WHEN MODEL COMPLEXITY MATTERS IN CROSS-SUBJECT FMRI PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Increasingly complex neural models predict blood-oxygen-level-dependent (BOLD) responses to naturalistic stimuli, but the predictive value of this added complexity remains unclear. We ask when pretrained stimulus representations help and how much downstream complexity is needed, with and without recent measured BOLD history. We compare linear diagnostic references and nonlinear predictors across HCP movie fMRI (154 subjects), BOLD Moments (10 subjects), and Algonauts 2025 (4 subjects). Under a shared offline protocol, models are fitted on training subjects and evaluated on held-out subjects viewing largely familiar stimuli, without refitting. Pretrained visual features generally improve prediction over raw pixels; for linear autoregression, their advantage grows as measured-history refreshes become less frequent. Measured BOLD history provides substantial gains that generally diminish with longer refresh intervals. Nonlinear stimulus-only predictors can improve on linear readouts, whereas linear autoregression provides a strong history-conditioned baseline. Controls examine context length, regularization, fitting procedures, and model capacity. Regional response consistency across subjects is associated with held-out prediction accuracy, while less consistent regions tend to gain more from measured history. In paired simulations with fixed linear neural dynamics, slower hemodynamic responses generally sustain high temporal prediction correlations over longer refresh intervals, illustrating how the observation process can contribute to predictive persistence. Together, these findings support evaluating stimulus representations and downstream complexity separately, using history-conditioned linear prediction as a diagnostic reference under matched inputs and measurement availability.

1 INTRODUCTION

Naturalistic fMRI prediction frameworks are becoming increasingly complex, but it remains unclear where that complexity actually pays off. Movies combine visual, auditory, and linguistic information over time, eliciting distributed and dynamic BOLD responses (Hasson et al., 2004). Recent brain foundation models and related neural architectures bring richer stimulus representations and increasingly sophisticated predictors to this problem (d'Ascoli et al., 2025; 2026; Gokce et al., 2026; Yin et al., 2026; Dong et al., 2026). We ask a simple question: *where and when is this additional model complexity actually useful?*

Cross-subject prediction makes this question especially meaningful. People watching the same movie share stimulus-locked responses, yet their BOLD trajectories differ across individuals (Hasson et al., 2004; Finn et al., 2020). Predicting an unseen subject therefore remains challenging (Wen et al., 2018a; d'Ascoli et al., 2026). Transfer learning and target-subject adaptation reduce data requirements but still require subject-specific fitting (Wen et al., 2018a; Mai et al., 2025), which may be impractical when paired naturalistic stimulus–fMRI data are scarce. We ask whether recent measured BOLD can inform predictions for held-out subjects without refitting. This is neuroscientifically plausible because spontaneous BOLD activity carries individual-specific structure and contributes to variability in task-evoked responses (Tavor et al., 2016; Fox et al., 2006).

Model complexity enters this problem in two distinct places: stimulus representation and downstream prediction (Figure 1). One line of work enriches the *stimulus representation*: TRIBE and TRIBE v2 combine pretrained video, audio, and language encoders, while MIRAGE uses a natively multimodal backbone with learned feature aggregation (d'Ascoli et al., 2025; 2026; Gokce et al., 2026). A

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

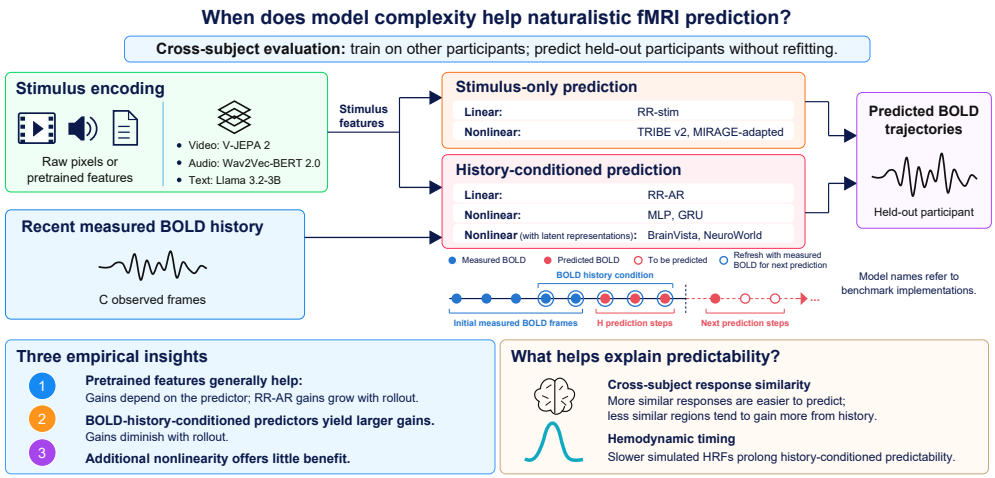

Figure 1: Overview of cross-subject naturalistic fMRI prediction and main findings.

second line develops the *downstream predictor*, whose models further differ in what they observe. *Stimulus-only* predictors, such as TRIBE v2 and MIRAGE, predict entirely from shared stimulus representations and therefore cannot condition on subject-specific deviations in the evoked trajectory. *History-conditioned* predictors additionally use recent BOLD from the target subject. BrainVista autoregressively predicts future brain tokens from observed fMRI under stimulus conditioning (Yin et al., 2026), whereas NeuroWorld propagates a learned latent brain state from an observed fMRI prefix (Dong et al., 2026). Although not designed specifically for no-refit cross-subject prediction, these formulations naturally admit recent BOLD as subject-specific input. Together with generic MLP and GRU predictors, they span a broad range of nonlinear history-conditioned approaches whose performance can vary as predictions are rolled farther from the last BOLD refresh.

These two sources of model complexity play different roles. Richer stimulus representations may better capture the sensory, semantic, and temporal structure of naturalistic experience (Huth et al., 2012; de Heer et al., 2017). Downstream prediction concerns a noisy, macroscopic BOLD signal shaped by neurovascular coupling and hemodynamic filtering (Logothetis et al., 2001; Nozari et al., 2024). The complexity of neural computation therefore does not directly determine the model complexity needed to predict BOLD trajectories (Nozari et al., 2024).

To determine where complexity pays off, we compare raw pixels with pretrained visual features and linear with nonlinear downstream predictors, with and without measured BOLD history. Our diagnostic references are stimulus-only ridge regression (RR-stim) and history-conditioned ridge autoregression (RR-AR). Nonlinear comparisons include stimulus-only TRIBE v2 and MIRAGE, and history-conditioned MLP, GRU, BrainVista, and NeuroWorld; the MLP and GRU match RR-AR's parameter count. We evaluate all models on HCP movie fMRI, BOLD Moments, and Algonauts 2025 under a shared cross-subject protocol without refitting on held-out subjects. We then relate regional prediction accuracy and history gains to cross-subject response similarity. We further construct controlled audiovisual simulations that transform neural activity into BOLD through an explicit biophysical hemodynamic model. These provide a mechanistic test bed for asking how the fMRI measurement itself shapes the persistence of history-conditioned predictability.

Across these controlled comparisons, **three key insights emerge**:

- **Pretrained stimulus representations generally improve prediction.** Gains over raw pixels are clearest for stimulus-only predictors and RR-AR, whose gains increase with rollout length.
- **BOLD history yields large prediction gains at shorter refresh intervals.** It provides predictive information beyond stimulus features alone, with the largest average gains under frequent BOLD refreshes and diminishing gains with rollout.
- **The value of downstream nonlinearity depends on history availability.** Nonlinear stimulus-only predictors can outperform linear readouts, whereas RR-AR generally leads the tested history-conditioned comparisons.

Regional analyses further relate predictability to cross-subject response similarity, while controlled simulations show that slower hemodynamic time scales can prolong the predictive value of recent

history. Together, these results clarify when model complexity improves cross-subject naturalistic fMRI prediction and illustrate how the hemodynamic observation process can prolong history-conditioned predictability.

2 Background and Related Work

Stimulus encoding. Visual encoding has progressed from hand-designed features to deep visual representations (Kay et al., 2008; Wen et al., 2018b). TRIBE and TRIBE v2 combine pretrained video, audio, and language encoders, whereas MIRAGE uses a natively multimodal backbone with learned feature aggregation (d'Ascoli et al., 2025; 2026; Gokce et al., 2026). Following TRIBE v2, we use V-JEPA 2, Wav2Vec-BERT 2.0, and Llama 3.2-3B for video, audio, and text, respectively (Assran et al., 2025; Seamless Communication et al., 2023; Meta AI, 2024). We additionally evaluate two popular visual encoders, DINOv2 (Oquab et al., 2023) and CLIP (using an OpenCLIP checkpoint) (Cherti et al., 2023), in visual-only comparisons on HCP movie fMRI. These comparisons test whether pretrained visual features improve prediction over raw pixels across downstream predictors; a comprehensive study of encoder architectures and pretraining strategies remains beyond our scope.

Downstream prediction. Downstream predictors differ in their inputs and mappings to BOLD. Stimulus-only predictors map stimulus features to responses using linear readouts or nonlinear temporal models, while history-conditioned predictors additionally use recent measured BOLD. BrainVista tokenizes BOLD for autoregressive prediction, whereas NeuroWorld learns stimulus-conditioned latent dynamics and trains a decoder on latent rollouts (Yin et al., 2026; Dong et al., 2026). These history-conditioned architectures complement stimulus-only temporal predictors such as TRIBE v2 and MIRAGE (Appendix Table 3).

Linear prediction of BOLD dynamics. Linear models have performed strongly in resting-state prediction (Nozari et al., 2024) and movie-fMRI prediction with history input alone (Paugam et al., 2024). Nozari et al. (2024) examine averaging, observation noise, and finite samples as possible explanations, while Paugam et al. (2024) find that linearized graph-based predictors outperform their nonlinear counterparts. We build on these findings by testing whether downstream nonlinearity adds predictive value when pretrained stimulus representations and measured BOLD history are combined for cross-subject prediction.

3 Experimental Setup

3.1 Predictors and diagnostic references

We compare eight predictors (Figure 1; Table 3): 3 stimulus-only models use stimulus features alone, and 5 history-conditioned predictors also use the latest C measured BOLD frames to predict the next H steps before refreshing their history input with newly measured BOLD. The comparison includes diagnostic references and recent neural architectures adapted to a shared cross-subject protocol.

