FISHBANK: A ZEBRAFISH BENCHMARK FOR CROSS-INDIVIDUAL WHOLE-BRAIN ACTIVITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting whole-brain neural activity provides a quantitative test of brain dynamics models. We investigate the predictability of future activity from recent observations across spatial and temporal scales, individuals, and stimulus conditions. To support this investigation, we introduce FishBank, a multi-individual zebrafish forecasting benchmark comprising calcium imaging recordings of 189 zebrafish spanning three broad sensory categories. Spatial realignment, blockwise aggregation, and activity standardization transform heterogeneous recordings into common block-level activity sequences without requiring one-to-one neuronal correspondence, enabling systematic evaluation across model families, spatiotemporal settings, and training-data compositions. Our evaluation reveals three main findings. First, all learned predictors outperform naive baselines, while five deep neural networks performs closely. Second, forecasting performance depends on the spatial and temporal definition of the target. Third, multi-individual training yields remarkably lower error than the average error of single-recording models, whereas mixed-stimulus settings achieve slight improvement. FishBank provides a unified framework for characterizing the predictability of whole-brain activity, its spatiotemporal dependence, and the benefits and limitations of learning across individuals. We will publicly release FishBank, together with the associated preprocessing and benchmarking code, to facilitate future research.

1 INTRODUCTION

Understanding how neural population activity evolves across space and time is central to building quantitative models of brain function (Churchland et al., 2012). Large-scale recordings reveal the coordinated activity of distributed neuronal populations, while forecasting provides a direct empirical test of the regularities learned from these observations: a model must predict unobserved future activity from its recent history (Lueckmann et al., 2025; Duan et al., 2025). Establishing the spatial detail, temporal range, and experimental conditions over which such predictions remain accurate helps characterize the capabilities and limitations of neural dynamics models. Larval zebrafish provide an important experimental setting for this investigation. Their small, optically accessible vertebrate brains enable brain-wide calcium imaging at cellular resolution, including recordings during controlled sensory stimulation and fictive behavior (Ahrens et al., 2013; Vladimirov et al., 2014). These capabilities make zebrafish well suited to studying the predictive structure of neural activity across distributed populations and individuals.

Recent studies have advanced zebrafish neural forecasting along several directions. ZAPBench establishes cellular-resolution, whole-brain forecasting in a single larval zebrafish, with diverse visual stimuli and evaluations of temporal and stimulus generalization (Lueckmann et al., 2025). POCO combines neuron-specific temporal prediction with population conditioning, supporting multi-session learning and adaptation to new recordings (Duan et al., 2025). The Sparse Brain Model links stimulus-conditioned forecasts of inferred neuronal spike probabilities to behavioral readouts (Vegas et al., 2025). Concurrent work on shared latent representations addresses cross-individual translation of instantaneous spontaneous activity (Dommanget-Kott et al., 2026). Despite these advances, systematic evaluation across heterogeneous recordings, multiple spatial resolutions, and distinct sensory categories remains limited. Constructing a unified evaluation is challenging because recordings differ in neuronal counts and locations, spatial coverage, acquisition rates, and activity scales. A neuronal index in one individual does not identify a corresponding neuron in an-

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.

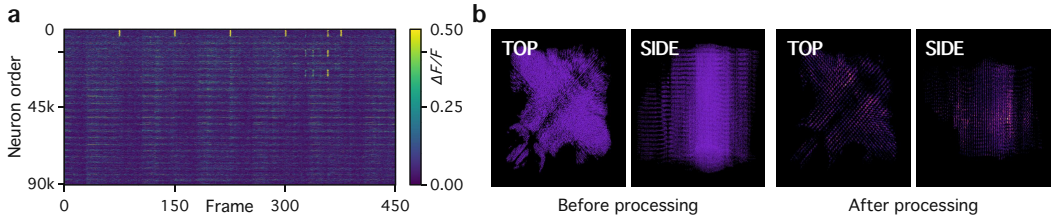

Figure 1: Whole-brain activity and spatial aggregation. (**a**) Calcium activity over 180 s (450 frames), with neurons ordered by calibrated spatial position and color indicating $\Delta F/F$. (**b**) TOP and SIDE views before (left pair) and after (right pair) spatial aggregation. After aggregation, marker size and brightness encode the number of neurons per spatial block. Point clouds illustrate the spatial structure of the resulting fixed-resolution activity sequences (3D space and time).

other, while incomplete coverage complicates spatial alignment (Duan et al., 2025; Dommanget-Kott et al., 2026). Extending forecasting across these sources therefore requires a common prediction target that accommodates different observed populations and makes the spatial scale of the predicted activity explicit.

We introduce a benchmark-construction pipeline that organizes heterogeneous zebrafish recordings into a shared, spatially indexed forecasting task. The pipeline converts neuronal coordinates to physical units using source-specific calibration, realigns recordings using bilateral spatial reference regions, averages neuronal activity within configurable three-dimensional blocks, and standardizes activity with one mean and standard deviation per recording. This construction produces compatible block-level activity sequences without requiring one-to-one neuronal correspondence, while allowing spatial resolution to be varied through the block dimensions. Using this pipeline, we construct ***FishBank***, a benchmark comprising 189 calcium-imaging recordings from nine studies spanning three broad sensory categories: visual, auditory, chemosensory, and vestibular. The collection draws on studies of brain-wide responses under these experimental paradigms (Marquez-Legorreta et al., 2022; Poulsen et al., 2021; Sy et al., 2023; Migault et al., 2018). Given C observed frames, the task is to predict the next P frames of spatially resolved population activity. FishBank supports within-recording and held-out-individual evaluation, together with comparisons across spatial resolutions, forecast horizons, and single- versus mixed-stimulus training compositions.

