# BrainVista: Modeling Naturalistic Brain Dynamics as Multimodal Next-Token Prediction

**Xuanhua Yin**[1] **Runkai Zhao**[1*] **Lina Yao**[2,3] **Weidong Cai**[1*]
[1]The University of Sydney, Australia
[2]University of New South Wales, Australia
[3]CSIRO's Data61
runkai.zhao@sydney.edu.au tom.cai@sydney.edu.au

## Abstract

Naturalistic fMRI offers a time-resolved window into cortical dynamics during continuous multimodal experience, where future brain states are jointly governed by endogenous neural context and exogenous sensory drive. Forecasting such dynamics, however, remains challenging because fast-changing stimulus streams must be reconciled with the sluggish hemodynamic BOLD response, while cortical activity is organized across functionally heterogeneous brain networks. To address these challenges, we introduce BrainVista, a multimodal autoregressive framework for observed-stimulus conditional brain-state forecasting. BrainVista predicts future fMRI states from the history of brain activity and stimulus tokens temporally aligned to each prediction query, enabling autoregressive prediction without access to future ground-truth fMRI states. Recursive rollout provides a within-run predictive test of brain-state evolution under continuing sensory input. Specifically, we combine Network-wise Tokenizers for functionally structured brain representations, a Spatial Mixer Head for cross-network interaction refinement, and Stimulus-to-Brain (S2B) masking to preserve temporally causal stimulus–brain conditioning while preventing future brain-state leakage and stimulus look-ahead. We evaluate BrainVista on Algonauts 2025, CineBrain, and HAD, three large-scale naturalistic fMRI benchmarks, under long-horizon conditional autoregressive forecasting. BrainVista consistently outperforms baselines, improving pattern correlation by 10.8% and 9.4% relative to the strongest baseline on Algonauts 2025 and CineBrain, respectively. Code is available at https://github.com/xuanhuayin/BrainVista.

## 1 Introduction

The Blood-Oxygen-Level-Dependent (BOLD) signal in functional magnetic resonance imaging (fMRI) is a temporally structured measurement of brain activity, where future observations are constrained by prior neural and hemodynamic states rather than arising as independent samples [1–3]. Beyond short-range autocorrelation, this temporal structure reflects the brain's continuous integration of intrinsic dynamics with incoming sensory evidence, a view closely related to predictive accounts of cortical processing [4, 5]. This coupling is particularly salient during naturalistic stimulation, such as movie viewing, where cortical responses evolve under continuous visual, auditory, and semantic input. We therefore view naturalistic fMRI as an observed, non-autonomous dynamical process: the future brain state depends on prior brain activity and an exogenous stimulus stream that is time-aligned to the fMRI acquisition grid.

*Corresponding authors: Runkai Zhao and Weidong Cai.
Xuanhua Yin: xuanhua.yin@sydney.edu.au; Lina Yao: lina.yao@unsw.edu.au.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The scientific question is whether a model can sustain the complementary information in brain history and observed stimuli when its own predictions replace measured brain states. We use within-run forecasting as a predictive assay: observed fMRI initializes the state, and the withheld continuation tests competing dynamical assumptions under identical input constraints. Input ablations test the contribution of each source; history occlusion tests predictive reliance on specific lags; and long-horizon rollout distinguishes one-step fitting from stable recursive prediction. This scope does not require treating forecasts as faithful brain simulations.

Despite the temporal nature of BOLD responses, many recent brain foundation models are trained with masked reconstruction, bidirectional attention, or global-context representation learning objectives [6–8]. These objectives are effective for self-supervised representation learning and downstream transfer, but they are not directly designed for autoregressive forecasting of future brain states. When bidirectional or masked-context models are used for forecasting, they may create a training–inference mismatch: during training, the model can exploit information from time points that would be unavailable during recursive rollout. This mismatch is especially problematic for long-horizon prediction, where future ground-truth fMRI states must be withheld and model predictions are recursively fed back as context. Thus, naturalistic fMRI forecasting requires an evaluation protocol and model architecture that distinguish observed exogenous stimulus inputs from unavailable future brain states.

Autoregressive models and other sequence architectures have advanced language modeling and time-series forecasting [9–13]. However, extending this paradigm to naturalistic fMRI is non-trivial. First, multimodal stimulus streams are sampled at substantially higher temporal resolution than fMRI, while BOLD responses are slow and hemodynamically delayed [14, 15]. A forecasting model must therefore align high-frequency sensory features to the fMRI grid without introducing look-ahead from stimulus tokens beyond the queried prediction step. In our setting the stimulus stream is treated as an observed exogenous driver: predicting the brain state at time $t+1$ may condition on the stimulus token aligned to that target time, but not on future stimulus tokens beyond $t+1$ or on the target and future fMRI states. Second, cortical responses are organized across heterogeneous functional systems. The cortex contains specialized networks with conserved functional organization [16–18], and hierarchical differences in circuit and temporal properties shape large-scale dynamics [19–21]. Generic sequence models that compress the whole brain into a single homogeneous representation can blur functional boundaries, reduce network-specific contrast, and obscure structured interactions across cortical systems.

To address these challenges, we propose BrainVista, a time-aligned multimodal framework that combines network-specific representations, directed stimulus–brain attention, and cross-network refinement for recursive fMRI forecasting. Our contributions are threefold:

- We formulate naturalistic fMRI modeling as observed-stimulus conditional autoregressive forecasting and develop BrainVista to predict future brain states from brain history and available stimuli. Its Stimulus-to-Brain (S2B) mask enforces temporally valid conditioning while keeping stimulus representations independent of brain-token history.
- We combine Network-wise Tokenizers with a Spatial Mixer Head to retain separate network representations before temporal prediction and refine predicted network components before decoding. Controlled ablations evaluate these design choices and their interaction during recursive forecasting.
- Experiments on Algonauts 2025 [22], CineBrain [23], and HAD [24] show consistent long-horizon improvements over matched-access baselines, including Transformer-ARX with the same frozen encoders and tokenizers. Additional experiments assess tokenizer objectives, data and model scaling, and the representational structure retained by forecasts.

## 2 Related Work

**fMRI Tokenization and Brain Foundation Models.** High-dimensional fMRI signals require compact representations for scalable sequence modeling. Existing approaches include discrete code sequences [25, 26], task-aware reconstruction [27], probabilistic latent tokens [28, 29], and masked-reconstruction pretraining [30, 31]. Recent brain foundation models learn transferable neural representations from large-scale neuroimaging data, including Brain-JEPA [7], Brain Harmony [8], and BrainLM [6], with related extensions to atlas-free, memory-efficient, concept-conditioned, and

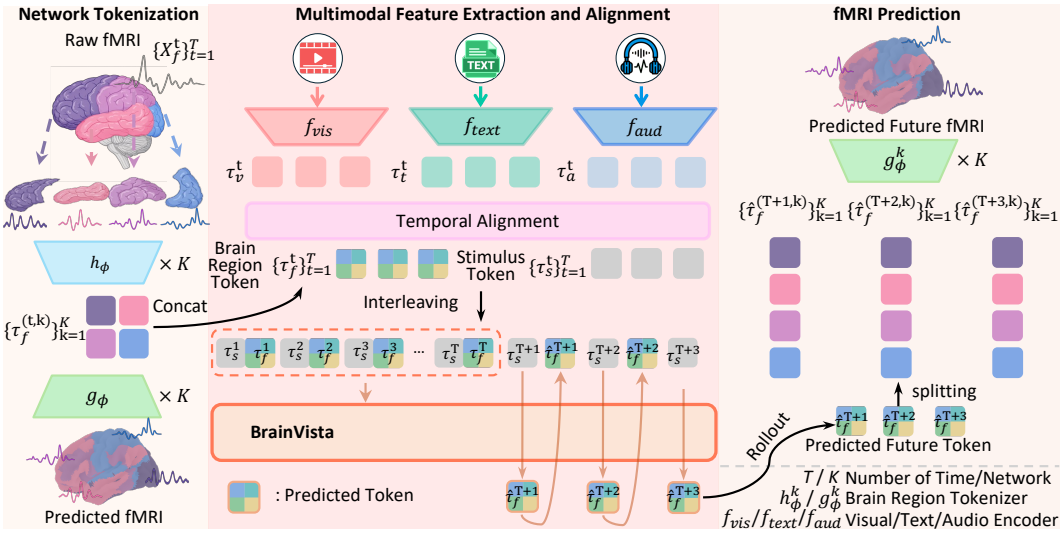


Figure 1: **Framework of BrainVista.** Time-aligned multimodal features are aggregated into stimulus tokens, while parcel-wise fMRI is encoded by network-specific tokenizers into functionally structured brain tokens. Given interleaved stimulus and brain tokens from the valid context, BrainVista predicts the next brain token and decodes it back to fMRI space.

graph-based modeling [32–34]. These methods mainly target reconstruction, representation learning, or task transfer, whereas BrainVista focuses on temporally constrained forecasting under observed naturalistic stimuli.

**Multimodal Stimulus–Brain Modeling.** Naturalistic movie paradigms motivate multimodal modeling of brain activity from temporally aligned visual, auditory, and language inputs. TRIBE predicts whole-brain responses from video, audio, and text features [35]; SynBrain models one-to-many stimulus-to-fMRI mappings [28]; dynamic subject-aware routing addresses inter-subject variability [36]; and Algonauts 2025 provides a large-scale benchmark for naturalistic fMRI prediction [22]. Related decoding studies reconstruct stimuli from brain activity using diffusion-based or unified multimodal models [37–40]. Bratrix explores language-anchored vision–brain alignment with uncertainty-aware weighting [41]. In contrast, BrainVista predicts future brain states from past brain activity and aligned sensory context, treating naturalistic fMRI as an evolving state process.

**Autoregressive Brain-State Forecasting.** Autoregressive modeling is well suited to temporally evolving brain activity, and prior work has studied individual fMRI time-series forecasting in large-scale datasets [42]. This direction is also supported by evidence for hierarchical temporal receptive windows and heterogeneous intrinsic timescales during naturalistic cognition [43–46, 15]. However, naturalistic fMRI forecasting must combine endogenous brain history with exogenous multimodal stimuli while avoiding leakage from mismatched stimulus and BOLD timescales [14, 15]. BrainVista addresses this setting through Stimulus-to-Brain masking, Network-wise Tokenizers, and Spatial Mixer-based cross-network refinement.

## 3 Methodology

BrainVista forecasts naturalistic fMRI as time-aligned multimodal next-token prediction without access to target or future fMRI. As shown in Fig. 1, it aligns observed stimulus features to the fMRI grid, encodes network-wise brain tokens, and predicts future tokens with a Transformer under Stimulus-to-Brain masking. The Spatial Mixer Head refines cross-network interactions before decoding to fMRI space.

### 3.1 Problem Setup and Notation

**Observed-stimulus Conditional Brain Forecasting.** For each subject and run, let $x^t \in \mathbb{R}^P$ denote the model-space fMRI state at time step $t$, where $P$ is the number of parcels or ROI features. Let

$\tau_s^t$ denote the stimulus token aligned to the same fMRI grid. The stimulus stream is treated as an observed external driver, whereas future fMRI states are latent variables to be predicted.

For a queried prediction time $u$, the valid conditioning set is:

$$\mathcal{C}_u = \left(x^{<u}, \tau_s^{\leq u}\right), \tag{1}$$

where the target-aligned stimulus token $\tau_s^u$ is observed, but the target and future fMRI states $x^{\geq u}$ are never used as inputs.

**One-step Transition.** At time $t$, BrainVista learns a one-step transition:

$$\hat{x}^{t+1} = f_\theta\left(x^{\leq t}, \tau_s^{\leq t+1}\right), \tag{2}$$

where $f_\theta$ is the forecasting model, $x^{\leq t} = (x^1, \ldots, x^t)$ denotes the ordered brain-state history, and $\tau_s^{\leq t+1} = (\tau_s^1, \ldots, \tau_s^{t+1})$ denotes the ordered stimulus sequence up to the queried prediction step.

**Multi-step Conditional Rollout.** Given an observed context window ending at time $T$, an $H$-step forecast is produced recursively:

$$\hat{x}^{T+h} = f_\theta\left(x^{\leq T}; \hat{x}^{T+1:T+h-1}, \tau_s^{\leq T+h}\right), \qquad h = 1, \ldots, H, \tag{3}$$

where $H$ is the rollout horizon and $\hat{x}^{T+1:T+h-1} = (\hat{x}^{T+1}, \ldots, \hat{x}^{T+h-1})$ denotes the ordered sequence of previously predicted brain states, which is empty when $h = 1$. The stimulus sequence $\tau_s^{\leq T+h}$ remains observed up to the queried prediction step. Thus, our setting is conditional brain-state forecasting under observed naturalistic stimuli; it neither models unconditional neural dynamics nor establishes biological causality.

### 3.2 Stimulus Tokens

**Frozen Encoders.** We employ fixed pretrained encoders $f_{\text{vis}}, f_{\text{aud}}, f_{\text{text}}$ to map raw video, audio, and text streams $s_{\text{vis}}, s_{\text{aud}}, s_{\text{text}}$ into time-indexed stimulus features, producing $\tau_{\text{vis}} \in \mathbb{R}^{d_{\text{v}}}$, $\tau_{\text{aud}} \in \mathbb{R}^{d_{\text{a}}}$, and $\tau_{\text{text}} \in \mathbb{R}^{d_{\text{t}}}$. Here, $d_{\text{v}}$, $d_{\text{a}}$, and $d_{\text{t}}$ denote the embedding dimensionalities of the video, audio, and text stimulus features, respectively.

**Temporal Alignment.** We aggregate stimulus features onto the native fMRI grid, treating each fMRI timestamp as the end of an acquisition bin. Video and audio features are eligible only if their extraction windows end no later than that bin boundary; text uses only the causal prefix available by the same boundary. For modality $m \in \{\text{vis}, \text{aud}, \text{text}\}$, let $\tau'_{m,i} \in \mathbb{R}^{d_m}$ denote the $i$-th high-rate feature, and let $\mathcal{B}_{m,t}$ denote the set of features assigned to fMRI bin $t$. We mean-pool the eligible features assigned to each bin:

$$\tau_m^t = \frac{1}{|\mathcal{B}_{m,t}|} \sum_{i \in \mathcal{B}_{m,t}} \tau'_{m,i}, \qquad t = 1, \ldots, T. \tag{4}$$

**Shared Stimulus Token.** At each fMRI time step $t$, we form a shared stimulus token by concatenating modality tokens and projecting them to the Transformer hidden space:

$$\tau_s^t = \text{Proj}_s\left([\tau_{\text{vis}}^t; \tau_{\text{aud}}^t; \tau_{\text{text}}^t]\right) \in \mathbb{R}^{d_{\text{model}}}, \tag{5}$$

where $\text{Proj}_s(\cdot)$ is a learned projection and $d_{\text{model}}$ denotes the token embedding dimensionality of the temporal Transformer. We add no separate video, audio, or text modality embeddings because these features form one stimulus token. Dataset-specific TRs, extraction windows, strides, and modality availability are detailed in Appendix A.2 and Table 7.

### 3.3 Brain Region Tokens

**Network-aware Grouping.** Because cortical systems exhibit heterogeneous temporal and functional properties [47, 16], we partition the $P$ fMRI dimensions into $K$ functional networks $\{\mathcal{V}_k\}_{k=1}^K$. This avoids compressing the whole brain into a single homogeneous representation, which can over-smooth network-specific structure [48]. Each $\mathcal{V}_k$ is a fixed, nonlearned index set: Algonauts 2025 and HAD use Schaefer-to-Yeo-7 assignments [47, 49], whereas CineBrain uses dataset-provided visual and

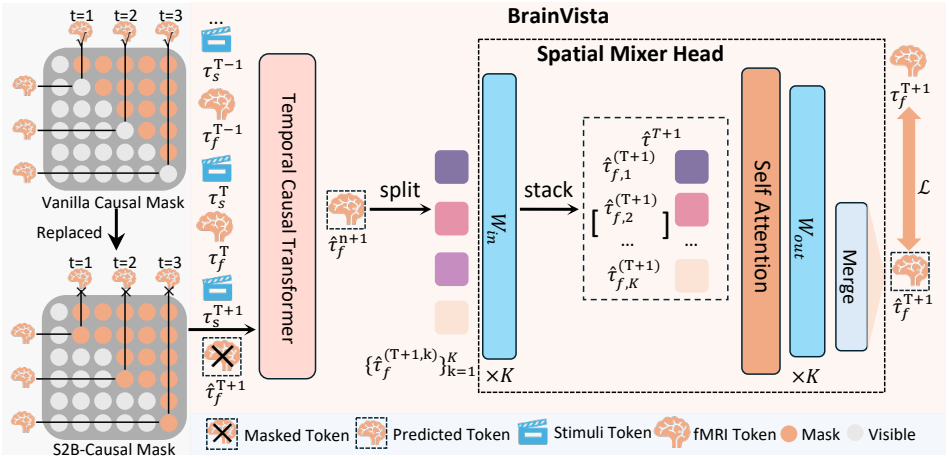


Figure 2: **BrainVista core.** For next-step prediction, BrainVista takes an interleaved sequence of stimulus tokens and available brain tokens from the recent context. The target brain-token position is decoded as the prediction target. The Stimulus-to-Brain mask keeps stimulus tokens in a stimulus-only causal stream. The temporal Transformer predicts a whole-brain latent token, which is refined by the Spatial Mixer Head and decoded to fMRI space.

auditory ROI groups. The selected signals remain multidimensional vectors, without averaging within groups. Given the native fMRI response $x^t_{\text{raw}}$, the model-space signal of group $k$ is:

$$x^{k,t} = \mathcal{A}_k\big(x^t_{\text{raw}}[\mathcal{V}_k]\big) \in \mathbb{R}^{P_k}, \tag{6}$$

where $\mathcal{A}_k$ is identity mapping for parcellated inputs and a training-set PCA projection for CineBrain. All preprocessing statistics are fitted on training runs only and fixed for validation, test, and rollout evaluation. The whole-brain representation is $x^t = [x^{1,t}; \ldots; x^{K,t}] \in \mathbb{R}^P$, where $P = \sum_{k=1}^K P_k$.

**Network-wise Tokenization.** We use independent network-wise tokenizer–detokenizer pairs to obtain compact brain tokens. For group $k$, a lightweight MLP tokenizer $h^k_{\phi_k}$ and detokenizer $g^k_{\psi_k}$ compute:

$$\tau^{t,k}_f = h^k_{\phi_k}(x^{k,t}) \in \mathbb{R}^{d_k}, \qquad \tilde{x}^{k,t} = g^k_{\psi_k}(\tau^{t,k}_f) \in \mathbb{R}^{P_k}, \tag{7}$$

where $\tau^{t,k}_f$ is the brain region token, $\tilde{x}^{k,t}$ is the reconstructed model-space signal, and $d_k$ is the network-specific token dimensionality. The parameters $\phi_k$ and $\psi_k$ are network-specific and are not shared across functional networks. Dataset-specific values of $P_k$, $d_k$, and $d_{\text{brain}}$ are reported in Appendix A.12.

**Projection to Transformer Space.** The group-wise brain region tokens are concatenated into a whole-brain token and projected to the Transformer hidden space before temporal modeling:

$$\tau^t_f = [\tau^{t,1}_f; \ldots; \tau^{t,K}_f] \in \mathbb{R}^{d_{\text{brain}}}. \tag{8}$$

A learned input projection maps this whole-brain token to the same Transformer token dimension as the stimulus token. For notational simplicity, we continue to denote the projected brain token by $\tau^t_f$ in the temporal Transformer.

### 3.4 BrainVista: Temporal Causal Transformer with Stimulus-to-Brain Mask

**Interleaved Time-aligned Input Sequence.** For a queried prediction time $u$, we interleave the recent brain and stimulus windows, whose exact index ranges are given in Appendix A.6. The stimulus token at the queried time is available, while the corresponding brain-token position is decoded as the prediction target. Fig. 2 illustrates this core computation, including the interleaved input sequence, the S2B-causal attention mask, target-token prediction, and Spatial Mixer refinement. Writing $N_u$ for the sequence length, the ordered projected tokens have the form:

$$S_u = \big(\ldots, \tau^{u-1}_s, \tau^{u-1}_f, \tau^u_s, \tau^u_f\big) \in \mathbb{R}^{N_u \times d_{\text{model}}}. \tag{9}$$

Here, the final brain-token position $\tau^u_f$ denotes the target position to be decoded, rather than an observed brain token encoded from $x^u$. This construction preserves the target-aligned stimulus token

$\tau_s^u$ while preventing access to the unavailable brain state $x^u$. Separate stimulus and brain input projections, fixed interleaved positions, positional embeddings, and S2B attention rules distinguish the two token types.

**Stimulus-to-Brain Mask (S2B).** We constrain self-attention with a Stimulus-to-Brain mask $M_u \in \{0,1\}^{N_u \times N_u}$ over $S_u$. Stimulus-token queries attend only to non-future stimulus tokens, forming a stimulus-only causal stream. Brain-token queries attend to causally available stimulus tokens and past brain tokens. The target brain-token position $\tau_f^u$ attends to $\tau_s^{\leq u}$ and $\tau_f^{<u}$, but never to the target or future fMRI states. This preserves the observed-stimulus conditional rollout graph without stimulus look-ahead beyond the queried step. Unlike S2B, vanilla causal attention allows stimulus queries to read earlier brain tokens, including predictions during rollout, so prediction errors can enter stimulus representations. S2B blocks this brain-to-stimulus path; its defining distinction is directed information flow, beyond target-aligned stimulus availability.