Linear and nonlinear diagnostic references. We construct two linear references. Following the linear readout approach used in stimulus encoding (Wen et al., 2018b), stimulus-only ridge regression (RR-stim) maps stimulus features directly to BOLD. Stimulus-conditioned ridge autoregression (RR-AR) additionally incorporates recent BOLD history. Let $\mathbf{y}_t \in \mathbb{R}^R$ denote the parcel-level BOLD vector and $\mathbf{x}_t \in \mathbb{R}^D$ the aligned stimulus representation. RR-AR predicts

$$\hat{\mathbf{y}}_t = \mathbf{b} + \sum_{\ell=1}^{C} A_\ell \tilde{\mathbf{y}}_{t-\ell} + B\mathbf{x}_t, \qquad A_\ell \in \mathbb{R}^{R\times R}, \quad B \in \mathbb{R}^{R\times D}. \tag{1}$$

Here C is the history length, $\tilde{\mathbf{y}}$ contains measured BOLD within the supplied context and predicted BOLD thereafter, and each A_ℓ allows cross-parcel prediction. RR-stim omits the history terms and is fitted separately. Both models minimize one-step squared error with an ℓ_2 penalty and an unpenalized intercept. Penalties are selected by multi-step validation BOLD MSE; fitting details are in Appendix B.3.

To test additional downstream nonlinearity, we replace the linear mapping with an MLP or GRU receiving the same stimulus features and BOLD-history window as RR-AR. Both directly predict parcel-level BOLD trajectories and feed their predictions back during rollout. Their parameter counts match RR-AR within 1%, and they are trained on multi-step BOLD prediction error.

Recent neural models and benchmark adaptations. We evaluate stimulus-only TRIBE v2 and MIRAGE, and history-conditioned BrainVista and NeuroWorld under the shared cross-subject protocol. TRIBE v2 and MIRAGE use released downstream code; BrainVista and NeuroWorld are implemented based on their published descriptions, retaining their learned BOLD representations and staged training procedures. All downstream neural models are initialized from scratch and trained on the common frozen stimulus features. For MIRAGE, we replace its native Qwen-Omni representations and subject-specific heads with the common features and a shared parcel readout. Model names throughout refer to these benchmark implementations. Model-specific pretraining objectives are retained where applicable, while final prediction checkpoints are selected using the same complete-coverage multi-step validation BOLD MSE used to select ridge penalties. Stimulus features are extracted offline, and TRIBE v2 and MIRAGE use bidirectional attention within the supplied stimulus windows. Feature extraction and temporal alignment are detailed in Appendix A.4, while architectures, adaptations, and training schedules are described in Appendix B.

3.2 Datasets and stimulus representations

We evaluate cross-subject prediction on three naturalistic fMRI datasets to test whether the predictive value of model complexity generalizes across subject sample sizes, stimulus types, and stimulus durations: HCP movie fMRI (Human Connectome Project, 2026), BOLD Moments (Lahner et al., 2024), and Algonauts 2025 (Gifford et al., 2025). These provide complementary settings, spanning four movie runs in a large subject cohort, short visual events, and extensive television/movie recordings from a small cohort. For HCP and BOLD Moments, BOLD time series are extracted using the Schaefer-Yeo 1,000-parcel atlas (Schaefer et al., 2018); Algonauts retains its native 1,000-parcel targets. Table 1 summarizes the dataset designs and cross-subject splits; Appendix Table 2 provides additional acquisition and preprocessing details.

Table 1: Dataset designs and cross-subject splits (HCP: train/validation/test; LOSO: leave-one-subject-out). Full details are in Appendix A.

Dataset	Subjects	Subject split	TR (s)	Unique movie duration	Train–test stimulus overlap	Scored window (min)
HCP four movies	154	107/15/32	1.00	49.4 min	100%	1.1–4.3 / clip
BOLD Moments	10	5 two-subject folds	1.75	5.1 min	100%	7.5 / run
Algonauts 2025	4	4 LOSO folds	1.49	63.1 h	89.7%	9.0–14.4 / clip

Following TRIBE's feature-extraction pipeline (d'Ascoli et al., 2025), all benchmark methods share cached features from frozen V-JEPA2 (Assran et al., 2025), Wav2Vec-BERT 2.0 (Seamless Communication et al., 2023), and Llama 3.2-3B (Meta AI, 2024) encoders. In our benchmark, features are aligned to the BOLD grid with a fixed three-step lag and standardized using training data. BOLD Moments subjects viewed silent videos: its audio and ASR-text features describe the source videos, not additional presented sensory inputs. Preprocessing, feature dimensions, missing-modality handling, and alignment are specified in Appendix A.

Downstream predictors are trained separately for each dataset using the shared stimulus representations, with parameters learned from training subjects and no refitting on held-out subjects. Stimulus content can recur across training and test subjects; the benchmark evaluates generalization to new subjects viewing familiar stimuli.

3.3 Controlled audiovisual simulation

To examine how the biophysical transformation from neural activity to BOLD shapes predictability, we adapt the DCM–Balloon framework of Smith et al. (2011) to a simplified 11-ROI audiovisual circuit (Fig. 6(a)). Synthetic visual and auditory inputs enter V1 and A1, respectively, while stochastic background inputs drive all ROIs throughout fixation and stimulation. A fixed network generates linear neural dynamics, which a nonlinear Balloon–Windkessel model transforms into BOLD (Friston et al., 2000). Each five-minute recording begins with 20 seconds of fixation, followed by sustained stimulation or intermittent stimulation at 5 s on/25 s off or 20 s on/20 s off. For each subject and schedule, the same neural trajectory generates BOLD under HRFs with target times to peak of 2, 6, and 10 seconds (Fig. 6(b)). Within each fast, medium, or slow condition, all ROIs share a common

nominal HRF. Neural dynamics are simulated with a 5-ms time step and passed through the Balloon–Windkessel model. The resulting BOLD signals are sampled at $\mathrm{TR} = 1$ s for prediction. The matched no-HRF neural reference is downsampled to 1 TR. We simulate 50 subjects per schedule, use a fixed 35/5/10 train/validation/test split, and give predictors the synthetic stimulus directly, without pretrained encoders. Further settings are in Appendix F.

3.4 Evaluation protocol and metrics

Prediction protocol. History-conditioned predictors start with $C = 8$ measured BOLD frames and predict one frame at a time, feeding each prediction into the rolling history. After H predictions, the history is refreshed with the latest C measured BOLD frames. For example, with $H = 4$, measured frames 1–8 initialize predictions for frames 9–12. Once frames 9–12 are measured, frames 5–12 form the history for predicting frames 13–16. Thus, H controls the number of prediction steps between measured-BOLD updates, corresponding to H TRs. We evaluate a range of refresh intervals and a no-refresh condition that uses measured BOLD only in the initial context. For each ROI within each clip or scan run, we correlate the predicted and measured BOLD time series over the scored samples. For HCP, preceding fixation supplies the initial history and all within-clip BOLD samples are scored; for BOLD Moments and Algonauts, scoring starts after the initial eight aligned BOLD frames (Appendix A). Each clip or run is predicted independently, and the same target samples are scored at every refresh interval. Stimulus-only predictions are independent of H. The full refresh grid and boundary conventions are in Appendix A.6. We report downstream predictor parameters or fitted ridge coefficients, including intercepts but excluding frozen stimulus encoders (Table 6). These counts describe model size rather than training or inference cost.

Evaluation metrics. To evaluate trajectory accuracy, our primary metric is the temporal Pearson correlation between predicted and measured BOLD for each parcel within each clip or scan run (Gifford et al., 2025). We further assess inter-parcel correlation structure using functional connectivity (FC) at $H = 32$. For predicted and measured time series, we compute pairwise parcel correlations within each record and average them across records within each subject using Fisher-z averaging. FC similarity is the Pearson correlation between the off-diagonal upper-triangular entries of the predicted and measured subject-level FC matrices. Seed-specific similarities are averaged within subject, and we report the mean and sample SD across held-out subjects. We also examine mean FC and intersubject FC variability, quantified as the SD of FC entries across subjects, with parcels grouped by the 7-Yeo networks (Yeo et al., 2011). Results include all designated folds and neural seeds 0/1/2; ridge fits are deterministic. Error-bar definitions are specified in the figure captions, with aggregation details in Appendices A.7 and E.

4 Results

4.1 Stimulus representation: pre-trained stimulus representations improve prediction

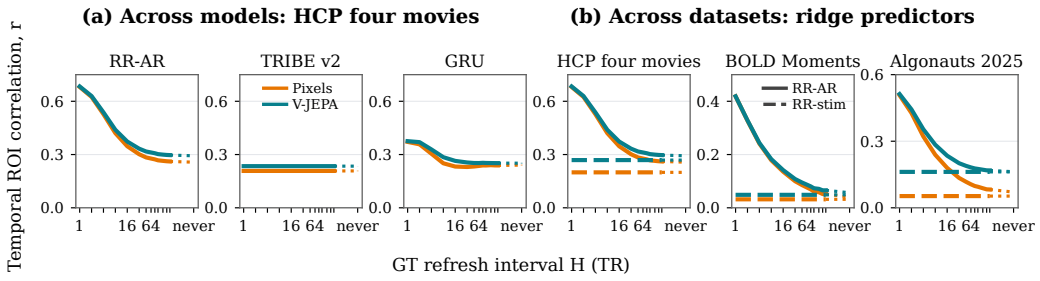

Figure 2: Pixels versus V-JEPA visual representations across representative models on HCP four movies (a) and ridge predictors across three datasets (b). Results include all designated folds and neural seeds 0/1/2; complete comparisons are provided in Appendix C.1.

Stronger visual representations benefit most of the tested predictors. Compared with raw pixels, V-JEPA 2 features improve the performance of most of the eight predictors across four HCP movies. Figure 2a presents three examples, with RR-stim and TRIBE v2 showing the largest relative gains on

$H = never$ (+35% and +13%, respectively). Results for the other models are shown in Figure 7: BrainVista changes little or declines slightly on HCP, while most comparisons on BOLD Moments and Algonauts 2025 show the same favorable trend. Separate comparisons using DINOv2 and CLIP encoders reveal a similar overall pattern (Appendix C.3).

