generalizable pre-trained models for OPhys data across visual areas.

3.3 Dissecting the source of EPhys gains based on regions

In the previous section we studied models that are pre-trained on OPhys data from a single-region at a time. We now study models that are exposed to OPhys data from all regions and measure how additional EPhys data from a given region influences downstream performance in each OPhys region. Figure 4 shows improvements in performance compared to a OPhys-only baseline. The effect of adding EPhys from a given region depends strongly on the target OPhys region and task. This is supported by the consistent trends along each column in the matrices in Figure 4. OPhys recordings in both VISp and VISpm generally benefit from additional EPhys data regardless of the task, whereas in VISam we see a slight decrease in performance. Other regions exhibit task-specific trends, for instance, VISal benefits from EPhys data collected during naturalistic stimulus presentations but not during artificial ones. These results highlight that EPhys does not always provide uniform gains across all regions and tasks, an effect also present in Figure 2 for drifting gratings. This underscores the value of strategically selecting complementary EPhys datasets when aiming to boost OPhys performance.

4 Discussion

We introduced a Perceiver-based universal encoder that integrates electrophysiology (Ephys) and optophysiology (Ophys) into a shared latent space without requiring simultaneously recorded multi-modal data. Our approach demonstrates that independently collected datasets can be combined to yield improved decoding, stronger transfer across brain regions. The analyses we presented: scaling trends, cross-region transfer, and region-by-region breakdowns, help clarify how multi-modal pre-training drives these improvements and can guide more effective strategies for combining modalities in the future.

References

[1] J. H. Siegle, X. Jia, S. Durand, S. Gale, C. Bennett, N. Graddis, G. Heller, T. K. Ramirez, H. Choi, J. A. Luviano, *et al.*, “Survey of spiking in the mouse visual system reveals functional hierarchy,” *Nature*, vol. 592, no. 7852, pp. 86–92, 2021.

[2] S. E. de Vries, J. A. Lecoq, M. A. Buice, P. A. Groblewski, G. K. Ocker, M. Oliver, D. Feng, N. Cain, P. Ledochowitsch, D. Millman, *et al.*, “A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex,” *Nature neuroscience*, vol. 23, no. 1, pp. 138–151, 2020.

[3] J. L. Stirman, I. T. Smith, M. W. Kudenov, and S. L. Smith, “Wide field-of-view, multi-region, two-photon imaging of neuronal activity in the mammalian brain,” *Nature Biotechnology*, vol. 34, pp. 857–862, Aug 2016.

[4] 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, pp. 44937–44956, 2023.

[5] 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, pp. 80352–80374, 2023.

[6] 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,” in *The Thirteenth International Conference on Learning Representations*, 2024.

[7] A. Antoniades, Y. Yu, J. Canzano, W. Wang, and S. L. Smith, “Neuroformer: Multimodal and multitask generative pretraining for brain data,” 2024.

[8] Y. Zhang, Y. Wang, M. Azabou, A. Andre, Z. Wang, H. Lyu, T. I. B. Laboratory, E. Dyer, L. Paninski, and C. Hurwitz, “Neural encoding and decoding at scale,” 2025.

[9] Y. Zhang, Y. Wang, D. J. Benetó, Z. Wang, M. Azabou, B. Richards, O. Winter, I. B. Laboratory, E. Dyer, L. Paninski, and C. Hurwitz, “Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution,” *arXiv preprint arXiv:2407.14668v2*, 2024.

[10] F. Yang, C. Feng, D. Wang, T. Wang, Z. Zeng, Z. Xu, H. Park, P. Ji, H. Zhao, Y. Li, and A. Wong, “Neurobind: Towards unified multimodal representations for neural signals,” 2024.

[11] C. Pandarinath, D. J. O’Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, and D. Sussillo, “Inferring single-trial neural population dynamics using sequential auto-encoders,” *Nature Methods*, vol. 15, no. 10, pp. 805–815, 2018.

[12] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Bransom, K. Ehsani, H. Ngo, Y. Chen, A. Patel, M. Yatskar, C. Callison-Burch, A. Head, R. Hendrix, F. Bastani, E. VanderBilt, N. Lambert, Y. Chou, A. Chheda, J. Sparks, S. Skjonsberg, M. Schmitz, A. Sarnat, B. Bischoff, P. Walsh, C. Newell, P. Wolters, T. Gupta, K.-H. Zeng, J. Borchardt, D. Groeneveld, C. Nam, S. Lebrecht, C. Wittlif, C. Schoenick, O. Michel, R. Krishna, L. Weihs, N. A. Smith, H. Hajishirzi, R. Girshick, A. Farhadi, and A. Kembhavi, “Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models,” 2024.

