CAPT: A MULTI-TASK CONTINUOUS AUTOREGRESSIVE TRANSFORMER ENABLING CROSS-DATASET AND CROSS-SPECIES TRANSFER FOR CALCIUM POPULATION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Calcium imaging has enabled large-scale recordings of neural population activity across diverse animals, laboratories, and experimental paradigms. However, these datasets remain highly fragmented: Existing approaches are often designed for individual datasets or tasks, making it unclear whether their learned representations are reusable for new calcium imaging trace datasets. To tackle this gap, we present **CAPT**, a **C**ontinuous **A**utoregressive **P**opulation **T**ransformer for calcium population dynamics. CAPT models continuous calcium traces directly through a continuous patch tokenization strategy and is trained autoregressively with a mean squared error objective, enabling end-to-end training and adaptation to diverse downstream tasks. We first pretrain CAPT on a large-scale mouse calcium imaging dataset and evaluate its transferability across independent mouse, larval zebrafish, and *C. elegans* datasets collected by different laboratories. In these transfer settings, the pretrained backbone is frozen and only adaptation modules, i.e., neuron and session embeddings or task-specific heads, are updated. Across neural population forecasting and behavior decoding tasks, CAPT consistently outperforms specialized and general-purpose baselines, demonstrating that a single continuous autoregressive backbone can learn representations that transfer beyond the original pretraining distribution. Beyond predictive performance, multimodal analyses using NeuroPAL annotations in *C. elegans* datasets show that CAPT embeddings form a shared functional space across datasets and capture cell-identity-related structure. Together, these results suggest that continuous autoregressive modeling provides an effective framework for learning reusable calcium population representations that transfer across datasets, experimental paradigms, and species. Code is available at `https://anonymous.4open.science/r/project_code_2026-0AC8/`

1 INTRODUCTION

Recent advances in neural recording technologies (Jun et al., 2017; Steinmetz et al., 2021; Ye et al., 2025; Stringer et al., 2019), especially large-scale calcium imaging (Barson et al., 2020; Sofroniew et al., 2016), have made it possible to observe neural population activity across many neurons, animals, brain regions, and model organisms (Zong et al., 2022; Kim et al., 2017; Musall et al., 2019). These recordings are rapidly accumulating across laboratories and species, including mice, larval zebrafish, and *C. elegans*. As the scale and diversity of calcium imaging data continue to grow, an important challenge is how to develop models whose learned representations can be reused across recordings rather than being restricted to individual datasets or experimental settings. Recent neural population models have begun to exploit large-scale or multi-session recordings to learn shared representations for tasks such as neural forecasting and behavior decoding (Duan et al., 2025; Azabou et al., 2025; Willeke et al., 2026). However, the extent to which such learned representations remain useful when transferred to independently collected calcium imaging datasets is still poorly understood.

This challenge is particularly pronounced for calcium imaging because datasets can differ substantially in species, brain regions, imaging conditions, sampling rates, neuron counts, and experimental paradigms (Grienberger et al., 2022; Ahrens et al., 2013; Kim et al., 2017). A model trained on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

one dataset may therefore face pronounced distribution shifts when applied to another due to this heterogeneity. While recent neural models have shown encouraging results within specific datasets, organisms, or task families, it remains less studied whether a single frozen pretrained backbone can be reused across independently collected calcium imaging datasets while supporting multiple downstream tasks, such as neural population forecasting and behavior decoding (Azabou et al., 2025). In particular, cross-species transferability remains largely underexplored in the current literature, leaving open how broadly representations learned from one neural population can remain useful across distinct experimental paradigms and biological systems.

To address this challenge, we propose **CAPT** (**C**ontinuous **A**utoregressive **P**opulation **T**ransformer), a continuous autoregressive model for calcium imaging population dynamics. **CAPT directly models continuous neural activity through a continuous patch tokenization strategy and is trained in an end-to-end, autoregressive manner with a mean squared error (MSE) objective, and is capable of performing diverse tasks.** We first pretrain CAPT on a large-scale mouse calcium imaging dataset. To systematically evaluate the transferability, we test CAPT on eight independent datasets beyond the pretraining data: three mouse datasets, three larval zebrafish datasets, and two *C. elegans* datasets (Table 1, Appendix A). In all transfer experiments, the pretrained backbone is frozen, and only adaptation components are updated: neuron and session embeddings for neural population forecasting and task-specific heads for behavior decoding. Following the aforementioned steps, we report that **CAPT consistently outperforms specialized and general-purpose baselines, including POYO+, POCO, and CalM, across both pretraining and transfer datasets**, and shows significantly better performance in most cases (Table 14), while POYO+ and CalM are less directly applicable or degrade substantially under cross-species transfer. As a long-term goal of neural representation learning is to provide tools that facilitate biological discovery beyond prediction, we further test whether CAPT can support multimodal biological analysis by linking functional neural embeddings with NeuroPAL annotations as cell-identity labels in the *C. elegans* datasets (Yemini et al., 2021). Interestingly, beyond forecasting and decoding performance, we find that **CAPT embeddings recover identity-related structure and support cross-dataset neuron identity classification using only a linear support vector machine (SVM)** (Cortes & Vapnik, 1995)**, suggesting that the model learns a shared functional embedding space across datasets and captures anatomical information.** By comparison, POCO embeddings show much weaker identity-related structure, with performance close to chance in the most challenging cross-dataset setting. These results indicate that CAPT not only improves predictive performance, but also provides embeddings that support biologically interpretable multimodal analysis.

In summary, our contributions are as follows:

Continuous autoregressive modeling for calcium traces. We introduce CAPT, a continuous autoregressive transformer for calcium population dynamics with continuous patch tokenization and neuro-aware QKV modulation, and show that it achieves superior performance on diverse neural tasks and datasets against strong specialized and general-purpose baselines.

Cross-dataset and cross-species transfer. We systematically evaluate the transferability of CAPT on multiple calcium imaging datasets spanning mice, larval zebrafish, and *C. elegans*. With the pretrained backbone frozen, CAPT generalizes successfully across datasets and species.

Biologically meaningful functional embeddings. We show that CAPT embeddings support multimodal analysis with NeuroPAL annotations in *C. elegans*, forming a shared functional embedding space that captures anatomical cell-identity-related structure across datasets.

2 Related Work

2.1 Neural dynamics and neuro-behavior analysis

A long line of classical work has modeled neural population activities to characterize neural dynamics and their relationships to behavior. Latent dynamical systems and state-space models have been used to uncover low-dimensional structure for neural forecasting, trajectory reconstruction, and trial-level interpretation (Macke et al., 2011; Paninski et al., 2010; Churchland et al., 2012). Gaussian-process-based methods extract smooth single-trial neural trajectories from population activity (Yu et al., 2008). In parallel, neural decoding and encoding models relate population activity to behavioral variables, stimuli, choices, or movement kinematics (Wu et al., 2006; Paninski et al., 2007; Pillow et al., 2008),

providing important tools for studying neural representations and behavior. However, these models are often developed and evaluated within individual sessions, animals, or experimental settings.

2.2 Transformer for neural data and neural foundation model

Recent work has begun to leverage Transformer architectures for large-scale representation modeling of invasive neural recordings. For spiking and electrophysiology data, NDT first showed that transformer architectures can model binned neural spiking activity with self-supervised sequence objectives (Ye & Pandarinath, 2021). This direction was extended by NDT2 for multi-context pretraining across spiking datasets (Ye et al., 2023), by POYO for scalable neural population decoding with spike-level tokenization (Azabou et al., 2023), and by Neuroformer for multimodal and multitask generative pretraining on spiking data (Antoniades et al., 2023). More recent models such as MtM and NEDS use multi-task masking to relate neural activity and behavior in large-scale Neuropixels recordings (Zhang et al., 2024; 2025), while NeuroPaint studies neural inpainting by inferring unrecorded brain-area dynamics from multi-animal electrophysiology datasets (Xia et al., 2025). These studies demonstrate the promise of scalable pretraining for neural data, but they are primarily developed for spiking or electrophysiology recordings, rather than continuous calcium traces.

For calcium imaging, POCO studies scalable neural forecasting through population conditioning and trains across calcium imaging datasets spanning mice, zebrafish, and *C. elegans* (Duan et al., 2025). POYO+ extends the POYO framework to multi-session decoding in large-scale mouse two-photon calcium imaging datasets (Azabou et al., 2025). POYO-CAP explores calcium pretraining with cell-pattern-aware neuron selection for visual decoding (Bae et al., 2025). CalM introduces a self-supervised model based on vector quantization and a dual-axis transformer, enabling both forecasting and behavior decoding on calcium population traces (Xu et al., 2026), while OmniMouse studies multi-modal, multi-task calcium modeling at scale in mouse visual cortex data with deconvoluted trace (Willeke et al., 2026). Together, these works demonstrate the growing potential of large-scale pretraining and reusable representation learning for calcium population activity.

2.3 Cross-dataset and cross-species neural modeling

An important question in neural representation learning is whether representations learned from one collection of recordings can remain useful across different datasets, animals, and experimental settings. In electrophysiology, MtM and NEDS demonstrate held-out-session or held-out-animal evaluation within standardized Neuropixels datasets (Zhang et al., 2024; 2025). POSSM further demonstrates cross-species transfer from non-human primates to humans using multi-dataset pretraining and full fine-tuning (Ryoo et al., 2025). These works exploit shared structure across sessions, animals, or brain areas, but they do not directly test frozen-backbone transfer across heterogeneous calcium imaging datasets and species. For calcium imaging, transfer remains less systematically studied. POCO trains across calcium imaging datasets from zebrafish, mice, and *C. elegans*, but mainly focuses on evaluation of forecasting performance within the datasets (Duan et al., 2025). While POYO+ demonstrates transfer across mouse calcium datasets, it is mainly a supervised decoding framework and does not demonstrate cross-species transfer (Azabou et al., 2025). Thus, whether a pretrained calcium backbone can transfer across tasks, datasets, and species remains underexplored.