Using FishBank, we compare six learned predictors spanning linear, fully connected, convolutional, recurrent, and attention-based models, and examine task settings and training composition using a default predictor. Three main findings emerge. ***First***, all six learned predictors outperform historical-mean and last-frame baselines in the evaluated cohort, while the five nonlinear models achieve closely spaced aggregate errors. ***Second***, in the default-model experiments, coarser spatial blocks yield lower absolute prediction errors, whereas improvements over the historical-mean baseline diminish as the forecast horizon increases. The spatial result concerns a more aggregated prediction target and therefore does not imply greater accuracy at finer neuronal scales. ***Third***, training composition reveals both opportunities and challenges: pooling 21 training recordings reduces aggregate mean absolute error on four held-out individuals by 9.5% relative to the average of 21 single-recording models, whereas mixed-stimulus configurations yield only 1.8–2.2% improvements over their corresponding historical-mean baselines. These findings characterize the benefits of learning and the conditions under which the tested predictors retain an advantage over simple temporal references.

Our main contributions can be summarized in threefold:

1. We introduce a construction pipeline that combines physical coordinate calibration, spatial realignment, blockwise aggregation, and recording-level standardization to define a common population-activity target across heterogeneous recordings.
2. We assemble a 189-recording benchmark covering three sensory categories, with evaluation settings for within-recording forecasting, held-out-individual prediction, and single- and mixed-stimulus training, with explicit spatial resolutions and forecast horizons.
3. We establish experimental results across six learned predictors and characterize the dependence of forecasting performance on spatial aggregation, forecast horizon, and train-

Table 1: Source composition of FishBank: the number of recordings, nominal recording length, acquisition frequency, spatial resolution, and stimulus type for each source study.

Article	Samples	Frames	Frequency (Hz)	Resolution (μm)	Stimulus type
Marquez-Legorreta et al. (2022)	40	3,600	2	$1 \times 1 \times 5$	Visual
Lamiré et al. (2023)	34	7,200	2	$0.6 \times 1 \times 10$	Visual
Légaré et al. (2025)	34	900	1	$0.9 \times 1.1 \times 12$	Visual
Favre-Bulle et al. (2020)	6	2,400	4	$1 \times 1 \times 10$	Auditory
Sy et al. (2023)	9	3,080	2	$1.2 \times 2.8 \times 7$	Chemosensory
Poulsen et al. (2021)	7	2,400	4	$1 \times 1 \times 10$	Auditory
Migault et al. (2018)	16	3,000	2.5	$0.8 \times 0.8 \times 10$	Visual
Tanaka & Portugues (2025)	18	3,840	2	$0.6 \times 0.6 \times 15$	Visual
Mancienne et al. (2026)	25	1,800	2	$1 \times 1 \times 5$	Visual
Total	**189**	–	–	–	–

ing composition, identifying both the gains from pooled training and the limited baseline-relative improvements in mixed-stimulus settings.

2 FishBank

2.1 Data Sources and Construction Challenges

Data Sources. FishBank integrates 189 calcium-imaging recordings from nine published studies, with the source composition summarized in Table 1. The collection spans three broad sensory categories: visual (Marquez-Legorreta et al., 2022; Tanaka & Portugues, 2025), auditory (Favre-Bulle et al., 2020; Poulsen et al., 2021), chemosensory (Sy et al., 2023), and vestibular (Migault et al., 2018). These categories describe the experimental settings represented by the source studies rather than a uniform set of stimulus protocols. Nominal recording lengths range from 900 to 7,200 frames, and acquisition rates range from 1 to 4 Hz. For each recording, the input consists of three-dimensional neuronal coordinates and their associated calcium-activity traces. We use the processed coordinates and traces supplied by the original authors where available; additional preprocessing follows the corresponding source study.

Construction challenges. Integrating these recordings requires a prediction target that remains comparable despite differences in acquisition and neuronal sampling. *First, spatial coordinates are not directly comparable across sources.* Coordinate units depend on acquisition-specific spatial calibration, while differences in brain morphology and imaging coverage affect spatial placement. In particular, uneven coverage of the forebrain can shift a reference center estimated from the entire neuronal cloud. *Second, the observed neuronal populations differ across recordings.* Neuron counts, locations, and sampling density vary, and neuronal indices do not establish correspondence between animals. A common representation must therefore accommodate different populations while specifying the spatial scale of the activity being predicted. *Third, activity scales differ across recordings.* Standardization must make overall intensity scales comparable without independently removing the activity differences between spatial regions or time points within a recording. Acquisition rates introduce a further temporal distinction: equal numbers of frames need not cover equal physical durations. These challenges motivate a construction pipeline with a common physical coordinate frame, configurable spatial aggregation, and recording-level activity standardization.

2.2 Benchmark Construction Pipeline

Let recording r contain N_r neurons observed over T_r frames, with neuronal coordinates $\mathbf{P}_r \in \mathbb{R}^{N_r \times 3}$ and activity traces $\mathbf{X}_r \in \mathbb{R}^{T_r \times N_r}$. We denote the coordinate of neuron n by $\mathbf{p}_{r,n}$ and its activity at frame t by $x_{r,t,n}$. The pipeline maps these variable-size observations to a shared spatial layout through three steps as follows.