[13] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “Imagebind: One embedding space to bind them all,” 2023.

[14] L. Zhao, N. B. Gundavarapu, L. Yuan, H. Zhou, S. Yan, J. J. Sun, L. Friedman, R. Qian, T. Weyand, Y. Zhao, R. Hornung, F. Schroff, M.-H. Yang, D. A. Ross, H. Wang, H. Adam, M. Sirotenko, T. Liu, and B. Gong, “Videoprism: A foundational visual encoder for video understanding,” 2025.

[15] L. Huang, P. Ledochowitsch, U. Knoblich, J. Lecoq, G. J. Murphy, R. C. Reid, S. E. J. de Vries, C. Koch, H. Zeng, M. A. Buice, J. Waters, and L. Li, "Relationship between simultaneously recorded spiking activity and fluorescence signal in gcamp6 transgenic mice," *eLife*, vol. 10, p. e51675, 2021.

[16] P. Ledochowitsch, L. Huang, U. Knoblich, M. Oliver, J. Lecoq, R. C. Reid, L. Li, H. Zeng, C. Koch, J. Waters, S. E. de Vries, and M. A. Buice, "On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging," *bioRxiv*, 2020.

[17] 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," *arXiv preprint arXiv:2107.14795*, 2021.

[18] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

Appendix

A Dataset

The Brain Observatory is the largest public collection of two-photon calcium imaging in mouse visual cortex (2). It contains recordings across six cortical areas (VISp, VISpm, VISam, VISrl, VISal, VISl) in response to a wide range of visual stimuli, including drifting gratings, static gratings, natural scenes, and natural movies. Each experiment consists of three one-hour imaging sessions per animal, spanning depths from 150–600 μm and covering layers L2/3 through L6. Recordings are genetically targeted using Cre driver lines: two pan-excitatory (Emx1, Slc17a7), eight excitatory sub-type-specific (Cux2, Rorb, Scnn1a, Nr5a1, Rbp4, Fezf2, Tlx3, Ntsr1), and three inhibitory (Vip, SST, PV). This provides diverse populations with distinct response properties and heterogeneous cell densities across regions. The Neuropixels survey dataset (1) provides complementary large-scale extracellular electrophysiology, recorded with high temporal precision across overlapping visual areas under similar stimulus conditions. This combination allows us to examine integration across modalities with distinct spatial and temporal resolutions: OPhys offering broad coverage and genetic specificity and EPhys providing precise spike timing.

B Training

During training, we applied neuron dropout, removing random subsets of neurons in each context window to encourage robustness. This transformation was removed during evaluation. Models were trained for 300 epochs with a batch size of 32, using a Lamb optimizer with a base learning rate of 3.125×10^{-5} (scaled linearly by batch size) and a weight decay of 1×10^{-4}. A OneCycleLR learning rate scheduler was used with a cosine annealing strategy. The learning rate was set to the initial value until halfway then decayed smoothly for the remaining steps. All models were set to train for 300 epochs with early stopping enabled. Model checkpointing and early stopping is decided by the best average metric between OPhys and EPhys. The majority of training parameters were based on the POYO and POYO+ training parameters available at the torch_brain repository.

C Finetuning

In each experiment, we held-out a subset of sessions. To evaluate the quality of these models we finetuned on the training split of these held-out sessions and reported accuracy on their test splits. Fine-tuning followed a gradual unfreezing schedule: unit and session embeddings were unfrozen first, followed by the decoder and encoder, although in practice convergence was often reached after unfreezing only the unit embeddings. Models were trained for 50 epochs with unfreezing occurring at 25 epochs.

D Ablation Study

We conducted an ablation study to confirm the effectiveness of our design choices for incorporating the two input modalities in our model. Specifically, we compare against three alternate architectures:

- **Common CA** uses a shared cross-attention encoder block at the input, instead of having a separate cross-attention block for each modality.
- **One-layer unique SA** uses separate cross-attention blocks for each modality and additionally adds a single layer of unique self-attention blocks for each modality.
- **Separate encoder** model has completely independent encoders for the both modality. In other words, each modality has its own cross-attention and self-attention transformers, and only the decoder is shared between the two modalities.