We feed $S_u$ into a GPT-style Transformer under $M_u$, take the hidden state at the target brain-token position, and apply a brain output projection to predict the preliminary whole-brain token:

$$H_u = \mathrm{Transformer}(S_u; M_u), \qquad \tilde{\tau}_f^u = \mathrm{Proj}_b^{\mathrm{out}}\big(H_u[\tau_f^u]\big) \in \mathbb{R}^{d_{\mathrm{brain}}}, \tag{10}$$

where $H_u \in \mathbb{R}^{N_u \times d_{\mathrm{model}}}$ is the sequence of hidden states, $H_u[\tau_f^u]$ denotes the hidden state at the target brain-token position, and $\tilde{\tau}_f^u$ is the preliminary whole-brain latent prediction.

**Spatial Mixer Head.** To refine cross-network interactions, we split $\tilde{\tau}_f^u$ into network-wise components $\{\tilde{\tau}_f^{u,k}\}_{k=1}^K$ and project them to a shared mixer dimension:

$$R^u = \mathrm{Stack}\Big(W_{\mathrm{in}}^1 \tilde{\tau}_f^{u,1}, \ldots, W_{\mathrm{in}}^K \tilde{\tau}_f^{u,K}\Big) \in \mathbb{R}^{K \times d_{\mathrm{mix}}}. \tag{11}$$

A multi-head self-attention layer models state-dependent interactions across the $K$ network tokens:

$$\tilde{R}^u = \mathrm{MHA}(R^u). \tag{12}$$

The outputs are mapped back to network-wise token dimensions and concatenated:

$$\hat{\tau}_f^{u,k} = W_{\mathrm{out}}^k \tilde{R}_k^u, \qquad \hat{\tau}_f^u = \big[\hat{\tau}_f^{u,1}; \ldots; \hat{\tau}_f^{u,K}\big] \in \mathbb{R}^{d_{\mathrm{brain}}}. \tag{13}$$

The attention matrix provides a diagnostic summary of pairwise network coupling, but is not interpreted as biological causality.

**Decoding to fMRI Space.** The refined network-wise tokens are decoded by the frozen detokenizers:

$$\hat{x}^{u,k} = g_{\psi_k}^k(\hat{\tau}_f^{u,k}), \qquad \hat{x}^u = \mathrm{Concat}_{k=1}^K \hat{x}^{u,k}, \tag{14}$$

where $\hat{x}^{u,k}$ is the predicted model-space signal for network $k$ and $\hat{x}^u$ is the predicted whole-brain fMRI state. When metrics are computed in the original signal scale, predictions are mapped back using the corresponding training-split preprocessing statistics.

### 3.5 Training Objectives and Optimization

**Tokenizer Reconstruction.** We pretrain deterministic network-wise autoencoders using reconstruction alone, with no additional latent prior. Their low-dimensional bottlenecks, network-specific parameters, full-signal reconstruction, and subsequent freezing constrain the learned representations; the tokens are not random projections. Appendix E.8 compares alternative latent regularizers. The tokenizer loss is:

$$\mathcal{L}_{\mathrm{tok}} = \sum_t \sum_{k=1}^K \Big\| x^{k,t} - g_{\psi_k}^k\Big(h_{\phi_k}^k(x^{k,t})\Big)\Big\|_2^2. \tag{15}$$

Here, $\phi_k$ and $\psi_k$ denote the network-specific tokenizer and detokenizer parameters, respectively.

**Next-token Prediction.** After tokenizer pretraining, we freeze the tokenizer–detokenizer pairs $\{h_{\phi_k}^k, g_{\psi_k}^k\}_{k=1}^K$ and train BrainVista with teacher forcing to predict the next whole-brain token. Let $\tau_f^{t+1} \in \mathbb{R}^{d_{\mathrm{brain}}}$ denote the target whole-brain token at time $t{+}1$, and let $\hat{\tau}_f^{t+1} \in \mathbb{R}^{d_{\mathrm{brain}}}$ be the model prediction after the Spatial Mixer Head. We minimize:

$$\mathcal{L}_{\mathrm{pred}} = \sum_t \Big\| \tau_f^{t+1} - \hat{\tau}_f^{t+1} \Big\|_2^2. \tag{16}$$

Table 1: **Main results for conditional autoregressive forecasting with horizon $H=20$ on Algonauts 2025, CineBrain, and HAD.** Non-ARX references are reported as references, while ARX baselines receive matched past fMRI states and aligned stimulus features. Entries are means; seed-level and subject-level statistics are reported in Tables 18 and 31. Higher is better. Best and second-best means are shown in red and blue.

| | Algonauts 2025 | | | CineBrain | | | HAD | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | $r$ | $p_{\text{corr}}$ | Cls@Top10 | $r$ | $p_{\text{corr}}$ | Cls@Top10 | $r$ | $p_{\text{corr}}$ | Cls@Top10 |
| *Non-ARX references* | | | | | | | | | |
| TRIBE [35] | 0.1379 | 0.1323 | 0.6823 | 0.4742 | 0.3012 | 0.8560 | 0.2685 | 0.1845 | 0.7012 |
| MindSimulator [33] | 0.1432 | 0.1478 | 0.6919 | 0.4856 | 0.3235 | 0.8615 | 0.2610 | 0.1960 | 0.7105 |
| BrainLM [6] | 0.1191 | 0.0829 | 0.6493 | 0.4633 | 0.2762 | 0.8478 | 0.2465 | 0.1450 | 0.6912 |
| BrainJEPA [7] | 0.1493 | 0.0916 | 0.6812 | 0.4941 | 0.2445 | 0.8455 | 0.2515 | 0.2353 | 0.7321 |
| Brain Harmony [8] | 0.1867 | 0.1466 | 0.7031 | 0.5397 | 0.3484 | 0.8642 | 0.2642 | 0.2688 | 0.7685 |
| *ARX baselines* | | | | | | | | | |
| Ridge-ARX | 0.1762 | 0.1584 | 0.7106 | 0.5483 | 0.4016 | 0.8678 | 0.2630 | 0.2327 | 0.7446 |
| LSTM-ARX | 0.1905 | 0.1718 | 0.7223 | 0.5614 | 0.4529 | 0.8735 | 0.2642 | 0.2605 | 0.7631 |
| TemporalCNN-ARX | 0.1968 | 0.1756 | 0.7268 | 0.5732 | 0.4724 | 0.8786 | 0.2654 | 0.2741 | 0.7762 |
| Transformer-ARX | 0.2026 | 0.1819 | 0.7315 | 0.5897 | 0.4978 | 0.8837 | 0.2678 | 0.2724 | 0.7836 |
| BrainHarmony-AR | **0.2065** | **0.1887** | **0.7392** | **0.5988** | **0.5156** | **0.8893** | **0.2685** | **0.2769** | **0.7894** |
| **BrainVista** | **0.2183** | **0.2090** | **0.7619** | **0.6315** | **0.5643** | **0.9012** | **0.2739** | **0.2950** | **0.7998** |

Table 2: **HAD forecasting results across horizons for matched-access autoregressive methods.** We report $H \in \{1, 10, 20\}$. Best and second-best values are shown in red and blue.

| | $H=1$ | | $H=10$ | | $H=20$ | |
|---|---|---|---|---|---|---|
| **Model** | $r$ | $p_{\text{corr}}$ | $r$ | $p_{\text{corr}}$ | $r$ | $p_{\text{corr}}$ |
| Ridge-ARX | 0.5740 | 0.5530 | 0.2860 | 0.2675 | 0.2630 | 0.2327 |
| LSTM-ARX | 0.5815 | 0.5740 | 0.3002 | 0.2940 | 0.2642 | 0.2605 |
| TemporalCNN-ARX | 0.5869 | 0.5831 | 0.3065 | 0.3072 | 0.2654 | 0.2741 |
| Transformer-ARX | 0.5890 | 0.5876 | 0.3112 | 0.3160 | 0.2678 | 0.2724 |
| BrainHarmony-AR | **0.5912** | **0.5925** | **0.3136** | **0.3215** | **0.2685** | **0.2769** |
| **BrainVista** | **0.6058** | **0.6214** | **0.3278** | **0.3390** | **0.2739** | **0.2950** |

# 4 Experiments and Results

## 4.1 Experimental Setup

**Datasets and Splits.** We evaluate BrainVista on Algonauts 2025 [22], CineBrain [23], and HAD [24], covering diverse stimulus paradigms, temporal resolutions, preprocessing pipelines, and brain-space definitions. For each dataset, we follow the official run-level train/validation/test split, use validation runs only for model selection, and report final results on held-out test runs. Forecasting windows are constructed within runs and never cross run boundaries. Algonauts 2025 and HAD use Schaefer-1000 parcels mapped to Yeo-7 systems, while CineBrain uses dataset-defined visual and auditory ROI groups with training-set PCA reduction before tokenization.

**Brain and Stimulus Representation.** The fMRI signal is represented on each dataset's native parcel, voxel, or ROI grid and grouped into functional systems for network-wise tokenization. Algonauts 2025 and HAD use Schaefer-to-Yeo assignments [49, 47], while CineBrain follows its dataset-defined sensory systems [23]. All normalization statistics are estimated from training runs only and fixed for validation, test, and rollout evaluation. When available, video, audio, and text features are extracted using frozen V-JEPA 2 [50], Wav2Vec2-BERT [51], and Llama 3.2 3B [52, 53], average-pooled within each fMRI bin, and projected to stimulus tokens.

**Baselines.** We compare with non-ARX references from stimulus-to-brain prediction and brain representation learning, including TRIBE [35], MindSimulator [33], BrainLM [6], BrainJEPA [7], and Brain Harmony [8]. We further evaluate matched-access autoregressive baselines with exogenous inputs (ARX): Ridge-ARX, LSTM-ARX, TemporalCNN-ARX, Transformer-ARX, and BrainHarmony-AR. Here, ARX denotes models that predict future fMRI states from both endogenous

Table 3: Input-source ablation under $H{=}20$ forecasting. Each entry reports $p_{\mathrm{corr}}$. The stimulus-only variant removes brain-history tokens and predicts future brain states from observed stimulus tokens alone.

| Variant | Algonauts 2025 | CineBrain | HAD |
|---|---|---|---|
| Stimulus only | 0.1391 | 0.3136 | 0.1908 |
| Past fMRI only | 0.1764 | 0.4872 | 0.2761 |
| Past fMRI + stimulus | **0.2090** | **0.5643** | **0.2950** |

brain history and exogenous observed stimulus tokens, while differing in the forecasting backbone. BrainHarmony-AR denotes an autoregressive adaptation of Brain Harmony, where its brain representations are used under the same causal forecasting protocol. All ARX baselines use past fMRI and observed stimuli through the queried bin, with target and future ground-truth fMRI withheld. Transformer-ARX additionally shares BrainVista's frozen stimulus encoders, pretrained Network-wise Tokenizer–detokenizer pairs, four-layer/256-hidden temporal backbone size, splits, context, optimization protocol, and tuning budget (Appendix B.5). Non-ARX references have different input access: TRIBE is stimulus-only, whereas BrainLM and BrainJEPA use past fMRI (Table 26).

**Training and Metrics.** Unless otherwise specified, models are trained for 40 epochs with AdamW, mixed precision, learning rate $\eta = 2 \times 10^{-4}$, weight decay $10^{-2}$, and batch size 32. The default context length and rollout horizon are $L = 40$ and $H = 20$, respectively. At test time, models are initialized with observed fMRI context and rolled out autoregressively while the stimulus stream remains observed up to each queried prediction step. We report vectorized Pearson correlation $r$, pattern correlation $p_{\mathrm{corr}}$, and Top-10 Identification Accuracy (Cls@Top10). Additional protocol and baseline details are provided in Appendices A and B. Model size, memory usage, inference time, and additional compute details are reported in Appendix J.

### 4.2 Main Results

All methods are evaluated under the same conditional autoregressive forecasting protocol when applicable, and all results are averaged over subjects. For models without a native forecasting interface, we train a forecasting head under the same input-access constraints, with details in Appendix B.

Table 1 shows that BrainVista performs best on all three datasets under $H{=}20$ forecasting. Non-ARX references provide useful references, while ARX baselines give stronger matched-access comparisons because they use past fMRI states and aligned stimulus features. BrainHarmony-AR is the strongest non-BrainVista baseline. BrainVista also exceeds the representation-matched Transformer-ARX by $0.0271$, $0.0665$, and $0.0226$ in $p_{\mathrm{corr}}$ on Algonauts, CineBrain, and HAD, respectively. Thus, identical stimulus encoders and tokenizers do not eliminate its advantage. The improvements are clearest in $p_{\mathrm{corr}}$, suggesting that BrainVista better preserves instantaneous cross-region brain-state structure rather than only improving global temporal similarity. Across the three datasets, the standard deviation of $p_{\mathrm{corr}}$ over five random seeds ranges from $0.0022$ to $0.0042$ for BrainVista and from $0.0026$ to $0.0048$ for BrainHarmony-AR (Table 18). Separately, paired subject-bootstrap 95% confidence intervals for their mean $p_{\mathrm{corr}}$ difference are $[0.0172, 0.0234]$, $[0.0449, 0.0523]$, and $[0.0171, 0.0191]$ on Algonauts, CineBrain, and HAD, with gains for $4/4$, $6/6$, and $30/30$ subjects, respectively (Table 31; Appendix C). Algonauts has only four subjects: despite its positive bootstrap interval, the one-sided exact sign test has $p = 0.0625$, limiting population-level inference.

Table 2 reports HAD results across forecasting horizons for matched-access autoregressive methods. All methods degrade as $H$ increases, reflecting accumulated rollout error under recursive prediction. BrainVista remains consistently strongest, with a more pronounced margin in $p_{\mathrm{corr}}$ than in vectorized $r$, suggesting that its network-wise tokenization and cross-network refinement better preserve spatial brain-state structure over long horizons. The advantage persists across the evaluated horizons; component ablations below examine how individual effects change with rollout length. Non-ARX stimulus-to-brain references are summarized in Appendix B.4, while full horizon results, subject-level stability, reliability diagnostics, and statistical analyses are provided in Appendix C.

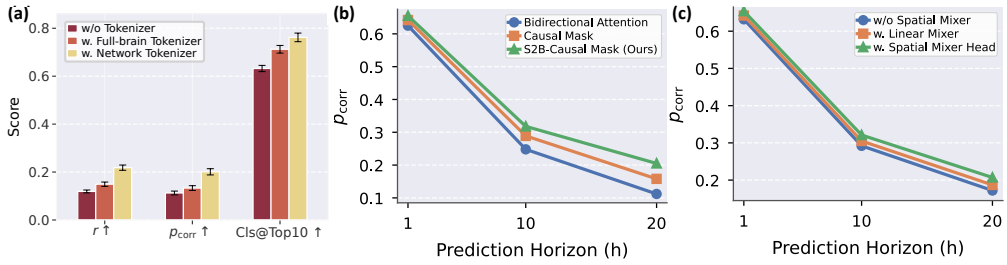


Figure 3: **Ablations on Algonauts 2025.** (a) Tokenization granularity: continuous fMRI, a shared full-brain tokenizer, and Network-wise Tokenizers. (b) Temporal-causality masking: bidirectional, standard causal, and S2B-causal attention. (c) Spatial Mixer Head: no mixer, linear mixer, and the proposed Spatial Mixer Head.

### 4.3 Ablation Studies

**Input-source Ablation.** Table 3 separates the effects of exogenous stimulus context and endogenous brain-state history. The stimulus-only variant corresponds to conventional stimulus-to-brain prediction, while the past-fMRI-only variant tests whether forecasting is dominated by temporal continuity. Combining past fMRI and aligned stimulus context yields the best $p_{\text{corr}}$ on all datasets, supporting our formulation of naturalistic fMRI as an evolving state process driven by both internal brain history and external sensory input.

**Design Ablations.** Fig. 3 evaluates the main design choices of BrainVista. Network-wise tokenization improves over continuous fMRI input and a shared full-brain tokenizer, indicating the value of preserving functional heterogeneity. S2B improves over vanilla causal masking by $0.008$, $0.023$, $0.040$, $0.056$, and $0.067$ at $H = 1, 5, 10, 15, 20$, respectively (Table 42). Both causal variants restrict future access; S2B additionally blocks brain-to-stimulus attention. The increasing margin is consistent with reduced propagation of prediction errors through stimulus states, but does not directly establish that mechanism. The Spatial Mixer Head further improves over no-mixer and linear-mixer variants, especially at longer horizons, indicating that explicit cross-network refinement helps reduce recursive error accumulation. Additional robustness, mechanism, multimodal, and cross-subject analyses are provided in Appendices D–F.

**Joint Mask–Mixer Control.** Table 4 jointly varies the temporal mask and attention mixer. At $H = 20$, adding the mixer changes Algonauts $p_{\text{corr}}$ by $-0.0153$ under vanilla causal masking and $+0.0278$ under S2B, giving an interaction contrast of $+0.0431$. This sign change shows that the benefit of cross-network refinement depends on the temporal mask in this setting (Appendix E.7).

Table 4: **Joint mask–mixer control on Algonauts 2025.** Entries are mean $p_{\text{corr}}$ at $H = 20$.

| Temporal mask | No mixer | Attention mixer |
|---|---|---|
| Vanilla causal | 0.1573 | 0.1420 |
| S2B | 0.1812 | **0.2090** |

**Tokenizer Objectives.** With network partition, latent dimensionality, stimulus encoders, predictor, and rollout protocol fixed, reconstruction-only tokenization reaches $H = 20$ $p_{\text{corr}} = 0.2090$ on Algonauts, compared with $0.1938$ for VQ-AE, $0.2031$ for $\beta$-VAE, and $0.2044$ for SoftCLIP. SoftCLIP is slightly better at $H = 1$ ($0.5613$ versus $0.5600$), but this advantage does not persist at longer horizons. These results support reconstruction-only training among the tested configurations (Appendix E.8).

**Stimulus Cutoff and Scaling.** Restricting stimuli to bins strictly before the queried target gives $p_{\text{corr}} = 0.2021$, $0.5520$, and $0.2904$ on Algonauts, CineBrain, and HAD (Appendix A.3); these scores still exceed Transformer-ARX under its default access. Increasing training data from 25% to 100% improves $H = 20$ from $0.1698$ to $0.2090$. Increasing model size from Base to Large gives only $+0.0028$ while compute rises from 2.1 to 4.9 GPU-hours per subject (Appendix I). These results show positive data scaling and diminishing capacity returns, not a general scaling law.

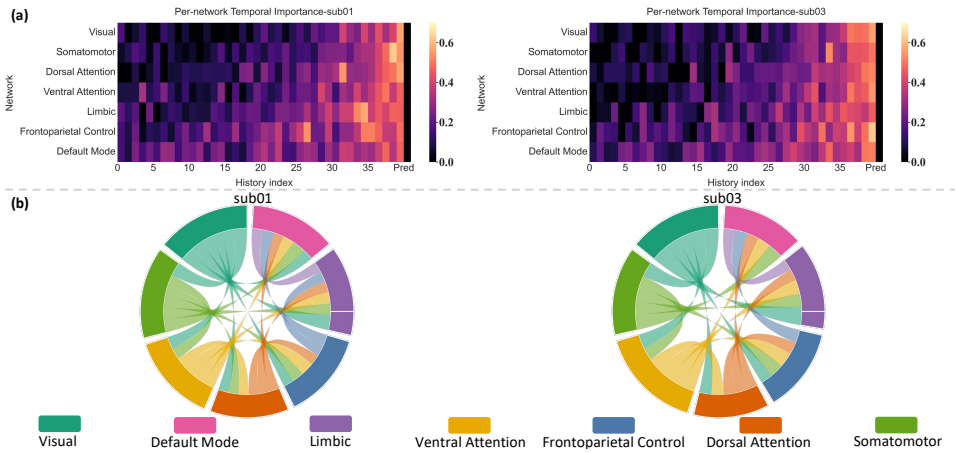


Figure 4: **Model-based analyses on Algonauts 2025 aggregated to Yeo-7.** Shown for sub01 and sub03. (a) History-token occlusion measures the increase in $L_2$ error after masking each history token. (b) Spatial Mixer attention is aggregated to Yeo-7 network pairs and averaged across time, with chord thickness indicating attention weight.

### 4.4 Visualization Analysis

**Temporal Reliance.** Occluding a history token (Fig. 4a) while retaining all other history and observed stimuli measures its incremental contribution to held-out prediction. Sensory networks rely more strongly on recent context, whereas association networks retain broader sensitivity to earlier history. This pattern is consistent with hierarchical intrinsic timescales in cortical processing [43, 44, 46, 45], and serves as a positive-control validation against established findings. It measures conditional predictive reliance rather than raw autocorrelation, but does not identify a new cortical law or biological mechanism.

**Cross-network Interaction.** Spatial Mixer attention exhibits non-uniform coupling across Yeo-7 systems rather than uniform global mixing. The Spatial Mixer Head thus selectively refines predicted brain tokens through structured network interactions. These attention patterns describe the fitted model and do not establish biological causality. Additional visualizations are provided in Appendix G.