3 Method

Calcium imaging provides cellular-resolution measurements of neural population dynamics, yet learning reusable representations remains challenging due to the heterogeneity of neural recordings across datasets. We hypothesize that directly modeling continuous calcium dynamics with autoregressive objectives can provide a unified representation of neural population activity. Based on this intuition, CAPT models calcium dynamics in continuous space through a patch-based autoregressive Transformer. Each neuron trace is segmented into temporal patches, projected with neuron and session embeddings, and processed by population-level attention. By predicting future calcium activity directly, CAPT learns representations aligned with neural dynamics while enabling efficient adaptation to new datasets through transfer modules. (Figure 1).

3.1 Continuous patch tokenization

Given a neural population segment, such as a trial, from session s, we denote it as

$$\mathbf{X} \in \mathbb{R}^{N_s \times T}, \tag{1}$$

where N_s is the number of recorded neurons and T is the number of time points. CAPT first divides each single-neuron trace into non-overlapping temporal patches of length L. We highlight the effectiveness of temporal compression, i.e., $L > 1$, through an ablation study in Appendix D. The p-th patch of neuron n is denoted as

$$\mathbf{x}_{n,p} = \left[x_{n,(p-1)L+1}, \ldots, x_{n,pL}\right]^\top \in \mathbb{R}^L. \tag{2}$$

Subsequently, each continuous patch is directly projected into the model space:

$$\mathbf{z}_{n,p} = \mathbf{W}_{\text{in}}\mathbf{x}_{n,p} + \mathbf{e}_n + \mathbf{g}_s. \tag{3}$$

Here, $\mathbf{W}_{\text{in}}$ is the continuous patch projection that maps the patch into the model dimension, $\mathbf{e}_n$ is the neuron embedding, and $\mathbf{g}_s$ is the session embedding. In practice, temporal order is encoded through rotary positional encoding (RoPE) in the temporal attention layers (Su et al., 2024).

The projected continuous patches are processed by the Dual-axis Transformer (DAT) backbone (Xu et al., 2026). The backbone factorizes population modeling into two complementary directions: causal temporal attention models the history of each neuron, while neuron-axis attention models population-level interactions at each patch position. The backbone outputs patch-level neural representations

$$\mathbf{h}_{n,p} \in \mathbb{R}^D. \tag{4}$$

3.2 Network architecture

To further improve the adaptability of population representation learning, we introduce two lightweight architectural components: dynamic residual gating and neuron-aware QKV modulation.

Dynamic residual gating. To adapt to heterogeneous neural dynamics, we replace fixed residual scaling with dynamic residual gating. Specifically, for a Transformer sub-layer $F(\cdot)$, the residual update is formulated as

$$\mathbf{h}' = \mathbf{h} + g(\mathbf{h})F(\text{LN}(\mathbf{h})), \tag{5}$$

where $g(\mathbf{h})$ is a token-dependent scalar gate generated by a lightweight two-layer MLP. Separate gates are applied to temporal attention, neuron-axis attention, and feed-forward branches. This adaptive residual mechanism allows different neural states to dynamically control the strength of representation updates while preserving the original residual structure.

Neuron-aware QKV modulation. Neuron identity can provide useful priors for population interaction modeling. We therefore modulate the query, key, and value projections in neuron-axis attention using neuron embeddings:

$$\mathbf{x}_n = W_x\mathbf{z}_n + U_x(\mathbf{e}_n), \qquad x \in \{q, k, v\}, \tag{6}$$

where U_x denotes a low-rank adapter for the corresponding projection. The modulation enables neuron-aware attention by allowing identity information to influence both interaction weights and aggregated representations. (Appendix D)

3.3 Autoregressive pretraining in continuous space

CAPT is pretrained to predict future continuous calcium activity from past activity. Given the patch-level representation $\mathbf{h}_{n,p}$, the prediction head outputs the future residual patch:

$$\widehat{\mathbf{r}}_{n,p+1} = \mathbf{W}_{\text{out}}\mathbf{h}_{n,p} \in \mathbb{R}^L. \tag{7}$$

To improve autoregressive forecasting stability, CAPT adopts residual-last prediction, where future activity is modeled as the change relative to the last observed value. Specifically, given the latest observed activity $x_{n,t}$, the residual target is defined as

$$r_{n,t+k} = x_{n,t+k} - x_{n,t}, \qquad k = 1, \ldots, L. \tag{8}$$

During training and inference, the predicted residual is converted back to the calcium space by:

$$\widehat{x}_{n,t+k} = x_{n,t} + \widehat{r}_{n,t+k}. \tag{9}$$

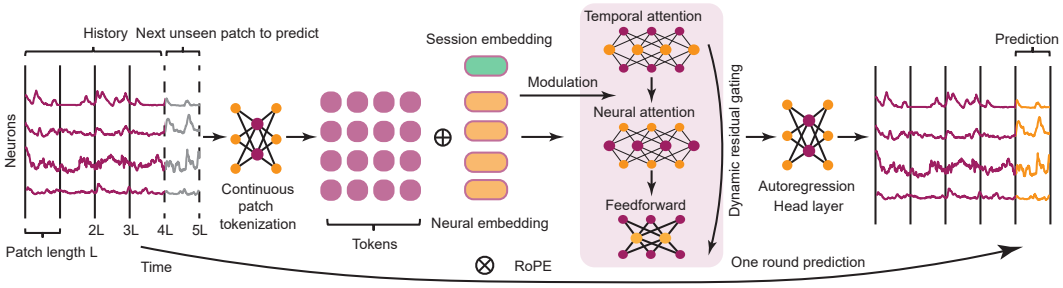


Figure 1: *Pretraining pipeline for CAPT.*

The autoregressive objective is defined as the mean squared error over future residual patches:

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{|\Omega|} \sum_{(n,p,\ell)\in\Omega} \left(\widehat{r}_{n,p+1,\ell} - r_{n,p+1,\ell}\right)^2 . \tag{10}$$

During training, future residual patches are predicted in parallel under teacher forcing, where previous patches are used to predict subsequent patches.

3.4 Behavior Decoding Head

In all decoding experiments, we freeze the pretrained CAPT backbone and train a task-specific decoder on top of the neural representations. Given the frozen backbone output H, the decoder predicts a behavior trace $\widehat{\mathbf{Y}} \in \mathbb{R}^{C_b \times T}$, where C_b is the number of behavior dimensions and $T = PL$.

We use a nonlinear low-rank per-neuron readout to aggregate information across neurons. For each neuron representation, a gated feature is computed as

$$q_{n,p,o,r} = \left(\mathbf{a}^u_{o,r}\right)^\top \mathbf{h}_{n,p} \cdot \mathrm{GELU}\left(\left(\mathbf{a}^v_{o,r}\right)^\top \mathbf{h}_{n,p}\right), \tag{11}$$

where $o = 1, \ldots, C_b L$ indexes the within-patch behavior outputs and $r = 1, \ldots, R_{\mathrm{dec}}$ is the low-rank index. The patch-level output is then obtained by a neuron-wise weighted aggregation:

$$\widehat{y}_{p,o} = \sum_{n=1}^{N_s} \sum_{r=1}^{R_{\mathrm{dec}}} q_{n,p,o,r} B_{o,\mathrm{id}(n),r} + b_o. \tag{12}$$

Here, $\mathrm{id}(n)$ denotes the global neuron index and $B_{o,\mathrm{id}(n),r}$ and b_o are learnable projection weights and bias. The $C_b L$ outputs are reshaped into C_b behavior channels at the original temporal resolution, followed by an optional depthwise temporal convolution for smoothing.

The decoder is trained with MSE:

$$\mathcal{L}_{\mathrm{dec}} = \frac{1}{|\Omega_y|} \sum_{(c,t)\in\Omega_y} \left(\widehat{y}_{c,t} - y_{c,t}\right)^2 , \tag{13}$$

where $|\Omega_y|$ is the number of entries used to compute the behavior loss.

3.5 Transfer Setup

CAPT is first pretrained on the source dataset using the self-supervised autoregressive objective:

$$\theta^\star = \arg\min_\theta \mathbb{E}_{\mathbf{X}\sim\mathcal{D}_{\mathrm{src}}} \left[\mathcal{L}_{\mathrm{AR}}(\mathbf{X};\theta)\right]. \tag{14}$$

For a target dataset, transfer is performed in two separate ways.

Forecasting adaptation by tuning new embeddings. We expand the neuron and session embedding tables with target-specific entries for new sessions and neurons. The patch projection, backbone,

Table 1: Summary of datasets used for pretraining and transfer evaluation.