In all cases, the number of layers in the model were adjusted to match the parameter count of the main model. Results are presented in table 1 and show that our presented architecture performs the best on OPhys datasets when the model is trained on both EPhys and OPhys data simultaneously.

	Static Gratings	Natural Scenes	Drifting Gratings
Main	**0.3297** ± 0.0026	**0.0833** ± 0.0028	**0.4212** ± 0.0011
Common CA	0.3289 ± 0.0014	0.0744 ± 0.0037	0.4038 ± 0.0037
One-layer unique SA	0.3288 ± 0.0049	0.0687 ± 0.0142	0.4130 ± 0.0016
Separate encoder	0.2741 ± 0.0570	0.0809 ± 0.0016	0.4109 ± 0.0006

Table 1: **Ablation study.** Test performance on the pretraining set for different architecture variations across three different behavior tasks. Bold indicates best performing variation. SEM is reported along with the mean performance across three seeds.

E Full Transfer Tables

The full transfer tables show the results of each model for each task averaged across three seeds. The variance is reported as the standard error of mean. The complete count of sessions in each pre-training set for each model is given in the leftmost column where 5x18 denotes 5 regions with 18 sessions per region. When a set of sessions is referenced e.g. EPhys (5x18) any repeats of this phrase in a set description e.g. EPhys (5x18) + OPhys VISl (18) refer to the same sessions with an additional 18 session of OPhys VISl added.