**Auxiliary Representational Utility.** On held-out Algonauts clips, emotion-category clustering of forecasted brain features improves over BrainHarmony-AR in ARI ($0.138 \rightarrow 0.214$), NMI ($0.214 \rightarrow 0.286$), and mean class purity ($0.39 \rightarrow 0.49$). Ground-truth brain features remain better (Appendix H). This tests retention of structure associated with stimulus-derived emotion labels, without assuming near-perfect simulation or validating clinical use.

## 5 Conclusion and Discussion

BrainVista models naturalistic fMRI as observed-stimulus conditional autoregressive forecasting. Its value is a within-run predictive assay for comparing how brain history, stimulus representations, and network organization support held-out recursive prediction. Under matched conditions, Network-wise Tokenizers outperform shared tokenization, S2B improves long-horizon forecasting over vanilla causal masking, and the Spatial Mixer helps under S2B. The positive mask–mixer interaction supports their combined use in the evaluated setting. Forecasted features also retain emotion-category structure in an auxiliary representational test.

Forecasting requires observed initial fMRI and stimuli available through each queried bin, and accuracy decreases with horizon. Cross-subject transfer remains weaker than within-subject prediction. Models are trained separately on each dataset, so the results do not establish zero-shot scanner transfer, arbitrary unseen-movie generalization, or accurate cross-session simulation. History occlusion and attention measure model behavior, not biological causality, connectivity, or a new neural mechanism. The emotion analysis uses stimulus-derived labels and does not validate clinical or individual-level prediction. Future work should test uncertainty calibration, additional data, and broader acquisition settings; further assumptions are discussed in Appendix K.

## Acknowledgments and Disclosure of Funding

This research received no specific external funding. The authors declare no competing interests.

## References

[1] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[2] John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982.

[3] Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001.

[4] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.

[5] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[6] Josue Ortega Caro, Antonio H. de O. Fonseca, Syed A. Rizvi, Matteo Rosati, et al. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

[7] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[8] Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, B. T. Thomas Yeo, and Juan Helen Zhou. Brain Harmony: A multimodal foundation model unifying morphology and function into 1D tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[9] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. TIMER-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

[10] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

[11] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

[12] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant R. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[14] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, et al. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15: 6241, 2024.

[15] Claire H. C. Chang, Samuel A. Nastase, and Uri Hasson. Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences*, 119(51):e2209307119, 2022.

[16] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[17] Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11): 1664–1671, 2015.

[18] Caterina Gratton, Timothy O. Laumann, Ashley N. Nielsen, Deanna J. Greene, Evan M. Gordon, Adrian W. Gilmore, Steven M. Nelson, Rebecca S. Coalson, Abraham Z. Snyder, Bradley L. Schlaggar, Nico U. F. Dosenbach, and Steven E. Petersen. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron*, 98 (2):439–452, 2018.

[19] Murat Demirtaş, Joshua B. Burt, Markus Helmer, Jie Lisa Ji, Brendan D. Adkinson, Matthew F. Glasser, David C. Van Essen, Stamatios N. Sotiropoulos, Alan Anticevic, and John D. Murray. Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron*, 101(6):1181–1194, 2019.

[20] Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640: 470–477, April 2025.

[21] Mohammad Javad Darvishi Bayazi, Hena Ghonia, Roland Riachi, Bruno Aristimunha, Arian Khorasani, Md Rifat Arefin, Amin Darabi, Guillaume Dumas, and Irina Rish. General-purpose brain foundation models for time-series neuroimaging data. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

[22] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The Algonauts Project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2025.

[23] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. CineBrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025.

[24] Ming Zhou, Zhengxin Gong, Yuxuan Dai, Yushan Wen, Youyi Liu, Zonglei Zhen, et al. A large-scale fMRI dataset for human action recognition. *Scientific Data*, 10(1):415, 2023.

[25] Yuto Nishimura, Masataka Sawayama, Ayumu Yamashita, Hideki Nakayama, and Kaoru Amano. BrainCodec: Neural fMRI codec for the decoding of cognitive brain states. *arXiv preprint arXiv:2410.04383*, 2024.

[26] Yanwu Yang and Thomas Wolfers. Hierarchical characterization of brain dynamics via state space-based vector quantization. *arXiv preprint arXiv:2506.22952*, 2025.

[27] Yunxi Zhao, Dong Nie, Geng Chen, Xia Wu, Daoqiang Zhang, and Xuyun Wen. TARDRL: Task-aware reconstruction for dynamic representation learning of fMRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

[28] Weijian Mai, Jiamin Wu, Yu Zhu, Zhouheng Yao, Dongzhan Zhou, Andrew F. Luo, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. SynBrain: Enhancing visual-to-fMRI synthesis via probabilistic representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[29] Yikang Li, Lyuan Xu, Lianrui Zuo, Yukie Chang, Zhaohua Ding, Adam W. Anderson, Kurt G. Schilling, John C. Gore, and Yurui Gao. BrainVAE: Exploring the role of white matter BOLD in preclinical Alzheimer's disease classification. *bioRxiv*, 2025.

[30] Youzhi Qu, Junfeng Xia, Xinyao Jian, Wendu Li, Kaining Peng, Zhichao Liang, Haiyan Wu, and Quanying Liu. Uncovering cognitive taskonomy through transfer learning in masked autoencoder-based fMRI reconstruction. In *Human Brain and Artificial Intelligence (HBAI 2024)*, volume 2438 of *Communications in Computer and Information Science*, pages 35–50. Springer, 2025. doi: 10.1007/978-981-96-4001-0_3.

[31] Weiran Xia, Xin Zhang, Dan Hu, Jiale Cheng, Xiaowei Yu, Weiyan Yin, Zhengwang Wu, Li Wang, Weili Lin, and Gang Li. Triplet longitudinal masked autoencoder for predicting individualized functional connectome development during infancy. *Medical Image Analysis*, 108:103860, 2026. doi: 10.1016/j.media.2025.103860.

[32] Mo Wang, Junfeng Xia, Wenhao Ye, Enyu Liu, Kaining Peng, Jianfeng Feng, Quanying Liu, and Hongkai Wen. SLIM-Brain: A data- and training-efficient foundation model for fMRI data analysis. *arXiv preprint arXiv:2512.21881*, 2025.

[33] Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. MindSimulator: Exploring brain concept localization via synthetic fMRI. In *International Conference on Learning Representations (ICLR)*, 2025.

[34] Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning for any atlas and disorder. *arXiv preprint arXiv:2506.02044*, 2025.

[35] Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. TRIBE: TRImodal brain encoder for whole-brain fMRI response prediction. In *International Conference on Learning Representations (ICLR)*, 2026.

[36] Xuanhua Yin, Runkai Zhao, and Weidong Cai. Improving multimodal brain encoding model with dynamic subject-awareness routing. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[37] Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *ICLR Workshop Re-Align*, 2024.

[38] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. UMBRAE: Unified multimodal brain decoding. In *European Conference on Computer Vision (ECCV)*, 2024.

[39] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[40] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[41] Zehui Feng, Chenqi Zhang, Mingru Wang, Minuo Wei, Shiwei Cheng, Cuntai Guan, and Ting Han. Unveiling deep semantic uncertainty perception for language-anchored multi-modal vision-brain alignment. *arXiv preprint arXiv:2511.04078*, 2025.

[42] François Paugam, Basile Pinsard, Guillaume Lajoie, and Pierre Bellec. A benchmark of individual auto-regressive models in a massive fMRI dataset. *Imaging Neuroscience*, 2:1–23, 2024.

[43] Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, 2008.

[44] Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31 (8):2906–2915, 2011.

[45] Christopher J. Honey, Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012.

[46] John D. Murray, Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12):1661–1663, 2014. doi: 10.1038/nn.3862.

[47] B. T. Thomas Yeo et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.

[48] Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[49] Alexander Schaefer et al. Local-Global Parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

[50] Mido Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[51] Seamless Communication, Loïc Barrault, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[52] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[53] Meta. Llama 3.2 model card, 2024. Released September 25, 2024.

[54] Oscar Esteban et al. fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019.

# A Evaluation Protocol and Reproducibility Details

This appendix specifies the experimental protocol used for all main results, including dataset splits, temporal alignment, conditional autoregressive rollout, training hyperparameters, evaluation metrics, baseline adaptation, dimensionality choices, and computational cost. We study conditional autoregressive forecasting: after rollout initialization, future ground-truth fMRI states are unavailable, while stimulus tokens are provided according to the same temporally causal interleaved context used by the main model.

## A.1 Dataset Metadata, Splits, and Forecasting Windows

Table 6 summarizes the datasets, raw brain-space representations, and model-space representations used by BrainVista. The three datasets differ in temporal resolution, stimulus paradigm, preprocessing pipeline, and brain-space definition. For each dataset, we follow the official train/validation/test distribution at the run or episode level. Validation runs are used only for model selection, and all final numbers are reported on held-out test runs. Forecasting windows are constructed within individual runs or episodes, and we never carry context tokens, hidden states, predicted states, or normalization statistics across boundaries.

For Algonauts 2025, we use the four challenge subjects: sub-01, sub-02, sub-03, and sub-05. Each fMRI frame is represented by 1,000 Schaefer parcels, which are mapped to Yeo-7 networks for network-wise tokenization and reporting. For CineBrain, we use the dataset-defined visual and auditory ROI groups. Because CineBrain provides high-dimensional voxel-level responses, we reduce each sensory ROI group using PCA fitted only on training runs before tokenization. For HAD, we use the released surface-based fMRI responses to human-action video clips and aggregate them to Schaefer-1000 parcels before Yeo-7 network grouping. For start positions with fewer than $L$ previous fMRI steps, the window is discarded rather than padded, ensuring a fixed context length for all evaluated samples.

## A.2 Temporal Alignment, Modality Availability, and Leakage Control

Stimulus streams are sampled at higher temporal resolution than fMRI and have modality-specific timestamps. For each dataset $\mathcal{D}$, we define an active modality set:

$$\mathcal{M}_{\mathcal{D}} \subseteq \{\text{vis}, \text{aud}, \text{text}\}. \tag{17}$$

Only modalities in $\mathcal{M}_{\mathcal{D}}$ are used to construct stimulus tokens for that dataset. If a modality is unavailable for an entire dataset, its modality ratio is set to zero and it is excluded from the concatenated stimulus representation.

For each active modality $m \in \mathcal{M}_{\mathcal{D}}$, let $\tau'_{m,i} \in \mathbb{R}^{d_m}$ denote the $i$-th high-rate stimulus token. For fMRI time step $t$, let:

$$\mathcal{B}_{m,t} = \{i : \tau'_{m,i} \text{ is assigned to fMRI bin } t\} \tag{18}$$

denote the modality-specific temporal bin. The fMRI-aligned modality token is computed by mean pooling:

$$\tilde{\tau}^t_m = \frac{1}{|\mathcal{B}_{m,t}|} \sum_{i \in \mathcal{B}_{m,t}} \tau'_{m,i}, \qquad m \in \mathcal{M}_{\mathcal{D}}. \tag{19}$$

The active modality tokens are concatenated in a fixed order and projected into the shared Transformer token space:

$$\tau^t_s = \text{Proj}_{\mathcal{M}_{\mathcal{D}}} \left( \underset{m \in \mathcal{M}_{\mathcal{D}}}{\text{Concat}} \tilde{\tau}^t_m \right) \in \mathbb{R}^{d_{\text{model}}}, \tag{20}$$

where:

$$\text{Proj}_{\mathcal{M}_{\mathcal{D}}} : \mathbb{R}^{\sum_{m \in \mathcal{M}_{\mathcal{D}}} d_m} \to \mathbb{R}^{d_{\text{model}}}. \tag{21}$$

Thus, dataset-level missing modalities are handled by reducing the projection input dimensionality rather than padding unavailable modalities. For example, when audio is unavailable, the audio ratio is set to zero and the projection operates on the remaining active modality dimensions.

For next-step prediction at queried time $u$, the valid conditioning set is:

$$\mathcal{C}_u = \left(x^{<u}, s^{\leq u}\right) = \left((x^1, \ldots, x^{u-1}), (s^1, \ldots, s^u)\right). \tag{22}$$

Table 5: **Stimulus modality availability.** Dataset-level unavailable modalities are excluded from the concatenation by setting their modality ratio to zero. The active input dimension is reported after modality-specific feature compression.

| Dataset | Visual | Audio | Text | Active input dim. | Projection output |
|---|---|---|---|---|---|
| Algonauts 2025 | Yes | Yes | Yes | $512 + 256 + 512 = 1280$ | $d_{\text{model}} = 256$ |
| CineBrain | Yes | Yes | Yes | $512 + 256 + 512 = 1280$ | $d_{\text{model}} = 256$ |
| HAD | Yes | No | Yes | $512 + 512 = 1024$ | $d_{\text{model}} = 256$ |

Table 6: **Dataset metadata and model-space representation.** Raw brain-space dimensionality follows the corresponding dataset release, while model-space dimensionality denotes the representation used by BrainVista after dataset-specific spatial preprocessing.

| Dataset | #Subjects | Native TR | Raw brain representation | Model-space representation | Split and rollout protocol |
|---|---|---|---|---|---|
| Algonauts 2025 | 4 | 1.49s | 1,000 Schaefer parcels in MNI space | 1,000 parcels mapped to Yeo-7 systems; $d_{\text{brain}} = 803$ | Official train/validation/test split; rollout within run; context reset at run boundaries. |
| CineBrain | 6 | 0.8s | HCP-MMP ROI voxels: 8,405 visual voxels and 10,541 auditory voxels | Training-set PCA to 512 visual + 512 auditory components; $d_{\text{brain}} = 820$ | Official train/validation/test split; audiovisual stimulus and fMRI aligned within run or episode. |
| HAD | 30 | 2.0s | Surface-based fMRI responses from participants viewing human-action clips | fsLR-to-Schaefer-1000 aggregation mapped to Yeo-7 systems; $d_{\text{brain}} = 803$ | Official train/validation/test split; rollout within run; context reset at run boundaries. |

The target-aligned stimulus token $s^u$ is available because it corresponds to the queried fMRI frame. However, the target brain state $x^u$, future brain states $x^{>u}$, and future stimulus tokens $s^{>u}$ are never provided as model inputs. During rollout, this rule is applied recursively: at step $u = T + h$, BrainVista may condition on previously predicted brain states $\hat{x}^{T+1:u-1}$ and observed stimulus tokens $s^{\leq u}$, but never on ground-truth future fMRI states.

The Stimulus-to-Brain mask implements this input-access rule in the interleaved sequence. Stimulus tokens remain in a stimulus-only causal stream, while brain-token queries may attend only to causally available stimulus tokens and available past or predicted brain tokens. This prevents future fMRI leakage and stimulus look-ahead beyond the queried prediction time.

### A.3 Temporal Validity and HRF Diagnostics

The main forecasting protocol treats the stimulus stream as an observed exogenous input, but all stimulus features must respect the causal cutoff used to construct the interleaved context. For a queried target time $u$, BrainVista uses brain states strictly before $u$ and stimulus features available by the end of bin $u$. During rollout, earlier predicted states replace unavailable ground truth; target and future ground-truth fMRI are never accessed. Because BOLD responses are delayed and temporally integrated, we further examine whether different stimulus cutoffs or HRF-aware stimulus constructions affect forecasting performance. We conduct four diagnostics: timestamp-boundary verification, stimulus lag sweep, HRF/FIR feature construction, and shifted-stimulus negative controls.

**Timestamp Convention and Feature Windows.** For all datasets, we define the stimulus context boundary as the end time of the queried fMRI acquisition bin. Stimulus features assigned to token $\tau_s^u$ must have feature-window end time no later than this boundary. Thus, stimulus conditioning uses content observed up to the queried bin boundary, not stimulus content after that boundary. For datasets whose metadata use center timestamps, we convert them to the same end-time convention by shifting the bin boundary by half a TR.

**Stimulus Lag Sweep.** We evaluate whether performance depends on the exact stimulus cutoff. For a cutoff $\Delta$ seconds relative to the queried fMRI bin boundary $t_u$, $\Delta = 0$ corresponds to the default causal context setting, $\Delta < 0$ restricts the model to earlier stimulus context, and $\Delta > 0$ is a diagnostic future-lookahead condition that is not used in the main results. Table 8 reports $p_{\text{corr}}$ at $H = 20$.

Table 7: **Timestamp convention and stimulus feature extraction.** All feature windows are end-aligned: a feature is assigned to $\tau_s^u$ only if its window end time is no later than the queried fMRI bin boundary $t_u$.

| Dataset | TR | fMRI boundary | Visual feature window | Audio feature window | Text feature source and boundary |
|---|---|---|---|---|---|
| Algonauts 2025 | 1.49s | Bin end | V-JEPA 2 clips sampled over 2.0s windows; window end $\leq t_u$ | Wav2Vec2-BERT windows of 1.0s with 0.5s stride; window end $\leq t_u$ | Dataset-provided subtitles or aligned transcripts; only text tokens ending before $t_u$ are used. |
| CineBrain | 0.8s | Bin end | V-JEPA 2 clips sampled over 2.0s windows; window end $\leq t_u$ | Wav2Vec2-BERT windows of 1.0s with 0.4s stride; window end $\leq t_u$ | Aligned movie transcript/subtitles; sentence embeddings use causal prefixes only. |
| HAD | 2.0s | Bin end | V-JEPA 2 clips sampled over 2.0s windows; window end $\leq t_u$ | Not used | Clip-level action descriptions and available annotations are aligned causally to clip boundaries. |

Table 8: **Stimulus lag sweep at $H = 20$.** Negative cutoffs restrict stimulus access to earlier content; $\Delta = 0$ is the default causal context setting; $\Delta = +2$s is a diagnostic lookahead condition and is not used in the main protocol.

| Dataset | $\Delta = -8$s | $\Delta = -6$s | $\Delta = -4$s | $\Delta = -2$s | $\Delta = 0$s | $\Delta = +2$s |
|---|---|---|---|---|---|---|
| Algonauts 2025 | 0.185 | 0.193 | 0.199 | 0.205 | **0.209** | 0.213 |
| CineBrain | 0.523 | 0.536 | 0.548 | 0.558 | **0.564** | 0.568 |
| HAD | 0.275 | 0.281 | 0.288 | 0.292 | **0.295** | 0.296 |

The lag sweep shows that performance improves as stimulus context approaches the queried fMRI bin boundary, consistent with the stimulus stream providing useful exogenous information. However, diagnostic future lookahead yields only marginal additional improvement, suggesting that the main gain is not driven by access to post-boundary stimulus content.

**HRF-convolved and FIR Stimulus Features.** We compare simple bin-wise average pooling with HRF-convolved stimulus features and finite impulse response (FIR) lag features. For the HRF setting, stimulus features are convolved with a canonical hemodynamic response kernel before being aggregated to the fMRI grid. For the FIR setting, we concatenate stimulus features from a causal lag bank ending at the queried bin boundary. No FIR feature uses content after the stimulus context boundary.

The HRF-convolved and FIR variants produce comparable performance to the default bin-wise pooling. This indicates that the main conclusions are not specific to a single alignment heuristic. We use bin-wise pooling in the main experiments because it is simple, model-agnostic, and avoids introducing additional HRF-shape assumptions that may differ across datasets and cortical systems.

**Target-boundary Stimulus Ablation.** We further test a stricter setting in which the stimulus token at the queried bin boundary is removed. The model may use only earlier stimulus tokens before the queried bin boundary. This ablation distinguishes the default causal stimulus context from a stricter history-only stimulus setting.

Removing the stimulus token at the queried bin boundary decreases performance but does not eliminate the gain over the past-fMRI-only baseline. Thus, BrainVista benefits from both endogenous brain history and temporally aligned exogenous stimulus context. The default setting should be interpreted as conditional autoregressive brain-state forecasting under causal stimulus context, not unconditional future brain simulation.

**Shifted-stimulus Negative Controls.** To test whether the stimulus stream provides meaningful time-aligned information rather than segment identity shortcuts, we shift the stimulus stream within each run while preserving its marginal distribution. We use circular shifts with offsets of at least 30 seconds and also evaluate cross-run stimulus replacement within the same dataset split. Both controls preserve stimulus statistics but break temporal correspondence between stimulus and fMRI.

Table 9: **Stimulus construction ablation at $H = 20$.** All settings are temporally causal with respect to the queried fMRI bin boundary.

| Stimulus construction | Algonauts 2025 | CineBrain | HAD |
|---|---|---|---|
| Bin-wise average pooling | 0.2090 | 0.5643 | 0.2950 |
| Canonical HRF-convolved features | 0.2075 | 0.5618 | 0.2938 |
| Causal FIR lag bank | **0.2112** | **0.5671** | **0.2965** |

Table 10: **Target-boundary stimulus ablation at $H = 20$.** The earlier-stimulus setting allows only $s^{<u}$, whereas the default causal context setting allows $s^{\leq u}$ for predicting $x^u$. Here $u$ is the queried target index, not the last observed brain-state index.

| Stimulus access | Algonauts 2025 | CineBrain | HAD |
|---|---|---|---|
| Past fMRI only | 0.1764 | 0.4872 | 0.2761 |
| Past fMRI + stimulus $s^{<u}$ | 0.2021 | 0.5520 | 0.2904 |
| Past fMRI + stimulus $s^{\leq u}$ | **0.2090** | **0.5643** | **0.2950** |
| Diagnostic stimulus lookahead $s^{\leq u+1}$ | 0.2131 | 0.5682 | 0.2961 |

The shifted-stimulus controls collapse close to the past-fMRI-only baseline. This supports that the improvement from stimulus conditioning depends on correct temporal alignment rather than retrieval of stimulus identity or segment-level shortcuts.