Dataset	Species	# Subjects	# Sessions	# Neurons	f_s	Ca^{2+} indicator
Tseng et al. (Tseng et al., 2022)	Mouse	8	286	273,770	6 Hz	GCaMP6s
Sun et al. (Sun et al., 2025)	Mouse	13	52	212,464	10 Hz	GCaMP6f
CaLiAli (Vergara et al., 2025)	Mouse	3	3	988	15.6 Hz	GCaMP6f
Park et al. (Park et al., 2024)	Mouse	3	3	9,619	7.8 Hz	GCaMP6s
Lavian et al. (Lavian et al., 2025)	Larval zebrafish	21	45	140,311	2-5 Hz	GCaMP6s
Brysch et al. (Brysch et al., 2019)	Larval zebrafish	3	29	4,026	2 Hz	GCaMP6f
Palieri et al. (Palieri et al., 2024)	Larval zebrafish	5	5	20,480	2 Hz	GCaMP6s
Atanas et al. (Atanas et al., 2023)	*C. elegans*	80	80	10,782	1.67 Hz	GCaMP7f
Copper boundary (Baskoylu & Flavell, 2026)	*C. elegans*	27	27	3,984	1.67 Hz	GCaMP7f

and autoregressive prediction head are frozen. Only the new target neuron and session embeddings are optimized using the same autoregressive forecasting objective:

$$\psi_{\mathrm{tgt}}^{\star} = \arg\min_{\psi_{\mathrm{tgt}}} \mathbb{E}_{\mathbf{X}\sim\mathcal{D}_{\mathrm{tgt}}} \left[\mathcal{L}_{\mathrm{AR}}\left(\mathbf{X}; \theta^{\star}, \psi_{\mathrm{tgt}}\right)\right]. \tag{15}$$

This step evaluates whether the frozen CAPT backbone can support neural population forecasting on a new dataset by learning only dataset-specific embeddings. Under this transfer protocol, increasing the number of target sessions does not provide additional shared trainable capacity, and fine-tuning performance reflects whether the frozen backbone can be reused through per-session adaptation.

Behavior decoding by tuning only the decoder. After the target neuron and session embeddings are adapted, we train only the behavior decoder:

$$\phi_{\mathrm{dec}}^{\star} = \arg\min_{\phi_{\mathrm{dec}}} \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{D}_{\mathrm{tgt}}} \left[\mathcal{L}_{\mathrm{dec}}\left(\mathbf{Y}, g_{\phi_{\mathrm{dec}}}\left(f_{\theta^{\star},\psi_{\mathrm{tgt}}^{\star}}\left(\mathbf{X}\right)\right)\right)\right]. \tag{16}$$

Here, $f_{\theta^{\star},\psi_{\mathrm{tgt}}^{\star}}$ denotes the frozen embedding-adapted CAPT backbone, and $g_{\phi_{\mathrm{dec}}}$ denotes the trainable per-neuron behavior decoder.

This two-step protocol separates neural population adaptation from task adaptation: forecasting transfer only updates new embeddings, while behavior decoding only updates the decoder. Therefore, the experiments directly test whether a single continuous autoregressive backbone can provide reusable representations across datasets, experimental paradigms, and species.

4 Benchmark and Evaluation

4.1 Datasets

For pretraining, we use a large-scale mouse calcium imaging dataset from Tseng et al., used to evaluate held-in and held-out performance within the pretraining distribution following CalM.

To evaluate cross-dataset and cross-species transfer, we further use eight independent multi-animal, multi-session calcium imaging datasets collected from different laboratories and model organisms, including three mouse datasets, three larval zebrafish datasets, and two *C. elegans* datasets. These datasets differ in species, imaging conditions, neuron counts, sampling rates, behavioral variables, and experimental paradigms, providing a broad benchmark for testing whether a single frozen CAPT backbone can be reused beyond the original pretraining distribution (Appendix A).

4.2 Tasks for performance measurement

As mentioned, we evaluate CAPT on two predictive tasks.

Neural population forecasting. Given a context window of calcium population activity, CAPT autoregressively predicts the horizon of future continuous traces (Figure 2A). We report Pearson correlation between predicted and ground-truth traces.

Behavior decoding. Given neural activity from a trial or pseudo-trial, CAPT decodes continuous behavioral variables (Figure 2D). We report R^2 for each behavior variable.

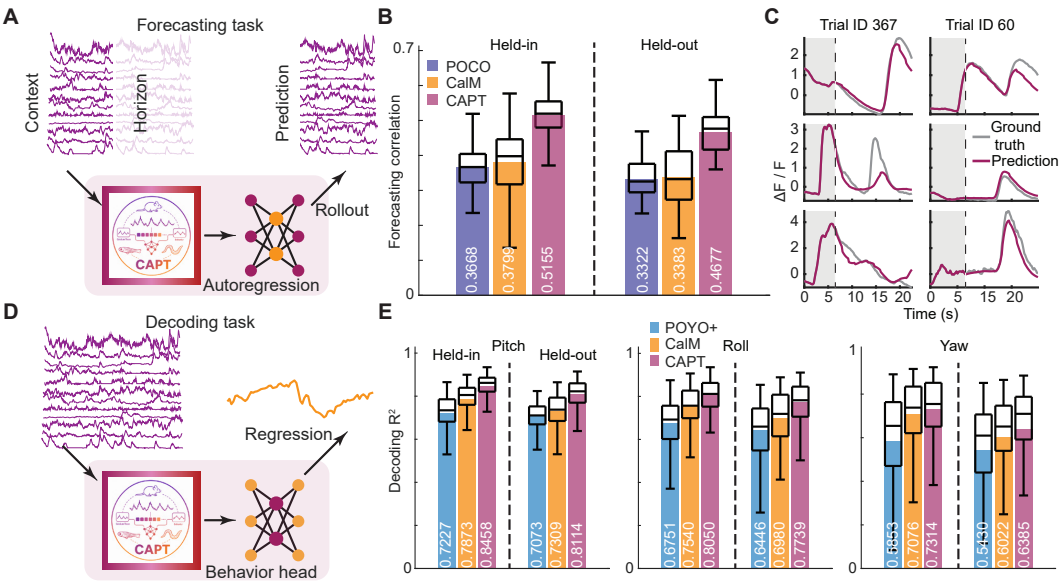


Figure 2: *Performance of CAPT on pretraining dataset.* (A) Illustration of CAPT on forecasting task. (B) Neural population forecasting performances on the pretraining mouse dataset. (C) Example trace visualization. (D) Illustration of CAPT on behavior decoding tasks. (E) Behavior decoding performances on the pretraining mouse dataset.

4.3 Baselines

For task-specialized methods, we use POCO for neural population forecasting and POYO+ for behavior decoding. In addition, we compare with CaIM for a general-purpose calcium modeling. The details of the baseline and CAPT setup can be found in the Appendix B and C.

5 Experiments and Results

For all datasets, we split trials or pseudo-trials into training, validation, and testing sets with a ratio of 70%, 15%, and 15%, respectively whenever applicable. The validation set is used for model selection, and all reported results are computed on the test set. All the transfer experiments, as well as held-out part from pretraining data, follow transfer setup, where the pretrained CAPT backbone is frozen and adaptation modules, i.e., session and neuron embedding and decoding head, are updated.

5.1 Performance on the pretraining dataset

We first evaluate whether CAPT can improve the source-domain performance. CAPT outperforms both POCO and CaIM on held-in and held-out sessions (Figure 2 B, C), indicating that continuous autoregressive modeling provides a strong alternative to discrete-token autoregression or POYO-augmented MLP method for calcium trace forecasting.

For behavior decoding, CAPT achieves stronger decoding performance than the baselines across behavioral variables, demonstrating the effectiveness of learned representations (Figure 2 E).

5.2 Performance on cross-lab mouse datasets

We next evaluate whether CAPT transfers to independent mouse datasets collected outside the pretraining distribution. This setting tests cross-dataset generalization under changes in experimental paradigm, neural population, and behavioral variables.

As shown in Figure 3A, B, CAPT consistently outperforms the applicable forecasting baselines on the cross-lab mouse datasets. Since the backbone is frozen and only new neuron and session embeddings are updated, these results show that CAPT can reuse the pretrained continuous backbone for new mouse calcium recordings and is more stable than CaIM.

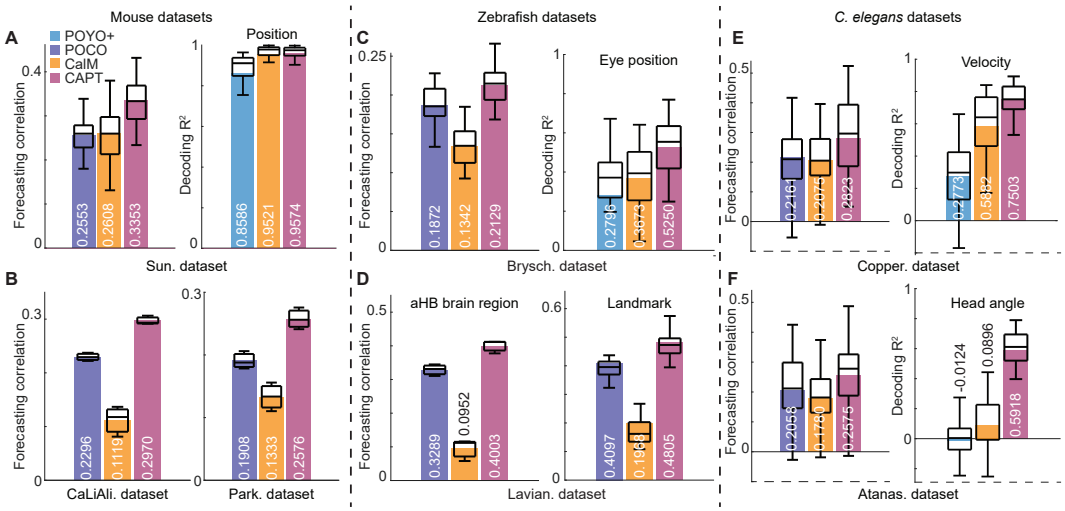


Figure 3: *Performance of CAPT on diverse transfer datasets.* (A) (B) Neural population forecasting and behavior decoding performances on Sun., CaLiAli. and Park. mouse datasets. (C) (D) Neural population forecasting and behavior decoding performance on Brysch. and Lavian. zebrafish datasets. (E) (F) Forecasting and decoding performance on Copper. and Atanas. *C. elegans* datasets.