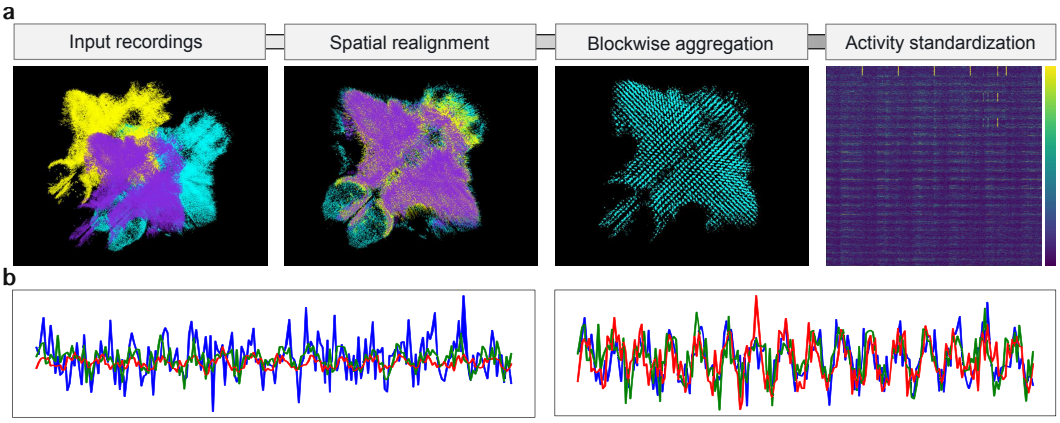

Figure 2: Construction pipeline of FishBank. **(a)** Heterogeneous recordings are organized into a common forecasting representation through spatial realignment, blockwise aggregation, and activity standardization, without requiring one-to-one neuronal correspondence across individuals. **(b)** Example activity traces before (left) and after (right) standardization. Rescaling reduces differences in activity scale while preserving the temporal patterns of each trace.

Physical coordinate calibration and spatial realignment. We first express neuronal coordinates in micrometers using the spatial calibration of each source. Let $\mathbf{D}_r = \mathrm{diag}(s_{r,x}, s_{r,y}, s_{r,z})$ contain the axis-specific conversion factors. Physical coordinates and their realigned positions are

$$\mathbf{p}_{r,n}^{\mathrm{phys}} = \mathbf{D}_r \mathbf{p}_{r,n}, \qquad \widetilde{\mathbf{p}}_{r,n} = \mathbf{p}_{r,n}^{\mathrm{phys}} - \mathbf{c}_r, \tag{1}$$

where $\mathbf{c}_r$ is the reference center estimated from the lateral one-third regions on the left and right sides of the brain. Their relatively consistent coverage and approximate bilateral symmetry motivate this choice over a centroid computed from all observed neurons. Realignment places these reference centers at a common origin while retaining recording-specific morphology. The coordinate conversion makes physical distances comparable; it does not equalize the optical resolution of the original acquisitions. Together, these operations establish the spatial frame in which population-level prediction targets are defined.

Multi-resolution spatial partitioning and aggregation. We construct a common bounding volume from the spatial extents of the realigned recordings and partition it into rectangular blocks with side lengths $\boldsymbol{\delta} = (\delta_x, \delta_y, \delta_z)$. For block b, let $V_b(\boldsymbol{\delta})$ denote its spatial region. Its neuronal membership and framewise population activity are

$$\mathcal{N}_{r,b}(\boldsymbol{\delta}) = \{n : \widetilde{\mathbf{p}}_{r,n} \in V_b(\boldsymbol{\delta})\}, \tag{2}$$

$$B_{t,b}^{(r)} = \frac{1}{|\mathcal{N}_{r,b}(\boldsymbol{\delta})|} \sum_{n \in \mathcal{N}_{r,b}(\boldsymbol{\delta})} x_{r,t,n}, \qquad |\mathcal{N}_{r,b}(\boldsymbol{\delta})| > 0. \tag{3}$$

Each grid location thus represents the mean activity of the recorded neurons within a fixed physical region. Shared block indices provide a consistent spatial layout without requiring identical neuron counts or one-to-one neuronal correspondence. The reference configuration uses $40 \times 40 \times 20\,\mu\mathrm{m}$ blocks on a $38 \times 30 \times 13$ grid. We retain the original point clouds and partitioning code so that alternative block sizes can be constructed from the same recordings. Changing block size changes the spatial detail retained in both the inputs and prediction targets. Coverage remains recording dependent: a block containing no recorded neuron represents missing observation rather than measured neural silence.

Recording-level activity standardization. After spatial aggregation, we standardize activity using one scalar mean μ_r and one scalar standard deviation σ_r per recording:

$$Y_{t,b}^{(r)} = \frac{B_{t,b}^{(r)} - \mu_r}{\sigma_r}. \tag{4}$$

The same affine transformation is applied across spatial blocks and time points within a recording. This places recordings on a standardized overall activity scale while preserving their internal spatial

and temporal differences up to a shared scaling factor. In contrast, independently standardizing each block would remove between-block differences in mean activity and variability. The resulting representation is a sequence of spatially indexed population-activity frames with a common grid across recordings. Spatial resolution remains controlled by $\boldsymbol{\delta}$, while temporal duration must be interpreted using the acquisition rate: at f_r frames per second, a P-frame forecast covers P/f_r seconds. FishBank thereby provides a common data interface for forecasting across individuals, spatial resolutions, and recording conditions.