### A.4 Dataset Heterogeneity and Training-set Normalization

We do not resample all datasets to a common absolute sampling rate. Instead, stimulus features are aligned to each dataset's native fMRI grid, and forecasting windows are defined on that grid. Consequently, one rollout step corresponds to different physical durations across datasets. For this reason, absolute scores should be interpreted within each dataset; cross-dataset comparisons are used to assess whether relative trends are robust across temporal resolutions and brain-space definitions.

Unless otherwise stated, all experiments use training-set normalization. For each subject, dataset, network or ROI group $k$, and model-space dimension $j$, normalization statistics are estimated only from training runs:

$$\mu_{k,j} = \frac{1}{|\mathcal{I}_{\text{train}}|} \sum_{(r,t)\in\mathcal{I}_{\text{train}}} x_{r,j}^{k,t}, \qquad \sigma_{k,j} = \sqrt{\frac{1}{|\mathcal{I}_{\text{train}}|} \sum_{(r,t)\in\mathcal{I}_{\text{train}}} \left(x_{r,j}^{k,t} - \mu_{k,j}\right)^2}. \tag{23}$$

The same fixed statistics are applied to training, validation, and test samples:

$$\bar{x}_{r,j}^{k,t} = \frac{x_{r,j}^{k,t} - \mu_{k,j}}{\sigma_{k,j} + \epsilon}. \tag{24}$$

No validation or test fMRI samples are used to estimate normalization statistics. Run-wise normalization is evaluated only as an offline diagnostic ablation and is not used as the default protocol.

### A.5 Conditional Autoregressive Rollout

Given an observed context window ending at time $T$, the model receives ground-truth fMRI tokens up to $T$ and stimulus tokens up to the queried prediction step. It predicts the next fMRI token, feeds the predicted token back into the context, and repeats this process for subsequent steps. The stimulus stream remains observed, but future ground-truth fMRI states are withheld after initialization.

For an $H$-step rollout, prediction is defined as:

$$\hat{x}^{T+h} = f_\theta\left([x^{\leq T}; \hat{x}^{T+1:T+h-1}], s^{\leq T+h}\right), \qquad h = 1, \ldots, H. \tag{25}$$

When $h = 1$, the predicted-token history $\hat{x}^{T+1:T+h-1}$ is empty. For $h > 1$, the model recursively consumes its own predicted brain states. This protocol evaluates long-horizon conditional forecasting under observed naturalistic stimuli without access to future fMRI targets.

Table 11: **Shifted-stimulus negative controls at $H = 20$.** The controls preserve stimulus statistics but disrupt stimulus–fMRI temporal correspondence.

| Setting | Algonauts 2025 | CineBrain | HAD |
|---|---|---|---|
| Past fMRI only | 0.1764 | 0.4872 | 0.2761 |
| Aligned stimulus | **0.2090** | **0.5643** | **0.2950** |
| Circular-shifted stimulus | 0.1788 | 0.4915 | 0.2774 |
| Cross-run replaced stimulus | 0.1775 | 0.4894 | 0.2769 |

**Training–evaluation distinction.** During training, BrainVista uses teacher-forced one-step prediction. For transition $t \to t{+}1$, the input contains brain tokens up to $t$ and stimulus tokens up to $t{+}1$; the target brain token at $t{+}1$ is used only as supervision and is never inserted into the input sequence. During evaluation, after initialization with the observed context, predicted brain tokens are recursively fed back for future predictions. This distinction keeps optimization stable while measuring error accumulation under autoregressive rollout.

### A.6 Dataset Statistics and Effective Forecasting Windows

Table 12 reports the processed data statistics used to construct forecasting anchors. All anchors are generated independently within each run or episode before aggregation across subjects. For a run $r$, let:

$$T_r = \mathrm{safeT}(V_r, A_r, W_r, Z_r) \tag{26}$$

denote the aligned valid length after truncating the pooled visual, audio, text, and brain-token streams to their common duration. Let $m = \max(L, L_b)$, where $L$ is the stimulus context length and $L_b$ is the brain-history context length. Unless otherwise specified, we use $L = 40$, $L_b = 40$, and rollout horizon $H = 20$.

For one-step teacher-forced training and validation, anchors are generated as:

$$\mathcal{A}_r^{\mathrm{1step}} = \{t \in \mathbb{Z} : m - 1 \le t \le T_r - 2\}. \tag{27}$$

Each anchor predicts the next brain token $z_{t+1}$ using brain history within $[t - L_b + 1, t + 1)$ and stimulus context up to the target-aligned stimulus bin. Specifically, the stimulus window includes $s_{t+1}$ and is indexed as $[t - L + 2, t + 2)$ under half-open interval notation. Thus, the number of one-step anchors in run $r$ is:

$$N_r^{\mathrm{1step}} = \max(0, T_r - m). \tag{28}$$

For fixed-horizon rollout evaluation, an anchor must additionally contain all $H$ future target brain states within the same run. The valid rollout-anchor set is:

$$\mathcal{A}_r^{H} = \{t \in \mathbb{Z} : m - 1 \le t \le T_r - H - 1\}, \tag{29}$$

with count:

$$N_r^{H} = \max(0, T_r - m - H + 1). \tag{30}$$

Therefore, boundary-crossing windows are not removed after concatenating runs; rather, they are never generated. The total number of valid anchors is computed as a run-wise sum:

$$N^{\mathrm{1step}} = \sum_r \max(0, T_r - m), \qquad N^{H} = \sum_r \max(0, T_r - m - H + 1). \tag{31}$$

Training and validation frame counts follow the same run-wise preprocessing protocol and are used only for model fitting and selection. The table focuses on test-frame and anchor counts because these determine the effective forecasting evaluation set. The number of valid anchors differs across datasets because of differences in TR, run duration, and split size. For every dataset, context states, predicted states, and rollout buffers are reset at run or episode boundaries. Thus, no forecasting anchor contains fMRI context, target states, or stimulus bins from a different run or stimulus segment.

### A.7 Per-subject Run and Window Summary

Table 13 summarizes the per-subject unit counts and valid forecasting anchors. For HAD, the split structure is identical across subjects after preprocessing, so the table reports the common per-subject configuration. As in Section A.6, all anchors are counted run-wise rather than after concatenating test runs.

Table 12: **Processed dataset statistics and valid run-wise forecasting anchors.** Frame counts are reported per subject after preprocessing. One-step anchors are computed as $\sum_r \max(0, T_r - m)$. Full-horizon rollout anchors are computed as $\sum_r \max(0, T_r - m - H + 1)$ with $L = L_b = 40$ and $H = 20$. All anchors are generated independently within each run or episode before aggregation across subjects.

| Dataset | #Subj. | Units/subj. | Test frames/subj. | 1-step anchors/subj. | $H$=20 anchors/subj. |
|---|---|---|---|---|---|
| Algonauts 2025 | 4 | 28 / 4 / 4 | $2{,}580 \pm 33$ | $2{,}420 \pm 33$ | $2{,}344 \pm 33$ |
| CineBrain | 6 | 18 / 3 / 3 | $8{,}100 \pm 67$ | $7{,}980 \pm 67$ | $7{,}923 \pm 67$ |
| HAD | 30 | 12 / 2 / 2 | $1{,}200 \pm 0$ | $1{,}120 \pm 0$ | $1{,}082 \pm 0$ |

*Note:* Units/subj. denotes train / validation / test units per subject. Standard deviations are computed across subjects.

Table 13: **Per-subject run/episode counts and valid forecasting anchors.** One-step anchors are computed as $\sum_r \max(0, T_r - m)$. Full-horizon rollout anchors are computed as $\sum_r \max(0, T_r - m - H + 1)$ with $L = L_b = 40$ and $H = 20$. All counts are obtained independently within each run or episode.

| Dataset | Subject | Train units | Val units | Test units | One-step anchors | $H$=20 anchors |
|---|---|---|---|---|---|---|
| Algonauts 2025 | sub-01 | 28 | 4 | 4 | 2,397 | 2,321 |
| | sub-02 | 28 | 4 | 4 | 2,459 | 2,383 |
| | sub-03 | 28 | 4 | 4 | 2,435 | 2,359 |
| | sub-05 | 28 | 4 | 4 | 2,389 | 2,313 |
| CineBrain | sub-01 | 18 | 3 | 3 | 7,951 | 7,894 |
| | sub-02 | 18 | 3 | 3 | 8,004 | 7,947 |
| | sub-03 | 18 | 3 | 3 | 7,883 | 7,826 |
| | sub-04 | 18 | 3 | 3 | 8,065 | 8,008 |
| | sub-05 | 18 | 3 | 3 | 8,033 | 7,976 |
| | sub-06 | 18 | 3 | 3 | 7,944 | 7,887 |
| HAD | 30 subjects | 12 | 2 | 2 | 1,120 per subject | 1,082 per subject |

### A.8 CineBrain PCA Retained Variance

CineBrain contains high-dimensional visual and auditory ROI voxel responses. To keep tokenizer and projection modules comparable across datasets, PCA is fitted separately within each ROI group using training runs only. The learned PCA basis is then fixed for validation and test runs. Table 14 reports retained variance for the selected dimensionality.

We retain 512 components per ROI group as a trade-off between variance preservation and model size. Increasing the retained dimensionality above 512 produces only marginal forecasting gains while increasing dataset-specific tokenizer and projection parameters.

### A.9 Stimulus Feature Extraction Details

Table 16 provides the stimulus feature extraction protocol. All features are assigned by feature-window end time, and no feature whose window ends after the queried fMRI boundary is used for prediction at that boundary.

For text features, we avoid using future sentence context. If a sentence crosses the queried fMRI boundary, only the prefix whose timestamp does not exceed the boundary is encoded. If only clip-level annotations are available, the annotation is assigned to the clip interval and becomes visible only after the corresponding clip start, with no access to future clip annotations.

### A.10 Cls@Top10 Candidate Pool Construction

Cls@Top10 evaluates whether a predicted fMRI segment can identify its matched ground-truth segment among distractors. For each predicted segment, the candidate pool contains one matched ground-truth segment and distractor segments sampled from the same subject and evaluation split. Distractors are non-overlapping with the query segment and are sampled from different runs or episodes when possible. The same candidate pools are used for all methods.

Table 14: **CineBrain PCA retained variance.** PCA is fitted on training runs only and then fixed for validation and test data.

| ROI group | Raw dimension $Q_k$ | Retained dimension $P_k$ | Explained variance | Latent dimension $d_k$ |
|---|---|---|---|---|
| Visual ROI | 8,405 | 512 | $93.6\% \pm 1.2\%$ | 410 |
| Auditory ROI | 10,541 | 512 | $92.1\% \pm 1.5\%$ | 410 |
| Total | 18,946 | 1,024 | $92.9\% \pm 1.3\%$ | 820 |

Table 15: **CineBrain PCA dimensionality sensitivity at $H = 20$.** Each row uses the same forecasting architecture after changing the retained PCA dimension per ROI group.

| Components per ROI | Model-space dim. | $d_{\text{brain}}$ | Explained variance | $p_{\text{corr}}$ |
|---|---|---|---|---|
| 256 | 512 | 410 | 87.1% | 0.5521 |
| 512 | 1,024 | 820 | 92.9% | **0.5643** |
| 768 | 1,536 | 1,230 | 95.2% | 0.5650 |

Using same-subject candidate pools prevents identification from being driven by subject-specific response scale. Sampling distractors from the same held-out split also reduces distribution mismatch between matched and distractor segments. Because shifted-stimulus controls in Table 11 collapse near the past-fMRI-only baseline, the identification results are unlikely to be explained by stimulus identity alone.

### A.11 Random Seed Variance

The main forecasting comparisons and Table 18 are averaged over five random seeds. The additional comparisons in Appendices E.7, E.8, and I report mean scores only; no across-run uncertainty estimate is claimed for those tables. The seeds affect model initialization, mini-batch order, dropout, and negative candidate sampling for identification. Table 18 reports seed-level mean and standard deviation for the strongest baseline and BrainVista at $H = 20$.

The seed-level variance is smaller than the improvement of BrainVista over the strongest baseline on all three datasets. This indicates that the reported gains are not driven by a favorable random initialization or candidate-pool sample.

### A.12 Training and Architecture Hyperparameters

Unless otherwise stated, all variants use context length $L = 40$ and rollout horizons $H \in \{1, 10, 20\}$. The temporal predictor is a Pre-LN Transformer with hidden size $d_{\text{model}} = 256$, 8 attention heads, and 4 layers. We use GELU activations, an FFN expansion ratio of $4\times$, absolute positional embeddings, dropout $0.1$, and gradient clipping at $0.5$. Models are trained for 40 epochs with AdamW, learning rate $2 \times 10^{-4}$, weight decay $0.01$, batch size 32, and PyTorch automatic mixed precision. Weight decay is not applied to LayerNorm parameters or bias terms.

**Network-wise tokenizer parameterization.** For each functional network $k$, BrainVista uses an independent tokenizer–detokenizer pair:

$$h^k_{\phi_k} : \mathbb{R}^{P_k} \to \mathbb{R}^{d_k}, \qquad g^k_{\psi_k} : \mathbb{R}^{d_k} \to \mathbb{R}^{P_k}. \tag{32}$$

The parameters are network-specific:

$$\Theta_{\text{tok}} = \{(\phi_k, \psi_k)\}_{k=1}^{K}. \tag{33}$$

No tokenizer or detokenizer parameters are shared across functional networks. This allows each network to learn a compact representation adapted to its own dimensionality and signal statistics.

The tokenizer–detokenizer pairs are pretrained by minimizing the network-wise reconstruction loss:

$$\mathcal{L}_{\text{tok}} = \sum_t \sum_{k=1}^{K} \left\| \bar{x}^{k,t} - g^k_{\psi_k}\left(h^k_{\phi_k}(\bar{x}^{k,t})\right) \right\|_2^2. \tag{34}$$

Table 16: **Stimulus feature extraction details.** Visual and audio features are end-aligned to the queried fMRI boundary. Text features use only tokens whose timestamps end before or at the boundary.

| Dataset | Video rate | Audio rate | Visual encoder | Audio encoder | Text source and context rule |
|---|---|---|---|---|---|
| Algonauts 2025 | 30 fps | 16 kHz | V-JEPA 2; 2.0s clips; 0.5s stride; end-aligned pooling | Wav2Vec2-BERT; 1.0s windows; 0.5s stride | Dataset-provided subtitles/transcripts; causal prefix ending at or before $t_u$. |
| CineBrain | 25 fps | 16 kHz | V-JEPA 2; 2.0s clips; 0.4s stride; end-aligned pooling | Wav2Vec2-BERT; 1.0s windows; 0.4s stride | Aligned transcript/subtitles; no sentence tokens after $t_u$ are used. |
| HAD | 30 fps | Not used | V-JEPA 2; 2.0s clips; 0.5s stride; end-aligned pooling | Not used | Clip-level action descriptions and annotations aligned to clip boundaries. |

Table 17: **Cls@Top10 candidate pool construction.** Chance level is $10/N$, where $N$ is the candidate pool size.

| Dataset | Pool size $N$ | Chance Top10 | Distractor source | Subject constraint |
|---|---|---|---|---|
| Algonauts 2025 | 200 | 5.0% | Same split, different run when available | Same subject |
| CineBrain | 400 | 2.5% | Same split, different episode when available | Same subject |
| HAD | 300 | 3.3% | Same split, different run or clip block | Same subject |

After tokenizer pretraining, the tokenizer parameters are fixed when training the autoregressive forecasting model. The forecasting model predicts latent brain tokens in the tokenizer space, and the frozen detokenizers map predicted latent tokens back to the normalized fMRI space.

**Brain-token Dimensionality.** For network-wise tokenization, we distinguish raw brain-space dimensionality $Q_k$ from tokenizer input dimensionality $P_k$. For parcellated inputs, $P_k = Q_k$. For high-dimensional voxel-level CineBrain inputs, each sensory ROI group is first reduced using PCA fitted on the training split only. The fitted PCA basis is then fixed and applied to validation and test runs. We set $d_k = \lceil 0.8 P_k \rceil$, and Table 21 lists the latent dimension of every group.

**CineBrain Voxel Reduction.** CineBrain provides high-dimensional visual and auditory ROI voxel responses. Directly tokenizing the raw voxel dimensions would produce substantially larger tokenizer and projection layers than the parcellated datasets. We therefore fit PCA separately within the visual and auditory ROI groups using only training runs, retaining 512 components per group. This yields $P_{\text{vis}} = 512$, $P_{\text{aud}} = 512$, and $d_{\text{brain}} = 820$. All validation and test responses are projected using the training-fitted PCA bases.

**Projection and Spatial Mixer Modules.** The concatenated brain latent $z_f^t \in \mathbb{R}^{d_{\text{brain}}}$ is mapped to the Transformer hidden space by $\text{Proj}_b^{\text{in}} : \mathbb{R}^{d_{\text{brain}}} \to \mathbb{R}^{256}$. The temporal Transformer output is mapped back by $\text{Proj}_b^{\text{out}} : \mathbb{R}^{256} \to \mathbb{R}^{d_{\text{brain}}}$ before Spatial Mixer refinement and detokenization. The Spatial Mixer operates over the $K$ functional-network components of a predicted whole-brain latent token. Given a preliminary latent prediction:

$$\bar{z}_f^u = [\bar{z}_f^{u,1}; \ldots; \bar{z}_f^{u,K}], \qquad \bar{z}_f^{u,k} \in \mathbb{R}^{d_k}, \tag{35}$$

each network component is first mapped to a shared mixer dimension using a network-specific projection:

$$R^u = \text{Stack}\left(W_{\text{in}}^1 \bar{z}_f^{u,1}, \ldots, W_{\text{in}}^K \bar{z}_f^{u,K}\right) \in \mathbb{R}^{K \times d_z}. \tag{36}$$

The Spatial Mixer first applies multi-head self-attention over the network-token stack:

$$\tilde{R}^u = \text{MHA}(R^u). \tag{37}$$

After self-attention, each mixed network token is mapped back to its original latent dimensionality using another network-specific projection $W_{\text{out}}^k$:

$$\hat{z}_f^{u,k} = W_{\text{out}}^k \tilde{R}_k^u, \quad \hat{z}_f^u = [\hat{z}_f^{u,1}; \ldots; \hat{z}_f^{u,K}] \in \mathbb{R}^{d_{\text{brain}}}. \tag{38}$$

Thus, the attention operation is applied over functional systems, but the input and output mappings remain network-specific. This prevents the mixer from treating the $K$ network tokens as exchangeable anonymous tokens and preserves the functional-network identity used by the tokenizers.

Table 18: **Random seed variance at $H = 20$.** Each entry reports mean $\pm$ standard deviation of $p_{\text{corr}}$ across five seeds.

| Dataset | Strongest baseline | Baseline $p_{\text{corr}}$ | BrainVista $p_{\text{corr}}$ |
|---|---|---|---|
| Algonauts 2025 | BrainHarmony-AR | $0.1887 \pm 0.0034$ | $\mathbf{0.2090 \pm 0.0031}$ |
| CineBrain | BrainHarmony-AR | $0.5156 \pm 0.0048$ | $\mathbf{0.5643 \pm 0.0042}$ |
| HAD | BrainHarmony-AR | $0.2769 \pm 0.0026$ | $\mathbf{0.2950 \pm 0.0022}$ |

Table 19: **BrainVista architecture configuration.** The same temporal forecasting backbone is used across datasets, while brain-token dimensionalities and active stimulus input dimensions are dataset-specific.

| Component | Setting |
|---|---|
| Brain grouping | Yeo-7 or dataset-defined ROI groups |
| Brain tokenizer | Independent two-layer MLP per network |
| Brain detokenizer | Independent two-layer MLP per network |
| Stimulus encoders | Frozen pretrained encoders |
| Stimulus projection output | $d_{\text{model}} = 256$ |
| Temporal backbone | Pre-LN causal Transformer |
| Transformer layers | 4 |
| Transformer attention heads | 8 |
| Feed-forward dimension | 1024 |
| Dropout | 0.1 |
| Context length | $L = 40$ |
| Brain-history length | $L_b = 40$ |
| Default rollout horizon | $H = 20$ |
| S2B mask | Directed stimulus-to-brain causal mask |
| Spatial Mixer | One-layer network-wise self-attention |
| Prediction target | Next whole-brain latent token |
| Optimizer | AdamW |
| Learning rate | $2 \times 10^{-4}$ |
| Weight decay | $10^{-2}$ |
| Batch size | 32 |
| Training epochs | 40 |

The $K \times K$ attention matrix from the Spatial Mixer is used only as a model-based diagnostic of learned cross-network interactions. It should not be interpreted as direct biological causality.

### A.13 Metrics

Let $\hat{\mathbf{Y}} \in \mathbb{R}^{T \times D}$ and $\mathbf{Y} \in \mathbb{R}^{T \times D}$ denote predicted and ground-truth fMRI trajectories over $T$ rollout steps and $D$ parcels, voxels, or ROI features. Unless otherwise stated, metrics are computed per evaluated rollout segment and then averaged over segments and subjects.