5.3 Performance on cross-species datasets

We further test CAPT in a more challenging cross-species transfer setting, including larval zebrafish and *C. elegans* datasets. Compared with cross-lab mouse transfer, these datasets introduce larger shifts in organism, recording scale, sampling rate, and experimental paradigm.

On larval zebrafish datasets, CAPT achieves better forecasting performance than the baselines across different recording settings (Figure 3C). In settings where pretrained POCO is unavailable, CAPT remains directly applicable and achieves strong performance (Appendix E).

On *C. elegans* datasets, CAPT also outperforms applicable baselines for neural forecasting and achieves superior performance on available behavioral variables, such as head-angle-related signals and velocity, where POYO+ and CalM show degradation (Figure 3E, F), which is not simply caused by hyperparameter selection as in-domain pretrained POYO+ performs much better (Appendix E).

These results suggest that a mouse-pretrained continuous autoregressive backbone can surprisingly provide useful representations for compact nervous systems with very different anatomical and functional organizations.

5.4 CAPT provides biologically meaningful embeddings

More importantly, we examine whether CAPT embeddings capture biologically meaningful structure in transfer datasets, based on NeuroPAL annotations of *C. elegans* datasets, which provide cell-identity labels for a subset of recorded neurons. We select POCO as the baseline due to the second-best performance of this self-supervised method.

We first visualize the learned neuron embeddings with UMAP (McInnes et al., 2018). As shown in Figure 4A, neurons with different NeuroPAL identities form structured distributions in the CAPT embedding space, suggesting that the model captures identity-related organization rather than dataset-specific variation, while the visualization of POCO embeddings appears to be more randomly distributed with the same UMAP parameter (Appendix E).

To further quantify this structure, we perform cell-identity classification using pooled neuron embeddings with abundant NeuroPAL labels (number > 200) with a linear SVM. CAPT clearly outperforms POCO (Figure 4B, C, Appendix E), showing that CAPT embeddings preserve stronger cell-identity information. We further test a more representative and challenging cross-dataset setting, where the classifier is trained on one *C. elegans* dataset and evaluated on the other. CAPT again achieves substantially stronger performance, indicating that its embedding space is shared across datasets rather

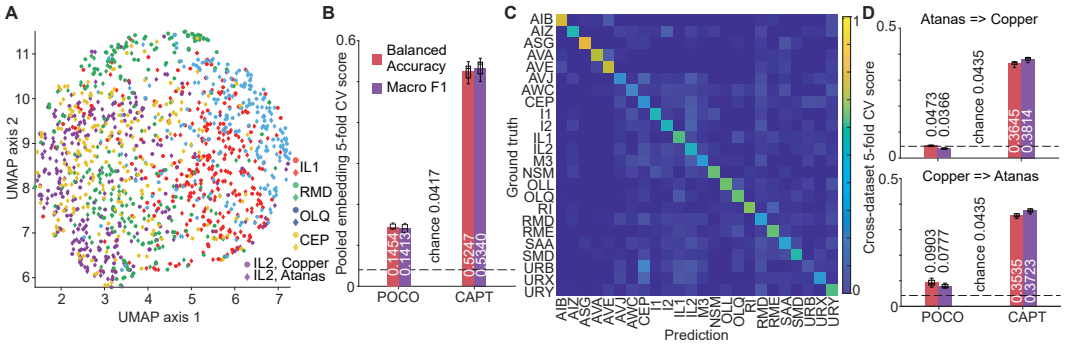


Figure 4: *Neural embedding analysis with NeuroPAL information.* (A) UMAP Visualization of top-5 most abundant NeuroPAL-labelled embeddings. (B) Classification performances on pooled embedding of Copper. and Atanas datasets. (C) Confusion matrix of CAPT. (D) Cross-dataset classification performances.

than being restricted to a single recording, while the performance of POCO appears to be chance level (Figure 4D). The shared space arises from adaptation into the frozen pretrained backbone as retraining CAPT independently on each dataset drops cross-dataset classification to chance (Table 15). We also demonstrate that this link cannot be discovered with solely trace statistics (Appendix E).

Together, these results show that CAPT not only improves forecasting and decoding performance, but also learns a shared functional embedding space that can be linked to biological cell identity within different transfer *C. elegans* datasets with even different data splitting setup (Appendix A).

6 Conclusion

In this work, we introduce CAPT, a continuous autoregressive transformer able to transfer to new datasets by updating only adaptation modules after pretraining on a large-scale mouse calcium imaging dataset. **Across neural population forecasting and behavior decoding tasks, CAPT consistently outperforms specialized and general-purpose baselines on mouse, larval zebrafish, and *C. elegans* datasets with distinct experimental paradigms.** Beyond predictive performance, in *C. elegans* datasets where NeuroPAL annotations provide cell identity information, **CAPT provides a shared functional embedding space across datasets and captures anatomical, identity-related structure.** These results suggest that CAPT provides an effective framework for reusing a single pretrained model across different neural populations, laboratories, experimental paradigms, and species, and may serve as a step toward future neural foundation models for calcium imaging datasets.

Several limitations remain. (1) CAPT is pretrained on a single mouse calcium imaging dataset. Multi-dataset pretraining across paradigms and species should improve the robustness and generality of the learned representations. (2) The performance gain under transfer is generally smaller than that on the pretraining dataset, suggesting that the backbone remains partially constrained by the source distribution. More diverse pretraining data may help close this gap. (3) The tested methods generally perform better on trial-aligned datasets than on recordings without clear trial structure, indicating that pseudo-trial segmentation may be insufficient for spontaneous activity. Stimulus or environment recordings may be needed in such settings. (4) The mechanism underlying cross-species transfer remains only partially understood. Further analyses of attention structure, embedding geometry, and biological modalities are needed to clarify how CAPT generalizes across species.

A broader goal of such transferable models is to enable scientific discovery beyond predictive performance. In this context, NeuroPAL-identified neurons can be linked to connectomic and transcriptomic resources, providing a unique opportunity to relate functional neural activity to circuit and molecular context. Our NeuroPAL analysis suggests that CAPT embeddings are not only useful for forecasting and decoding, but may also provide a functional coordinate system that aligns neural activity with cell identity across datasets. Together, these observations suggest that future CAPT-style models may provide a bridge from neural computation to biological structures across species, which remain fundamental questions in the neuroscience community.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and summarization, code development and debugging, interpretation and organization of experimental results, and language and LaTeX editing of the manuscript. All AI-assisted suggestions, analyses, and code were critically reviewed by the authors; relevant scientific claims were checked against the original literature, and AI-assisted code was verified through inspection and empirical testing. We take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only previously published, publicly available animal calcium imaging datasets; no new animal experiments were performed and no human subjects or human neural data are involved. All datasets are used in accordance with their original licenses and access terms.

This work proposes a continuous autoregressive model for calcium-imaging population activity to learn transferable neural representations across datasets and species, and to improve downstream analyses such as forecasting and behavior decoding. It may reduce task-specific modeling and support more unified and data-efficient neuroscience research. Potential risks include over-interpretation or misuse of learned representations, and privacy concerns if similar approaches are applied to human neural data. This paper makes no clinical claims; any extension to human recordings should require ethical review, informed consent, strong de-identification and access control, and careful reporting of limitations and uncertainty.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available and are described in Appendix A, including per-dataset preprocessing, causal EMA smoothing, normalization fitted on the training split only, and the exact trial / pseudo-trial construction and 70/15/15 split procedure for each of the nine datasets. Architecture and hyperparameters for CAPT are given in Appendix C (Tables 7 and 8), and baseline configurations and fine-tuning protocols for POCO, POYO+, and CaIM in Appendix B (Tables 2–6). The transfer protocol is specified in Section 3.5. Evaluation metrics and the hierarchical neuron/trial/session aggregation used for all reported numbers are defined in Appendix A; statistical comparisons and exact p-values are in Table 14. Compute is reported in Appendix F. Anonymized code for preprocessing, pretraining, transfer, and the NeuroPAL analyses is available at `https://anonymous.4open.science/r/project_code_2026-0AC8/`.

REFERENCES

Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5): 413–420, 2013.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. *arXiv preprint arXiv:2311.00136*, 2023.

Adam A Atanas, Jungsoo Kim, Ziyu Wang, Eric Bueno, McCoy Becker, Di Kang, Jungyeon Park, Talya S Kramer, Flossie K Wan, Saba Baskoylu, et al. Brain-wide representations of behavior spanning multiple timescales and states in *C. elegans*. *Cell*, 186(19):4134–4151, 2023.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sangyoon Bae, Mehdi Azabou, Jiook Cha, and Blake Richards. Decoding dynamic visual experience from calcium imaging via cell-pattern-aware ssl. *arXiv preprint arXiv:2510.18516*, 2025.

Daniel Barson, Ali S Hamodi, Xilin Shen, Gyorgy Lur, R Todd Constable, Jessica A Cardin, Michael C Crair, and Michael J Higley. Simultaneous mesoscopic and two-photon imaging of neuronal activity in cortical circuits. *Nature methods*, 17(1):107–113, 2020.