3 BENCHMARK

FishBank defines a forecasting benchmark for predicting future spatially resolved whole-brain population activity from recent observations. Following the prediction-oriented formulation of prior whole-brain forecasting work (Lueckmann et al., 2025), each model receives a sequence of block-level activity frames and predicts the immediately following frames. The common spatial representation introduced in Section 2 makes the same input–output interface applicable across recordings with different neuronal identities and sampling densities. We consider two complementary forms of generalization: forecasting held-out temporal segments from recordings represented during training, and forecasting complete held-out individuals. We additionally vary the spatial resolution, forecast horizon, and composition of the training recordings. Performance is measured by mean absolute error (MAE) and calibrated against two parameter-free temporal references, *Mean* and *Copy*. Our initial benchmark evaluates six learned predictors spanning linear, fully connected, convolutional, recurrent, and attention-based approaches.

3.1 SETUP

Let $\mathbf{Y}^{(r)} \in \mathbb{R}^{T_r \times D_x \times D_y \times D_z}$ denote the standardized block-level activity sequence for recording r, where T_r is the number of frames and (D_x, D_y, D_z) are the dimensions of the spatial grid. Given C observed frames, the forecasting task is to predict the next P frames:

$$f_\theta\left(\mathbf{Y}^{(r)}_{t:t+C-1}\right) = \widehat{\mathbf{Y}}^{(r)}_{t+C:t+C+P-1}. \tag{5}$$

Thus, the input and target occupy disjoint but consecutive intervals. For a contiguous temporal block of length L, we construct unit-stride forecasting windows

$$\mathcal{W}^{(r)} = \left\{ \left(\mathbf{Y}^{(r)}_{t:t+C-1}, \mathbf{Y}^{(r)}_{t+C:t+C+P-1} \right) : \; t = 1, \ldots, L - C - P + 1 \right\}, \tag{6}$$

with all context and target frames kept inside the same assigned temporal block.

We use two evaluation settings. **Within-recording forecasting** holds out temporal blocks from recordings that are also represented in training. Following the experimental protocol used in our current benchmark, 80% of the available temporal blocks are used for training and 20% for testing, and overlapping windows are constructed only after this block-level partition. This setting measures prediction on unseen time intervals while retaining the same recording identities. **Held-out-individual forecasting** instead separates complete recordings by individual: all temporal blocks from the test individuals are reserved for evaluation. In the cross-individual experiment, 21 recordings from one source study form the training cohort and four individuals are held out for testing. A pooled model is trained on all 21 training recordings, while the corresponding single-recording models are trained separately on each recording and evaluated on the same held-out targets.

Unless otherwise stated, the model-comparison setting uses $C = P = 4$ and spatial blocks of $40 \times 40 \times 20\,\mu\mathrm{m}$, corresponding to a $38 \times 30 \times 13$ grid. SimVPv2 is used as the default predictor for the task-setting analyses. The temporal study fixes $C = 32$ and varies the forecast horizon over $P \in \{1, 2, 4, 8, 16, 32\}$, whereas the spatial study varies the block dimensions while keeping the forecasting formulation unchanged. Because acquisition rates differ across source studies, a fixed number of frames can correspond to different physical durations; temporal results are therefore reported in frames and interpreted together with the source sampling rate. Sensory-category labels are used to define training compositions rather than supplied as forecasting covariates, preserving a common activity-only interface across heterogeneous studies.

3.2 MODELS

We evaluate six learned predictors spanning linear, fully connected, convolutional, recurrent, and attention-based approaches. All models predict the same standardized block-level activity, but differ in how they represent and process the observed context. To describe their interfaces, let $D = D_x D_y D_z$ and define $\mathbf{Z}^{(r)} = \mathcal{R}_{\mathrm{ts}}(\mathbf{Y}^{(r)}) \in \mathbb{R}^{T_r \times D}$, where $\mathcal{R}_{\mathrm{ts}}$ flattens the spatial dimensions while preserving the temporal axis. For video-oriented models, we use $\mathbf{V}^{(r)} = \mathcal{R}_{\mathrm{vid}}(\mathbf{Y}^{(r)})$, where $\mathcal{R}_{\mathrm{vid}}$ folds the depth dimension into a planar axis. Both operations are reversible rearrangements: predictions are mapped back to the original spatial grid before evaluation. Below, θ denotes the parameters of the model under consideration, and hatted quantities denote its predictions.

The simple **Linear** predictor applies a learned linear transformation to the observed context:

$$f_\theta^{\mathrm{Linear}}\left(\mathbf{Z}^{(r)}_{t:t+C-1}\right) = \mathcal{L}_\theta\left(\mathbf{Z}^{(r)}_{t:t+C-1}\right) = \widehat{\mathbf{Z}}^{(r)}_{t+C:t+C+P-1}, \tag{7}$$

where $\mathcal{L}_\theta$ maps C context frames to P future frames. This model provides a learned reference for assessing the predictive structure captured by linear transformations, beyond the fixed averaging and persistence rules of the naive baselines. We also consider **FC** (fully connected network), which replaces the linear transformation with a nonlinear mapping. Together with Linear, this model assesses prediction with and without nonlinear transformations, without explicitly introducing convolutional locality, recurrent states, or attention mechanisms.