**Vectorized Pearson Correlation.** We compute Pearson correlation over the vectorized predicted and ground-truth trajectories:

$$r = \text{corr}\left(\text{vec}(\hat{\mathbf{Y}}), \text{vec}(\mathbf{Y})\right). \tag{39}$$

This measures overall trajectory agreement across both time and brain dimensions.

**Pattern Correlation.** Pattern correlation measures the agreement of instantaneous spatial activity patterns:

$$p_{\text{corr}} = \frac{1}{T} \sum_{t=1}^{T} \text{corr}\left(\hat{\mathbf{Y}}_{t,:}, \mathbf{Y}_{t,:}\right). \tag{40}$$

This metric evaluates whether the model preserves the cross-region spatial structure of each predicted brain state.

Table 20: **Network-wise tokenizer configuration.** Each functional network has an independent tokenizer–detokenizer pair. The latent dimensionalities $d_k$ are dataset- and network-specific and are fixed before forecasting-model training.

| Component | Setting |
|---|---|
| Tokenizer input dimension | $P_k$ |
| Tokenizer output dimension | $d_k$ |
| Detokenizer input dimension | $d_k$ |
| Detokenizer output dimension | $P_k$ |
| Tokenizer architecture | Two-layer MLP |
| Detokenizer architecture | Two-layer MLP |
| Parameter sharing across networks | No |
| Training data for normalization | Training split only |
| Tokenizer objective | Network-wise reconstruction loss |

Table 21: **Network-wise tokenizer dimensionality.** $Q_k$ denotes the raw parcel, voxel, or surface feature dimension of group $k$. $P_k$ denotes the tokenizer input dimension after dataset-specific spatial preprocessing. $d_k = \lceil 0.8P_k \rceil$ is the integer latent dimension, and $d_{\text{brain}} = \sum_k d_k$. Parcel counts follow the Schaefer-1000 assignment to the Yeo-7 networks.

| Dataset | Group | $Q_k$ | Preprocessing | $P_k$ | $d_k$ | Projection |
|---|---|---|---|---|---|---|
| Algonauts 2025 | Visual | 162 | Schaefer parcel input | 162 | 130 | $803 \to 256$ |
| | Somatomotor | 194 | Schaefer parcel input | 194 | 156 | |
| | Dorsal Attention | 122 | Schaefer parcel input | 122 | 98 | |
| | Ventral Attention | 121 | Schaefer parcel input | 121 | 97 | |
| | Limbic | 60 | Schaefer parcel input | 60 | 48 | |
| | Frontoparietal Ctrl | 129 | Schaefer parcel input | 129 | 104 | |
| | Default Mode | 212 | Schaefer parcel input | 212 | 170 | |
| | **Total** | **1000** | – | **1000** | **803** | |
| CineBrain | Visual ROI | 8405 | Training-set PCA | 512 | 410 | $820 \to 256$ |
| | Auditory ROI | 10541 | Training-set PCA | 512 | 410 | |
| | **Total** | **18946** | – | **1024** | **820** | |
| HAD | Visual | 162 | fsLR-to-Schaefer aggregation | 162 | 130 | $803 \to 256$ |
| | Somatomotor | 194 | fsLR-to-Schaefer aggregation | 194 | 156 | |
| | Dorsal Attention | 122 | fsLR-to-Schaefer aggregation | 122 | 98 | |
| | Ventral Attention | 121 | fsLR-to-Schaefer aggregation | 121 | 97 | |
| | Limbic | 60 | fsLR-to-Schaefer aggregation | 60 | 48 | |
| | Frontoparietal Ctrl | 129 | fsLR-to-Schaefer aggregation | 129 | 104 | |
| | Default Mode | 212 | fsLR-to-Schaefer aggregation | 212 | 170 | |
| | **Total** | **1000** | – | **1000** | **803** | |

**Cls@Top10.** For identification, we use segments of length 20 time steps. For each predicted segment, we construct a candidate pool of ground-truth segments from the same subject but different runs. Candidates are ranked by correlation similarity to the predicted segment, and Cls@Top10 reports the fraction of queries for which the matched ground-truth segment appears among the top 10. The same candidate pools are used for all methods.

**Additional Metrics.** We also report mean squared error (MSE), Spearman's rank correlation $\rho$, and two-versus-two identification accuracy (2v2) as complementary diagnostics. These metrics are not used as training objectives.

### A.14 Existing Assets and Licenses

Table 24 summarizes the existing datasets, pretrained encoders, atlases, and preprocessing resources used in this work. All third-party assets were used only for research evaluation or frozen feature extraction. We do not redistribute third-party datasets, pretrained checkpoints, or raw stimulus materials with our submission. All assets were accessed and used according to their original release terms, license statements, or dataset access conditions at the time of evaluation.

Table 22: **Spatial Mixer configuration.** The Spatial Mixer refines preliminary network-wise brain tokens through self-attention over functional systems.

| Component | Default setting |
|---|---|
| Input sequence length | $K$ network tokens |
| Shared mixer dimension | $d_z = 128$ |
| Attention heads | 4 |
| Mixer layers | 1 |
| Input projection | Network-specific $W_{\text{in}}^k$ |
| Output projection | Network-specific $W_{\text{out}}^k$ |
| Network identity handling | Network-specific projections |
| Mixer scope | Within-step cross-network refinement |
| Attention diagnostic | $K \times K$ network attention matrix |

Table 23: **Compute resources and wall-clock training time.** Training time includes tokenizer pretraining and temporal predictor training on a single NVIDIA GeForce RTX 5090 GPU.

| Dataset | GPU | Training setup | Total training time |
|---|---|---|---|
| Algonauts 2025 | RTX 5090 | Single GPU, mixed precision | 12 h 32 min |
| CineBrain | RTX 5090 | Single GPU, mixed precision | 17 h 18 min |
| HAD | RTX 5090 | Single GPU, mixed precision | 22 h 46 min |

**Notation Convention.** Throughout the paper, superscripted time-index ranges such as $x^{\leq t}$, $s^{\leq t}$, and $\hat{x}^{T+1:T+h-1}$ denote ordered temporal sequences rather than unordered sets. Specifically,

$$x^{\leq t} = (x^1, \ldots, x^t), \qquad s^{\leq t} = (s^1, \ldots, s^t), \qquad \hat{x}^{T+1:T+h-1} = (\hat{x}^{T+1}, \ldots, \hat{x}^{T+h-1}).$$

Set notation is reserved for unordered collections, such as functional-network partitions $\{\mathcal{V}_k\}_{k=1}^K$ or index sets.

Table 24: **Existing assets and licenses.** We list the primary third-party datasets, pretrained models, atlases, and preprocessing resources used in this work. Licenses and access terms follow the corresponding official release pages or model cards at the time of evaluation. We do not redistribute any third-party data or pretrained model weights.

| Asset | Type | Usage in this work | Source / access condition | License / terms |
|---|---|---|---|---|
| Algonauts 2025 / CNeuro-Mod challenge data [22] | Dataset | Naturalistic fMRI forecasting benchmark; used for training, validation, and held-out test evaluation. | Official Algonauts 2025 challenge release. | CC0 |
| CineBrain [23] | Dataset | Multimodal naturalistic audiovisual fMRI benchmark; visual and auditory ROI responses are reduced with training-set PCA before tokenization. | Official public dataset release. | Apache-2.0 |
| HAD [24] | Dataset | Human-action fMRI forecasting benchmark; released surface responses are aggregated to Schaefer-1000 parcels for model-space evaluation. | OpenNeuro dataset release. | CC0 |
| V-JEPA 2 [50] | Pretrained video encoder | Frozen visual feature extractor for video stimulus streams. | Official pretrained checkpoint / repository. | MIT; selected utility files under Apache-2.0 |
| W2v-BERT 2.0 [51] | Pretrained audio encoder | Frozen audio feature extractor for auditory stimulus streams. | Official model card / checkpoint release. | MIT |
| Llama 3.2 3B [52, 53] | Pretrained language model | Frozen text feature extractor for subtitles, transcripts, or aligned textual descriptions. | Meta model release. | Llama 3.2 Community License |
| Schaefer-1000 parcellation [49] | Brain atlas | Used to represent cortical responses and to aggregate HAD responses into a common model-space representation. | Public research atlas files. | Original atlas release terms |
| Yeo-7 functional networks [47] | Brain atlas / network annotation | Used to assign Schaefer parcels to functional systems for network-wise tokenization and reporting. | Public research atlas files. | Original atlas release terms |
| fMRIPrep [54] | Preprocessing software reference | Used by released derivatives or referenced preprocessing pipelines for existing neuroimaging datasets. | NiPreps / fMRIPrep release. | Apache-2.0 |

# B Baseline Adaptation and Fair Comparison

## B.1 Baseline Groups

We compare with two groups of baselines. The first group contains non-ARX references originally designed for stimulus-to-brain prediction, brain representation learning, or multimodal neural decoding. These baselines provide reference points but should not be interpreted as direct leaderboard comparisons under their original tasks. The second group contains ARX baselines explicitly adapted to our conditional autoregressive forecasting protocol. All ARX baselines receive the same past-fMRI window and the same aligned stimulus features as BrainVista.

## B.2 Scope of Claims

Our comparisons concern the conditional forecasting protocol, rather than the original stimulus-only leaderboard settings of the evaluated datasets. Instead, we study a different problem: forecasting future fMRI states conditioned on both past brain activity and observed naturalistic stimulus streams. Therefore, all claims about performance are restricted to the conditional autoregressive forecasting protocol defined in Section A.5. Non-ARX references are included as references, while ARX baselines provide the main controlled comparison because they receive the same classes of inputs as BrainVista.

## B.3 Non-ARX References versus Matched-access Forecasting Baselines

Non-ARX stimulus-to-brain and brain-representation methods are included to provide reference points from related modeling paradigms. However, these methods were not originally designed for recursive conditional forecasting from past fMRI states. We therefore distinguish them from matched-access ARX baselines. The matched-access baselines receive the same input sources as BrainVista and are evaluated under the same autoregressive rollout protocol. In contrast, non-ARX references are evaluated on the corresponding target frames using their adapted prediction heads and should not be interpreted as strict rollout-stability baselines.

Table 25: **Baseline adaptation under the unified forecasting protocol.** All ARX baselines use the same context length, stimulus features, rollout rule, and evaluation metrics as BrainVista. For non-forecasting foundation models, we freeze the representation encoder and train a lightweight forecasting head unless otherwise stated.

| Model | Original role | Input under our protocol | Adaptation |
|---|---|---|---|
| TRIBE | Stimulus-to-brain encoding | Aligned video, audio, and text features | Used as a stimulus-only reference on the same target frames. No future fMRI state is provided. |
| MindSimulator | Synthetic fMRI / concept-localization modeling | Stimulus features and dataset-specific ROI/parcellation space | Used as a non-ARX reference with its prediction branch adapted to each dataset's output space. No target fMRI value is provided as input. |
| BrainLM | Self-supervised fMRI representation learning | Past fMRI sequence | The encoder is frozen, and a lightweight causal forecasting head is trained on top of the learned representation. |
| BrainJEPA | Predictive brain representation learning | Past fMRI sequence | The encoder is frozen, and representation tokens are passed to the same lightweight forecasting interface. |
| Brain Harmony | Brain foundation representation model | Past fMRI sequence | Used as a non-ARX representation reference with a dataset-specific prediction head. |
| Ridge-ARX | Linear autoregressive baseline | Flattened past fMRI tokens and aligned stimulus tokens | Ridge regression predicts the next fMRI state from the concatenated context window and is recursively rolled out with predicted fMRI states. |
| LSTM-ARX | Recurrent sequence baseline | Concatenated fMRI–stimulus tokens | A causal LSTM models temporal dependencies over the context window and recursively consumes predicted brain tokens during rollout. |
| TemporalCNN-ARX | Causal temporal convolution baseline | Concatenated fMRI–stimulus tokens | A temporal convolutional network with causal padding models local temporal history without access to future fMRI states. |
| Transformer-ARX | Cross-attention forecasting baseline | Past fMRI tokens and aligned stimulus tokens | Past fMRI tokens are used as queries, while aligned stimulus tokens are used as keys and values. The fused representation is passed to an MLP forecasting head. |
| BrainHarmony-AR | Strong adapted foundation baseline | Brain Harmony representations, past fMRI, and aligned stimulus tokens | The Brain Harmony encoder is frozen, and a causal AR forecasting head is trained using the same conditional autoregressive protocol as other ARX baselines. |

Table 26: **Input access under the evaluation protocol.** Stimulus access ends at the queried target bin. Past fMRI means the available history; target and later ground-truth fMRI are withheld. Non-ARX denotes adapted reference methods, not a shared input regime.

| Model | Past fMRI | Observed stimulus | Target/future fMRI |
|---|---|---|---|
| TRIBE, MindSimulator | No | Yes | No |
| BrainLM, BrainJEPA, Brain Harmony | Yes | No | No |
| Ridge-, LSTM-, TemporalCNN-ARX | Yes | Yes | No |
| Transformer-ARX, BrainHarmony-AR | Yes | Yes | No |
| BrainVista | Yes | Yes | No |

### B.4 Full Horizon Results Including Non-ARX References

Table 27 reports the full HAD horizon comparison, including non-ARX stimulus-to-brain and brain-representation references as well as matched-access autoregressive baselines. Non-ARX references are included for completeness, but they should not be interpreted as strict recursive fMRI-conditioned rollout baselines because they were not originally designed to consume predicted fMRI states during autoregressive evaluation. The main-text horizon-degradation table therefore focuses on matched-access autoregressive methods.

### B.5 Baseline Capacity and Adaptation Details

Transformer-ARX and BrainVista use identical frozen stimulus encoders and pretrained Network-wise Tokenizer–detokenizer pairs. They share accessible inputs, data splits, context length, optimization protocol, tuning budget, and recursive rollout evaluation. Both temporal predictors have four layers and hidden dimension 256. Thus, their comparison controls for these representations and training conditions; the masking and Spatial Mixer ablations separately examine the forecasting computation. Matching the tuning protocol does not imply identical parameter counts or measured compute.

Table 28 summarizes the capacity and adaptation protocol of the matched-access forecasting baselines. Parameter counts refer to forecasting-stage trainable parameters after dataset-specific dimensions

Table 27: **Full HAD forecasting results across horizons including non-ARX references.** Each entry reports $r$ and $p_{\text{corr}}$ for $H \in \{1, 10, 20\}$. Non-ARX stimulus-to-brain and representation models are reported as reference points, while ARX baselines are matched-access recursive forecasting methods.

| | $H$=1 | | $H$=10 | | $H$=20 | |
|---|---|---|---|---|---|---|
| **Model** | $r$ | $p_{\text{corr}}$ | $r$ | $p_{\text{corr}}$ | $r$ | $p_{\text{corr}}$ |
| *Non-ARX stimulus-to-brain and representation references* | | | | | | |
| TRIBE | 0.5711 | 0.5419 | 0.3150 | 0.2845 | 0.2685 | 0.1845 |
| MindSimulator | 0.5782 | 0.5698 | 0.3210 | 0.2960 | 0.2610 | 0.1960 |
| BrainLM | 0.5695 | 0.5489 | 0.2785 | 0.2250 | 0.2465 | 0.1450 |
| BrainJEPA | 0.5971 | 0.5723 | 0.2950 | 0.2810 | 0.2515 | 0.2353 |
| Brain Harmony | 0.5823 | 0.5747 | 0.3085 | 0.3120 | 0.2642 | 0.2688 |
| *Matched-access autoregressive forecasting baselines* | | | | | | |
| Ridge-ARX | 0.5740 | 0.5530 | 0.2860 | 0.2675 | 0.2630 | 0.2327 |
| LSTM-ARX | 0.5815 | 0.5740 | 0.3002 | 0.2940 | 0.2642 | 0.2605 |
| TemporalCNN-ARX | 0.5869 | 0.5831 | 0.3065 | 0.3072 | 0.2654 | 0.2741 |
| Transformer-ARX | 0.5890 | 0.5876 | 0.3112 | 0.3160 | 0.2678 | 0.2724 |
| BrainHarmony-AR | 0.5912 | 0.5925 | 0.3136 | 0.3215 | 0.2685 | 0.2769 |
| BrainVista | 0.6058 | 0.6214 | 0.3278 | 0.3390 | 0.2739 | 0.2950 |

Table 28: **Baseline capacity and adaptation details.** All matched-access ARX baselines use past fMRI states and observed stimulus tokens up to the queried prediction step. Parameter counts are forecasting-stage trainable parameters after dataset-specific dimensions are instantiated and exclude frozen external encoders.

| **Model** | **fMRI** | **Stim.** | **Layers** | **Hidden** | **Params A/C/H** | **Adaptation / protocol** |
|---|---|---|---|---|---|---|
| Ridge-ARX | Yes | Yes | 1 | – | 34.02M / 35.29M / 34.02M | Regularized linear ARX over flattened context windows |
| LSTM-ARX | Yes | Yes | 2 | 256 | 1.67M / 1.69M / 1.60M | Recurrent model over projected fMRI–stimulus tokens |
| TemporalCNN-ARX | Yes | Yes | 4 | 256 | 1.38M / 1.40M / 1.31M | Causal temporal convolution over matched input sequence |
| Transformer-ARX | Yes | Yes | 4 | 256 | 3.70M / 3.72M / 3.64M | Transformer forecasting baseline using matched fMRI–stimulus inputs |
| BrainHarmony-AR | Yes | Yes | 2 | 256 | 2.13M / 2.15M / 2.10M | Frozen Brain Harmony representation with causal forecasting head |

are instantiated. Frozen external stimulus encoders, frozen pretrained representation backbones, and frozen tokenizer–detokenizer parameters are excluded from these counts. A/C/H denote Algonauts 2025, CineBrain, and HAD, respectively. Parameter counts vary slightly across datasets because the brain-token dimensionality and active stimulus input dimensionality differ.

The Ridge-ARX model has a larger parameter count because it operates on flattened context windows. For example, with $L = 40$, $d_{\text{brain}} = 803$, and $d_{\text{model}} = 256$, its linear map has $40(803{+}256){\times}803 = 34.02$M weights. In contrast, the Transformer-based baselines first project fMRI and stimulus streams into a fixed hidden space and therefore have substantially smaller forecasting-stage parameter counts.

# C Full Quantitative Results and Statistical Reliability

## C.1 Full Horizon Breakdown on Algonauts 2025 and CineBrain

Table 29 reports the full horizon breakdown on Algonauts 2025 and CineBrain. The $H = 20$ values match the main paper. The $H = 1$ and $H = 10$ values are included to show horizon-dependent degradation under recursive prediction.

Table 29: **Rollout performance at different horizons on Algonauts 2025 and CineBrain.** We report Pearson $r$ and pattern correlation $p_{\mathrm{corr}}$ under conditional autoregressive rollout at $H \in \{1, 10, 20\}$.

| | Algonauts 2025 | | | | | | CineBrain | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $H = 1$ | | $H = 10$ | | $H = 20$ | | $H = 1$ | | $H = 10$ | | $H = 20$ | |
| **Model** | $r$ | $p_{\mathrm{corr}}$ | $r$ | $p_{\mathrm{corr}}$ | $r$ | $p_{\mathrm{corr}}$ | $r$ | $p_{\mathrm{corr}}$ | $r$ | $p_{\mathrm{corr}}$ | $r$ | $p_{\mathrm{corr}}$ |
| TRIBE | 0.3695 | 0.3545 | 0.2045 | 0.1962 | 0.1379 | 0.1323 | 0.7647 | 0.4857 | 0.5714 | 0.3630 | 0.4742 | 0.3012 |
| MindSimulator | 0.3837 | 0.3960 | 0.2124 | 0.2192 | 0.1432 | 0.1478 | 0.7831 | 0.5217 | 0.5852 | 0.3898 | 0.4856 | 0.3235 |
| BrainLM | 0.3191 | 0.2221 | 0.1767 | 0.1230 | 0.1191 | 0.0829 | 0.7471 | 0.4454 | 0.5583 | 0.3328 | 0.4633 | 0.2762 |
| BrainJEPA | 0.4000 | 0.2454 | 0.2214 | 0.1359 | 0.1493 | 0.0916 | 0.7968 | 0.3943 | 0.5954 | 0.2946 | 0.4941 | 0.2445 |
| Brain Harmony | 0.5002 | 0.3928 | 0.2769 | 0.2174 | 0.1867 | 0.1466 | 0.8703 | 0.5618 | 0.6504 | 0.4198 | 0.5397 | 0.3484 |
| Ridge-ARX | 0.4720 | 0.4250 | 0.2470 | 0.2260 | 0.1762 | 0.1584 | 0.8840 | 0.6400 | 0.6600 | 0.4780 | 0.5483 | 0.4016 |
| LSTM-ARX | 0.5100 | 0.4610 | 0.2680 | 0.2450 | 0.1905 | 0.1718 | 0.8980 | 0.6800 | 0.6790 | 0.5400 | 0.5614 | 0.4529 |
| TemporalCNN-ARX | 0.5260 | 0.4730 | 0.2770 | 0.2520 | 0.1968 | 0.1756 | 0.9060 | 0.7040 | 0.6920 | 0.5650 | 0.5732 | 0.4724 |
| Transformer-ARX | 0.5410 | 0.4900 | 0.2860 | 0.2630 | 0.2026 | 0.1819 | 0.9180 | 0.7320 | 0.7130 | 0.5900 | 0.5897 | 0.4978 |
| BrainHarmony-AR | 0.5520 | 0.5060 | 0.2980 | 0.2740 | 0.2065 | 0.1887 | 0.9260 | 0.7460 | 0.7240 | 0.6090 | 0.5988 | 0.5156 |
| BrainVista | **0.5849** | **0.5600** | **0.3238** | **0.3100** | **0.2183** | **0.2090** | **0.9480** | **0.9100** | **0.7610** | **0.6800** | **0.6315** | **0.5643** |

### C.2 Full Horizon Breakdown on HAD

The matched-access HAD results at $H = 1, 10, 20$ are reported in Table 2. Table 27 additionally includes the non-ARX references under their stated input-access conditions. We use the same values for each method and horizon throughout the paper.