Saba Baskoylu and Steven Flavell. Whole-brain calcium imaging in freely-moving *C. elegans* during aversive copper boundary encounters, 2026. URL `https://datadryad.org/dataset/doi:10.5061/dryad.w9ghx3g4v`.

Christian Brysch, Claire Leyden, and Aristides B. Arrenberg. Functional architecture underlying binocular coordination of eye position and velocity in the larval zebrafish hindbrain. *BMC Biology*, 17(1):110, 2019.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Yu Duan, Hamza Tahir Chaudhry, Misha B Ahrens, Christopher D Harvey, Matthew G Perich, Karl Deisseroth, and Kanaka Rajan. Poco: Scalable neural forecasting through population conditioning. *ArXiv*, pp. arXiv–2506, 2025.

Christine Grienberger, Andrea Giovannucci, William Zeiger, and Carlos Portera-Cailliau. Two-photon calcium imaging of neuronal activity. *Nature Reviews Methods Primers*, 2(1):67, 2022.

James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

Dal Hyung Kim, Jungsoo Kim, João C Marques, Abhinav Grama, David G C Hildebrand, Wenchao Gu, Jennifer M Li, and Drew N Robson. Pan-neuronal calcium imaging with cellular resolution in freely swimming zebrafish. *Nature Methods*, 14(11):1107–1114, 2017.

Hagar Lavian, Ot Prat, Luigi Petrucco, Vilim Štih, and Ruben Portugues. Visual motion and landmark position align with heading direction in the zebrafish interpeduncular nucleus. *Nature Communications*, 16(1):9924, 2025.

Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural information processing systems*, 24, 2011.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Simon Musall, Matthew T. Kaufman, Ashley L. Juavinett, Steven Gluf, and Anne K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22 (10):1677–1686, 2019.

Virginia Palieri, Emanuele Paoli, You Kure Wu, Martin Haesemeyer, Ilona C Grunwald Kadow, and Ruben Portugues. The preoptic area and dorsal habenula jointly support homeostatic navigation in larval zebrafish. *Current Biology*, 34(3):489–504, 2024.

Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, 165:493–507, 2007.

Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. A new look at state-space models for neural data. *Journal of computational neuroscience*, 29(1):107–126, 2010.

Seungbin Park, Megan Lipton, and Maria C Dadarlat. Decoding multi-limb movements from two-photon calcium imaging of neuronal activity using deep learning. *Journal of neural engineering*, 21(6):066006, 2024.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Avery Hee-Woon Ryoo, Nanda H Krishna, Ximeng Mao, Mehdi Azabou, Eva L Dyer, Matthew G Perich, and Guillaume Lajoie. Generalizable, real-time neural decoding with hybrid state-space models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *elife*, 5:e14472, 2016.

Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539): eabf4588, 2021.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Weinan Sun, Johan Winnubst, Maanasa Natrajan, Chongxi Lai, Koichiro Kajikawa, Arco Bast, Michalis Michaelos, Rachel Gattoni, Carsen Stringer, Daniel Flickinger, James E. Fitzgerald, and Nelson Spruston. Learning produces an orthogonalized state machine in the hippocampus. *Nature*, 640(8057):165–175, 2025.

Shih-Yi Tseng, Selmaan N. Chettih, Charlotte Arlt, Roberto Barroso-Luque, and Christopher D. Harvey. Shared and specialized coding across posterior cortical areas for dynamic navigation decisions. *Neuron*, 110(15):2484–2502.e16, 2022.

Pablo Vergara, Yuteng Wang, Sakthivel Srinivasan, Zhe Dong, Yu Feng, Iyo Koyanagi, Deependra Kumar, Yoan Chérasse, Toshie Naoi, Yuki Sugaya, et al. A comprehensive suite for extracting neuron signals across multiple sessions in one-photon calcium imaging. *Nature communications*, 16(1):3443, 2025.

Konstantin F Willeke, Polina Turishcheva, Alex Gilbert, Goirik Chakrabarty, Hasan A Bedel, Paul G Fahey, Yongrong Qiu, Marissa A Weis, Michaela Vystrčilová, Taliah Muhammad, et al. Omni-mouse: Scaling properties of multi-modal, multi-task brain models on 150b neural tokens. *arXiv preprint arXiv:2604.18827*, 2026.

Wei Wu, Yun Gao, Elie Bienenstock, John P Donoghue, and Michael J Black. Bayesian population decoding of motor cortical activity using a kalman filter. *Neural computation*, 18(1):80–118, 2006.

Ji Xia, Yizi Zhang, Shuqi Wang, Genevera I Allen, Liam Paninski, Cole Lincoln Hurwitz, and Kenneth D Miller. Inpainting the neural picture: Inferring unrecorded brain area dynamics from multi-animal datasets. *arXiv preprint arXiv:2510.11924*, 2025.

Xinhong Xu, Yimeng Zhang, Qichen Qian, and Yuanlong Zhang. Self-supervised foundation model for calcium-imaging population dynamics. *arXiv preprint arXiv:2604.04958*, 2026.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. *arXiv preprint arXiv:2108.01210*, 2021.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36: 80352–80374, 2023.

Zhiwen Ye, Andrew M Shelton, Jordan R Shaker, Julien Boussard, Jennifer Colonell, Daniel Birman, Sahar Manavi, Susu Chen, Charlie Windolf, Cole Hurwitz, et al. Ultra-high-density neuropixels probes improve detection and identification in neuronal recordings. *Neuron*, 113(23):3966–3982, 2025.

Eviatar Yemini, Albert Lin, Amin Nejatbakhsh, Erdem Varol, Ruoxi Sun, Gonzalo E. Mena, Aravinthan D.T. Samuel, Liam Paninski, Vivek Venkatachalam, and Oliver Hobert. NeuroPAL: A multicolor atlas for whole-brain neuronal identification in *C. elegans*. *Cell*, 184(1):272–288.e11, 2021.

Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems*, 21, 2008.

Yizi Zhang, Yanchen Wang, Donato M Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, Eva Dyer, Liam Paninski, et al. Towards a" universal translator" for neural dynamics at single-cell, single-spike resolution. *Advances in Neural Information Processing Systems*, 37:80495–80521, 2024.

Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Neural encoding and decoding at scale. *ArXiv*, pp. arXiv–2504, 2025.

Weijian Zong, Horst A. Obenhaus, Emilie R. Skytøen, Hanna Eneqvist, Nienke L. de Jong, Ruben Vale, Marina R. Jorge, May-Britt Moser, and Edvard I. Moser. Large-scale two-photon calcium imaging in freely moving mice. *Cell*, 185(7):1240–1256.e30, 2022.

A DATASET

In this work, we use a collection of calcium imaging datasets spanning three species (mouse, larval zebrafish, and *C. elegans*) and covering diverse experimental paradigms. In total, the datasets comprise 676,424 neurons and 530 sessions. All neural traces are smoothed and z-score normalized, while behavioral variables are z-score normalized before training. To prevent leakage across train/validation/test splits and temporal leakage within trials or pseudo-trials, neural smoothing is performed with a causal exponential moving average (EMA) with $\alpha = 0.15$, and the EMA state is not propagated across split boundaries. For normalization, the mean and standard deviation are estimated only from the training split and the fitted parameters are directly applied to the corresponding validation and test splits. All reported metrics are aggregated hierarchically to avoid domination by large sessions. For neural population forecasting, Pearson correlation is computed per neuron within each trial over prediction time points, averaged across neurons to obtain a trial-level score, then averaged across trials within each session and finally across sessions. For behavior decoding, R^2 is computed separately for each behavior channel within each trial, followed by the same trial-level and session-level aggregation.

Mouse

Tseng. dataset: The dataset used for pretrain from (Tseng et al., 2022) comprises calcium recordings of 286 sessions from 8 mice, totaling 273,770 neurons recorded from 6 brain regions and 2 cortical layers during a decision-making task. In the task, mice are placed in a Y-maze, where they receive a visual cue (black or white) in the stem section and subsequently make a left or right turn in the arm section according to the cue. The cue-choice mapping rule is switched every over 100 trials. The original multi-plane imaging data are acquired at 30 Hz. In our analysis, neurons from different planes are treated as being sampled simultaneously, resulting in an effective sampling rate of 6 Hz. The three-dimensional velocity from the spherical treadmill (pitch, roll, and yaw) is used as the behavioral variable for the decoding task. The dataset is partitioned into held-in and held-out subset by mouse subjects. The held-in dataset is used for pretrain (6 subjects, 189 sessions, 197,704 neurons) and held-out for within-dataset finetune (2 subjects, 97 sessions, 76,066 neurons). Within each session, trials are split into training, validation, and test sets with a ratio of 70:15:15. This splitting ratio is consistently applied across all datasets used in this paper.

Sun. dataset: The mouse dataset from (Sun et al., 2025) consists of calcium imaging recordings of the hippocampus with a sampling rate of 10 Hz. It is collected from multiple mice across days during a navigation task. In this task, mice navigate a linear track and encounter sequences of sensory cues. Depending on the cue sequence, mice are required to infer the trial type and produce licking behavior at specific positions. To facilitate model training, trials shorter than 64 frames are excluded, while those longer than 200 frames are truncated to 200 frames. Neurons from different sessions of the same subject are treated as distinct. The final dataset comprised 52 sessions from 13 subjects, totaling 212,464 neurons and 7,564 trials. Each trial contains an average of approximately 133 frames. The position of the mouse was used as behavior for decoding task.