SimVPv2 (Tan et al., 2025) predicts future frames using a spatial encoder, a temporal translator, and a spatial decoder:

$$f_\theta^{\mathrm{SimVPv2}}\left(\mathbf{V}^{(r)}_{t:t+C-1}\right) = \widehat{\mathbf{V}}^{(r)}_{t+C:t+C+P-1}, \qquad f_\theta^{\mathrm{SimVPv2}} = \mathcal{D}_\theta \circ \mathcal{T}_\theta \circ \mathcal{E}_\theta, \tag{8}$$

where $\mathcal{E}_\theta$ extracts spatial features, $\mathcal{T}_\theta$ models their temporal evolution, and $\mathcal{D}_\theta$ reconstructs the predicted frames. Its convolutional architecture incorporates gated spatiotemporal attention to model dependencies in the encoded sequence. We use SimVPv2 as the default predictor for the spatial-resolution, forecast-horizon, and training-composition experiments.

PredRNNv3 (Guo & Zhou, 2026) implements the forecasting map through a recurrent spatiotemporal network. The model combines convolutional processing and transformer-style modeling with recurrent state updates. Information is carried across time through the recurrent states, providing a different temporal modeling mechanism from the feed-forward encoder–translator–decoder structure of SimVPv2. **Predformer** (Tang et al., 2026) uses gated transformer blocks to map the observed video sequence to future frames. It models spatial and temporal dependencies without recurrence or convolution, providing an attention-based counterpart to the convolutional and recurrent video predictors. The prediction target remains the same block-activity sequence despite these differences in internal computation. **Crossformer** (Zhang & Yan, 2023) treats forecasting as a multivariate time-series problem. Its architecture explicitly models dependencies across temporal and variable dimensions. In FishBank, each variable corresponds to the activity of a spatial block rather than an individual neuron. Crossformer therefore provides a sequence-modeling approach to learning temporal and cross-block dependencies, complementing the video-oriented predictors.

3.3 EVALUATION

We use mean absolute error as the primary forecasting metric, consistent with prior neural-activity forecasting benchmarks (Lueckmann et al., 2025). Let Ω denote the set of evaluated recording, window, future-step, and spatial-block indices. The aggregate error of predictor f_θ is

$$\mathrm{MAE}(f_\theta) = \frac{1}{|\Omega|} \sum_{(r,t,p,b)\in\Omega} \left| \widehat{Y}^{(r)}_{t+C+p,b} - Y^{(r)}_{t+C+p,b} \right|, \tag{9}$$

where b indexes a spatial block and $p \in \{0, \ldots, P-1\}$. MAE is computed in the standardized activity space and directly measures absolute error in the block-level prediction target.

To calibrate the benefit of learning, we define two parameter-free temporal references. The **Mean** baseline repeats the temporal average of each spatial block over the observed context,

$$\widehat{Y}^{\mathrm{Mean},(r)}_{t+C+p,b} = \frac{1}{C} \sum_{j=0}^{C-1} Y^{(r)}_{t+j,b}, \qquad p = 0, \ldots, P-1, \tag{10}$$

whereas the **Copy** baseline repeats the last observed frame,

$$\widehat{Y}^{\mathrm{Copy},(r)}_{t+C+p,b} = Y^{(r)}_{t+C-1,b}. \tag{11}$$

Mean measures improvement over a recent activity level, while Copy measures improvement over temporal persistence. The two references are particularly useful in FishBank because absolute MAE changes when the spatial aggregation, forecast horizon, or evaluation cohort changes.

For comparisons across such settings, we additionally report the relative error reduction with respect to a target-matched baseline $B \in \{\mathrm{Mean}, \mathrm{Copy}\}$:

$$\mathrm{Gain}_B(f_\theta) = 100 \left(1 - \frac{\mathrm{MAE}(f_\theta)}{\mathrm{MAE}(B)}\right). \tag{12}$$

This quantity is interpreted only within a fixed prediction setting, where the model and baseline share the same targets and scoring support. In particular, changing the spatial block size changes the prediction target itself, so raw MAEs across resolutions should not be interpreted as a scale-independent measure of forecasting difficulty. We therefore report both absolute MAE and baseline-relative improvement in the subsequent experiments.

4 Experimental Results and Analysis

We examine the predictability of whole-brain activity, its dependence on spatial and temporal scales, and forecasting performance under mixed stimulus categories. Unless otherwise stated, we use $C = P = 4$ and spatial blocks of $40 \times 40 \times 20\,\mu\mathrm{m}$, following Section 3.1. SimVPv2 is the default predictor for all analyses beyond the six-model comparison.

Table 2: Forecasting performance across model families. Default description: $C = P = 4$ and $40 \times 40 \times 20\,\mu\mathrm{m}$ blocks. Linear is learned; Mean and Copy are parameter-free. Errors are the archived aggregate values associated with MAE evaluation (Section 4).

Predictor	Forecast error $L \downarrow$	G_{Mean} (%)	G_{Copy} (%)
Mean	0.4552	–	–
Copy	0.5393	–	–
Linear	0.4236	6.9	21.5
FC	0.3748	17.7	30.5
Predformer	0.3753	17.6	30.4
Crossformer	0.3775	17.1	30.0
SimVPv2	0.3784	16.9	29.8
PredRNN	0.3835	15.8	28.9

4.1 Predictability and the Benefits of Multi-Individual Training

We first evaluate prediction on held-out temporal segments from recordings represented during training, and then examine generalization to entirely held-out individuals. The first setting assesses the predictive structure accessible to different model families; the second compares learning from one individual with learning jointly from multiple individuals.