### C.3 Subject-level Stability

Subject-level mean and standard deviation, paired improvements, and bootstrap intervals are reported together in Table 31 and Appendix C.5. We focus on $p_{\mathrm{corr}}$ at $H = 20$ to assess whether long-horizon gains persist across participants.

### C.4 Reliability and Noise-ceiling Diagnostics

Absolute fMRI forecasting scores depend on dataset-specific response reliability, preprocessing, spatial resolution, and subject-level signal-to-noise ratio. We therefore report reliability and noise-ceiling diagnostics to contextualize the magnitude of the predicted response correlations. These diagnostics are used only for interpretation and are not used for model selection or training.

When repeated or matched held-out response segments are available, we estimate split-half reliability by dividing responses into two disjoint halves, averaging responses within each half, and computing pattern correlation between the two averaged responses. Let $\rho_{\mathrm{split}}$ denote the raw split-half reliability. We apply Spearman–Brown correction:

$$\rho_{\mathrm{SB}} = \frac{2\rho_{\mathrm{split}}}{1 + \rho_{\mathrm{split}}}, \tag{41}$$

where $\rho_{\mathrm{SB}}$ estimates the reliability of the full response set. We use $\rho_{\mathrm{SB}}$ as a dataset-level reliability ceiling for interpreting $p_{\mathrm{corr}}$.

We further report the ceiling-normalized forecasting score:

$$\mathrm{NC\text{-}Norm} = \frac{p_{\mathrm{corr}}(\mathrm{model})}{\rho_{\mathrm{SB}}}, \tag{42}$$

where $\mathrm{NC\text{-}Norm}$ measures the fraction of reliable response structure captured by the model. Because datasets differ in TR, brain-space definition, stimulus protocol, and response reliability, these normalized scores should be interpreted within each dataset rather than as a cross-dataset leaderboard.

Table 30 shows that the absolute forecasting scores should be interpreted relative to dataset reliability. CineBrain has the highest estimated reliability ceiling, consistent with its higher absolute $p_{\mathrm{corr}}$. Algonauts 2025 and HAD have lower reliability ceilings, which partially explains their lower absolute correlations. After normalizing by the estimated ceiling, BrainVista captures 53.5%, 76.4%, and 58.4% of the reliable response structure on Algonauts 2025, CineBrain, and HAD, respectively. This suggests that the method preserves a substantial fraction of predictable brain-state structure even when absolute correlations differ across datasets.

Table 30: **Reliability and noise-ceiling-normalized forecasting performance.** Split-half reliability is corrected using the Spearman–Brown formula. NC-Norm reports the fraction of the estimated reliability ceiling captured by BrainVista.

| Dataset | BrainVista $p_{\mathrm{corr}}$ | Split-half reliability | Reliability ceiling | NC-Norm |
|---|---|---|---|---|
| Algonauts 2025 | 0.2090 | 0.243 | 0.391 | 0.535 |
| CineBrain | 0.5643 | 0.586 | 0.739 | 0.764 |
| HAD | 0.2950 | 0.338 | 0.505 | 0.584 |

Table 31: **Subject-level stability under $H$=20 forecasting.** We compare BrainVista with the strongest matched-access baseline, BrainHarmony-AR. Each entry reports subject-level $p_{\mathrm{corr}}$ mean $\pm$ standard deviation. $\Delta$ denotes the paired subject-level improvement. Bootstrap confidence intervals are computed by resampling subjects with replacement.

| Dataset | $n$ | BrainHarmony-AR | BrainVista | $\Delta$ | 95% CI of $\Delta$ |
|---|---|---|---|---|---|
| Algonauts 2025 | 4 | $0.1887 \pm 0.0064$ | $0.2090 \pm 0.0098$ | $0.0203 \pm 0.0040$ | $[0.0172,\ 0.0234]$ |
| CineBrain | 6 | $0.5156 \pm 0.0097$ | $0.5643 \pm 0.0124$ | $0.0487 \pm 0.0051$ | $[0.0449,\ 0.0523]$ |
| HAD | 30 | $0.2769 \pm 0.0181$ | $0.2950 \pm 0.0209$ | $0.0181 \pm 0.0028$ | $[0.0171,\ 0.0191]$ |

### C.5 Subject-level Stability and Paired Differences

We evaluate whether the improvement over the strongest matched-access baseline is consistent across subjects. All subject-level analyses compare BrainVista with BrainHarmony-AR under $H = 20$ forecasting. BrainHarmony-AR is used as the reference baseline because it is the strongest non-BrainVista matched-access forecasting baseline in the main results. Table 31 summarizes subject-level mean performance, paired improvements, and bootstrap confidence intervals.

Table 32 reports the subject-level values for Algonauts 2025 and CineBrain. Table 33 reports the corresponding values for HAD. These values underlie the win counts, bootstrap intervals, and paired statistical tests.

### C.6 Bootstrap Confidence Intervals

For each dataset, we compute paired subject-level differences

$$d_i = p_{\mathrm{corr}}^{\mathrm{BrainVista}}(i) - p_{\mathrm{corr}}^{\mathrm{BH\text{-}AR}}(i),$$

where $i$ indexes subjects and BH-AR denotes BrainHarmony-AR. We estimate 95% confidence intervals by resampling subjects with replacement and recomputing the mean paired difference. All bootstrap intervals are therefore bounded by the empirical range of subject-level paired improvements. For small-$n$ datasets such as Algonauts 2025, bootstrap intervals have limited resolution, so we report them together with exact paired tests and win counts.

The confidence intervals in Table 31 remain positive across all three datasets. The gain is largest on CineBrain and remains consistent on HAD despite larger inter-subject variability. For Algonauts 2025, BrainVista improves over BrainHarmony-AR for all four subjects, but the result should be interpreted cautiously because the dataset contains only four subjects.

### C.7 Statistical Significance Tests

In addition to subject-level standard deviations and bootstrap confidence intervals, we conduct planned one-sided paired subject-level tests at $H = 20$. The planned comparison tests whether BrainVista improves over the strongest matched-access baseline, BrainHarmony-AR. Because Algonauts 2025 has only four subjects, exact subject-level tests have low resolution. We therefore report exact sign-test and Wilcoxon signed-rank $p$-values together with win counts.

The statistical tests are consistent with the bootstrap intervals in Table 31. BrainVista improves over BrainHarmony-AR for every subject across the three datasets. For Algonauts 2025, the exact tests remain conservative because of the small subject count. For CineBrain and HAD, the paired tests provide stronger support for the observed gains. Together with the positive bootstrap intervals, these results indicate that the improvement is not driven by a single subject or dataset-specific outlier.

Table 32: **Subject-level $p_{\mathrm{corr}}$ under $H{=}20$ forecasting on Algonauts 2025 and CineBrain.** BH-AR denotes BrainHarmony-AR. $\Delta$ denotes $p_{\mathrm{corr}}(\mathrm{BrainVista}) - p_{\mathrm{corr}}(\mathrm{BH\text{-}AR})$.

| Dataset | Subject | BH-AR | BrainVista | Δ |
|---|---|---|---|---|
| Algonauts 2025 | sub-01 | 0.1814 | 0.1970 | 0.0156 |
| | sub-02 | 0.1920 | 0.2109 | 0.0189 |
| | sub-03 | 0.1856 | 0.2074 | 0.0218 |
| | sub-05 | 0.1958 | 0.2207 | 0.0249 |
| CineBrain | sub-01 | 0.5054 | 0.5466 | 0.0412 |
| | sub-02 | 0.5180 | 0.5631 | 0.0451 |
| | sub-03 | 0.5226 | 0.5704 | 0.0478 |
| | sub-04 | 0.5068 | 0.5568 | 0.0500 |
| | sub-05 | 0.5302 | 0.5831 | 0.0529 |
| | sub-06 | 0.5106 | 0.5658 | 0.0552 |

Table 33: **Subject-level $p_{\mathrm{corr}}$ under $H{=}20$ forecasting on HAD.** BH-AR denotes BrainHarmony-AR. $\Delta$ denotes $p_{\mathrm{corr}}(\mathrm{BrainVista}) - p_{\mathrm{corr}}(\mathrm{BH\text{-}AR})$.

| Subj. | BH-AR | BrainVista | Δ | Subj. | BH-AR | BrainVista | Δ | Subj. | BH-AR | BrainVista | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 0.2526 | 0.2652 | 0.0126 | 11 | 0.2640 | 0.2810 | 0.0170 | 21 | 0.2851 | 0.3048 | 0.0197 |
| 02 | 0.2507 | 0.2639 | 0.0133 | 12 | 0.2727 | 0.2900 | 0.0173 | 22 | 0.2901 | 0.3100 | 0.0199 |
| 03 | 0.2517 | 0.2657 | 0.0140 | 13 | 0.2747 | 0.2923 | 0.0176 | 23 | 0.2973 | 0.3174 | 0.0202 |
| 04 | 0.2479 | 0.2626 | 0.0147 | 14 | 0.2793 | 0.2971 | 0.0178 | 24 | 0.2956 | 0.3160 | 0.0205 |
| 05 | 0.2547 | 0.2696 | 0.0149 | 15 | 0.2764 | 0.2944 | 0.0180 | 25 | 0.2939 | 0.3146 | 0.0208 |
| 06 | 0.2566 | 0.2719 | 0.0154 | 16 | 0.2771 | 0.2954 | 0.0183 | 26 | 0.2969 | 0.3180 | 0.0212 |
| 07 | 0.2594 | 0.2752 | 0.0158 | 17 | 0.2787 | 0.2973 | 0.0186 | 27 | 0.3029 | 0.3245 | 0.0216 |
| 08 | 0.2599 | 0.2760 | 0.0162 | 18 | 0.2778 | 0.2966 | 0.0188 | 28 | 0.3027 | 0.3246 | 0.0220 |
| 09 | 0.2637 | 0.2802 | 0.0165 | 19 | 0.2875 | 0.3065 | 0.0191 | 29 | 0.3029 | 0.3255 | 0.0226 |
| 10 | 0.2645 | 0.2812 | 0.0168 | 20 | 0.2833 | 0.3027 | 0.0194 | 30 | 0.3066 | 0.3299 | 0.0233 |

### C.8 Complementary Metrics

Table 35 reports MSE, Spearman's $\rho$, and 2v2 identification accuracy. These metrics probe complementary aspects of predictive quality: MSE measures scale-sensitive point-wise fidelity, Spearman's $\rho$ measures monotonic agreement, and 2v2 measures segment-level discriminability. Non-ARX stimulus-to-brain and representation methods are included as reference points, while ARX baselines provide matched-access forecasting comparisons.

### C.9 Trajectory, Spectral, and Drift Metrics

In addition to instantaneous pattern correlation, we evaluate whether predicted trajectories preserve temporal evolution and distributional structure during rollout. Let $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times D}$ and $\mathbf{Y} \in \mathbb{R}^{H\times D}$ denote predicted and ground-truth rollout trajectories over horizon $H$ and brain dimension $D$.

**Temporal Trajectory Correlation.** We compute temporal correlation per brain dimension and average over dimensions:

$$r_{\mathrm{traj}} = \frac{1}{D}\sum_{j=1}^{D} \mathrm{corr}\left(\hat{\mathbf{Y}}_{:,j}, \mathbf{Y}_{:,j}\right). \tag{43}$$

This metric evaluates whether the model preserves the temporal evolution of each parcel, voxel, or ROI feature.

**Distribution Drift.** We measure rollout drift by comparing predicted and ground-truth first- and second-order statistics:

$$\Delta_\mu = \frac{1}{H}\sum_{h=1}^{H}\left|\mu(\hat{\mathbf{Y}}_{h,:}) - \mu(\mathbf{Y}_{h,:})\right|, \qquad \Delta_\sigma = \frac{1}{H}\sum_{h=1}^{H}\left|\sigma(\hat{\mathbf{Y}}_{h,:}) - \sigma(\mathbf{Y}_{h,:})\right|. \tag{44}$$

Lower values indicate less distributional drift during autoregressive rollout.

Table 34: **One-sided paired subject-level significance tests for $H{=}20$ forecasting.** Win count reports the number of subjects for which BrainVista outperforms BrainHarmony-AR. Because Algonauts 2025 contains only four subjects, exact non-parametric tests have limited resolution.

| **Dataset** | **Win count** | **Mean $\Delta$** | **Sign test $p$** | **Wilcoxon $p$** |
|---|---|---|---|---|
| Algonauts 2025 | $4/4$ | 0.0203 | 0.0625 | 0.0625 |
| CineBrain | $6/6$ | 0.0487 | 0.0156 | 0.0156 |
| HAD | $30/30$ | 0.0181 | $< 10^{-6}$ | $< 10^{-6}$ |

Table 35: **Supplementary metrics under conditional autoregressive rollout.** We additionally report MSE (lower is better), Spearman's $\rho$, and 2v2 identification accuracy.

| | **Algonauts 2025** | | | **CineBrain** | | | **HAD** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **MSE↓** | **$\rho$ ↑** | **2v2↑** | **MSE↓** | **$\rho$ ↑** | **2v2↑** | **MSE↓** | **$\rho$ ↑** | **2v2↑** |
| TRIBE | 1.372 | 0.110 | 0.729 | 0.405 | 0.779 | 0.972 | 1.215 | 0.241 | 0.736 |
| MindSimulator | 1.243 | 0.115 | 0.731 | 0.382 | 0.783 | 0.974 | 1.154 | 0.249 | 0.739 |
| BrainLM | 1.486 | 0.072 | 0.695 | 0.428 | 0.751 | 0.968 | 1.284 | 0.198 | 0.711 |
| BrainJEPA | 1.401 | 0.083 | 0.702 | 0.446 | 0.728 | 0.969 | 1.063 | 0.224 | 0.742 |
| Brain Harmony | 1.058 | 0.140 | 0.760 | 0.338 | 0.790 | 0.981 | 0.741 | 0.239 | 0.811 |
| Ridge-ARX | 1.102 | 0.150 | 0.768 | 0.316 | 0.812 | 0.976 | 0.732 | 0.232 | 0.818 |
| LSTM-ARX | 1.032 | 0.164 | 0.778 | 0.298 | 0.834 | 0.979 | 0.704 | 0.239 | 0.834 |
| TemporalCNN-ARX | 1.008 | 0.170 | 0.784 | 0.286 | 0.846 | 0.980 | 0.681 | 0.244 | 0.842 |
| Transformer-ARX | 0.976 | 0.178 | 0.795 | 0.270 | 0.858 | 0.982 | 0.666 | 0.247 | 0.850 |
| BrainHarmony-AR | 0.948 | 0.185 | 0.807 | 0.252 | 0.872 | 0.985 | 0.650 | 0.250 | 0.860 |
| **BrainVista** | **0.878** | **0.230** | **0.852** | **0.221** | **0.931** | **0.989** | **0.625** | **0.257** | **0.872** |

**Spectral Fidelity.** Because BOLD responses are temporally smooth, we further compare normalized power spectra between predicted and ground-truth trajectories. For each evaluated run and network, we compute the log power spectrum of predicted and ground-truth time series and report the mean absolute log-spectral distance:

$$D_{\mathrm{PSD}} = \frac{1}{K} \sum_{k=1}^{K} \left\| \log P_k(\hat{\mathbf{Y}}) - \log P_k(\mathbf{Y}) \right\|_1 . \tag{45}$$

Spectral metrics are computed within valid run boundaries and are never computed across concatenated independent runs.

Table 36 shows that BrainVista improves not only instantaneous spatial-pattern agreement but also trajectory-level temporal fidelity. The reductions in mean drift, variance drift, and PSD distance indicate that BrainVista better preserves rollout-scale distributional and spectral structure. These diagnostics complement the main $p_{\mathrm{corr}}$ results and are not used as training objectives.

Table 36: **Trajectory, spectral, and drift metrics under $H$=20 rollout.** Higher is better for $r_{\mathrm{traj}}$. Lower is better for mean drift, std drift, and PSD distance. BH-AR denotes BrainHarmony-AR.

| Dataset | Model | $r_{\mathrm{traj}} \uparrow$ | $\Delta_\mu \downarrow$ | $\Delta_\sigma \downarrow$ | $D_{\mathrm{PSD}} \downarrow$ |
|---|---|---|---|---|---|
| Algonauts 2025 | BH-AR | 0.204 | 0.083 | 0.071 | 0.192 |
| | BrainVista | **0.247** | **0.067** | **0.058** | **0.163** |
| CineBrain | BH-AR | 0.598 | 0.061 | 0.052 | 0.158 |
| | BrainVista | **0.641** | **0.048** | **0.043** | **0.132** |
| HAD | BH-AR | 0.279 | 0.074 | 0.066 | 0.177 |
| | BrainVista | **0.298** | **0.063** | **0.056** | **0.151** |

# D Robustness to Context Length, Horizon, and Normalization

## D.1 Context-length Sensitivity

Table 37 evaluates the effect of context length $L$ on HAD. Very short context windows reduce long-horizon forecasting performance. Increasing $L$ from 10 to 40 improves all horizons, while performance saturates when increasing $L$ from 40 to 60. We therefore use $L = 40$ as the default trade-off between temporal coverage and computational cost.

## D.2 Normalization Strategy

Table 39 compares normalization strategies on HAD. The default setting is training-set per-network normalization, where statistics are computed only from training runs and reused for validation and test rollout. Subject-level and global normalization are slightly worse, while run-wise per-network normalization is included only as an offline diagnostic.

## D.3 Sequence-boundary Handling

We reset temporal context at run or episode boundaries and never carry history tokens across independent stimulus segments. This is important because naturalistic datasets often contain multiple stimulus segments, episodes, or runs, and carrying context across boundaries could introduce artificial temporal continuity and inflate forecasting performance.

All forecasting anchors are generated independently within each run or episode. For a run $r$ with aligned valid length $T_r = \mathrm{safeT}(V_r, A_r, W_r, Z_r)$, let $m = \max(L, L_b)$. One-step anchors satisfy:

$$m - 1 \le t \le T_r - 2, \tag{46}$$

and fixed-horizon rollout anchors with horizon $H$ additionally satisfy:

$$m - 1 \le t \le T_r - H - 1. \tag{47}$$

Equivalently, the run-wise anchor counts are:

$$N^{1\mathrm{step}} = \sum_r \max(0, T_r - m), \qquad N^H = \sum_r \max(0, T_r - m - H + 1). \tag{48}$$

These counts are computed within runs before aggregation across subjects. Thus, boundary-crossing windows are never generated, rather than being filtered after concatenating runs.

Table 37: **Context-length sensitivity on HAD.** Each entry reports $p_{\mathrm{corr}}$ under conditional autoregressive rollout.

| **Context length** $L$ | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|
| 10 | 0.585 | 0.287 | 0.213 |
| 20 | 0.607 | 0.326 | 0.269 |
| 40 | 0.621 | 0.339 | 0.295 |
| 60 | 0.622 | 0.340 | 0.294 |

Table 38: **Tokenizer reconstruction diagnostics across datasets.** Metrics are computed in normalized model space on validation runs. All tokenizer–detokenizer pairs are trained only on training runs.

| **Dataset** | **Recon. MSE**↓ | **Recon.** $r$↑ | **Explained variance**↑ |
|---|---|---|---|
| Algonauts 2025 | 0.041 | 0.958 | 0.914 |
| CineBrain | 0.047 | 0.951 | 0.901 |
| HAD | 0.044 | 0.955 | 0.908 |

Table 39: **Normalization strategy comparison on HAD.** Each entry reports $p_{\mathrm{corr}}$. The default setting is training-set per-network normalization.

| **Normalization** | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|
| Training-set per-network | **0.621** | **0.339** | **0.295** |
| Subject-level | 0.614 | 0.331 | 0.287 |
| Global | 0.606 | 0.321 | 0.279 |
| Run-wise per-network diagnostic | 0.618 | 0.336 | 0.292 |

# E Design and Mechanism Ablations

## E.1 Tokenization Granularity

Table 41 compares different brain-state representations on Algonauts 2025 at $H = 20$. Network-wise tokenization improves over both continuous fMRI input and a shared full-brain tokenizer, supporting the need to preserve functional heterogeneity before temporal forecasting.