CaLiAli dataset: The mouse dataset CaLiAli from (Vergara et al., 2025) consists of one-photon calcium imaging recordings of hippocampal CA1 place cells from freely moving mice on a linear track. The dataset contains 3 sessions from 3 subjects and 988 neurons in total. For each subject, the four daily recordings are first separated according to the frame counts provided in the dataset, ensuring that pseudo-trials did not cross day boundaries. Within each day, the continuous recording is divided into non-overlapping temporal blocks of 1600 neural frames. In each block, continuous temporal splits which contain the first 70% of frames are assigned to training, the following 15% to validation, and the final 15% to testing. Within each split, pseudo-trials are generated using sliding windows of 80 frames with a stride of 10 frames, ensuring that no pseudo-trial crossed split or day boundaries. This procedure yields 11176 pseudo-trials in total, with each pseudo-trial containing 80 frames.

Park. dataset: The mouse dataset from (Park et al., 2024) consists of two-photon calcium imaging recordings collected from head-fixed mice freely running on an air-lifted spherical treadmill. Neural activity was recorded from somatosensory cortex at a sampling rate of 7.8 Hz. This dataset comprises 3 sessions from 3 subjects, totaling 9,619 neurons. Pseudo-trials are constructed from its continuous recordings. Each session is divided into non-overlapping temporal blocks of 800 neural frames, and each block is randomly assigned to training, validation, and test splits. Within each split, pseudo-trials

are generated using sliding windows of 80 frames with a stride of 20 frames. This procedure yields 1,221 pseudo-trials in total, with each pseudo-trial containing 80 frames.

Larval zebrafish

Lavian. dataset: The larval zebrafish dataset (Lavian et al., 2025) contains three parts: whole brain, aHB and landmark.

The whole-brain subset consists of light-sheet calcium imaging recordings from head-fixed larval zebrafish during a whole-field visual-motion stimulation paradigm. Visual stimuli consisted of translational motion patterns presented along multiple directions, and neural activity is recorded across the whole brain at 2 Hz. This subset contains 15 sessions from 12 subjects, comprising 493,695 raw neurons in total. To facilitate model training, only representative neurons are chosen by K-means clustering to reduce each session to 4,096 neurons, resulting in 61,440 neurons in total. Stimulus-aligned trials with a 16-frame context followed by a 16-frame horizon are reconstructed. This procedure produced 1,200 stimulus-aligned trials.

The aHB subset is collected under the same visual-motion stimulation paradigm, with imaging restricted to the anterior hindbrain at 5 Hz. This subset contains 3 sessions from 3 subjects, comprising 4,839 neurons in total. Stimulus-aligned trials with a 40-frame context followed by a 24-frame horizon are reconstructed. This procedure produced 204 stimulus-aligned trials.

The landmark subset consists of two-photon calcium imaging recordings at 3 Hz from head-fixed larval zebrafish during a landmark-position stimulation task. In this task, a visual landmark was presented at different azimuthal positions in the frontal visual field. This subset contains 27 sessions from 6 subjects, comprising 74,032 neurons in total. Landmark-aligned trials with a 40-frame context followed by a 24-frame horizon are constructed according to landmark onset. This procedure produced 1,920 landmark-aligned trials.

Brysch. dataset: The larval zebrafish dataset from (Brysch et al., 2019) consists of calcium imaging recordings collected during an optokinetic stimulation task, in which visual stimuli are used to evoke controlled eye movements. The visual stimulus is discretized into position bins, and within each bin, velocity is systematically varied. We retain recording sessions with more than 100 neurons. Each session is divided into non-overlapping temporal blocks of 1000 neural frames, and each block is randomly assigned to training, validation, and test splits. Within each split, pseudo-trials are generated using sliding windows of 64 frames with a stride of 16 frames, resulting in 29 sessions from 3 subjects, with 4,026 neurons and 4,635 pseudo-trials in total. The binocularly averaged eye position is used as the behavioral variable.

Palieri. dataset: The larval zebrafish dataset from (Palieri et al., 2024) consists of light-sheet whole-brain calcium imaging recordings from head-restrained larval zebrafish during a homeostatic stimulation paradigm. We use the short-protocol temperature task in which repeated temperature stimulation blocks are delivered. This subset contains 5 sessions from 5 subjects, comprising 313,319 raw neurons in total. Neurons are downsampled using the same clustering-based strategy, resulting in 20,480 in total and 4,096 neurons per session. Each continuous session was divided into three non-overlapping blocks of 1600 frames. Within each block, temporal continuous splits are assigned to training, validation and testing. Within each split, pseudo-trials are generated using sliding windows of 80 frames with a stride of 20 frames. This procedure resulted in 1,065 pseudo-trials in total.

C. elegans

Atanas. dataset: The *C. elegans* dataset from (Atanas et al., 2023) consists of calcium imaging recordings from freely moving *C. elegans* during spontaneous behavior. Neural activity is recorded at an effective sampling rate of approximately 1.67 Hz. The dataset contains 80 sessions, totaling 10,782 neurons. Pseudo-trials are constructed by first dividing each session into non-overlapping blocks of 800 frames. Within each block, continuous temporal splits are assigned to training, validation, and test splits. Within each split, pseudo-trials were generated using sliding windows of 80 frames with a stride of 10 frames. This procedure yielded 9,145 pseudo-trials in total. The head angle is used as the behavioral variable for decoding task.

Copper boundary dataset: The *C. elegans* Copper dataset from (Baskoylu & Flavell, 2026) consists of calcium imaging recordings from freely moving *C. elegans* during an aversive copper-boundary encounter paradigm. *C. elegans* expressing the NeuroPAL transgene are recorded while exploring

circular arenas surrounded by copper-containing agar. Each recording contains 1,600 confocal calcium-imaging frames, corresponding to a sampling rate of 1.67 Hz. The dataset contains 27 sessions from 27 animals, totaling 3,984 neurons. For each session, the continuous recording is divided into 20 non-overlapping pseudo-trials of 80 frames, resulting in 540 pseudo-trials in total. This setup is used to evaluate the transfer ability of CAPT under limited-trial settings. The velocity is used as the behavioral variable for decoding task.

Both *C. elegans* datasets contain NeuroPAL-based cell-identity annotations for a subset of recorded neurons (Yemini et al., 2021). In our processed data, NeuroPAL identities are available for 2,159 of 10,782 neurons in the Atanas dataset and for all 3,984 neurons in the Copper dataset, resulting in 6,143 NeuroPAL-labeled neurons across the two datasets. These annotations are used only for downstream multimodal biological analyses.

B Baseline and training details

POCO

For POCO (Duan et al., 2025) on the multi-session forecasting task, the model requires trials to have a fixed context and forecasting horizon. Accordingly, all trials are truncated to 64 time steps (40 for context and 24 for forecasting), and trials shorter than this length are discarded. This fixed-length configuration is kept consistent between pretraining and fine-tuning. We adopt the sliding window hyperparameter in POCO with a stride of 64 to avoid splitting or concatenating trials during training and evaluation. Hyperparameters are selected via a broad search based on the validation Pearson correlation coefficient. For each experiment, models are trained for 200 epochs, and the checkpoint with the best validation correlation is selected. Detailed hyperparameter settings for pretraining are provided in Table 2. During fine-tuning, we freeze the POYO backbone and the conditioning MLP, and only update the neural and session embeddings.

Table 2: Hyperparameters used for training POCO model

Hyperparameter	Value
MLP hidden size	1024
Decoder layers	1
Decoder hidden size	128
Number of heads	16
Latent tokens	8
FFN hidden size	1024
Compression Factor	8

POYO+

For POYO+ (Azabou et al., 2025) on the multi-session decoding task, we conduct a broad hyperparameter search on a small subset of 9 sessions from the pretraining dataset. The best configuration selected based on validation R^2 is then applied to the full pretrain dataset. Only the learning rate is adjusted to ensure stable and effective training. Detailed hyperparameters are provided in Table 3. For cross-dataset fine-tuning, we only adjust the sequence length and latent step based on real time to match sampling rates across different datasets. The pretrain backbone is kept frozen across all fine-tuning experiments. For within-dataset fine-tuning, the decoder head is frozen, and only the neural and session embeddings are updated. For cross-dataset fine-tuning, a new decoder head is initialized to match the corresponding behavioral variables, and both the decoder head and embeddings are jointly updated.

CaIM

For CaIM (Xu et al., 2026) on multi-session forecasting and decoding tasks, we first perform hyperparameter selection on a 12-session dataset, choosing the configuration that achieves the best validation correlation for the forecasting task. The NQ module is pretrained on the pretraining dataset and subsequently used to tokenize other fine-tuning datasets. The DAT module is pretrained on the same pretraining dataset. For forecasting fine-tuning, we freeze the pretrained DAT backbone and update only neuron and session embeddings. For decoding tasks that introduce new behavioral

Table 3: Hyperparameters used for training POYO+ model

Hyperparameter	Value
Embedding Dimension	128
Head Dimension	64
Number of Latents	32
Sequence length	2.0
Latent Step	0.125
Depth	24
Self-attention Heads	8
Cross-attention Heads	4
FFN Dropout	0.2
Linear Dropout	0.4
Attention Dropout	0.2

variables, we initialize and train only a new decoding head while keeping the backbone fixed. For within-dataset held-out decoding fine-tuning, we update only the new neuron-specific rows in the decoding head corresponding to the held-out neurons. We train for 200 epochs during pretraining and within-dataset fine-tuning of DAT, and for all decoding tasks. For cross-dataset fine-tuning of DAT, we extend training to 400 epochs to ensure convergence. Detailed hyperparameters are provided in Table 4, 5 and 6.