All learned predictors outperform the temporal baselines. Table 2 compares six predictors trained on the pooled model-comparison cohort. ***First***, we can see that all six models achieve lower MAE than baselines such as Mean and Copy. These results correspond to error reductions of 6.9–17.7% relative to Mean and 21.5–30.5% relative to Copy. Even Linear improves over both references, indicating that the common block-level representation retains predictive structure beyond historical averaging and persistence of the last observed frame. ***Second***, the five nonlinear predictors have closely spaced aggregate errors. Their full range spans only 0.0087 MAE, whereas Linear has a higher error of 0.4236. Thus, a fully connected network provides a competitive reference alongside convolutional, recurrent, and attention-based approaches. The clearest aggregate distinction is between Linear and the nonlinear predictors; the smaller differences within the nonlinear group do not establish a statistically reliable architecture ranking.

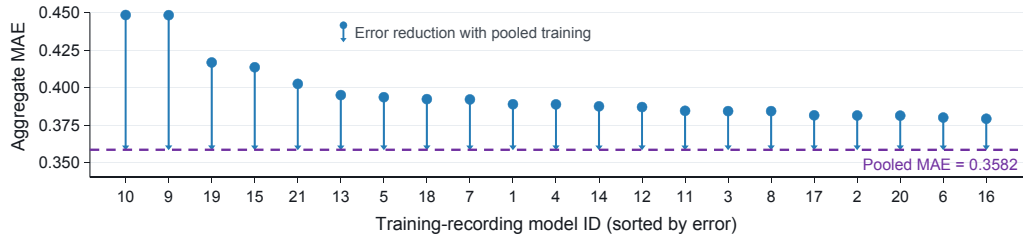

Figure 3: Single-recording versus pooled training. Circles show the MAE of 21 single-recording models. Arrows indicate reductions to the pooled model's MAE.

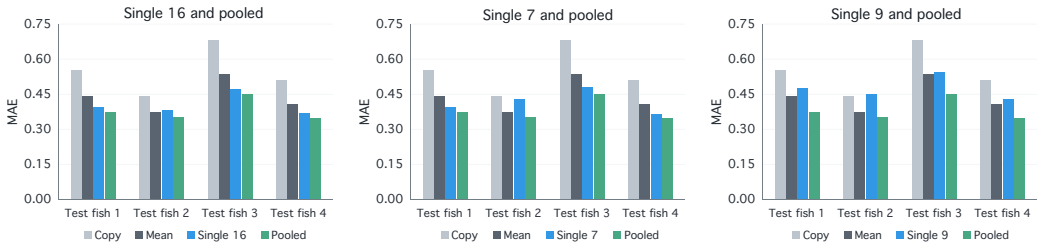

Figure 4: Per-target plots for some case of single-recording models (16, 7, and 9), each against the pooled model and temporal references.

Multi-individual training improves forecasting on held-out individuals. We compare pooled and single-recording SimVPv2 models on the same four held-out individuals from one study. Figure 3 shows that pooling outperforms every single-recording model in aggregate, achieving 0.3582 MAE and reducing error by 9.5% relative to the single-recording average (0.3958). We observe that single-recording performance varies substantially with the training individual. The examples in Fig. 4 illustrate different transfer patterns: Single 7 improves over Mean on most test fish but loses this advantage on another, whereas Single 9 performs worse than Mean throughout. In contrast, the pooled model maintains lower error than both temporal baselines and all three selected models on every test fish. These examples suggest that single-recording transfer depends on the particular training–test pairing, whereas pooled training supports more consistent baseline-relative improvements across the evaluated individuals. A plausible explanation is that pooling provides broader coverage of activity patterns, helping the model learn transferable regularities while reducing reliance on patterns specific to one recording. Both the larger training set and its greater individual diversity may contribute; their respective effects are not isolated in this comparison.

4.2 Spatial and Temporal Dependence of Predictability

We vary forecast horizon and spatial block size using SimVPv2 on the 25-recording cohort from one study, holding out temporal blocks within each recording.

Longer horizons reduce the advantage over historical averaging. Across $P \in \{1, 2, 4, 8, 16, 32\}$, SimVPv2 outperforms both baselines, but its gain over Mean falls from 23.7% at $P = 1$ to 8.1% at $P = 32$ (Fig. 5(a)). Copy deteriorates more strongly, leaving a 42.2% gain over Copy at the longest horizon. We speculate that transient departures from the recent activity level are easier to anticipate over short windows, whereas extending the last observed state increasingly misrepresents later activity. Averaging over the context is less tied to a single transient state, which may explain why Mean remains competitive as the model's short-range advantage diminishes.

Coarser blocks lower absolute error, while gains persist across scales. Reducing block dimensions from $40 \times 40 \times 20$ to $10 \times 10 \times 10\,\mu$m increases SimVPv2 MAE from 0.3648 to 0.4720, with both baselines showing the same trend (Fig. 5(b)). Larger blocks average over broader neuronal populations and suppress within-block variation, providing a plausible explanation for the lower errors on these more aggregated targets. Finer blocks retain more local fluctuations, yet the model's improvement over Mean remains 17.4–20.2% across all four resolutions. Useful predictive structure therefore remains accessible at finer scales despite higher absolute error. Unlike extending the forecast horizon, spatial refinement does not substantially reduce the gain over historical averag-