Table 2: Static Gratings Transfer Table

Pretraining set	**EPhys**					**OPhys**				
	VISal	VISp	VISpm	VISam	VISl	VISal	VISp	VISpm	VISam	VISl
EPhys (5x18)	0.4299 ± 0.0060	0.6257 ± 0.0031	0.4657 ± 0.0037	0.5472 ± 0.0125	0.3331 ± 0.0030	–	–	–	–	–
EPhys (5x18) + OPhys VISal (18)	0.4500 ± 0.0052	0.6083 ± 0.0036	0.4743 ± 0.0039	0.5366 ± 0.0063	0.3336 ± 0.0067	0.3514 ± 0.0003	0.3508 ± 0.0023	0.3523 ± 0.0061	0.2212 ± 0.0036	0.2800 ± 0.0058
EPhys (5x18) + OPhys VISp (18)	0.4385 ± 0.0063	0.6218 ± 0.0020	0.4606 ± 0.0103	0.5472 ± 0.0069	0.3247 ± 0.0025	0.3431 ± 0.0051	0.3165 ± 0.0136	0.3419 ± 0.0030	0.2102 ± 0.0048	0.2826 ± 0.0069
EPhys (5x18) + OPhys VISpm (18)	0.4360 ± 0.0064	0.6192 ± 0.0041	0.4735 ± 0.0020	0.5399 ± 0.0069	0.3437 ± 0.0080	0.3092 ± 0.0072	0.3180 ± 0.0044	0.3383 ± 0.0080	0.2074 ± 0.0039	0.2596 ± 0.0070
EPhys (5x18) + OPhys VISam (18)	0.4372 ± 0.0066	0.6162 ± 0.0119	0.4821 ± 0.0017	0.5310 ± 0.0020	0.3426 ± 0.0074	0.3285 ± 0.0168	0.2935 ± 0.0243	0.3207 ± 0.0143	0.2168 ± 0.0015	0.2581 ± 0.0055
EPhys (5x18) + OPhys VISl (18)	0.4419 ± 0.0030	0.6067 ± 0.0097	0.4696 ± 0.0027	0.5382 ± 0.0107	0.3213 ± 0.0124	0.3603 ± 0.0103	0.3397 ± 0.0163	0.3613 ± 0.0031	0.2147 ± 0.0019	0.2780 ± 0.0055
OPhys (5x18)	–	–	–	–	–	0.3674 ± 0.0016	0.3424 ± 0.0048	0.3455 ± 0.0075	0.2252 ± 0.0046	0.2860 ± 0.0033
OPhys (5x18) + EPhys VISal (18)	0.4265 ± 0.0067	0.6019 ± 0.0063	0.4651 ± 0.0034	0.5363 ± 0.0029	0.3373 ± 0.0081	0.3562 ± 0.0033	0.3591 ± 0.0066	0.3553 ± 0.0071	0.2211 ± 0.0009	0.2885 ± 0.0034
OPhys (5x18) + EPhys VISp (18)	0.4352 ± 0.0063	0.6304 ± 0.0068	0.4572 ± 0.0053	0.5307 ± 0.0051	0.3037 ± 0.0189	0.3488 ± 0.0060	0.3519 ± 0.0024	0.3753 ± 0.0051	0.2241 ± 0.0040	0.2759 ± 0.0010
OPhys (5x18) + EPhys VISpm (18)	0.4372 ± 0.0008	0.6050 ± 0.0089	0.4821 ± 0.0104	0.5274 ± 0.0037	0.3241 ± 0.0092	0.3565 ± 0.0062	0.3556 ± 0.0091	0.3809 ± 0.0101	0.2298 ± 0.0046	0.2823 ± 0.0035
OPhys (5x18) + EPhys VISam (18)	0.4307 ± 0.0039	0.6064 ± 0.0044	0.4587 ± 0.0063	0.5374 ± 0.0047	0.3317 ± 0.0051	0.3754 ± 0.0095	0.3505 ± 0.0048	0.3791 ± 0.0021	0.2244 ± 0.0055	0.2834 ± 0.0027
OPhys (5x18) + EPhys VISl (18)	0.4260 ± 0.0034	0.5994 ± 0.0053	0.4631 ± 0.0054	0.5296 ± 0.0049	0.3563 ± 0.0052	0.3642 ± 0.0077	0.3479 ± 0.0141	0.3610 ± 0.0060	0.2241 ± 0.0070	0.2844 ± 0.0042
OPhys (5x18) + EPhys (5x18)	0.4461	0.6078	0.4779	0.5492	0.3286	0.3591	0.3444	0.3868	0.2236	0.2817
OPhys VISal (18)	–	–	–	–	–	0.3256 ± 0.0040	0.3038 ± 0.0083	0.3210 ± 0.0100	0.2017 ± 0.0033	0.2581 ± 0.0030
OPhys VISp (18)	–	–	–	–	–	0.3059 ± 0.0139	0.2997 ± 0.0114	0.3248 ± 0.0105	0.2104 ± 0.0066	0.2546 ± 0.0016
OPhys VISpm (18)	–	–	–	–	–	0.3256 ± 0.0113	0.2905 ± 0.0105	0.3252 ± 0.0081	0.2089 ± 0.0082	0.2590 ± 0.0041
OPhys VISam (18)	–	–	–	–	–	0.3202 ± 0.0049	0.2898 ± 0.0188	0.3050 ± 0.0110	0.2074 ± 0.0049	0.2569 ± 0.0065
OPhys VISl (18)	–	–	–	–	–	0.3342 ± 0.0091	0.3042 ± 0.0106	0.3410 ± 0.0048	0.2077 ± 0.0034	0.2702 ± 0.0028
EPhys VISal (18)	0.4349 ± 0.0049	0.6042 ± 0.0093	0.4626 ± 0.0024	0.5346 ± 0.0095	0.3124 ± 0.0070	–	–	–	–	–
EPhys VISp (18)	0.4352 ± 0.0055	0.6094 ± 0.0065	0.4659 ± 0.0014	0.5357 ± 0.0051	0.3163 ± 0.0063	–	–	–	–	–
EPhys VISpm (18)	0.4277 ± 0.0060	0.6044 ± 0.0110	0.4642 ± 0.0031	0.5307 ± 0.0030	0.3264 ± 0.0128	–	–	–	–	–
EPhys VISam (18)	0.4304 ± 0.0031	0.6028 ± 0.0021	0.4559 ± 0.0036	0.5299 ± 0.0112	0.3224 ± 0.0035	–	–	–	–	–
EPhys VISl (18)	0.4237 ± 0.0042	0.5977 ± 0.0037	0.4567 ± 0.0060	0.5156 ± 0.0040	0.3191 ± 0.0121	–	–	–	–	–