Table 4: Hyperparameters used for CalM (NQ) model

Hyperparameter	Value
Embedding Dimension	512
Codebook Size	128
Encoder / Decoder Layers	4
Attention Heads	4
Discretization Window / Overlap	4
EMA Decay	0.99
Max AR Horizon	4

Table 5: Hyperparameters used for CalM (DAT) model

Hyperparameter	Value
Model Dimension	512
Layers	6
Attention Heads	8
FFN Dimension	2048
Scheduled sampling probability	0.6
Neighborhood replacement probability	0.1

Table 6: Hyperparameters used for CalM (downstream decoder head) model

Hyperparameter	Value
Rank	16
Head dropout	0.6
Temporal convolution kernel	9

C MODEL AND HYPERPARAMETERS

For CAPT, we pretrain a continuous patch-based autoregressive backbone on the multi-session forecasting task. Continuous calcium traces are divided into non-overlapping temporal patches, and the model predicts the next continuous patch directly in the original trace space. We use mean squared error as the forecasting objective. During pretraining, we use teacher-forced next-patch prediction. During evaluation, CAPT is evaluated by true autoregressive rollout from a fixed context window.

Following the same protocol as CalM, we perform hyperparameter selection on a 12-session subset and choose the configuration with the best validation Pearson correlation. The selected hyperparameters are reported in Table 7 and 8. For forecasting and decoding fine-tuning, we adopt the same fine-tuning strategy as described above for CalM. We train for 200 epochs during pretraining and all fine-tuning tasks. Checkpoints are selected based on validation Pearson correlation for neural forecasting and validation R^2 for behavior decoding.

Table 7: Hyperparameters used for CAPT backbone

Hyperparameter	Value
Patch length	8
Model dimension	512
Layers	6
Attention heads	8
FFN dimension	2048
Dropout	0.10
Embedding dropout	0.10
Attention dropout	0.10
Sampled neurons per batch	512

Table 8: Hyperparameters used for CAPT downstream decoder head

Hyperparameter	Value
Decoder rank	16
Head dropout	0.6
Temporal smoothing kernel size	9

D ABLATION STUDIES

We study the effect of patch length via ablation experiments on a 12-session pretraining subset. As shown in the Table 9, increasing the patch length from 2 to 8 consistently improves forecasting performance. While a further increase to 10 yields comparable results, it does not lead to additional gains. Overall, these results suggest that a moderately large patch length is beneficial for model performance, and we adopt 8 as the default setting in all main experiments.

Table 9: Ablation results for patch length.

Patch Length	AR Corr
2	0.2201 ± 0.1073
4	0.3689 ± 0.0899
8	0.4421 ± 0.0680
10	0.4406 ± 0.0651

In addition, we conduct ablation experiments on a 3-session subset to demonstrate the advantages of applying the residual learning, dynamic gating and QKV modulation for predictive performance and NeuroPAL classification under transferring setup in Table 10 and 11.

Table 10: Ablation study of CAPT components on the forecasting task. Results are reported as session-level mean ± standard deviation across three test sessions.

Component setting	AR Corr
Baseline	0.4916 ± 0.0553
+ Dynamic gating	0.5109 ± 0.0509
+ QKV modulation	0.4980 ± 0.0491
+ Residual-last	0.5026 ± 0.0595
+ All components	$\mathbf{0.5166 \pm 0.0618}$

Table 11: Ablation study of QKV modulation on NeuroPAL cell identity classification after embedding-only transfer. Results are balanced accuracy (%), reported as mean ± standard deviation across five folds.

Evaluation task	Baseline	+ QKV modulation
Atanas internal	56.81 ± 4.19	$\mathbf{61.16 \pm 2.47}$
Copper internal	65.32 ± 4.27	$\mathbf{67.06 \pm 6.44}$
Pooled	38.63 ± 2.24	$\mathbf{43.96 \pm 1.69}$
Atanas → Copper	27.03 ± 0.99	$\mathbf{28.25 \pm 1.46}$
Copper → Atanas	28.23 ± 0.97	$\mathbf{32.56 \pm 1.22}$

E ADDITIONAL RESULTS

Extra datasets

As additional evidence supporting the main forecasting results in Fig. 3, Fig. 5 reports performance on two further zebrafish datasets: the Palieri dataset and the whole-brain recordings from Lavian et al. Across both datasets, CAPT continues to outperform the compared baselines. In particular, the whole-brain dataset contains pseudo-trials with a 16-frame context and a 16-frame prediction horizon. Pretrained POCO is not directly applicable for fine-tuning because its downstream forecasting setup requires the context and prediction horizon to remain consistent with the configuration used by the pretrained backbone. By contrast, the autoregressive designs of CalM and CAPT are naturally compatible with different context and rollout horizons, allowing them to adapt to this shorter pseudo-trial setting.

Further investigation on POYO+ and POCO

To further examine whether the reason fine-tuned POYO+ degrades on the *C. elegans* datasets in Fig. 3 is due to suboptimal hyperparameter choices, we train POYO+ directly on each target *C. elegans* dataset using the same model configuration and hyperparameters as in the main experiments. As shown in Table 12, in-domain POYO+ training substantially improves decoding performance compared with cross-species fine-tuning, indicating that the degradation observed in Fig. 3 is unlikely to be explained simply by inappropriate hyperparameters or insufficient model capacity.

Table 12: Pretraining POYO+ results for *C. elegans* decoding.

Dataset (behavior)	POYO+ fine-tuning	POYO+ pretraining	CAPT fine-tuning
Atanas (head angle)	-0.0124 ± 0.1349	0.4968 ± 0.1715	0.5918 ± 0.1522
Copper (velocity)	0.2773 ± 0.2149	0.6814 ± 0.1131	0.7503 ± 0.0948

As additional evidence supporting the main embedding analysis results in Fig 4, Fig 6 shows that neural embeddings learned by POCO exhibit less structured organization in the UMAP visualization and weaker cell-type separability in the corresponding confusion matrix. Compared with CAPT embeddings, POCO embeddings show less coherent clustering by NeuroPAL-defined cell-type, and the classifier produces a less accurate and more diffuse confusion pattern.

Statistical test

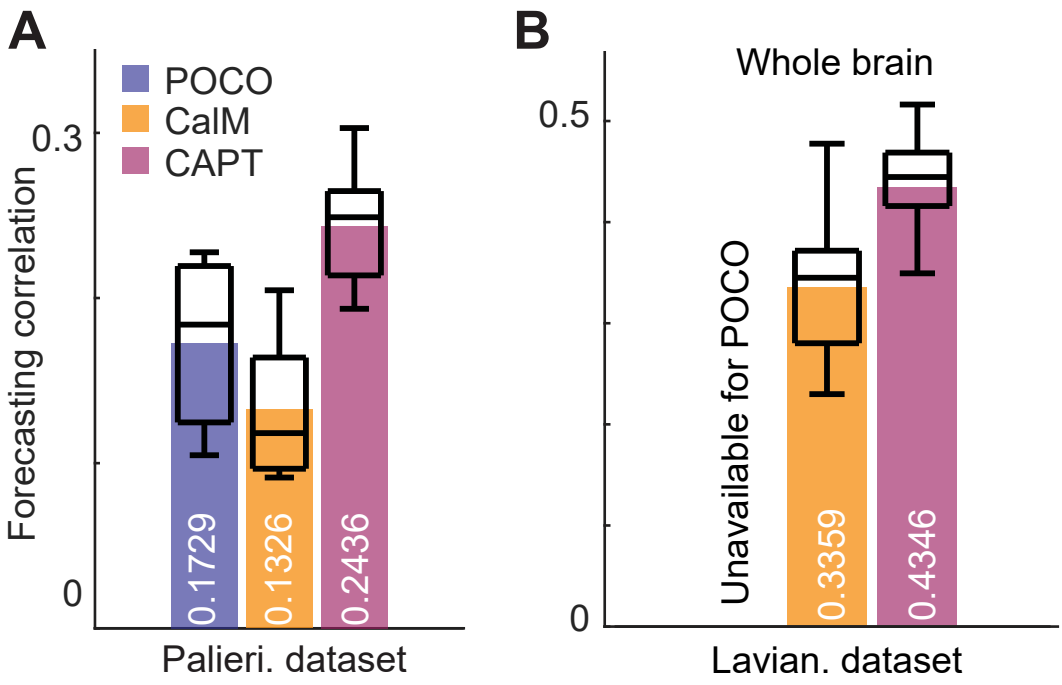


Figure 5: *Performance of CAPT on additional transfer datasets.* (A) Neural population forecasting on Palieri. zebrafish datasets. (B) Neural population forecasting on the whole brain subset from Lavian. zebrafish datasets.

Table 13: NeuroPAL classification using handcrafted calcium trace statistics.

Setup	Balanced Accuracy
Pooled	0.1663 ± 0.0107
Atanas $\rightarrow$ Copper	0.0479 ± 0.0060
Copper $\rightarrow$ Atanas	0.0722 ± 0.0157

We report the mean ± standard deviation and exact p-values in the experiments in Table 14. All methods are evaluated on identical session-level splits and compared using two-sided paired Wilcoxon signed-rank tests. CAPT significantly outperforms POCO/POYO+ in 17 of 20 comparisons and CaIM in 16 of 20; the nonsignificant cases involve three-session datasets, where the paired Wilcoxon test cannot reach $p < 0.25$, or closely matched performance.