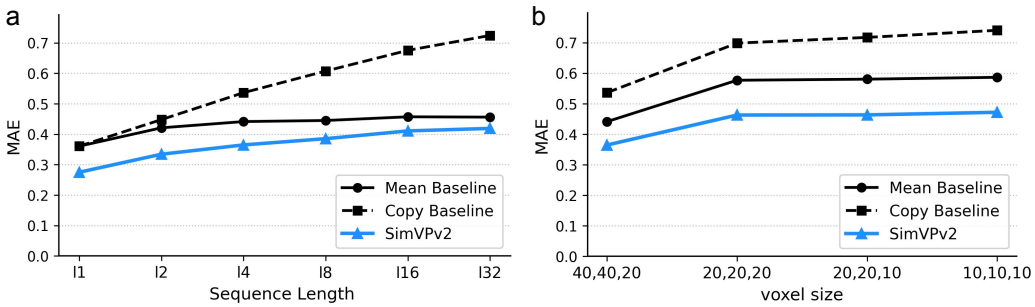

Figure 5: Forecasting across temporal and spatial scales. **(a)** MAE for output-length settings with $C = 32$ and $40 \times 40 \times 20\,\mu\text{m}$ blocks, averaged over each complete forecast window. **(b)** MAE at spatial resolutions with $C = P = 4$. Each block size defines a different population-average target.

ing. Lower coarse-grid MAE should consequently be interpreted together with the spatial detail discarded by aggregation.

4.3 Mixed-Stimulus Forecasting

We examine forecasting when recordings from different sensory categories are combined. The mixed-stimulus study draws on a 27-recording subset spanning visual, auditory, and chemosensory conditions. We compare the visual configuration with three mixtures: visual–chemosensory, visual–auditory, and three-category configurations under the default spatial and temporal settings.

Mixed-stimulus forecasting remains challenging. SimVPv2 outperforms both naive baselines in all three mixed configurations, but its gains over Mean are only 2.2%, 2.1%, and 1.8%, compared with 16.9% in the visual configuration (Fig. 6). Alongside the benefits of within-study multi-individual training, these results highlight the additional challenge of forecasting across heterogeneous sensory conditions. We speculate that differences in stimulus-dependent activity patterns, acquisition rates, and spatial coverage make it harder for a shared activity-only predictor to learn predictive regularities across recordings. Their respective contributions remain unresolved, however, because the evaluation populations also differ across configurations. The comparison therefore does not establish that mixing stimulus categories degrades performance on a fixed target set. Rather, the limited gains identify mixed-stimulus forecasting as an important opportunity for improvement, motivating condition-aware models and fixed-target comparisons of single-category and mixed training.

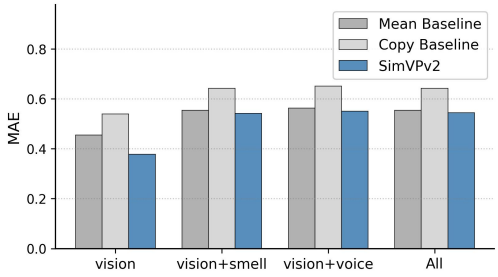

Figure 6: Mixed-stimulus forecasting results. MAE for visual recordings and mixtures with chemosensory (smell) and auditory (voice) recordings. "All" combines all three categories.

5 Conclusion

We introduced FishBank, a benchmark for cross-individual whole-brain activity forecasting comprising 189 calcium-imaging recordings across three sensory categories. Its common block-level representation enables learning across recordings without one-to-one neuronal correspondence. All six learned predictors outperform temporal baselines, while the five nonlinear models achieve similar aggregate errors. Coarser spatial blocks yield lower errors on more aggregated targets, whereas longer forecast horizons reduce gains over historical averaging. Pooled training reduces held-out-individual MAE by 9.5% relative to the average single-recording model, while mixed-stimulus settings yield only modest improvements over Mean. These findings characterize the spatial and temporal dependence of forecasting performance and the benefits and limitations of learning across individuals. Future work should separate the effects of training-set size and individual diversity and evaluate condition-aware models on fixed target sets.

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. doi: 10.1038/nmeth.2434.

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. doi: 10.1038/nature11129.

Mattéo Dommanget-Kott, Jorge Fernandez-de Cossio-Diaz, Guillaume Faye-Bédrin, Georges Debrégeas, and Volker Bormuth. Cross-individual translation of spontaneous zebrafish brain activity through a shared latent representation. *Proceedings of the National Academy of Sciences*, 123(20):e2529064123, 2026. doi: 10.1073/pnas.2529064123.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0798. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/21fbf654963f9e455e8eec48e45bdea0-Abstract-Conference.html.

Itia A. Favre-Bulle, Michael A. Taylor, Emmanuel Marquez-Legorreta, Gilles Vanwalleghem, Rebecca E. Poulsen, Halina Rubinsztein-Dunlop, and Ethan K. Scott. Sound generation in zebrafish with Bio-Opto-Acoustics. *Nature Communications*, 11(1):6120, 2020. doi: 10.1038/s41467-020-19982-5.

Gaihui Guo and You Zhou. PredRNNv3: A CNN-transformer collaborative recurrent neural network for spatiotemporal prediction. *Neurocomputing*, 664:132078, 2026. doi: 10.1016/j.neucom.2025.132078.

Jan-Matthis Lueckmann, Alexander Immer, Alex Bo-Yuan Chen, Peter H. Li, Mariela D. Petkova, Nirmala A. Iyer, Luuk Willem Hesselink, Aparna Dev, Gudrun Ihrke, Woohyun Park, Alyson Petruncio, Aubrey Weigel, Wyatt Korff, Florian Engert, Jeff W. Lichtman, Misha B. Ahrens, Michał Januszewski, and Viren Jain. ZAPBench: A benchmark for whole-brain activity prediction in zebrafish. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/668563ef18fbfef0b66af491ea334d5f-Abstract-Conference.html.