The size of the V-JEPA gain varies across predictors and datasets. At the no-refresh endpoint, all eight predictors show their largest absolute gain over pixels on Algonauts; MIRAGE gains 0.084 in correlation there, versus 0.007 on HCP and 0.005 on BOLD Moments (Appendix Figure 7). For RR-AR, the advantage becomes more apparent as measured-BOLD refreshes become less frequent (Figure 2b). Separate modality controls also evaluate video, audio, and text alone and in combination (Appendix C.2). These results support the predictive value of the tested visual representations across datasets. Feature dimensions and preprocessing differ across representations, so these comparisons evaluate representation pipelines without isolating the effect of pretraining under dimension-matched inputs (Appendices C.1 and C.3).

4.2 Predictor Performance

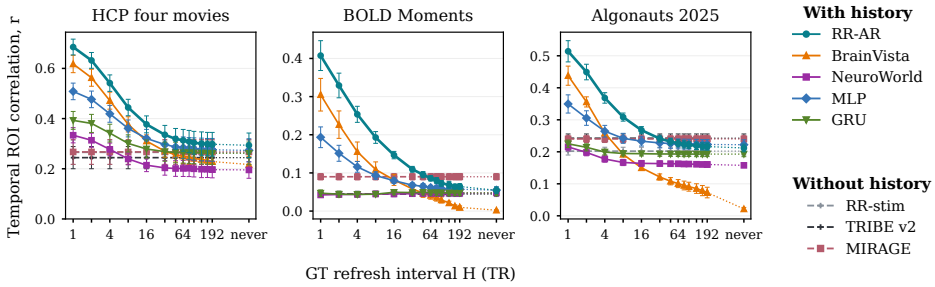

Figure 3: Comparison of eight models for cross-subject fMRI prediction across three datasets. Error bars show standard deviations across held-out subjects.

History-conditioned prediction generally leads at short refresh intervals. Most history-conditioned predictors outperform stimulus-only predictors at short refresh intervals across the three datasets (Fig. 3). RR-AR's large gains over RR-stim provide a clear example within the same linear ridge framework. Exceptions include NeuroWorld and GRU, which perform below MIRAGE on BOLD Moments and Algonauts even with frequent refreshes. As H increases, history-conditioned prediction accuracy generally declines, whereas stimulus-only predictions are independent of H. For example, RR-AR maintains its lead as refresh intervals approach $H = 64$—approximately 1.6 minutes on Algonauts and 1.9 minutes on BOLD Moments—where MIRAGE becomes competitive. On HCP, RR-AR remains the strongest predictor throughout the evaluated intervals. The advantage of history-conditioned prediction therefore generally narrows and may reverse at longer intervals, indicating that the benefit of incorporating measured BOLD depends on both refresh frequency and how effectively the model uses this input.

Nonlinearity can improve stimulus-only prediction. Among stimulus-only predictors, MIRAGE performs similarly to RR-stim on HCP, differing by about 1%, but clearly outperforms it on BOLD Moments and Algonauts. These results show that nonlinear stimulus-to-BOLD mappings can improve prediction when models use stimulus features alone, although the gains vary across datasets.

Linear prediction leads most history-conditioned comparisons. With comparable stimulus features and BOLD-history inputs, RR-AR generally outperforms the nonlinear alternatives, including the parameter-matched MLP and GRU controls. On one fixed diagnostic split per dataset, varying the history length C and training interval K largely preserves the broad comparison, although individual model rankings can change (Appendix D). Absolute accuracy is sensitive to ridge regularization (Appendix D.4). On the same diagnostic splits, mini-batch optimization of the one-step ridge objective closely reproduces the closed-form model's prediction accuracy (Appendix D.3). Reduced-rank variants, evaluated on all designated held-out subjects, retain much of RR-AR's performance with substantially fewer coefficients (Appendix D.5). Together, these controls support the strength of linear history-conditioned prediction across the tested fitting procedures and capacity choices, while retaining sensitivity to hyperparameters and evaluation conditions.

4.3 Regional analysis: where does measured history help linear prediction?

We next examine regional differences in prediction accuracy and whether gains from incorporating measured BOLD are greater where cross-subject response similarity is low.

Measured BOLD history supports regional and network-level prediction. At $H = 32$, the displayed models predict visual and auditory cortex comparatively well, while RR-AR retains higher accuracy in dorsal frontoparietal and medial association cortex, broadly overlapping Dorsal Attention, Control, and Default networks (Figure 4, top). On HCP, BrainVista also predicts these association regions more accurately than MIRAGE and GRU. This pattern is consistent with the longer temporal integration windows of association cortex (Hasson et al., 2008), suggesting that recent BOLD may supply accumulated stimulus context and ongoing response information that stimulus features alone miss. Across all three datasets, incorporating measured BOLD history improves FC similarity within the linear ridge framework, with RR-AR exceeding RR-stim by 0.14, 0.19, and 0.09 on HCP, BOLD Moments, and Algonauts, respectively (Appendix Figure 16 and Table 7). RR-AR achieves the highest mean FC similarity on HCP and BOLD Moments and remains close to BrainVista and MLP on Algonauts. However, preserving FC structure does not guarantee accurate BOLD trajectories: on Algonauts at $H = 32$, BrainVista achieves the highest mean FC similarity but the lowest mean temporal prediction accuracy (Figure 3 and Appendix Table 7). On HCP, history-conditioned predictors produce appreciable intersubject FC variability, whereas stimulus-only predictors show little to no variability across subjects receiving the same stimulus inputs (Figure 4, bottom, and Appendix Figure 16).

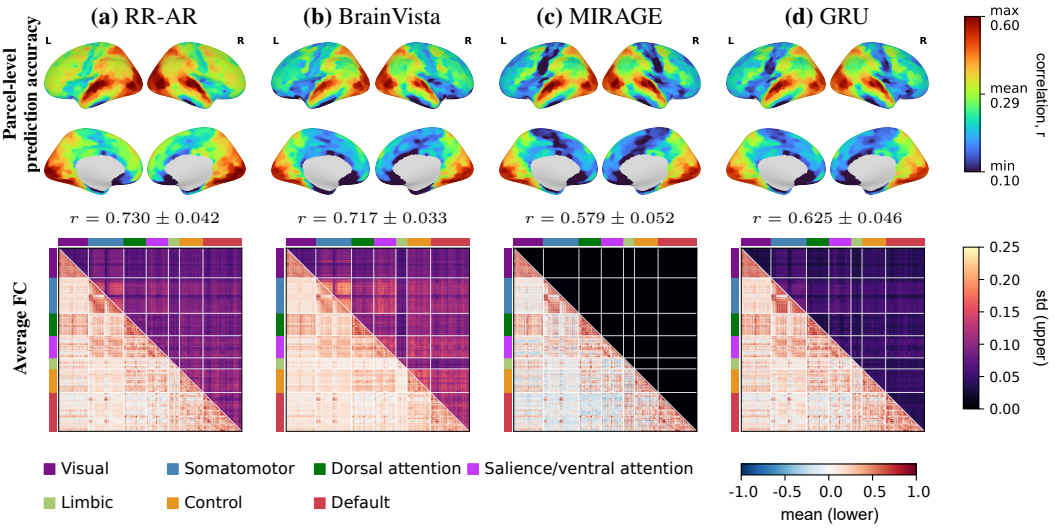

Figure 4: Regional prediction and functional connectivity (FC) on HCP at $H = 32$. Top: parcelwise temporal prediction correlations. Bottom: mean predicted FC (lower triangles) and intersubject SD of Fisher-z FC (upper triangles). Annotations report predicted–measured FC similarity (mean $\pm$ SD across held-out subjects). Full three-dataset results: Appendix E.1.

Measured BOLD history is particularly valuable in regions where responses differ more across individuals. We ask whether regions that respond similarly across individuals are easier to predict, and whether regions with less similar responses benefit more from incorporating recent BOLD. We quantify response consistency across training subjects using temporal intersubject correlation (ISC) for HCP and Algonauts and event-aligned response-pattern correlation for BOLD Moments; higher values indicate more similar responses to the same stimulus (Appendix E.2). For each parcel, we compare this consistency with RR-AR's prediction accuracy (Figure 5, left) and with the gain from incorporating BOLD history, defined as the difference in prediction correlation between RR-AR and RR-stim (Figure 5, right). Across all three datasets, regions with more consistent responses are generally predicted more accurately. However, Control and Default networks are less accurately predicted by RR-stim than visual cortex, yet show larger gains from incorporating BOLD history. On HCP, these gains are 0.0445 in Control and 0.0417 in Default, nearly double the 0.0226 gain in the visual network. The same qualitative pattern persists at $H = 96$: across parcels, response consistency

is positively associated with prediction accuracy for both RR-AR and RR-stim and negatively associated with the gain from measured BOLD history in all three datasets (Appendix Table 8). This complementary profile is consistent with greater interindividual functional-connectivity variability in association cortex (Mueller et al., 2013) and sensitivity to narrative context (Yeshurun et al., 2017), suggesting that recent BOLD is particularly valuable where shared stimulus features provide a less complete account of the measured response. Because the consistency–accuracy relationship also holds for RR-stim, it characterizes general regional predictability rather than uniquely explaining the linear–neural performance gap (Appendix E.2).

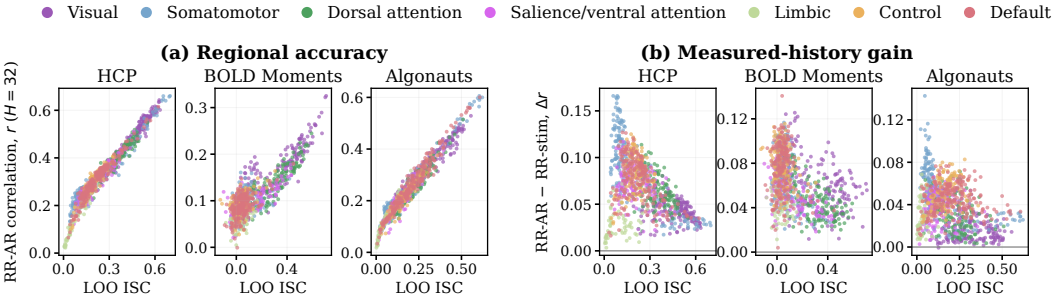

Figure 5: Regional prediction accuracy and BOLD-history gains ($H = 32$) versus training-subject response consistency. Each point represents a cortical parcel, with prediction performance evaluated on held-out subjects and response consistency (x-axis) computed across training subjects. (a) Absolute prediction accuracy for RR-AR. (b) Prediction gain of RR-AR over RR-stim.