Table 3: Drifting Gratings Transfer Table

Pretraining set	**EPhys**					**OPhys**				
	VISal	VISp	VISpm	VISam	VISl	VISal	VISp	VISpm	VISam	VISl
EPhys (5x18)	0.8456 ± 0.0207	0.8745 ± 0.0099	0.7359 ± 0.0062	0.6842 ± 0.0169	0.7359 ± 0.0265	–	–	–	–	–
EPhys (5x18) + OPhys VISal (18)	0.8281 ± 0.0123	0.8940 ± 0.0031	0.7340 ± 0.0071	0.6912 ± 0.0123	0.7219 ± 0.0099	0.4580 ± 0.0044	0.5442 ± 0.0181	0.4427 ± 0.0117	0.4016 ± 0.0199	0.2910 ± 0.0198
EPhys (5x18) + OPhys VISp (18)	0.8316 ± 0.0061	0.8870 ± 0.0047	0.7375 ± 0.0036	0.6559 ± 0.0106	0.7185 ± 0.0047	0.4488 ± 0.0049	0.5359 ± 0.0062	0.4133 ± 0.0125	0.4087 ± 0.0091	0.3057 ± 0.0180
EPhys (5x18) + OPhys VISpm (18)	0.8333 ± 0.0213	0.8834 ± 0.0170	0.7358 ± 0.0031	0.6594 ± 0.0123	0.7256 ± 0.0185	0.4217 ± 0.0284	0.5267 ± 0.0078	0.3844 ± 0.0212	0.4032 ± 0.0192	0.3093 ± 0.0089
EPhys (5x18) + OPhys VISam (18)	0.8316 ± 0.0061	0.8871 ± 0.0155	0.7251 ± 0.0081	0.6770 ± 0.0133	0.7220 ± 0.0062	0.4549 ± 0.0183	0.5120 ± 0.0211	0.3978 ± 0.0102	0.3770 ± 0.0207	0.2775 ± 0.0088
EPhys (5x18) + OPhys VISl (18)	0.8667 ± 0.0262	0.8780 ± 0.0153	0.7429 ± 0.0064	0.6683 ± 0.0127	0.7395 ± 0.0077	0.4643 ± 0.0167	0.5387 ± 0.0076	0.3938 ± 0.0101	0.4115 ± 0.0103	0.3273 ± 0.0085
OPhys (5x18)	–	–	–	–	–	0.4223 ± 0.0053	0.5113 ± 0.0206	0.3931 ± 0.0102	0.4001 ± 0.0118	0.3090 ± 0.0057
OPhys (5x18) + EPhys VISal (18)	0.8649 ± 0.0088	0.8693 ± 0.0127	0.7340 ± 0.0047	0.6858 ± 0.0063	0.6990 ± 0.0052	0.4616 ± 0.0071	0.5613 ± 0.0158	0.4008 ± 0.0295	0.3931 ± 0.0169	0.3118 ± 0.0134
OPhys (5x18) + EPhys VISp (18)	0.8351 ± 0.0137	0.8605 ± 0.0093	0.7323 ± 0.0093	0.6613 ± 0.0152	0.7061 ± 0.0071	0.4248 ± 0.0196	0.5505 ± 0.0072	0.3838 ± 0.0125	0.3925 ± 0.0322	0.3116 ± 0.0018