Emmanuel Marquez-Legorreta, Lena Constantin, Marielle Piber, Itia A. Favre-Bulle, Michael A. Taylor, Ann S. Blevins, Jean Giacomotto, Dani S. Bassett, Gilles C. Vanwalleghem, and Ethan K. Scott. Brain-wide visual habituation networks in wild type and fmr1 zebrafish. *Nature Communications*, 13(1):895, 2022. doi: 10.1038/s41467-022-28299-4.

Geoffrey Migault, Thijs L. van der Plas, Hugo Trentesaux, Thomas Panier, Raphaël Candelier, Rémi Proville, Bernhard Englitz, Georges Debrégeas, and Volker Bormuth. Whole-brain calcium imaging during physiological vestibular stimulation in larval zebrafish. *Current Biology*, 28(23):3723–3735.e6, 2018. doi: 10.1016/j.cub.2018.10.017.

Rebecca E. Poulsen, Leandro A. Scholz, Lena Constantin, Itia Favre-Bulle, Gilles C. Vanwalleghem, and Ethan K. Scott. Broad frequency sensitivity and complex neural coding in the larval zebrafish auditory system. *Current Biology*, 31(9):1977–1987.e4, 2021. doi: 10.1016/j.cub.2021.01.103.

Samuel K. H. Sy, Danny C. W. Chan, Roy C. H. Chan, Jing Lyu, Zhongqi Li, Kenneth K. Y. Wong, Chung Hang Jonathan Choi, Vincent C. T. Mok, Hei-Ming Lai, Owen Randlett, Yu Hu, and Ho Ko. An optofluidic platform for interrogating chemosensory behavior and brainwide neural representation in larval zebrafish. *Nature Communications*, 14(1):227, 2023. doi: 10.1038/s41467-023-35836-2.

Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z. Li. SimVPv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Transactions on Multimedia*, 27:5170–5184, 2025. doi: 10.1109/TMM.2025.3543051.

Ryosuke Tanaka and Ruben Portugues. Algorithmic dissection of optic flow memory in larval zebrafish. *Current Biology*, 35(20):4870–4881.e4, 2025. doi: 10.1016/j.cub.2025.08.047.

Yujin Tang, Lu Qi, Xiangtai Li, Chao Ma, and Ming-Hsuan Yang. Video prediction transformers without recurrence or convolution. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Afvhu9Id8m.

Sam Fatehmanesh Vegas, Matt Thomson, James Gornet, and David Prober. A sensing whole brain zebrafish foundation model for neuron dynamics and behavior, 2025. URL https://arxiv.org/abs/2510.27366.

Nikita Vladimirov, Yu Mu, Takashi Kawashima, Davis V. Bennett, Chao-Tsung Yang, Loren L. Looger, Philipp J. Keller, Jeremy Freeman, and Misha B. Ahrens. Light-sheet functional imaging in fictively behaving zebrafish. *Nature Methods*, 11(9):883–884, 2014. doi: 10.1038/nmeth.3040.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

A Appendix

A.1 Limitations

The results in our study are specific to the evaluated zebrafish recordings, preprocessing pipeline, model families, and forecasting settings. Blockwise aggregation enables cross-individual evaluation but limits the assessment of neuron-level dynamics. Broader datasets, controlled training-data comparisons, and condition-aware models remain future work.

A.2 Models and Training Details

We used SimVPv2 with the gSTA translator. The encoder and decoder each comprised four ConvSC blocks with a kernel size of 3. The encoder applied stride-2 downsampling in alternating blocks, yielding an overall downsampling factor of 4. We trained the model for 100 epochs using Adam with an initial learning rate of 5×10^{-3}, no weight decay, a cosine annealing schedule, and a batch size of 32. The checkpoint with the lowest validation loss was selected for evaluation.

We used the TSST variant of PredFormer, which employs quadruplet spatiotemporal factorization. Input frames were resized to 32×392 and divided into 8×8 patches, yielding 196 tokens per frame. We trained the model for 50 epochs using AdamW with an initial learning rate of 10^{-3}, a weight decay of 10^{-2}, a cosine annealing schedule, and a batch size of 8. Training minimized the mean absolute error (MAE), and the checkpoint with the lowest validation loss was selected for evaluation.

We used PredRNNv3 with the dual-branch LSTM (DB-LSTM) backbone. Input frames were resized to 32×392 and patchified with a patch size of 8. We trained the model for 100 epochs using AdamW with a batch size of 8. The learning rate was scaled linearly with batch size from a per-sample value of 6.25×10^{-6}, resulting in an initial learning rate of 5×10^{-5}. We used a MultiStepLR schedule with milestones at epochs 40 and 80 and a decay factor of 0.1. Gradient clipping was applied with a threshold of 1, and reverse-exponential scheduled sampling was used during training. Training minimized MAE, and the checkpoint with the lowest validation loss was selected for evaluation.

We configured Crossformer with a segment length of 1, a segment merging window size of 2, and 10 routers in the cross-dimension stage of each two-stage attention layer. We trained the model for up to 50 epochs using Adam with an initial learning rate of 2×10^{-4} and a batch size of 32. If the validation loss did not improve for 5 consecutive epochs, the learning rate was reduced by a factor of 10 and the patience counter was reset.

Both fully connected and linear networks comprised an input layer, a single hidden layer with $100{,}000$ units, and an output layer. The linear network used the same architecture without activation functions. We trained both networks for 100 epochs using Adam with an initial learning rate of 5×10^{-3}, a cosine annealing schedule, and a batch size of 32. For each network, the checkpoint with the lowest validation loss was selected for evaluation.