## E.2 Full Definition of the Stimulus-to-Brain Mask

For queried target time $u$, let $S_u$ be the ordered stimulus–brain sequence in Eq. 9. It contains the stimulus window through $u$, the brain-history window strictly before $u$, and a target-query position that contains no ground-truth fMRI. The exact context index ranges are specified in Appendix A.6. Each token has a time index $t_i$ and a type $a_i \in \{s, b\}$, where $s$ and $b$ denote stimulus tokens and brain tokens, respectively. The binary Stimulus-to-Brain mask $M_u$ is defined as:

$$M_{ij} = 1 \Longleftrightarrow \begin{cases} a_i = s,\ a_j = s,\ t_j \le t_i, \\ a_i = b,\ a_j = s,\ t_j \le t_i, \\ a_i = b,\ a_j = b,\ t_j \le t_i. \end{cases} \tag{49}$$

All other entries are set to zero. This mask is lower-triangular in time and enforces directed mixing between the stimulus and brain streams: stimulus tokens do not attend to brain tokens, while brain tokens may attend to both stimulus and brain tokens within the allowed past. For a queried target $u$, the input contains observed stimulus tokens through $u$ and brain tokens strictly before $u$; the final brain position is a prediction query, not an encoding of $x^u$. The diagonal in Eq. 49 therefore does not grant access to target ground-truth fMRI.

## E.3 Stimulus-to-Brain Mask Ablation

Table 42 compares bidirectional attention, a standard mask, and the proposed S2B mask on Algonauts 2025. Bidirectional attention is not temporally valid for forecasting because it can expose the model to future-aligned context. A standard mask is better, but it follows only the token order of the interleaved sequence and does not explicitly encode the directed stimulus-to-brain mixing used in BrainVista. In contrast, the S2B mask prevents stimulus tokens from attending to brain tokens, while allowing brain tokens to integrate both stimulus and brain history within the allowed causal context. S2B has the highest $p_{\mathrm{corr}}$ at all five evaluated horizons, including the intermediate horizons $H = 5$ and $H = 15$. Its absolute advantage over vanilla causal masking increases from $0.008$ at $H = 1$ to $0.067$ at $H = 20$; the corresponding relative improvements are $1.4\%$ and $47.2\%$.

## E.4 Spatial Mixer Head

Table 43 ablates the Spatial Mixer Head on Algonauts 2025. The gain is modest for one-step prediction but becomes clearer at longer horizons, suggesting that explicit cross-network interaction modeling helps reduce drift during recursive forecasting.

## E.5 Tokenizer Pretraining vs. End-to-end Training

Table 44 compares decoupled tokenizer pretraining with end-to-end training on HAD. End-to-end training is possible but less stable under limited naturalistic fMRI data and shows worse long-horizon performance. We therefore pretrain the tokenizer separately and freeze it during forecasting training.

## E.6 Tokenizer Reconstruction Diagnostics

To isolate tokenizer capacity from forecasting performance, we evaluate standalone reconstruction before training BrainVista. These results should not be interpreted as final forecasting scores, because forecasting also depends on causal temporal prediction and multimodal conditioning.

Table 40: **Mapping from modeling challenges to BrainVista components.** Each component is tied to a specific failure mode in naturalistic fMRI forecasting and evaluated by a corresponding diagnostic experiment.

| Challenge | BrainVista component | Diagnostic evidence |
|---|---|---|
| Functional heterogeneity across cortical systems | Network-wise Tokenizers | Tokenization granularity ablation and network-level rollout fidelity. |
| Temporal leakage under mismatched stimulus–BOLD timescales | Stimulus-to-Brain mask | Mask ablation comparing bidirectional, vanilla causal, and S2B-causal attention. |
| Error accumulation in recursive forecasting | Conditional autoregressive rollout with past brain-state feedback | Horizon breakdown and context-length sensitivity. |
| Cross-network interaction without over-smoothing | Spatial Mixer Head | Spatial Mixer ablation and model-based attention diagnostics. |
| Dependence on multimodal sensory drive | Time-aligned multimodal stimulus tokens | Input-source and modality ablations. |

Table 41: **Tokenization granularity ablation on Algonauts 2025 at $H = 20$.**

| Variant | $r$ | $p_{\text{corr}}$ | Cls@Top10 |
|---|---|---|---|
| Continuous fMRI input | 0.151 | 0.118 | 0.673 |
| Shared full-brain tokenizer | 0.184 | 0.166 | 0.721 |
| Network-wise tokenizer | **0.218** | **0.209** | **0.762** |

### E.7 Interaction between Temporal Masking and Spatial Mixing

We jointly vary the temporal mask and the presence of the attention-based Spatial Mixer on Algonauts 2025 at $H = 20$. Table 4 reports the four controlled configurations. The attention mixer decreases $p_{\text{corr}}$ by $0.0153$ under vanilla causal masking but increases it by $0.0278$ under S2B. The corresponding difference-in-differences is

$$(0.2090 - 0.1812) - (0.1420 - 0.1573) = 0.0431.$$

Thus, the benefit of mixing depends on the temporal information-flow constraint in this experiment. These measurements support the combined design in the evaluated setting; they do not establish a universal advantage of either operator or directly measure the proposed error-propagation pathway.

### E.8 Tokenizer Objective and Latent Regularization

The default tokenizer is a deterministic network-specific autoencoder trained with reconstruction loss and frozen before forecasting. It has no additional distributional prior: the bottleneck, network partition, reconstruction objective, and fixed tokenizer–detokenizer mapping constrain the representation. Absence of a latent prior does not mean that the tokens are random projections.

We compare this reconstruction-only objective with three additional constraints on Algonauts 2025. VQ-AE adds codebook and commitment losses and uses discrete latent codes. The $\beta$-VAE adds a KL penalty toward an isotropic Gaussian prior. AE + SoftCLIP adds alignment to frozen stimulus embeddings. The network partition, latent dimensionality, stimulus encoders, forecasting model, and rollout protocol are held fixed. Table 46 reports reconstruction and forecasting results for these evaluated configurations.

Reconstruction-only training has the highest reconstruction fidelity and the highest mean forecasting score at $H = 10$ and $H = 20$. SoftCLIP has a slightly higher mean at $H = 1$ ($+0.0013$), which does not persist at longer horizons. These comparisons support the default objective among the tested configurations; they do not imply that latent regularization is generally ineffective or that other regularization settings cannot improve forecasting.

Table 42: **Temporal mask ablation on Algonauts 2025.** Scores report $p_{\text{corr}}$ under conditional autoregressive rollout; the last row gives the relative improvement of S2B over vanilla causal masking.

| **Mask** | $H = 1$ | $H = 5$ | $H = 10$ | $H = 15$ | $H = 20$ |
|---|---|---|---|---|---|
| Bidirectional attention | 0.545 | 0.361 | 0.245 | 0.169 | 0.113 |
| Vanilla causal mask | 0.552 | 0.380 | 0.270 | 0.195 | 0.142 |
| S2B mask | **0.560** | **0.403** | **0.310** | **0.251** | **0.209** |
| S2B relative gain | 1.4% | 6.1% | 14.8% | 28.7% | 47.2% |

Table 43: **Spatial Mixer Head ablation on Algonauts 2025.** Each entry reports $p_{\text{corr}}$.

| **Variant** | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|
| w/o Spatial Mixer | 0.548 | 0.294 | 0.181 |
| Linear Mixer | 0.553 | 0.302 | 0.195 |
| Spatial Mixer Head | **0.560** | **0.310** | **0.209** |

Table 44: **Tokenizer pretraining vs. end-to-end training on HAD.** Each entry reports $p_{\text{corr}}$.

| **Training strategy** | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|
| End-to-end training | 0.574 | 0.246 | 0.173 |
| Pretrained tokenizer + frozen encoder | **0.621** | **0.339** | **0.295** |

Table 45: **Tokenizer reconstruction diagnostics.** Reconstruction correlation $r$ is reported. This is a standalone tokenizer diagnostic, not end-to-end forecasting performance.

| **Family** | **Objective** | $r$ |
|---|---|---|
| MLP | AE | **0.834** |
| MLP | VAE | 0.507 |
| BrainVAE | AE | 0.814 |
| BrainVAE | VAE | 0.797 |

Table 46: **Tokenizer regularization on Algonauts 2025.** Reconstruction correlation and MSE measure signal fidelity; forecasting columns report mean $p_{\text{corr}}$. Higher correlation and lower MSE are better.

| Tokenizer | Recon. $r$ | Recon. MSE | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|---|---|
| VQ-AE | 0.916 | 0.080 | 0.5384 | 0.2887 | 0.1938 |
| $\beta$-VAE | 0.934 | 0.064 | 0.5521 | 0.3021 | 0.2031 |
| AE + SoftCLIP | 0.947 | 0.052 | **0.5613** | 0.3062 | 0.2044 |
| Reconstruction only | **0.958** | **0.041** | 0.5600 | **0.3100** | **0.2090** |

# F Multimodal Fusion and Cross-subject Transfer

## F.1 Modality Ablation

Table 47 evaluates the contribution of each stimulus modality on Algonauts 2025. The full multimodal setting performs best across horizons, indicating that visual, auditory, and textual streams provide complementary information for conditional brain-state forecasting. Removing video causes the largest degradation, suggesting that visual context contributes strongly to preserving cross-ROI spatial patterns under long-horizon rollout. Single-modality variants further show that multimodal conditioning is more stable than relying on any individual input stream.

## F.2 Frozen Stimulus Encoder Comparison

Table 48 evaluates the effect of replacing the frozen stimulus encoders while keeping the forecasting architecture fixed on Algonauts 2025 at $H = 20$. The main encoder setting performs best, suggesting that temporally specific stimulus representations are important for conditional rollout. The performance drop under alternative encoders also indicates that BrainVista's gains are not solely due to the autoregressive forecasting architecture, but also depend on the quality and temporal alignment of the stimulus representation.

## F.3 Cross-subject Generalizability

Cross-subject transfer evaluates whether the learned forecasting rule generalizes across individuals. We conduct this analysis on Algonauts 2025, where all subjects share the same Schaefer-1000 parcel space and Yeo-7 network assignment. We compare BrainVista with BrainHarmony-AR, the strongest matched-access baseline in the main results. For single-source cross-subject transfer, the model is trained on one source subject and evaluated on a different target subject without target-subject fine-tuning. For leave-one-subject-out transfer, the model is trained on three subjects and evaluated on the held-out subject.

Table 49 shows that BrainVista retains a positive margin over BrainHarmony-AR under both single-source transfer and leave-one-subject-out transfer. However, cross-subject performance remains lower than within-subject training, indicating that atlas-level anatomical alignment does not fully remove inter-subject functional variability. This suggests that future work could combine BrainVista with hyperalignment, subject-specific functional alignment, or lightweight subject-adaptation layers.

The source–target matrices in Tables 50 and 51 show that transfer performance varies across subject pairs. The diagonal entries are highest, as expected, because within-subject training avoids inter-subject functional misalignment. Across off-diagonal entries, BrainVista consistently improves over BrainHarmony-AR, suggesting that the proposed network-wise tokenization and temporally valid conditioning are beneficial even when the source and target subjects differ. Nevertheless, the reduction from within-subject to cross-subject performance highlights a remaining limitation of using a shared atlas-level representation alone.

Table 47: **Modality ablation on Algonauts 2025.** Each entry reports $p_{\mathrm{corr}}$. The full multimodal setting uses visual, audio, and text features.

| **Variant** | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|
| All (V+A+T) | **0.560** | **0.310** | **0.209** |
| w/o Video (A+T) | 0.510 | 0.270 | 0.184 |
| w/o Audio (V+T) | 0.540 | 0.295 | 0.200 |
| w/o Text (V+A) | 0.538 | 0.293 | 0.198 |
| Video only (V) | 0.485 | 0.255 | 0.175 |
| Audio only (A) | 0.455 | 0.235 | 0.165 |
| Text only (T) | 0.440 | 0.228 | 0.160 |

Table 48: **Frozen stimulus encoder ablation on Algonauts 2025 at $H = 20$.** The forecasting architecture is fixed, and only the frozen stimulus encoders are changed.

| **Frozen stimulus encoder(s)** | $r$ | $p_{\mathrm{corr}}$ | Cls@Top10 |
|---|---|---|---|
| Main paper encoders | **0.2183** | **0.2090** | **0.7619** |
| ImageBind | 0.2110 | 0.1990 | 0.7420 |
| Qwen Omni 2.5 | 0.2020 | 0.1880 | 0.7240 |

Table 49: **Cross-subject transfer summary on Algonauts 2025 under $H$=20 forecasting.** Each entry reports subject-level $p_{\mathrm{corr}}$ mean $\pm$ standard deviation. BH-AR denotes BrainHarmony-AR.

| **Training / evaluation setting** | **BH-AR** | **BrainVista** | **Δ** |
|---|---|---|---|
| Within-subject training | $0.1887 \pm 0.0064$ | $\mathbf{0.2090 \pm 0.0098}$ | $+0.0203$ |
| Single-source cross-subject | $0.1583 \pm 0.0108$ | $\mathbf{0.1802 \pm 0.0119}$ | $+0.0219$ |
| Leave-one-subject-out | $0.1698 \pm 0.0106$ | $\mathbf{0.1912 \pm 0.0118}$ | $+0.0214$ |

Table 50: **BrainHarmony-AR cross-subject transfer matrix on Algonauts 2025.** Columns index the training subject, rows index the evaluation subject, and each entry reports $p_{\mathrm{corr}}$. Diagonal entries correspond to within-subject training.

| **Test subject** | **Training subject** | | | |
|---|---|---|---|---|
| | Sub-01 | Sub-02 | Sub-03 | Sub-05 |
| Sub-01 | 0.1814 | 0.1628 | 0.1660 | 0.1542 |
| Sub-02 | 0.1605 | 0.1920 | 0.1640 | 0.1558 |
| Sub-03 | 0.1650 | 0.1586 | 0.1856 | 0.1516 |
| Sub-05 | 0.1533 | 0.1560 | 0.1518 | 0.1958 |

Table 51: **BrainVista cross-subject transfer matrix on Algonauts 2025.** Columns index the training subject, rows index the evaluation subject, and each entry reports $p_{\mathrm{corr}}$. Diagonal entries correspond to within-subject training.

| **Test subject** | **Training subject** | | | |
|---|---|---|---|---|
| | Sub-01 | Sub-02 | Sub-03 | Sub-05 |
| Sub-01 | 0.1970 | 0.1850 | 0.1881 | 0.1717 |
| Sub-02 | 0.1824 | 0.2109 | 0.1859 | 0.1761 |
| Sub-03 | 0.1875 | 0.1813 | 0.2074 | 0.1738 |
| Sub-05 | 0.1762 | 0.1790 | 0.1758 | 0.2207 |

# G Model-based Neuroscience Analyses

## G.1 Network-level Rollout Fidelity

Table 52 reports Yeo-7 network-averaged $p_{\mathrm{corr}}$ under $H = 10$ rollout. Visual and default-mode systems show higher fidelity, whereas limbic regions remain more difficult to predict. This pattern is consistent across subjects and suggests that prediction difficulty varies systematically across functional systems.

## G.2 History-token Occlusion

To estimate temporal reliance, we mask the history tokens corresponding to a specific lag while keeping all other inputs unchanged. We then measure the increase in prediction loss relative to the unoccluded setting. Larger loss increases indicate stronger reliance on that lagged context. This analysis is a model-based diagnostic, not a causal neuroscience intervention. It is intended to measure how the trained model distributes temporal dependence across the available context window.

## G.3 Spatial Mixer Attention

We extract attention weights from the Spatial Mixer Head and aggregate them over time and heads to obtain a compact network-to-network interaction summary. The resulting attention patterns should be interpreted as model-based summaries of learned cross-network interaction structure, not as direct biological causality. We therefore describe the observed pattern as *heterogeneous and selective* rather than as a binary statement that some networks are coupled and others are uncoupled.

## G.4 Additional Subject-level Visualizations

Figure 5–Figure 7 provide additional subject-level diagnostics on Algonauts 2025. These visualizations extend the main-text analysis by applying the same diagnostic pipeline to additional subjects and diagnostic views. The goal is to verify that the temporal-reliance and cross-network interaction patterns discussed in the main paper are not driven by a single favorable subject.

**Parcel-wise Rollout Fidelity.** For each subject, we compute parcel-wise rollout fidelity by correlating the predicted and ground-truth trajectories at each parcel under conditional autoregressive rollout. The resulting maps show that forecasting quality is spatially heterogeneous rather than uniformly distributed across cortex. Sensory and default-mode regions generally show higher fidelity, while limbic and some association regions are harder to predict. This pattern is consistent with the network-level results in Table 52.

**Network-level Aggregation.** We aggregate parcel-wise scores to Yeo-7 functional systems to obtain a compact summary of system-level predictability. Across subjects, visual and default-mode networks tend to preserve higher $p_{\mathrm{corr}}$, whereas limbic regions remain consistently lower. This supports the interpretation that BrainVista captures stable system-level differences in forecasting difficulty, rather than only improving global average scores.

**History-token Occlusion.** For each lag in the context window, we mask the corresponding history token and measure the increase in prediction loss relative to the unoccluded setting. The resulting heatmaps show that the model relies strongly on recent history, while several association networks retain broader sensitivity over longer temporal lags. This analysis is diagnostic only: it measures model reliance on historical tokens and should not be interpreted as direct evidence of biological causality.

**Spatial Mixer Attention.** We further visualize the Spatial Mixer attention after aggregating attention weights over time and heads. The resulting chord diagrams show heterogeneous and selective cross-network interaction patterns. The analysis does not imply that the learned attention weights are biological connectivity estimates. Instead, they provide a model-based summary of which network pairs are emphasized by the forecasting model when producing future brain-region tokens.

Table 52: **Yeo-7 network-level $p_{\text{corr}}$ under $H = 10$ rollout.**

| **Subject** | Visual | SomMot | DorsAttn | VentrAttn | Limbic | FPN/Control | Default |
|---|---|---|---|---|---|---|---|
| sub01 | 0.3512 | 0.2830 | 0.3131 | 0.2673 | 0.1807 | 0.2690 | 0.3237 |
| sub02 | 0.3381 | 0.2605 | 0.2977 | 0.2505 | 0.1695 | 0.2595 | 0.3092 |
| sub03 | 0.3420 | 0.2726 | 0.3020 | 0.2581 | 0.1715 | 0.2592 | 0.3127 |
| sub05 | 0.3275 | 0.2520 | 0.2835 | 0.2415 | 0.1585 | 0.2465 | 0.3015 |

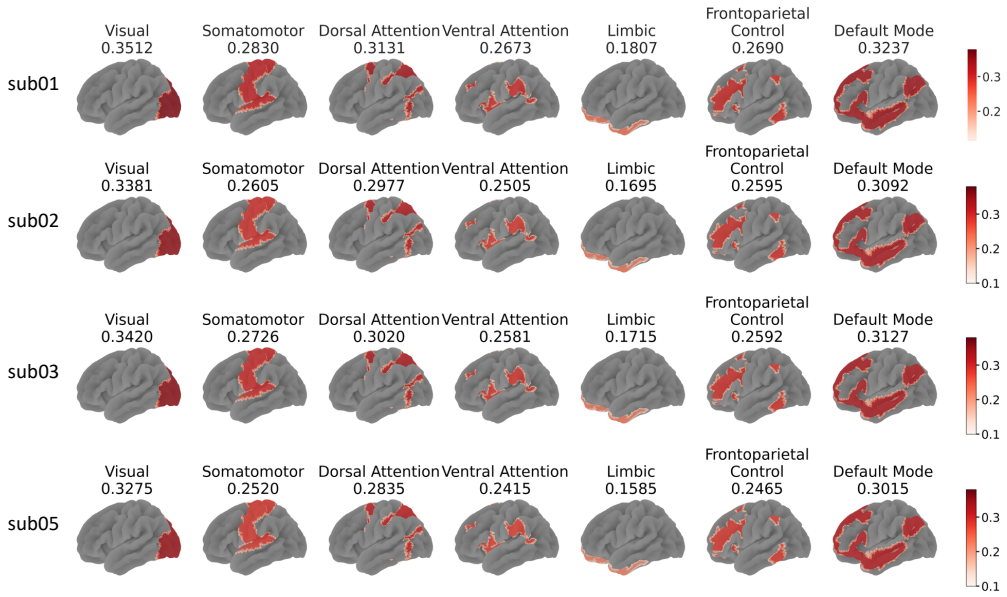


Figure 5: **Additional subject-level model diagnostics on Algonauts 2025.** We visualize parcel-wise rollout fidelity and Yeo-7 network-level aggregation under conditional autoregressive rollout. The results show spatially structured prediction quality and stable differences across functional systems.

Overall, the additional visualizations show that the main qualitative findings persist across subjects: rollout fidelity is spatially structured, temporal reliance is not uniform across functional systems, and the Spatial Mixer learns non-uniform cross-network interactions. These diagnostics support the claim that BrainVista improves forecasting while preserving interpretable system-level structure.