4.4 Hemodynamic time scale and history-conditioned predictability

History-conditioned prediction is generally strongest when measured BOLD is refreshed frequently (Figure 3). We test whether hemodynamic filtering can contribute to this predictive persistence by varying the observation time scale while holding simulated neural trajectories fixed. RR-AR receives the same synthetic stimulus inputs in each condition, and the untransformed neural state provides a matched no-HRF reference (Section 3.3).

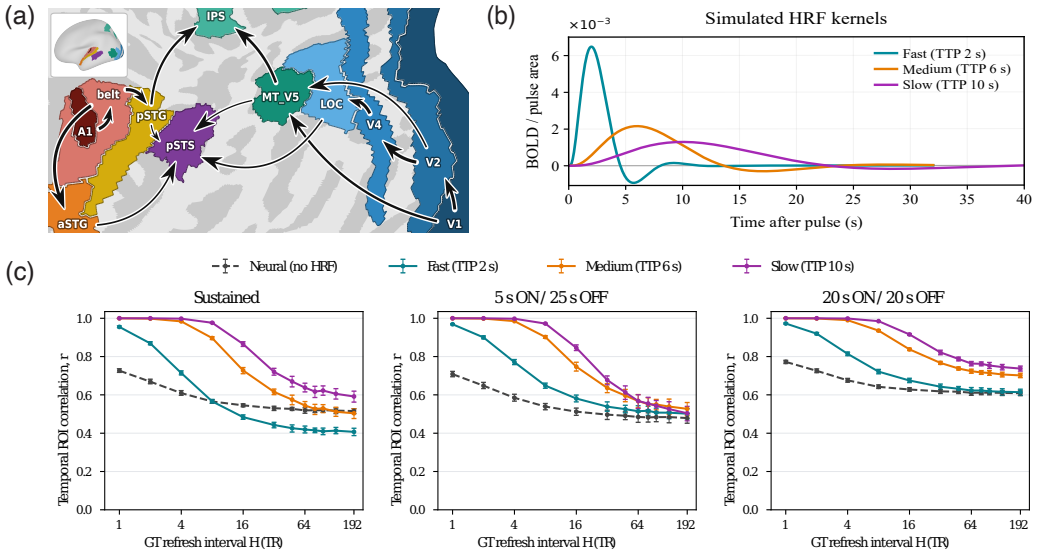

Figure 6: Effect of hemodynamic time scale on RR-AR prediction. (a) Audiovisual circuit. (b) Fast, medium, and slow Balloon–Windkessel HRFs. (c) Prediction across refresh intervals and stimulus schedules for the no-HRF neural reference and HRF conditions. Conditions share neural trajectories; TR = 1 s. Error bars: paired-subject bootstrap 95% CIs. Details: Appendix F.

Slower HRFs generally sustain higher temporal prediction correlations over longer refresh intervals (Figure 6). Medium and slow HRFs produce a short-interval plateau with correlations near one,

whereas the no-HRF reference does not exhibit this plateau. Under sustained stimulation at $H = 32$ (32 s), correlation increases from 0.443 with the fast HRF to 0.720 with the slow HRF, compared with 0.530 for the no-HRF reference. The effect depends on stimulus scheduling: under 5 s on/25 s off stimulation, the medium and slow HRF curves converge and cross at longer intervals. These comparisons isolate an effect of the hemodynamic observation time scale within the simulated system. Because the latent neural dynamics are linear, they do not establish how hemodynamic filtering affects nonlinear neural dynamics or explain the empirical performance gap between linear and nonlinear predictors.

5 Discussion and Conclusion

Measured BOLD history changes the prediction task. A stimulus-only predictor infers responses from stimulus features, whereas a history-conditioned predictor also observes recent responses from the target subject. These observations may summarize accumulated stimulus effects and ongoing fluctuations that shared stimulus representations do not capture. This distinction offers an interpretation of why nonlinear mappings can improve stimulus-only prediction while linear autoregression remains strong when measured history is available. The regional analyses further suggest that measured history is particularly useful where responses are less consistent across subjects. Strong linear predictability of BOLD does not establish that the underlying neural computations are linear.

Hemodynamic filtering shapes predictive persistence. In the paired simulations, slower hemodynamic responses generally sustain high prediction correlations over longer refresh intervals while neural trajectories remain fixed. This demonstrates an effect of the observation time scale within the simulated system and illustrates a plausible contributor to sustained history-conditioned BOLD predictability. The effect varies with stimulus scheduling, so the benefit of slower filtering is not uniform across conditions.

Implications for predictive modeling. Pretrained visual representations remain useful with linear downstream predictors; for RR-AR, their advantage over pixels generally grows as measured-history refreshes become less frequent. Stimulus representation and downstream prediction should therefore be evaluated separately. RR-AR provides a diagnostic reference for assessing additional complexity under comparable stimulus inputs, history availability, and refresh intervals. Regularization and reduced-rank parameterizations provide further control over its capacity. The relevant question is whether added complexity contributes predictive information under the intended measurement conditions.

Limitations. Our conclusions concern the tested benchmark adaptations and training procedures with shared frozen stimulus features. They do not evaluate the complete original pretrained systems. The benchmark primarily evaluates new subjects viewing familiar stimuli and uses offline preprocessing of complete recordings; generalization to unseen stimuli and performance with online preprocessing remain to be established. The small subject cohorts in BOLD Moments and Algonauts also limit population-level conclusions. Finally, the simplified simulations assume linear neural dynamics and do not establish the mechanism underlying the empirical performance gap between linear and nonlinear predictors.

Conclusion. Across three naturalistic fMRI datasets, pretrained stimulus representations generally improve prediction, and measured BOLD history provides substantial gains that usually diminish as refresh intervals lengthen. Nonlinear stimulus-only predictors can improve on linear readouts, while linear autoregression remains a strong history-conditioned baseline. These findings support judging additional model complexity by its predictive contribution under matched inputs and measurement availability.

AI use statement

We used generative AI tools to help refine hypotheses and diagnostics, assist with figure preparation, improve the organization and wording of the manuscript, and support code organization, verification, and reproducibility documentation. The authors reviewed the AI-assisted code and checked the resulting derivations, references, figures, and claims against the cited sources and experimental results. The authors take responsibility for the final content of the paper and all AI-assisted artifacts.

Ethics statement

This work performs secondary analysis of existing human neurophysiological data and includes no new participant recruitment or data collection. The HCP movie-fMRI data were collected with informed consent and institutional approval (Human Connectome Project, 2026); we used the open-access release under its data-use terms, did not use restricted identifying variables, and made no attempt to re-identify participants. The experiments study forecasting methodology and are not intended for clinical decision making or subject-level interpretation.

Reproducibility statement

The appendices document dataset preprocessing and splits, feature extraction and temporal alignment, model adaptations, training hyperparameters, validation and checkpoint-selection procedures, evaluation protocols, numerical handling, and robustness analyses. We report all designated folds and random seeds used in the main benchmark.

References

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction. *arXiv preprint arXiv:2507.22229*, 2025.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience. *arXiv preprint arXiv:2605.04326*, 2026.

Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37 (27):6539–6557, 2017. doi: 10.1523/JNEUROSCI.3267-16.2017.

Zijian Dong, Jianxiong Zhou, Kwun Kei Ng, Jan Paolo Macapinlac Balagtas, Zhizhou Li, Zijiao Chen, and Juan Helen Zhou. NeuroWorld: A latent brain world model for stimulus-conditioned human brain dynamics. *arXiv preprint arXiv:2608.01773*, 2026.

Emily S. Finn, Enrico Glerean, Arman Y. Khojandi, Dylan Nielson, Peter J. Molfese, Daniel A. Handwerker, and Peter A. Bandettini. Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, 215:116828, 2020.

Michael D Fox, Abraham Z Snyder, Jeffrey M Zacks, and Marcus E Raichle. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature neuroscience*, 9(1):23–25, 2006.

Karl J. Friston, Andrea Mechelli, Robert Turner, and Cathy J. Price. Nonlinear responses in fMRI: The balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000. doi: 10.1006/nimg.2000.0630.

Alessandro Thomas Gifford et al. The Algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2025.

Abdulkadir Gokce, Badr AlKhamissi, and Martin Schrimpf. MIRAGE: Adaptive multimodal gating for whole-brain fMRI encoding. *arXiv preprint arXiv:2605.29850*, 2026.

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303, 2004. doi: 10.1126/science.1089506.

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. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5487-07.2008. URL https://www.jneurosci.org/content/28/10/2539.

Human Connectome Project. HCP protocols: Young adult 7t imaging, 2026. Accessed September 2026.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. doi: 10.1016/j.neuron.2012.10.014.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

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.

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

Weijian Mai et al. SynBrain: Enhancing visual-to-fMRI synthesis via probabilistic representation learning. In *Advances in Neural Information Processing Systems*, 2025.

Meta AI. Llama-3.2-3B: Model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B.

Sophia Mueller, Danhong Wang, Michael D. Fox, B. T. Thomas Yeo, Jorge Sepulcre, Mert R. Sabuncu, Rebecca Shafee, Jie Lu, and Hesheng Liu. Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3):586–595, 2013. doi: 10.1016/j.neuron.2012.12.028.

Erfan Nozari, Maxwell A. Bertolero, Jennifer Stiso, et al. Macroscopic resting-state brain dynamics are best described by linear models. *Nature Biomedical Engineering*, 8:68–84, 2024.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

François Paugam, Basile Pinsard, Guillaume Lajoie, and Pierre Bellec. A benchmark of individual auto-regressive models in a massive fMRI dataset. *Imaging Neuroscience*, 2:imag–2–00228, 2024.

Alexander Schaefer, Ru Kong, Evan M. Gordon, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

Seamless Communication et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Ido Tavor, O Parker Jones, Rogier B Mars, Stephen M Smith, Timothy E Behrens, and Saad Jbabdi. Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage*, 176:152–163, 2018a.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12): 4136–4160, 2018b.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, J. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106 3:1125–65, 2011. URL https://api.semanticscholar.org/CorpusID:263439953.

Yaara Yeshurun, Stephen Swanson, Erez Simony, Janice Chen, Christina Lazaridi, Christopher J. Honey, and Uri Hasson. Same story, different story: The neural representation of interpretive frameworks. *Psychological Science*, 28(3):307–319, 2017. doi: 10.1177/0956797616682029.