Trace-statistic control

To examine whether the NeuroPAL classification performance can be explained by simple calcium trace properties, we construct handcrafted trace-statistic features and evaluate them using the same linear SVM classifier and validation protocol as the embedding analysis. For each neuron, we extract a 45-dimensional feature vector from the full calcium trace, including signal statistics, first-order variability, temporal autocorrelation, autocorrelation decay, variance decomposition, Hjorth parameters, noise and SNR estimates, and frequency-domain characteristics. These features are computed without using cell-identity labels.

Table 13 reports the balanced accuracy. Trace statistics achieve above-chance performance only in the pooled setting, while cross-dataset transfer remains close to chance, suggesting that simple signal properties alone are insufficient to explain the transferable cell-identity information captured by CAPT embeddings.

Role of adaptation into the frozen backbone

Table 14: Comparison of CAPT with POCO / POYO+ and CaIM across forecasting and decoding tasks. Values are reported as mean ± standard deviation.

Result	POCO / POYO+	CaIM	CAPT	p (vs. POCO / POYO+)	p (vs. CaIM)
Tseng held-in forecasting corr ($n = 189$)	0.3668 ± 0.0605	0.3799 ± 0.0978	$\mathbf{0.5155 \pm 0.0655}$	9.136×10^{-33} (***)	9.136×10^{-33} (***)
Tseng held-out forecasting corr ($n = 97$)	0.3322 ± 0.0511	0.3383 ± 0.0860	$\mathbf{0.4677 \pm 0.0556}$	1.218×10^{-17} (***)	1.218×10^{-17} (***)
Sun forecasting corr ($n = 52$)	0.2553 ± 0.0378	0.2608 ± 0.0660	$\mathbf{0.3353 \pm 0.0578}$	3.504×10^{-10} (***)	3.504×10^{-10} (***)
Brysch forecasting corr ($n = 29$)	0.1872 ± 0.0239	0.1342 ± 0.0248	$\mathbf{0.2129 \pm 0.0249}$	4.098×10^{-7} (***)	3.725×10^{-9} (***)
Copper forecasting corr ($n = 27$)	0.2161 ± 0.1070	0.2075 ± 0.1142	$\mathbf{0.2823 \pm 0.1298}$	1.843×10^{-4} (***)	7.290×10^{-5} (***)
CaLiAli forecasting corr ($n = 3$)	0.2296 ± 0.0076	0.1119 ± 0.0282	$\mathbf{0.2970 \pm 0.0082}$	0.2500 (n.s.)	0.2500 (n.s.)
Park forecasting corr ($n = 3$)	0.1908 ± 0.0141	0.1333 ± 0.0228	$\mathbf{0.2576 \pm 0.0173}$	0.2500 (n.s.)	0.2500 (n.s.)
Landmark forecasting corr ($n = 27$)	0.4097 ± 0.0694	0.1968 ± 0.0979	$\mathbf{0.4805 \pm 0.0576}$	1.490×10^{-8} (***)	1.490×10^{-8} (***)
aHB forecasting corr ($n = 3$)	0.3289 ± 0.0172	0.0952 ± 0.0324	$\mathbf{0.4003 \pm 0.0197}$	0.2500 (n.s.)	0.2500 (n.s.)
Atanas forecasting corr ($n = 80$)	0.2058 ± 0.1075	0.1780 ± 0.0846	$\mathbf{0.2575 \pm 0.1062}$	3.992×10^{-10} (***)	3.619×10^{-11} (***)
Tseng held-in decoding pitch R^2 ($n = 189$)	0.7227 ± 0.0840	0.7873 ± 0.0737	$\mathbf{0.8458 \pm 0.0623}$	9.135×10^{-33} (***)	9.136×10^{-33} (***)
Tseng held-in decoding roll R^2 ($n = 189$)	0.6751 ± 0.1132	0.7540 ± 0.0883	$\mathbf{0.8050 \pm 0.0759}$	9.135×10^{-33} (***)	1.054×10^{-32} (***)
Tseng held-in decoding yaw R^2 ($n = 189$)	0.5853 ± 0.2321	0.7076 ± 0.1691	$\mathbf{0.7314 \pm 0.1649}$	2.076×10^{-30} (***)	4.948×10^{-17} (***)
Tseng held-out decoding pitch R^2 ($n = 97$)	0.7073 ± 0.0611	0.7309 ± 0.0785	$\mathbf{0.8114 \pm 0.0656}$	1.296×10^{-17} (***)	1.218×10^{-17} (***)
Tseng held-out decoding roll R^2 ($n = 97$)	0.6446 ± 0.1299	0.6980 ± 0.1200	$\mathbf{0.7739 \pm 0.0897}$	1.218×10^{-17} (***)	1.218×10^{-17} (***)
Tseng held-out decoding yaw R^2 ($n = 97$)	0.5430 ± 0.2240	0.6022 ± 0.2172	$\mathbf{0.6385 \pm 0.2348}$	5.331×10^{-11} (***)	4.636×10^{-9} (***)
Sun decoding R^2 ($n = 52$)	0.8586 ± 0.1438	0.9521 ± 0.0658	$\mathbf{0.9574 \pm 0.0466}$	3.504×10^{-10} (***)	0.4335 (n.s.)
Brysch decoding R^2 ($n = 29$)	0.2796 ± 0.3976	0.3673 ± 0.1696	$\mathbf{0.5250 \pm 0.1503}$	1.144×10^{-6} (***)	1.193×10^{-5} (***)
Copper decoding velocity R^2 ($n = 27$)	0.2773 ± 0.2149	0.5882 ± 0.2213	$\mathbf{0.7503 \pm 0.0948}$	1.490×10^{-8} (***)	1.879×10^{-5} (***)
Atanas decoding head angle R^2 ($n = 80$)	-0.0124 ± 0.1349	0.0896 ± 0.2729	$\mathbf{0.5918 \pm 0.1522}$	7.850×10^{-15} (***)	7.850×10^{-15} (***)

Table 15: Effect of adaptation into the frozen pretrained backbone on cross-dataset NeuroPAL cell-identity classification. Embedding-only adaptation updates target neuron and session embeddings against the frozen mouse-pretrained backbone, while independent retraining trains CAPT from scratch on each *C. elegans* dataset separately. Results are mean ± standard deviation across five folds; chance level is 0.0435.

Adaptation	Direction	Balanced accuracy	Macro F1
Embedding-only (frozen backbone)	Atanas → Copper	$\mathbf{0.3645 \pm 0.0091}$	$\mathbf{0.3814 \pm 0.0084}$
Embedding-only (frozen backbone)	Copper → Atanas	$\mathbf{0.3535 \pm 0.0079}$	$\mathbf{0.3723 \pm 0.0084}$
Independent retraining	Atanas → Copper	0.0373 ± 0.0034	0.0199 ± 0.0025
Independent retraining	Copper → Atanas	0.0388 ± 0.0020	0.0152 ± 0.0037

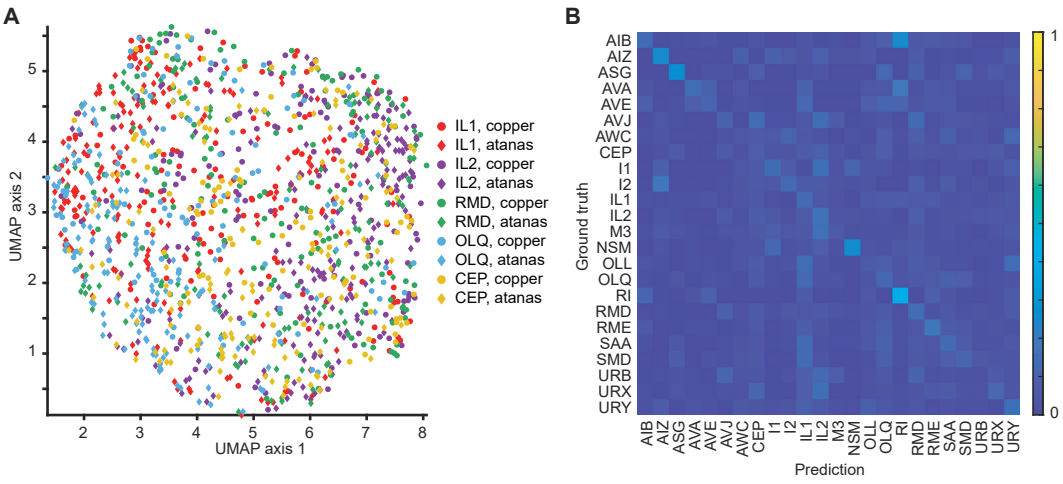


Figure 6: *Embedding analysis for POCO baseline.* (A) 2d UMAP visualization for POCO neural embeddings (B) Confusion matrix of cell-type classification from POCO neural embeddings

The cross-dataset results in Fig. 4D compare embeddings that were adapted into a single frozen pretrained backbone. To test whether this shared coordinate system, rather than the architecture or the per-neuron embedding capacity, is what makes embeddings comparable across datasets, we retrain CAPT from scratch independently on each *C. elegans* dataset and repeat the identical cross-dataset classification protocol. As shown in Table 15, cross-dataset transfer collapses to chance in both directions, even though each model is trained on the same traces with the same architecture and the same number of embedding parameters. Independently learned embedding spaces are therefore not

mutually comparable, indicating that adaptation into a shared frozen backbone, and not the embedding parameterization itself, is what renders neuron embeddings commensurable across datasets.

F COMPUTATIONAL RESOURCE

Experiments were conducted using one GPU node equipped with 8× NVIDIA A100 GPUs (40 GB memory each). Models were trained using distributed data parallelism across available GPUs.