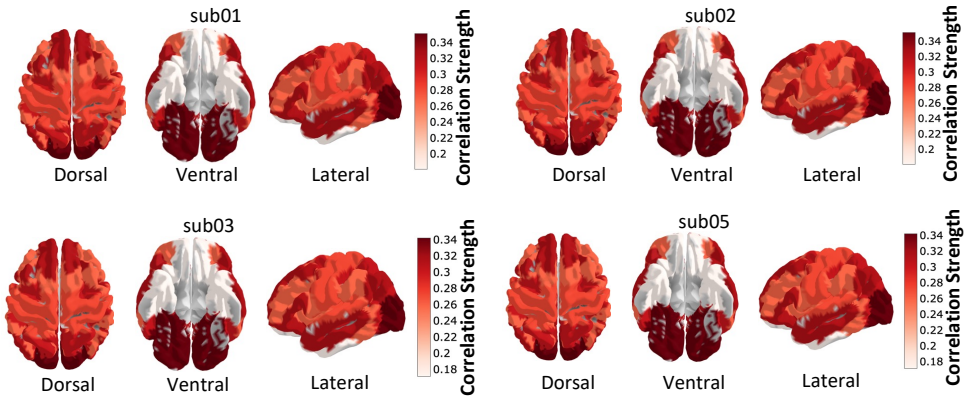


Figure 6: **Additional temporal-reliance diagnostics on Algonauts 2025.** History-token occlusion measures the increase in prediction loss after masking lagged brain-state tokens. The results show that recent context is important across systems, while higher-order systems retain broader sensitivity to longer temporal histories.

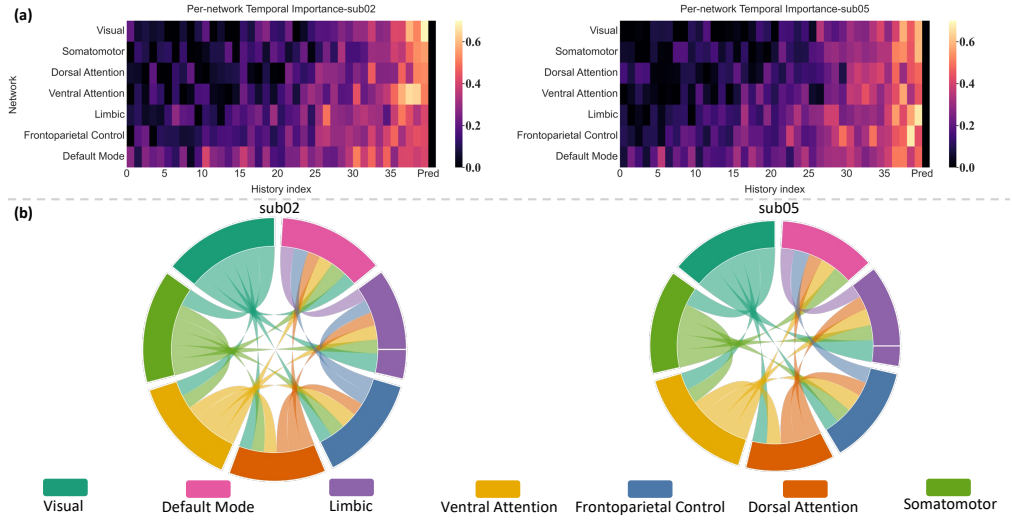


Figure 7: **Additional Spatial Mixer attention diagnostics on Algonauts 2025.** Spatial Mixer attention is aggregated over time and heads to summarize model-based cross-network interaction patterns. The chord diagrams indicate heterogeneous and selective network interactions rather than uniform global mixing.

## H Downstream Utility: Emotion-category Structure

To evaluate whether predicted brain features preserve behaviorally meaningful structure beyond trajectory-level metrics, we conduct an auxiliary emotion-category clustering analysis on Algonauts 2025 held-out movie segments. This analysis is intended as a representational diagnostic rather than a clinical, demographic, or individual-level prediction task.

**Protocol.** We divide held-out movie runs into non-overlapping 10-second clips. For each clip, we obtain an emotion-category label from the corresponding multimodal stimulus content using the same external emotion-labeling pipeline for all methods. We retain nine emotion categories with sufficient clip counts: happy, sad, neutral, angry, worried, surprise, fear, contempt, and doubt. Clips with low-confidence or ambiguous labels are excluded from this auxiliary analysis. For each method, predicted brain features are averaged over the fMRI frames corresponding to each 10-second clip. The resulting clip-level predicted brain vector is reduced with PCA before clustering. Unless otherwise specified, we retain 50 principal components.

Table 53: **Emotion-category clustering from predicted brain features on Algonauts 2025.** Higher ARI and NMI indicate better agreement between predicted brain-feature clusters and stimulus-derived emotion categories.

| **Feature** | ARI | NMI |
|---|---|---|
| Ground-truth brain feature | 0.318 | 0.392 |
| BrainHarmony-AR | 0.138 | 0.214 |
| BrainVista | **0.214** | **0.286** |

Table 54: **Per-class purity in emotion-category clustering on Algonauts 2025.** Purity is computed between K-means clusters from predicted brain features and stimulus-derived emotion categories.

| **Emotion category** | BrainHarmony-AR | BrainVista |
|---|---|---|
| Happy | 0.46 | **0.58** |
| Sad | 0.39 | **0.50** |
| Neutral | 0.58 | **0.67** |
| Angry | 0.42 | **0.54** |
| Worried | 0.34 | **0.44** |
| Surprise | 0.39 | **0.49** |
| Fear | 0.36 | **0.46** |
| Contempt | 0.26 | **0.35** |
| Doubt | 0.29 | **0.38** |
| Mean | 0.39 | **0.49** |

**Clustering Metrics.** We apply K-means clustering with $K = 9$, matching the number of retained emotion categories. K-means is run with 10 random initializations, and the best solution by within-cluster inertia is used. We evaluate cluster–label agreement using adjusted Rand index (ARI), normalized mutual information (NMI), and per-class purity. ARI and NMI measure global agreement between clusters and emotion labels, while per-class purity measures whether clips from each emotion category concentrate in a dominant cluster.

BrainVista improves ARI, NMI, and mean per-class purity over BrainHarmony-AR, suggesting that its predicted brain features preserve more stimulus-related affective structure. The ground-truth brain-feature upper reference remains higher, indicating that the clustering analysis is not saturated and that additional information is still lost during forecasting. We therefore treat this result as an auxiliary analysis of representational utility rather than as a primary evaluation objective.

**Relation to Downstream Prediction.** Some downstream tasks, such as demographic prediction, disease diagnosis, or clinical progression modeling, would require subject-level labels that are not consistently available in the naturalistic multimodal fMRI datasets used here. We therefore do not claim clinical, demographic, or individual-level predictive validity. The emotion-category clustering analysis instead provides a lightweight representational utility test, evaluating whether predicted brain features preserve affect-related structure associated with naturalistic stimuli.

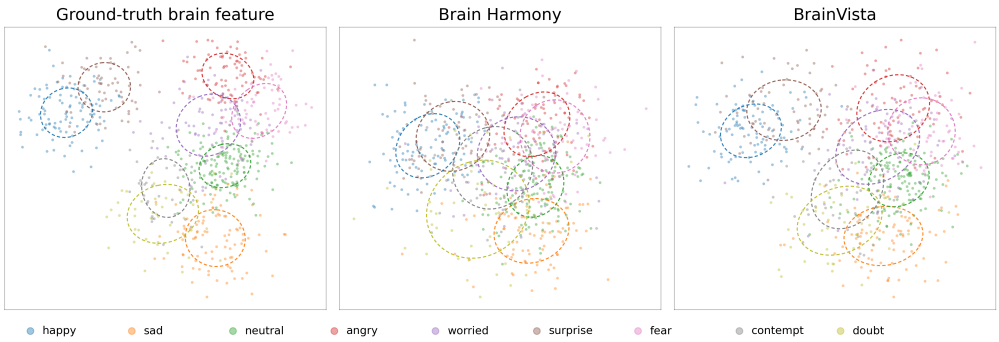


Figure 8: **Emotion-category structure from predicted brain features on Algonauts 2025.** Predicted brain features are averaged over 10-second clips, reduced with PCA, and clustered with K-means. Compared with BrainHarmony-AR, BrainVista produces clusters that better align with stimulus-derived emotion-category labels, indicating stronger preservation of affect-related representational structure.

# I Data and Model Scaling

We study scaling on Algonauts 2025 with fixed pretrained Network-wise Tokenizer–detokenizer pairs, frozen stimulus encoders, validation and test splits, optimization protocol, and rollout evaluation. The purpose is to distinguish additional training data from increased temporal-predictor capacity. The Base model is the configuration used in the main experiments.

## I.1 Training-data Scaling

We subsample complete training units, preserving run boundaries rather than sampling overlapping windows across split boundaries. A unit is a complete run or episode used by the training pipeline; an anchor is a valid within-unit context–target position as defined in Appendix A.6. The retained unit and anchor counts per subject are given in Table 55. Validation and test sets remain fixed across fractions.

Table 55: **Training-data scaling on Algonauts 2025.** Forecasting columns report $p_{\mathrm{corr}}$. Relative performance is the $H = 20$ score divided by the full-data score. Anchor counts are rounded.

| Data | Units/subject | Anchors/subject | $H = 1$ | $H = 10$ | $H = 20$ | Relative |
|---|---|---|---|---|---|---|
| 25% | 7 | 2.9k | 0.5174 | 0.2589 | 0.1698 | 81.2% |
| 50% | 14 | 5.9k | 0.5408 | 0.2818 | 0.1874 | 89.7% |
| 75% | 21 | 8.8k | 0.5528 | 0.2978 | 0.2003 | 95.8% |
| 100% | 28 | 11.8k | **0.5600** | **0.3100** | **0.2090** | 100% |

Performance increases monotonically across the four data fractions at all three horizons. At $H = 20$, $p_{\mathrm{corr}}$ increases from $0.1698$ to $0.2090$. This is evidence for a positive data trend over the tested range, not an estimate of a power-law exponent.

## I.2 Model Capacity and Training–Validation Gap

Table 56 compares Small, Base, and Large configurations. Only the predictor capacity is varied; the pretrained tokenizers and stimulus encoders are fixed.

Increasing capacity from Small to Base improves forecasting. Increasing it further to Large improves $H = 20$ by only $0.0028$, while reported computation increases from $2.1$ to $4.9$ GPU-hours per subject. Table 57 examines training and validation performance at the selected checkpoint.

The Large model fits training data better, but its validation improvement is small, its training–validation gap is larger, and its best epoch occurs earlier. These observations are consistent with diminishing returns and emerging overfitting on this dataset. Four data fractions and three model sizes

Table 56: **Model-capacity scaling on Algonauts 2025.** Scores are $p_{\mathrm{corr}}$. Compute is reported in GPU-hours per subject.

| Model | GPU-h/subject | $H = 1$ | $H = 10$ | $H = 20$ |
|---|---|---|---|---|
| Small | 1.4 | 0.5417 | 0.2913 | 0.1938 |
| Base | 2.1 | 0.5600 | 0.3100 | 0.2090 |
| Large | 4.9 | **0.5641** | **0.3134** | **0.2118** |

Table 57: **Training–validation gaps in the capacity comparison.** Scores are $p_{\mathrm{corr}}$ at $H = 20$; the gap is training minus validation performance. Best epoch denotes the selected epoch out of 40.

| Model | Train | Validation | Gap | Best epoch/40 | Test |
|---|---|---|---|---|---|
| Small | 0.2178 | 0.1902 | 0.0276 | 38 | 0.1938 |
| Base | 0.2611 | 0.2047 | 0.0564 | 34 | 0.2090 |
| Large | 0.3094 | 0.2069 | 0.1025 | 27 | 0.2118 |

on one dataset do not establish a general neural scaling law. The results motivate testing additional naturalistic fMRI data before assuming that larger predictors will continue to improve performance.

# J Computational Cost and Reproducibility

Table 58 summarizes model dimensions, training settings, and computational cost. Because $d_{\mathrm{brain}}$ depends on the model-space brain representation, the brain projection layers are dataset-dependent.

Table 58: **Model configuration and computational cost by dataset.** The temporal predictor uses the same hidden dimension across datasets; brain input/output projection dimensions follow the dataset-specific latent representation.

| Component / metric | Algonauts 2025 | CineBrain | HAD |
|---|---|---|---|
| Raw brain dimension $\sum_k Q_k$ | 1,000 | 18,946 | fsLR surface |
| Model-space input dimension $\sum_k P_k$ | 1,000 | 1,024 | 1,000 |
| Concatenated brain latent $d_{\mathrm{brain}}$ | 803 | 820 | 803 |
| Transformer hidden size $d_{\mathrm{model}}$ | 256 | 256 | 256 |
| Brain projection | $803 \leftrightarrow 256$ | $820 \leftrightarrow 256$ | $803 \leftrightarrow 256$ |
| Peak training memory | 24.0 GB | 26.8 GB | 24.3 GB |
| Inference time for $H = 20$ rollout | 0.18 s | 0.20 s | 0.18 s |
| Default context length | $L = 40$ | $L = 40$ | $L = 40$ |
| Default rollout horizon | $H = 20$ | $H = 20$ | $H = 20$ |
| Optimizer | AdamW | AdamW | AdamW |
| Learning rate | $2 \times 10^{-4}$ | $2 \times 10^{-4}$ | $2 \times 10^{-4}$ |
| Weight decay | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ |
| Batch size | 32 | 32 | 32 |
| Training epochs | 40 | 40 | 40 |
| Mixed precision | PyTorch AMP | PyTorch AMP | PyTorch AMP |

All experiments were conducted on a single NVIDIA GeForce RTX 5090 GPU using PyTorch automatic mixed precision. The reported wall-clock training time includes both network-wise tokenizer pretraining and temporal predictor training. Table 23 summarizes the compute resources used for each dataset. All models fit within a single-GPU training setup without model parallelism or distributed training.

# K Limitations and Assumptions

BrainVista assumes that the exogenous stimulus stream is observed during forecasting. Therefore, the task is conditional brain-state forecasting rather than unconditional brain simulation. The model also assumes that past fMRI context is available for rollout initialization. Forecasting accuracy decreases as the horizon grows because prediction errors accumulate recursively. In addition, association systems remain harder to predict than sensory systems, and the present analyses do not isolate the causes of this difference.

The Spatial Mixer Head provides useful model-based diagnostics of learned cross-network interactions, but its attention weights should not be interpreted as direct biological causality. All main results use training-set normalization to avoid evaluation-time leakage from validation or test runs. CineBrain uses training-set PCA to reduce high-dimensional voxel-level responses before tokenization, and the resulting PCA basis is fixed for validation and test runs. Finally, the three evaluated datasets differ in stimulus protocol, temporal resolution, and preprocessing pipeline. Models are trained separately on each dataset. These results evaluate held-out runs or episodes under the official splits, not arbitrary unseen movies, and do not establish zero-shot transfer across scanners or protocols. Cross-subject transfer in Appendix F has lower absolute accuracy than within-subject training.

BrainVista is a within-run predictive assay for comparing dynamical assumptions and stimulus representations against withheld fMRI. It does not establish accurate cross-session trajectory simulation, pre-scan stimulus evaluation, online experimental control, clinical diagnosis, or individual-level decision-making validity. History occlusion measures conditional predictive reliance, not a biological intervention; attention weights do not establish direct connectivity. The emotion analysis uses stimulus-derived labels and is an auxiliary representational diagnostic, not a measure of participants' subjective emotional states. Neuroimaging data remain subject to their original access and privacy conditions; forecasted responses should not be used to infer sensitive individual attributes.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction define the scope as observed-stimulus conditional autoregressive forecasting of naturalistic fMRI responses from past brain states and observed multimodal stimuli. The claimed improvements are supported by the main experiments in Section 4, with additional horizon, robustness, and statistical analyses in Appendix C.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: The limitations are discussed in Section 5 and Appendix K. We state that BrainVista assumes observed exogenous stimuli and an available past fMRI context, that long-horizon error accumulation remains unavoidable, that attention-based mixer diagnostics should not be interpreted as biological causality, and that dataset heterogeneity limits claims of distribution-free generalization.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions.
   - The authors should reflect on the scope of the claims made.
   - The authors should reflect on the factors that influence the performance of the approach.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

   Justification: The paper does not claim formal theoretical guarantees or present theorem-style results. The mathematical expressions in the method section define the forecasting objective, tokenization procedure, S2B masking, and conditional autoregressive rollout protocol rather than proving theoretical bounds.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper?

   Answer: [Yes].

   Justification: Section 4 specifies the evaluated datasets, baselines, metrics, training protocol, and conditional autoregressive rollout setup. Appendix A provides dataset metadata, temporal alignment, rollout rules, hyperparameters, and metric definitions, while Appendix B describes how baselines are adapted to the unified forecasting protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility.

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

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

   Answer: [Yes].

   Justification: The evaluated datasets are existing de-identified fMRI datasets cited in the paper and remain subject to their original access conditions. The BrainVista code is publicly available at `https://github.com/xuanhuayin/BrainVista`, including the core implementation and the data-preparation, training, evaluation, and rollout-inference entry points, with shell run scripts whose command-line arguments specify the main hyperparameters. During execution, the training scripts save the resolved runtime settings, such as `predictor_cfg.json` and `tokenizer_cfg.json`, into the checkpoint directory as reproducibility snapshots. Baseline adaptation details are documented in the paper and appendices rather than released as separate scripts. The repository provides instructions for dataset preparation and running the reported forecasting protocol, follows dataset-specific access requirements, and does not redistribute restricted raw fMRI data or individual identifiers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer.
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - The authors should provide instructions on data access and preparation.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines.

- At submission time, to preserve anonymity, the authors should release anonymized versions if applicable.
- Providing as much information as possible in supplemental material is recommended.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes].

   Justification: Section 4 describes datasets, modalities, baselines, metrics, and training/evaluation protocol. Appendix A further specifies dataset metadata, run-level splitting, temporal alignment, hyperparameters, optimizer settings, context length, rollout horizon, sequence-boundary handling, and metric definitions.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes].

   Justification: Section 4.2 summarizes five-seed variability, paired subject-bootstrap intervals, and win counts. Table 18 reports the five-seed standard deviations of $p_{\mathrm{corr}}$ for BrainHarmony-AR and BrainVista; Appendix C reports the complete subject-level statistics and paired significance tests at $H{=}20$. The text specifies that paired comparisons are computed at the subject level and that results on datasets with few subjects should be interpreted cautiously.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated.
   - The method for calculating the error bars should be explained.
   - The assumptions made should be given.
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it.
   - For asymmetric distributions, the authors should be careful not to show symmetric error bars that would yield results that are out of range.
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix J reports model dimensions, peak training memory, inference time for $H{=}20$ rollout, optimizer settings, training epochs, batch size, and mixed-precision configuration. These details provide the compute scale required to reproduce the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes].

   Justification: The work analyzes existing de-identified fMRI datasets and does not collect new human-subject data. Appendix K states that BrainVista is not intended for clinical diagnosis, demographic inference, disease prediction, or individual-level decision making. It also cautions against interpreting attention-based model diagnostics as biological causality.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity.

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: The potential positive impact is improved modeling of naturalistic brain dynamics and better tools for studying stimulus-driven fMRI responses. Appendix K also discusses negative or risky interpretations, including overinterpreting attention-based diagnostics as biological causality, using the method for clinical or demographic prediction, and privacy concerns associated with neuroimaging data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses, fairness considerations, privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, when it gives incorrect results, and from misuse.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A].

    Justification: The paper does not release a new high-risk pretrained model, scraped dataset, or neuroimaging dataset. The method is evaluated on existing datasets under their original access conditions, and we do not redistribute raw fMRI data or individual identifiers.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards.
- Datasets that have been scraped from the Internet could pose safety risks.
- We recognize that providing effective safeguards is challenging, and many papers do not require this.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited, and are the license and terms of use explicitly mentioned and respected?

    Answer: [Yes].

    Justification: Appendix A.14 summarizes the existing datasets, pretrained encoders, atlases, and preprocessing tools used in this work, including their usage and reported license or access terms. We cite the original sources and do not redistribute third-party datasets, pretrained checkpoints, or raw stimulus materials.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license should be included for each asset.
    - For scraped data from a particular source, the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes].

    Justification: The code at `https://github.com/xuanhuayin/BrainVista` is released under the MIT License, documents BrainVista training, evaluation, and rollout inference in its README, and provides shell scripts that encode the reported settings through CLI arguments, and the training scripts dump resolved runtime configuration snapshots into the checkpoint directory. Baseline adaptation details are documented in the paper and appendices rather than released as separate scripts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets if applicable.

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

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

    Answer: [N/A].

    Justification: The paper does not conduct new crowdsourcing experiments or recruit new human participants. It uses existing de-identified fMRI datasets and cites the original dataset papers for acquisition protocols and participant procedures.

Guidelines:

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

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed, and whether Institutional Review Board approvals or equivalent approvals were obtained?

    Answer: [N/A].

    Justification: The paper does not collect new human-subject data. It analyzes existing de-identified fMRI datasets, and the original dataset publications document the data-collection protocols and participant-related approvals or consent procedures where applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval or equivalent may be required for any human subjects research.
    - We recognize that the procedures for this may vary significantly between institutions and locations.
    - For initial submissions, do not include any information that would break anonymity.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes].

    Justification: The paper uses Llama 3.2 3B as a frozen text encoder for constructing text stimulus tokens, as described in Section 4. Its role in the forecasting model is to provide frozen text features for multimodal stimulus alignment. Appendix H separately describes externally generated stimulus-derived labels in the auxiliary emotion analysis.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to the NeurIPS LLM policy in the handbook for what should or should not be described.