Xuanhua Yin, Runkai Zhao, Lina Yao, and Weidong Cai. BrainVista: Modeling naturalistic brain dynamics as multimodal next-token prediction. *arXiv preprint arXiv:2602.04512*, 2026.

APPENDIX CONTENTS

A Dataset and Protocol Details **14**

A.1 HCP four movies . . . 14
A.2 BOLD Moments . . . 14
A.3 Algonauts 2025 . . . 15
A.4 Stimulus representations and temporal alignment . . . 15
A.5 Normalization and information separation . . . 16
A.6 Prediction protocol . . . 16
A.7 Metrics, aggregation, and uncertainty . . . 17

B Implementation Details **17**

B.1 Model taxonomy and benchmark scope . . . 17
B.2 Configuration, model classes, and sampling . . . 17
B.3 Optimization settings and configuration sources . . . 18
B.4 Parameter-matched controlled transitions . . . 20
B.5 Allocated parameter counts . . . 21

C Stimulus Representation and Modality Controls **21**

C.1 Visual representations across models and datasets . . . 22
C.2 Modality combinations . . . 22
C.3 Generalization across stimulus encoders . . . 25

D Robustness to Context, Training and Model Capacity **26**

D.1 Context length . . . 26
D.2 Training interval . . . 26
D.3 Optimization controls: mini-batch versus closed-form fitting . . . 26
D.4 Regularization strength . . . 28
D.5 Reduced-rank controls . . . 29

E Regional Predictability and Cross-Subject Analyses **30**

E.1 Preservation of functional organization . . . 30
E.2 Training-subject response consistency and held-out regional prediction . . . 31

F Simulation Design for Separating Neural and Hemodynamic Time Scales **33**

A Dataset and Protocol Details

Table 2: Dataset designs and cross-subject splits. A stimulus unit denotes a unique movie clip for HCP and BOLD Moments, and a unique movie segment for Algonauts. Unique duration excludes fixation and repeated presentations. BOLD Moments additionally includes 4-s null trials. Full details are provided in Appendix A.

	HCP four movies	BOLD Moments	Algonauts 2025
Subjects	154	10	4
Subject split	107/15/32	Five two-subject folds	Four LOSO folds
TR	1.00 s	1.75 s	1.49 s
Movie duration	49.4 min	5.1 min	63.1 h
Unique clips	15	102	331
Clip duration	1.1–4.3 min	3 s	∼10–12 min
Between-clip interval	20 s fixation	1 s fixation	Between-run break
Clips / run	4–5	85	1
Train–test stimulus overlap	100%	100%	89.7%
Scored prediction window	(1.1–4.3 min) / clip	(7.5 min) / run	(9.0–14.4 min) / clip

A.1 HCP four movies

We use four movie-viewing scans from each of 154 HCP subjects, giving 616 scan recordings. The four scans contain 5, 4, 5, and 4 clip presentations, respectively. These 18 presentations correspond to 15 unique clips because some content is repeated. The analysis therefore contains 2,772 clip presentations, but repeated presentations are not counted as distinct stimuli. When movie versions differ across subjects, stimulus features correspond to the version actually presented to each subject.

Subjects are partitioned into 107 training, 15 validation, and 32 test subjects, yielding 1,926, 270, and 576 clip records, respectively. Each clip presentation is treated as a separate prediction and scoring unit. The preceding fixation supplies the initial measured BOLD history, and every BOLD sample within the clip is included as a prediction target. Prediction blocks never cross clip boundaries.

We extract regional time series using the 1,000-parcel, seven-network Schaefer atlas. The 2-mm atlas label image is resampled onto the 1.6-mm BOLD sampling grid using nearest-neighbor interpolation, which preserves discrete parcel labels. Voxel signals are then averaged within each parcel. We use the supplied hp2000-clean BOLD data without applying an additional temporal filter or per-record target normalization. Stimulus features sampled at 2 Hz are averaged in consecutive pairs to match the one-second BOLD grid. The stimulus-to-BOLD delay and training-set normalization are described in Appendix A.4.

A.2 BOLD Moments

We use the version-B, fMRI Prep-preprocessed volumetric BOLD data from ten subjects in MNI152NLin2009cAsym space. Our analysis uses the continuous test-task scans from sessions 02–05, with three scans per session, giving 120 scan recordings. These scans represent 102 unique video clips, each lasting 3 seconds, with 1-second fixation intervals and additional 4-second null trials. The prediction targets are continuous BOLD time series.

Because stimulus order differs across subjects, each scan is aligned using its own event identities and onset times. An entire scan is treated as one prediction record. A prediction block may therefore include several stimulus events, but it cannot cross a scan boundary. The evaluation measures generalization to unseen subjects, rather than generalization to a held-out set of video identities.

Regional signals are obtained by resampling the FSLMNI152 Schaefer-1000 label image onto each BOLD grid with nearest-neighbor interpolation and averaging voxel signals within parcels. This operation aligns the sampling grids using their spatial coordinates.

Nuisance regression is performed separately for each scan. Available regressors include the six motion parameters, their temporal derivatives, their squares and squared derivatives, white-matter and

CSF signals, cosine regressors, and non-steady-state indicators. An intercept and linear trend are also included, and missing confound values are interpolated. Ordinary least squares residuals are centered and divided by their sample standard deviation within each complete scan. This offline preprocessing is applied to training, validation, and test scans alike, including the full duration of held-out scans.

Video features are aligned to scanner sampling times relative to each event onset using interpolation of the 2-Hz feature sequence. Queries outside a clip's available feature times use its nearest endpoint value. Intervals without represented stimulus content retain their BOLD targets, with unavailable input modalities marked as missing. In particular, oddball events without corresponding movie features are treated as unmodeled inputs, not as known fixation.

Subjects viewed silent videos. Audio features and ASR-derived text, when included in a multimodal comparison, are extracted from the original video soundtracks. They are therefore auxiliary descriptions of stimulus content, not auditory or linguistic signals actually presented to subjects. ASR text is not a visual description of the video.

Applying the three-sample stimulus-to-BOLD delay to each 268-TR scan leaves 265 aligned samples. The first eight aligned BOLD samples provide the initial history, and the remaining 257 samples are scored at every refresh interval. The five test-subject pairs are (01,06), (02,07), (03,08), (04,09), and (05,10). For each fold, the other eight subjects each contribute eleven training scans and one validation scan, namely the third scan of session 05. Each fold thus contains 88 training, eight validation, and 24 test records. No test subject contributes to predictor fitting or to the shared training-set normalization statistics.

A.3 ALGONAUTS 2025

We use the publicly labeled recordings from subjects 01, 02, 03, and 05, retaining the native 1,000-parcel targets and a TR of 1.49 seconds. Each subject contributes 348 recorded segments, giving 1,392 records across subjects and 331 unique stimulus feature timelines. Each recording is processed and evaluated separately.

We construct four leave-one-subject-out folds. In each fold, all 348 records of one subject are reserved for testing, while the remaining three subjects contribute 927 training and 117 validation records. Training and validation are partitioned by grouped stimulus identities, using Friends episode and movie groups, rather than independently assigning temporal windows. This partition is selected once with split seed 0 and is shared by all models and training seeds. These folds define our cross-subject benchmark and should not be interpreted as a reproduction of the original challenge evaluation protocol.

Stimulus features follow the released event-based temporal alignment: feature samples assigned to the same rounded time bin are averaged. The aligned data already associate each BOLD sample with stimulus features three TRs earlier, corresponding to 4.47 seconds, so no additional temporal shift is applied. We retain the source whole-record standardization and subsequently apply the shared training-set normalization described below. The first eight aligned BOLD samples initialize prediction, and all remaining samples are scored without crossing recording boundaries.

A.4 STIMULUS REPRESENTATIONS AND TEMPORAL ALIGNMENT

The primary multimodal comparisons use frozen Llama 3.2-3B, Wav2Vec-BERT 2.0, and V-JEPA2 ViT-g extractors for text, audio, and video, respectively. Two group-averaged representations from each extractor yield feature shapes of 2×3072, 2×1024, and 2×1408. Flattening and concatenating these representations gives 11,008 input dimensions: 6,144 for text, 2,048 for audio, and 2,816 for video. ASR and feature extraction are performed before predictor fitting, and the stimulus extractors are not fine-tuned.

Stimulus representations are first aligned to each dataset's BOLD sampling grid using the dataset-specific procedures above. We then pair a BOLD observation at sample t with the stimulus representation at sample $t - 3$:

$$\widetilde{\mathbf{x}}_t = \mathbf{x}_{t-3}.$$

The shift is applied once, whether it is already present in the released aligned data or introduced during our preprocessing. It corresponds to 3.00 seconds for HCP, 5.25 seconds for BOLD Moments, and 4.47 seconds for Algonauts.

All models receive stimulus features aligned to the same target times, although their temporal receptive fields within the supplied stimulus interval may differ. Stimulus extraction is offline and is not required to be causal with respect to the stimulus timeline.

Records with missing modalities are retained. Unavailable or experimentally ablated modalities are set to zero after normalization. A modality-presence mask determines which entries participate in normalization and which are zeroed, but the mask is not supplied as an additional learned input channel.

A.5 Normalization and information separation

For each fold, we estimate feature means and population standard deviations from available training features only. An identical stimulus feature sequence used by multiple subjects is counted once when estimating these statistics. Target means and population standard deviations are estimated from the training BOLD records. Normalization is performed independently for each feature dimension and each target ROI, with standard deviations floored at 10^{-4}.

The resulting transformation is shared by all models within a fold, including TRIBE v2 trained from scratch, and is applied unchanged to validation and test data. Validation data are used for model selection, while test data are used only for final evaluation. Neither validation nor test observations contribute to the shared normalization statistics.

This training-set transformation is separate from dataset-level offline preprocessing. In particular, the complete-scan nuisance regression and standardization used for BOLD Moments, and the retained source standardization for Algonauts, precede predictor fitting and also operate on held-out recordings. Consequently, the benchmark evaluates prediction of offline preprocessed BOLD; it is not an end-to-end online preprocessing benchmark. Within this setting, predictors receive measured BOLD only according to the refresh protocol below.

A.6 Prediction protocol

We distinguish three temporal quantities. The context length C is the number of measured BOLD samples provided when prediction is initialized or refreshed. The training rollout length K is the number of targets in a training prediction window. The test refresh interval H is the maximum number of consecutive targets predicted before measured BOLD is supplied again. We use $C = 8$; for Algonauts, $K = 92$, so the context and training targets together span 100 samples. This training-window length does not define the test refresh interval.