OPhys (5x18) + EPhys VISpm (18)	0.8386 ± 0.0220	0.8709 ± 0.0047	0.7340 ± 0.0018	0.6753 ± 0.0127	0.6885 ± 0.0106	0.4494 ± 0.0203	0.5209 ± 0.0113	0.3783 ± 0.0062	0.3931 ± 0.0119	0.3237 ± 0.0154
OPhys (5x18) + EPhys VISam (18)	0.8421 ± 0.0161	0.8834 ± 0.0092	0.7324 ± 0.0169	0.6770 ± 0.0082	0.6955 ± 0.0283	0.4311 ± 0.0091	0.5541 ± 0.0043	0.4017 ± 0.0185	0.4059 ± 0.0134	0.3312 ± 0.0184
OPhys (5x18) + EPhys VISl (18)	0.8018 ± 0.0063	0.8835 ± 0.0141	0.7412 ± 0.0061	0.6964 ± 0.0063	0.7008 ± 0.0077	0.4244 ± 0.0162	0.5106 ± 0.0125	0.4101 ± 0.0168	0.3897 ± 0.0064	0.3086 ± 0.0091
OPhys (5x18) + EPhys (5x18)	0.8474	0.8675	0.7782	0.7090	0.7045	0.4341	0.5584	0.3773	0.4197	0.2857
OPhys VISal (18)	–	–	–	–	–	0.4298 ± 0.0085	0.5287 ± 0.0192	0.4272 ± 0.0147	0.4047 ± 0.0060	0.2804 ± 0.0133
OPhys VISp (18)	–	–	–	–	–	0.4356 ± 0.0407	0.5389 ± 0.0280	0.4217 ± 0.0082	0.3627 ± 0.0248	0.3107 ± 0.0106
OPhys VISpm (18)	–	–	–	–	–	0.4320 ± 0.0079	0.4916 ± 0.0012	0.4061 ± 0.0151	0.4101 ± 0.0149	0.2946 ± 0.0153
OPhys VISam (18)	–	–	–	–	–	0.3994 ± 0.0029	0.5291 ± 0.0054	0.4310 ± 0.0226	0.3648 ± 0.0088	0.3116 ± 0.0041
OPhys VISl (18)	–	–	–	–	–	0.4295 ± 0.0141	0.5252 ± 0.0039	0.4086 ± 0.0103	0.3807 ± 0.0144	0.2949 ± 0.0119
EPhys VISal (18)	0.8456 ± 0.0035	0.8551 ± 0.0035	0.7500 ± 0.0107	0.6735 ± 0.0179	0.7095 ± 0.0000	–	–	–	–	–
EPhys VISp (18)	0.8508 ± 0.0140	0.8816 ± 0.0174	0.7287 ± 0.0124	0.6508 ± 0.0109	0.7149 ± 0.0309	–	–	–	–	–
EPhys VISpm (18)	0.7895 ± 0.0299	0.8552 ± 0.0018	0.7483 ± 0.0064	0.6701 ± 0.0124	0.7186 ± 0.0168	–	–	–	–	–
EPhys VISam (18)	0.8439 ± 0.0195	0.8604 ± 0.0117	0.7270 ± 0.0138	0.6983 ± 0.0185	0.7218 ± 0.0097	–	–	–	–	–
EPhys VISl (18)	0.8333 ± 0.0017	0.8994 ± 0.0081	0.7325 ± 0.0155	0.6541 ± 0.0184	0.7218 ± 0.0046	–	–	–	–	–