Each recording begins with C measured BOLD samples. A history-conditioned predictor then generates successive prediction blocks containing at most H targets. At the start of each block, its context is refreshed using the latest C measured BOLD samples strictly preceding that block. No measured BOLD from within the block is supplied to the predictor. Thus, H measures the interval between observations used to refresh the model, rather than the lead time of one forecast endpoint.

We evaluate the finite refresh intervals

$$H \in \{1, 2, 4, 8, 16, 32, 48, 64, 80, 96, 128, 160, 192\}.$$

These intervals are expressed in samples on the dataset-specific TR grids listed in Table 2. Predictions are assembled in temporal order and cover the same target samples for every H, including a final block shorter than H when necessary. No prediction block crosses a recording boundary. History-free predictors instead use fixed K-step stimulus windows independent of H; their predictions therefore do not change with the measured-BOLD refresh interval.

No-refresh evaluation. The *never* condition supplies only the initial measured BOLD context of each recording and performs no subsequent refresh. Its prediction duration depends on the finite recording length; it is not an infinite-time limit. Finite-H and no-refresh evaluations use the same selected model parameters and score the same target samples. Changing H does not involve retraining.

A.7 METRICS, AGGREGATION, AND UNCERTAINTY

Prediction accuracy is measured by temporal Pearson correlation. For subject s, recording c, ROI i, and refresh interval H, we correlate the predicted and observed time series over the scored target samples $\mathcal{T}_c$:

$$r^{(H)}_{s,c,i} = \mathrm{corr}_{t\in\mathcal{T}_c}\left(y_{s,c,t,i}, \widehat{y}^{(H)}_{s,c,t,i}\right).$$

We first average valid ROI correlations within each recording, then average recordings within each subject, and finally average test subjects with equal weight:

$$r^{(H)} = \frac{1}{|\mathcal{S}_{\mathrm{test}}|}\sum_{s\in\mathcal{S}_{\mathrm{test}}}\frac{1}{|\mathcal{C}_s|}\sum_{c\in\mathcal{C}_s}\frac{1}{|\mathcal{R}_{s,c}|}\sum_{i\in\mathcal{R}_{s,c}} r^{(H)}_{s,c,i}.$$

Here, $\mathcal{C}_s$ is the set of evaluated recordings for subject s, and $\mathcal{R}_{s,c}$ is the set of valid target ROIs for that recording. Recordings receive equal weight within a subject even when their durations differ; longer clips do not receive additional weight merely because they contain more samples. We neither concatenate different clips to calculate a single correlation nor substitute spatial correlation across ROIs for temporal correlation.

ROIs with constant target time series are excluded using the same target-derived mask for every model. A constant prediction for a nonconstant target is assigned a correlation of zero. Nonfinite predictions are treated as evaluation failures rather than silently omitted.

Seed averaging and between-subject variability. Each designated test subject appears in exactly one test fold per seed, and dataset-level results include all designated folds. For each neural model, we first compute a subject-level score for each training seed using the ROI and recording aggregation described above, then average these scores across seeds 0, 1, and 2. The reported dataset mean is the equally weighted average of these seed-averaged subject scores. Error bars show the sample standard deviation across held-out subjects after seed averaging, describing between-subject variability rather than variability across training seeds or a confidence interval for the population mean. Ridge models contribute one deterministic score per subject; their repeated evaluations across seeds are not treated as independent observations. All comparisons use the same designated test subjects and, for neural models, the complete seed set.

Regional comparisons likewise include all designated test subjects and neural seeds 0, 1, and 2. Model differences are paired within seed before averaging, while absolute ridge profiles reuse the deterministic fits. No ROI is selected on the basis of its test-set prediction gain.

B IMPLEMENTATION DETAILS

B.1 MODEL TAXONOMY AND BENCHMARK SCOPE

Table 3 summarizes representative prediction approaches by their inputs and downstream architectures. It describes the cited methods; the benchmark adaptations, diagnostic references, and training procedures are specified below.

B.2 CONFIGURATION, MODEL CLASSES, AND SAMPLING

The benchmark uses $C = 8$ measured context frames and training/validation rollout length $K = 64$ on HCP and BOLD Moments or $K = 92$ on Algonauts. These window lengths are distinct from the test refresh interval H. Each training record contributes $\lceil (T - C)/K \rceil$ sampled windows per epoch. Starts are drawn uniformly from legal C+K windows using a generator keyed by training seed, epoch, and sample index. Therefore different models have the same candidate record/start identities at the same seed and epoch even when batch sizes differ. Windows never cross a record boundary. Hidden state is reset per window; the true prefix warms the recurrent state, and no true BOLD is injected within its K-step prediction.

Validation replaces random starts with the deterministic partition

$$a_j = C + jK, \qquad b_j = \min(T, a_j + K), \qquad a_j < T.$$

Table 3: Representative downstream prediction approaches, organized by input and prediction algorithm. Linear and nonlinear refer to the downstream predictor, independently of the stimulus encoder. Benchmark adaptations and diagnostic references are described in Section 3.1.

Predictor input	Mapping	Prediction algorithm	Representative work
Stimulus features	Linear	Ridge-regression voxel-wise readout	Wen et al. (Wen et al., 2018b)
Stimulus features	Nonlinear	Temporal Transformer	TRIBE; TRIBE v2 (d'Ascoli et al., 2025; 2026)
Stimulus features	Nonlinear	Adaptive layer aggregation and temporal brain encoder	MIRAGE (Gokce et al., 2026)
BOLD history	Linear	Linear autoregressive models; linearized graph-based predictors	Nozari et al.; Paugam et al. (Nozari et al., 2024; Paugam et al., 2024)
Stimulus features and BOLD history	Nonlinear	Brain tokenizer and autoregressive Transformer	BrainVista (Yin et al., 2026)
Stimulus features and BOLD history	Nonlinear	Latent dynamics model and rollout decoder	NeuroWorld (Dong et al., 2026)

Block j receives true BOLD $[a_j - C, a_j)$ and predicts $[a_j, b_j)$. All target indices C through T-1 occur exactly once. Variable-length tails are evaluated at their true length and grouped into homogeneous sub-batches, with no padded attention tokens or discarded targets. Validation is global squared-error sum divided by the global number of scalar targets, not the mean of per-batch means.

Validation runs every epoch. A strict loss improvement saves the checkpoint; ties keep the earlier checkpoint. Patience is 10 unsuccessful validations after the native warmup; the counter restarts at each stage. An epoch-1 checkpoint remains eligible even during warmup. The best model is restored before testing or the next stage. Native stage ceilings remain in effect. No old-protocol checkpoints are imported.

B.3 Optimization settings and configuration sources

Table 4: Declared optimization ceilings and schedules. Every neural stage has patience 10; final prediction stages select by complete-coverage BOLD MSE. Pretraining uses the indicated stage loss. Numbers are not a claim that every stage runs to its ceiling.

Model/stage	Max epochs	Batch	Optimizer	LR	WD	Schedule / clip norm
BrainVista tokenizer	50	32	AdamW	$2 \cdot 10^{-5}$	.01	Constant / .5
BrainVista predictor	40	32	AdamW	$2 \cdot 10^{-4}$	.01	5% warmup, cosine / .5
NeuroWorld LDL	50	16	AdamW	$2 \cdot 10^{-4}$	.01	100-update warmup, cosine / .5
NeuroWorld decoder	200	16	AdamW	10^{-3}	.01	500-update warmup, cosine / .5
TRIBE v2 scratch	15	8	Adam	10^{-4}	0	10% warmup, OneCycle / none
MIRAGE	15	16	AdamW	10^{-4}	.01	10% warmup, OneCycle / 1.0
Ridge models	Closed form	All training rows	Eigendecomposition	–	–	Validation-selected λ

All Adam-family optimizers use $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. BrainVista excludes bias and one-dimensional normalization parameters from weight decay. Neural operations use bf16 autocast, except for the explicitly documented NeuroWorld FP32 recovery fits in Appendix C; Main-benchmark ridge fitting uses float64, with coefficients and intercepts cast to float32 for prediction; exceptions for representation controls are specified in Appendix C.1. TF32 is disabled; cuDNN deterministic mode is enabled and benchmarking disabled. Dataset-specific loader workers are 2/4/2; this is an I/O setting, not a different sampling or evaluation rule.

MIRAGE architecture and scope. We use the released downstream model classes with hidden width 3,072, eight Transformer layers, eight heads and fourfold feed-forward expansion. The layer-gating module uses four heads, 24 queries and attention dropout 0.2; modality dropout is 0.3. The Transformer attention, feed-forward and layer dropout settings are zero. Linear stimulus projectors and the upstream group-linear ROI head are retained. Allocated downstream sizes for the current atlas are reported in Table 6.

Table 5: Configuration provenance and deliberate comparison adaptations. Scientific choices are declared in the shared configuration and, for MIRAGE, its explicit adapter settings, not runtime paper defaults.

Component	Source and scope
Shared input/test protocol	Study design: C=8, dataset K, HRF=3 TR, fixed three-modality slots, full validation MSE, patience10 and temporal ROI corr. These replace model-native evaluation conventions.
BrainVista	Published two-stage structure and nominal training budgets; our implementation supplies dimension reconciliation, mask/filter choices, atlas-label grouping and shared autonomous BOLD-MSE training.
NeuroWorld	Published LDL/decoder structure and nominal budgets; latent widths, SIGReg settings and shared decoder are specified implementation choices; the adapter supplies all C observed frames.
TRIBE v2	Released FmriEncoder class and architecture/config-derived optimizer schedule; scratch initialization, atlas-sized output, shared normalization, common input interval and val-loss selection are comparison adaptations.
Ridge	Explicit affine design, covariance divided by training-row count, intercept unpenalized; adaptive boundary search uses common K-step val MSE.
MIRAGE	Released downstream code and declared width/depth; common cached features, shared ROI readout, no test-subject head fitting, and common validation-MSE selection.

This is a matched-feature downstream adaptation, not a reproduction of native pretrained MIRAGE scores. The inputs are the common two-group Llama/w2v-BERT/V-JEPA cached features, rather than full-layer Qwen-Omni representations. Subject embeddings are disabled and one shared ROI head is trained on the training subjects; no unseen-subject head is fitted. Temporal output pooling is disabled to preserve the aligned TR grid. Bidirectional stimulus attention is retained within fixed C+K tiles; no measured BOLD vector enters the model. The same fixed tiles are used for every H, including never. All downstream weights are initialized from scratch at seeds 0/1/2 and trained using the shared sampler, AdamW, 15-epoch ceiling, batch size 16, learning rate 10^{-4}, weight decay 0.01, OneCycle schedule with 10% warmup and gradient clipping at 1.0. Checkpoint selection uses complete-coverage validation BOLD MSE, not the upstream validation-Pearson criterion. The common BF16/TF32-disabled policy and all held-out subjects are retained. These adaptations isolate the released downstream architecture under common inputs; they do not evaluate the complete native pretrained system.

BrainVista architecture and stages. Actual atlas network labels partition all ROI columns exactly once into seven groups; consecutive equal-size fallback groups are not used. Each network has a two-linear-layer GELU autoencoder with token width $\max(1, \mathrm{round}(0.8n_g))$ and intermediate width $\max(n_g, \mathrm{tokenwidth})$. The stitched tokens project to width 256. Depthwise brain/stimulus convolutions have widths 3/5 and padding 1/2. Token-type and learned position embeddings precede four pre-normalized Transformer layers, eight heads, a 1,024-wide GELU feed-forward layer and dropout 0.1. A four-head spatial mixer operates at width 256. The model predicts a token residual added to the last token.

The configured attention-mask mode is named `released` in our code; this name is not a claim of complete official-code reproduction. It imposes causal token order and blocks brain-query/brain-key attention, while stimulus queries may attend earlier brain keys. Temporal convolutions stay within the supplied history/prediction window. All eight observed brain frames are tokenized. Generated latent tokens are recursively appended; the active attention window retains C tokens.

Tokenizer training minimizes reconstruction MSE on the C+K measured frames. The best tokenizer is frozen for predictor training. Predictor training differentiates through autonomous K-step latent rollout and the frozen decoder, using decoded BOLD MSE rather than the legacy teacher-forced latent-token loss. No test-subject decoder is fitted.

NeuroWorld architecture and stages. Atlas-network MLPs concatenate into a LayerNorm-normalized 256-dimensional latent; widths are split across the seven networks. A Linear–GELU–LayerNorm stimulus encoder maps 11,008 features to 256. Four causal adaptive-LayerNorm transition blocks use eight heads, fourfold feed-forward expansion, and dropout 0.1. The decoder is a $256 \to 256$ Linear–GELU–LayerNorm trunk and a $256 \to R$ linear head. Despite its internal name `subject_head`, this adapter uses one shared head, not a fitted head for each unseen test subject.

LDL learns next-latent prediction over the K target transitions with teacher forcing and SIGReg weight 0.01. SIGReg uses 256 fixed normalized random directions with projection seed1701. This is a latent-pretraining objective, not a reported BOLD metric. The best world model is then frozen in

evaluation mode, including its dropout, while the decoder is trained on autonomous K-step latent rollouts by BOLD MSE. Rollout starts from the encoded sequence of all C observations and retains generated latent history; it does not initialize from only the last measured frame.

TRIBE v2. The released FmriEncoder uses hidden width 1,152, eight Transformer layers, eight heads and fourfold feed-forward expansion. The configuration retains rotary positions, time-position embeddings, ScaleNorm and residual scaling, with maximum sequence length 1,024 and low-rank readout width 2,048. Attention is bidirectional within the supplied stimulus interval, not over an entire scanner run. Modality dropout is 0.3, subject dropout 0.1; ordinary, attention, feed-forward and temporal dropout are zero.

Fixed-slot grouped features pass through the released projection and subject-conditioned readout. Training uses the training subject's head; validation/testing use the average training heads. No held-out subject embedding or readout is fitted. All predictor weights start from scratch; frozen feature extractors remain pretrained. OneCycle uses division factors 25 and 10,000 and momentum limits 0.85/0.95. Its official-style 15-epoch ceiling is retained, but checkpoint selection is common validation MSE, not the native validation Pearson criterion. At test, fixed K-step stimulus tiles are independent of H. Bidirectional access within a known stimulus tile is distinguished from access to future measured BOLD, which is never supplied.

Ridge design, solver, and selected regularization. With normalized aligned features $\mathbf{x}_t$, the designs are

$$\mathbf{z}_t^{\mathrm{stim}} = \mathbf{x}_t, \qquad \mathbf{z}_t^{\mathrm{AR}} = [\mathbf{y}_{t-8}; \ldots; \mathbf{y}_{t-1}; \mathbf{x}_t].$$

All ROI lags can affect all predicted ROIs. Every training row with a complete eight-frame history is used; one-step fitting is closed form, not BPTT. Let $Z \in \mathbb{R}^{N \times d}$ contain the centered training design rows, where $d = CR + D$, and $Y \in \mathbb{R}^{N \times R}$ the centered next-step targets. With this oldest-to-newest history ordering, $W = [A_C \ \cdots \ A_1 \ B]^\top$, matching Equation 1; the column-vector prediction is $\mathbf{b} + W^\top \mathbf{z}_t^{\mathrm{AR}}$. Writing $G = Z^\top Z/N$ and $Q = Z^\top Y/N$, the ridge estimator is

$$W_\lambda = \arg\min_W N^{-1}\|Y - ZW\|_F^2 + \lambda\|W\|_F^2 = (G + \lambda I)^{-1} Q. \tag{2}$$

The unpenalized intercept is $\mathbf{b} = \bar{\mathbf{y}} - W_\lambda^\top \bar{\mathbf{z}}$, where the training means are column vectors. At $\lambda = 0$, the numerical minimum-norm solution replaces the inverse. No further private feature scaling or stimulus-lag expansion is added. Training sufficient statistics are accumulated in chunks of 256 rows in float64. The centered Gram matrix is symmetrized and eigendecomposed in float64 to obtain the ridge solutions. The fitted coefficients and unpenalized intercept are cast to float32 for prediction. Candidate values begin at $\{0.01, 0.1, 1, 10, 100\}$; a boundary optimum expands by a factor of ten in that direction, with a maximum of 12 expansions and an explicit failure if no interior optimum is found. Selection evaluates autonomous K-step validation MSE, not the one-step fitting residual or test correlation.

In BOLD fold0–4, stimulus Ridge selects $(10, 10, 10, 10, 10)$; RR-AR selects $(10, 10, 1, 10, 1)$. All are interior optima of the initial grid.

B.4 Parameter-matched controlled transitions

The MLP and GRU receive the current stimulus vector and flattened C=8 measured/predicted BOLD history. The MLP uses an affine hidden layer, GELU and an affine ROI readout; the GRU uses a GRUCell and an affine ROI readout, resetting its hidden state at each measured-BOLD refresh. Both feed predicted BOLD back during rollout. Hidden widths match the dense RR coefficient budget, including offsets, within 1%, without test selection. Neural weights use default random initialization, not RR weights or teachers.

Training retains the existing AdamW, K-step MSE, BF16 computation, 80-epoch budget and patience-10 full-validation protocol. Each prescribed learning rate $(10^{-4}, 3 \times 10^{-4}, 10^{-3})$ is assessed separately; minimum finite full-validation MSE selects the checkpoint. Numerical failure stops that candidate, retaining its best finite validated checkpoint and explicitly recording non-convergence; remaining candidates are still assessed. Test outcomes never trigger another checkpoint selection. Standard and extended evaluation use the same dataset-specific batch size and retain the H=96 replay check. All designated folds and seeds 0/1/2 are included in the main benchmark (Figure 3). Runs span the declared GPU/software environments; this is not an identical-runtime comparison or a certificate of globally optimal neural fitting.

Numerical handling. Seventeen endpoints encountered a nonfinite gradient or prediction in at least one learning-rate candidate. The full prescribed candidate grid was retained, and the validation-selected candidate in each case completed normally. This denotes completion under the budget or early-stopping rule, not certified convergence. Test scores were never used to reselect a candidate. Evaluation replays used the same selected weights and a consistent batch size for standard and extended refresh intervals; no retraining or test-based checkpoint selection was introduced.

B.5 Allocated parameter counts

Table 6: Downstream model size in millions of allocated parameters or fitted coefficients, excluding frozen stimulus extractors. MLP and GRU are omitted from the table because their parameter counts match dense RR-AR within 1% in each dataset. Rank64 counts the stored factors and output offsets of one prespecified reduced-rank setting; the full rank sweep is shown in Figure 15.

Model	HCP	BOLD	Algonauts
RR-AR	19.009	19.009	19.009
RR-AR (rank64)	1.282	1.282	1.282
RR-stim	11.009	11.009	11.009
TRIBE v2 scratch	356.526	153.675	143.430
BrainVista	7.649	7.649	7.649
NeuroWorld	7.613	7.613	7.613
MIRAGE	1097.679	1097.679	1097.679

Counts include predictor parameters frozen between training stages and fitted ridge coefficients, but exclude all frozen feature extractors. Ridge has $(D+1)R$ stimulus-only coefficients including the intercept, or $(D+CR+1)R$ autoregressive coefficients. TRIBE counts depend on the number of training-subject readouts and atlas size. These counts do not imply equal effective capacity, training time, FLOPs, or memory consumption. The cross-model counts are reported in Table 6. For rank r, storing the factorized map requires $(D+CR+R)r+R$ coefficients, including the output offset. This formula applies to every prespecified rank.

Reproducibility. Training, validation selection and evaluation use the declared configurations and folds. Tables and plots are aggregated from saved results, without changing model weights or selecting by test performance. Reporting retains complete declared fold/seed sets and distinguishes fixed-fold controls from all-fold benchmarks. Source manifests, configurations and numerical failure records are retained with the experiments.

C Stimulus Representation and Modality Controls

Every reported neural input combination includes all designated held-out subjects and seeds 0/1/2. Within each seed, records are averaged within subject and subjects equally; error bars show the sample SD of the resulting seed means. Ridge fits are deterministic. Each input combination is refitted and selected using validation data, rather than masking a trained full-input model at test. The same frozen feature slots, target grids and scoring samples are retained. Full V/A/T inputs are reported in Figure 3; the six reduced combinations and pixels are reported below.

BOLD Moments was presented silently. Its audio and ASR-text inputs describe the unpresented source soundtrack and are auxiliary feature controls, not evidence of sensory integration. ASR text is derived from audio, not an independent sensory channel. TRIBE, MIRAGE and RR-stim receive no measured BOLD. Internal neural projections, tokenizers and transitions remain in the pixel-input controls.