Table 4: Natural Scenes Transfer Table

Pretraining set	**EPhys**					**OPhys**				
	VISal	VISp	VISpm	VISam	VISl	VISal	VISp	VISpm	VISam	VISl
EPhys (5x18)	0.1935 ± 0.0076	0.5081 ± 0.0066	0.2592 ± 0.0066	0.2949 ± 0.0178	0.2386 ± 0.0018	–	–	–	–	–
EPhys (5x18) + OPhys VISal (18)	0.1530 ± 0.0209	0.5315 ± 0.0061	0.2734 ± 0.0303	0.2602 ± 0.0124	0.2076 ± 0.0236	0.0754 ± 0.0050	0.0573 ± 0.0065	0.0554 ± 0.0048	0.0204 ± 0.0030	0.0513 ± 0.0020
EPhys (5x18) + OPhys VISp (18)	0.1776 ± 0.0169	0.5427 ± 0.0123	0.2798 ± 0.0127	0.2594 ± 0.0250	0.2386 ± 0.0081	0.0584 ± 0.0084	0.0607 ± 0.0058	0.0577 ± 0.0028	0.0236 ± 0.0013	0.0439 ± 0.0011
EPhys (5x18) + OPhys VISpm (18)	0.1784 ± 0.0126	0.5545 ± 0.0020	0.3038 ± 0.0102	0.3041 ± 0.0076	0.1970 ± 0.0063	0.0529 ± 0.0024	0.0327 ± 0.0008	0.0387 ± 0.0034	0.0224 ± 0.0010	0.0340 ± 0.0034
EPhys (5x18) + OPhys VISam (18)	0.1927 ± 0.0095	0.5030 ± 0.0284	0.2748 ± 0.0288	0.2245 ± 0.0543	0.2344 ± 0.0135	0.0469 ± 0.0040	0.0298 ± 0.0036	0.0370 ± 0.0029	0.0190 ± 0.0013	0.0323 ± 0.0011
EPhys (5x18) + OPhys VISl (18)	0.2083 ± 0.0119	0.5324 ± 0.0207	0.3181 ± 0.0148	0.3013 ± 0.0019	0.2305 ± 0.0214	0.0728 ± 0.0040	0.0747 ± 0.0048	0.0684 ± 0.0035	0.0210 ± 0.0016	0.0537 ± 0.0034
OPhys (5x18)	–	–	–	–	–	0.0814 ± 0.0024	0.0726 ± 0.0085	0.0719 ± 0.0052	0.0250 ± 0.0030	0.0539 ± 0.0010
OPhys (5x18) + EPhys VISal (18)	0.1687 ± 0.0067	0.4668 ± 0.0118	0.2318 ± 0.0106	0.2466 ± 0.0086	0.2001 ± 0.0050	0.0757 ± 0.0091	0.0923 ± 0.0015	0.0757 ± 0.0064	0.0244 ± 0.0015	0.0577 ± 0.0008
OPhys (5x18) + EPhys VISp (18)	0.1469 ± 0.0039	0.5777 ± 0.0015	0.2041 ± 0.0063	0.2399 ± 0.0143	0.1838 ± 0.0123	0.0771 ± 0.0031	0.0909 ± 0.0029	0.0768 ± 0.0054	0.0242 ± 0.0010	0.0571 ± 0.0017
OPhys (5x18) + EPhys VISpm (18)	0.1427 ± 0.0109	0.4182 ± 0.0292	0.2952 ± 0.0073	0.2106 ± 0.0162	0.1643 ± 0.0116	0.0901 ± 0.0021	0.0801 ± 0.0062	0.0728 ± 0.0048	0.0216 ± 0.0026	0.0531 ± 0.0006
OPhys (5x18) + EPhys VISam (18)	0.1695 ± 0.0048	0.4699 ± 0.0128	0.2220 ± 0.0109	0.2474 ± 0.0017	0.1931 ± 0.0066	0.0823 ± 0.0018	0.0923 ± 0.0019	0.0762 ± 0.0005	0.0262 ± 0.0025	0.0580 ± 0.0023
OPhys (5x18) + EPhys VISl (18)	0.1368 ± 0.0025	0.4450 ± 0.0169	0.1969 ± 0.0049	0.2251 ± 0.0090	0.2400 ± 0.0057	0.0809 ± 0.0032	0.0833 ± 0.0022	0.0780 ± 0.0015	0.0242 ± 0.0000	0.0560 ± 0.0054
OPhys (5x18) + EPhys (5x18)	0.1382	0.4684	0.1860	0.2572	0.1978	0.0812	0.0703	0.0841	0.0190	0.0571
OPhys VISal (18)	–	–	–	–	–	0.0621 ± 0.0043	0.0497 ± 0.0025	0.0574 ± 0.0010	0.0195 ± 0.0008	0.0430 ± 0.0033
OPhys VISp (18)	–	–	–	–	–	0.0593 ± 0.0016	0.0419 ± 0.0024	0.0410 ± 0.0056	0.0199 ± 0.0048	0.0364 ± 0.0018
OPhys VISpm (18)	–	–	–	–	–	0.0567 ± 0.0032	0.0339 ± 0.0013	0.0422 ± 0.0019	0.0173 ± 0.0010	0.0332 ± 0.0013
OPhys VISam (18)	–	–	–	–	–	0.0509 ± 0.0043	0.0301 ± 0.0020	0.0367 ± 0.0028	0.0176 ± 0.0017	0.0309 ± 0.0025
OPhys VISl (18)	–	–	–	–	–	0.0596 ± 0.0048	0.0443 ± 0.0035	0.0618 ± 0.0055	0.0204 ± 0.0013	0.0476 ± 0.0040
EPhys VISal (18)	0.1564 ± 0.0063	0.4084 ± 0.0087	0.1857 ± 0.0215	0.2223 ± 0.0047	0.1735 ± 0.0075	–	–	–	–	–
EPhys VISp (18)	0.1424 ± 0.0093	0.5559 ± 0.0056	0.2145 ± 0.0090	0.2156 ± 0.0067	0.1617 ± 0.0085	–	–	–	–	–
EPhys VISpm (18)	0.1382 ± 0.0065	0.4182 ± 0.0094	0.2285 ± 0.0172	0.1851 ± 0.0050	0.1646 ± 0.0100	–	–	–	–	–
EPhys VISam (18)	0.1391 ± 0.0038	0.4084 ± 0.0138	0.1910 ± 0.0042	0.2374 ± 0.0106	0.1844 ± 0.0014	–	–	–	–	–
EPhys VISl (18)	0.1343 ± 0.0087	0.3916 ± 0.0172	0.1918 ± 0.0149	0.1832 ± 0.0041	0.1733 ± 0.0133	–	–	–	–	–