Numerical precision. Eleven NeuroWorld input-control endpoints required FP32 fitting after nonfinite-gradient failures under BF16: HCP V and VT at seed1; Algonauts V at seed0 (sub-01/02/03), seed1 (sub-02/03/05), and seed2 (sub-02/03); and Algonauts T at seed0 (sub-01). Seeds, learning rates, architectures, batches, stages and validation criteria were unchanged; other endpoints retain BF16. Recovery results were not selected by test scores. MLP/GRU learning-rate selection uses finite completed candidates on the prescribed validation grid.

C.1 VISUAL REPRESENTATIONS ACROSS MODELS AND DATASETS

Pixel construction and ridge representation controls. All designated HCP, BOLD and Algonauts folds compare two inputs: AR plus 768 raw RGB values and AR plus 2,816 V-JEPA features. Each arm is jointly refitted with the same BOLD-history input and the same complete-coverage validation selector. Raw frames are resized to 16×16 at 2 Hz, with area interpolation, scaled by 255, then averaged in the same dataset-specific TR bins as the cached video representation. HCP reads the correct version's full movie and slices the already declared stimulus TRs; Algonauts reads source paths and offsets from the audited official event table. No additional HRF shift is applied. Pixel features use training-only normalization; the resulting BOLD normalizer must exactly match the source. All arms share the same target references, context, K, validation origins, test H, and ROI reduction. Audio and text are absent in all three arms. This is a representation contrast, not a dimension-matched attribution of the entire encoder effect to pretraining. Original pilot coefficients are preserved: HCP uses FP64 sufficient statistics and a symmetric eigensolve; BOLD fold0 and Algonauts sub-01 retain FP32 Cholesky fits. Newly completed BOLD folds 1–4 and Algonauts sub-02/sub-03/sub-05 use FP64 fitting and FP32 inference, with normalization recomputed from each fold's training subjects. The objective and validation-only penalty-selection rule are shared, but numerical solvers are not identical across all folds. HCP's pixel arm selects a penalty near 10^{-11} after prescribed boundary expansion. All arms cover the same 14 H, including BOLD's extended horizons. Results average subjects, not folds when their sizes differ; deterministic fits are not counted as three stochastic seed replications.

Figures 7 gives the full eight-model comparisons underlying Figure 2. No model is omitted based on the size or direction of its encoder benefit.

C.2 MODALITY COMBINATIONS

The following figures compare the same stimulus inputs across predictors. Full video/audio/text results are shown in Figure 3; Figures 8 and 9 show the single- and paired-modality controls, respectively.

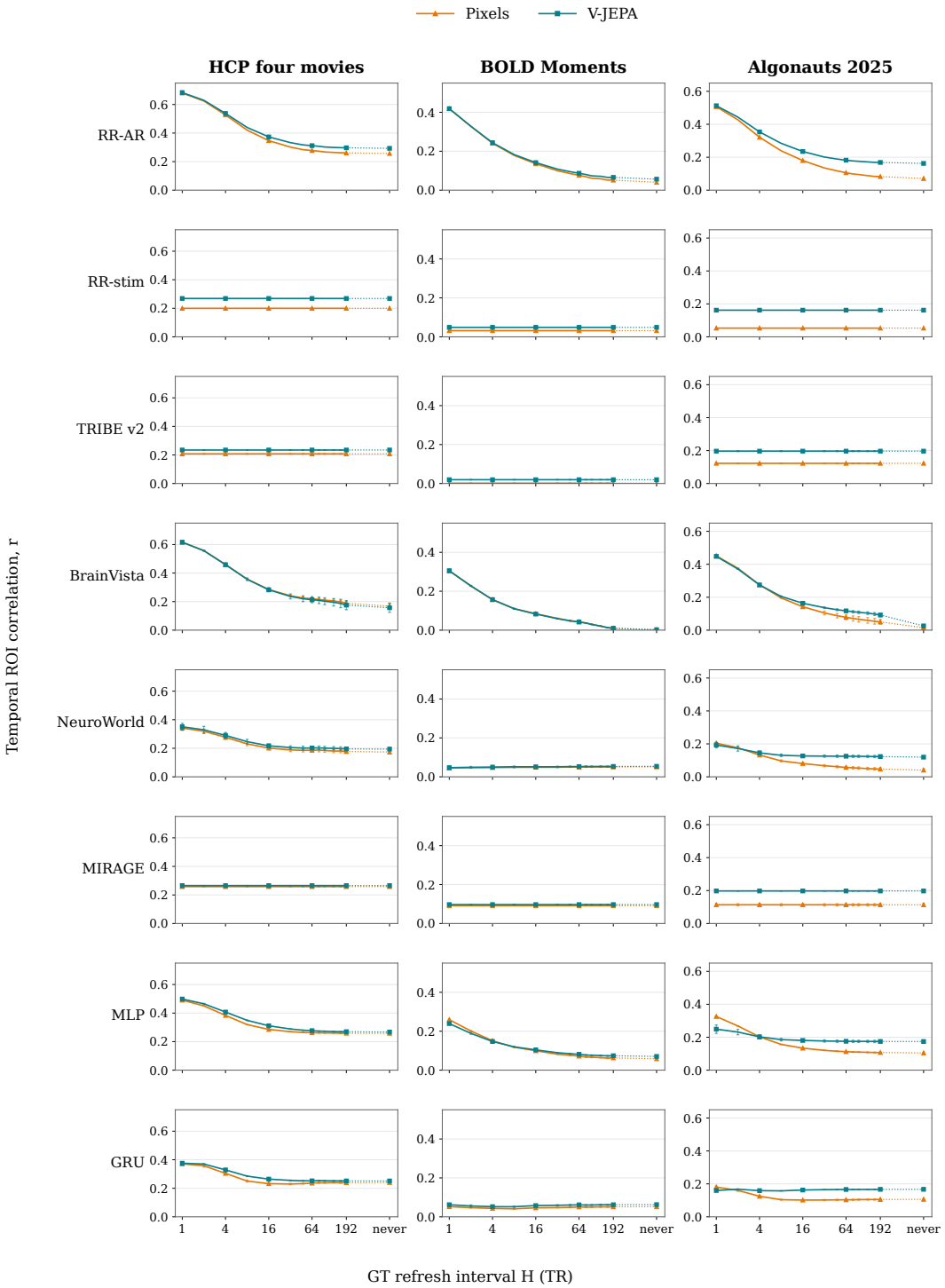

Figure 7: Complete eight-model pixels versus V-JEPA comparison on HCP, BOLD Moments and Algonauts 2025 Datasets. All designated held-out subjects and neural seeds 0/1/2 are included. Bars show seed SD; RR-AR and RR-stim are deterministic. Dotted connectors lead to the categorical no-refresh endpoint (*never*).

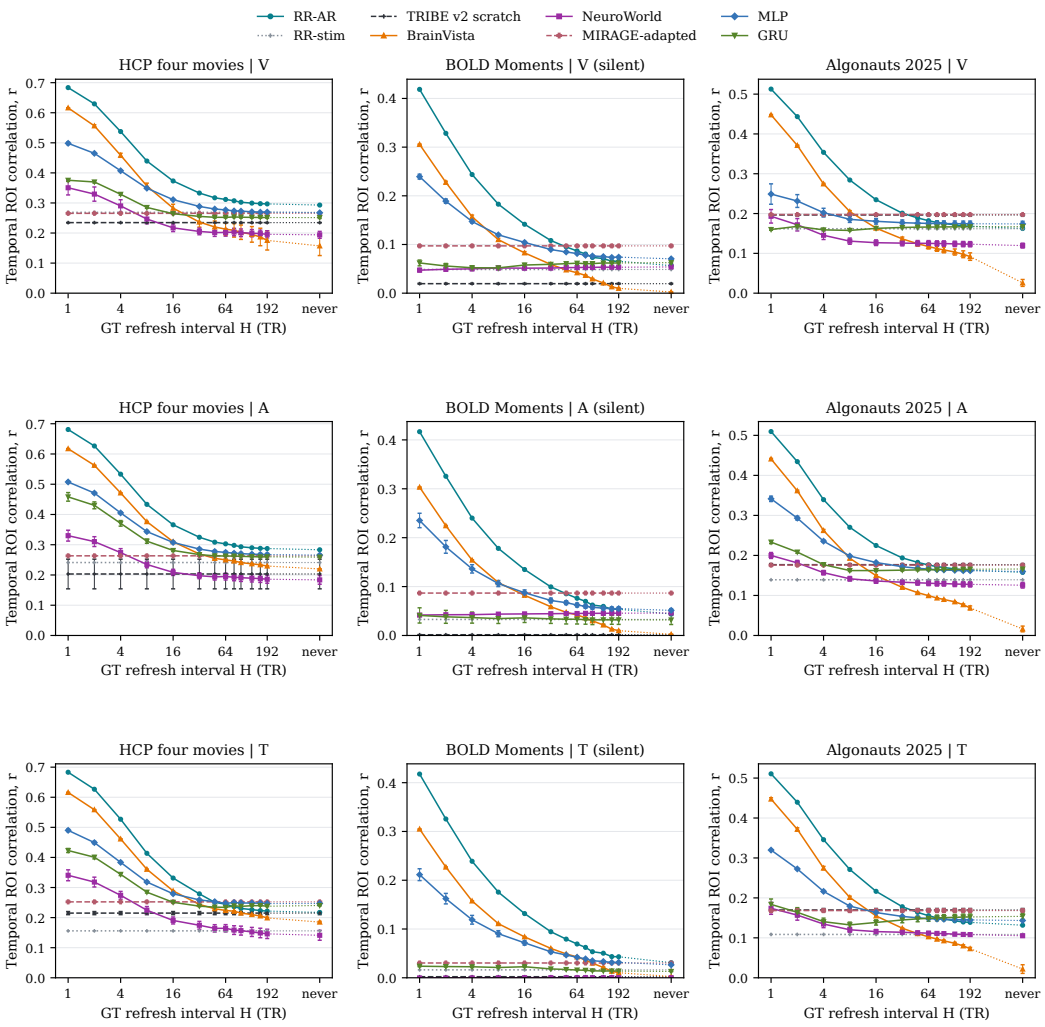

Figure 8: Matched V, A and T inputs across datasets and models. Every displayed neural curve includes all designated held-out subjects and seeds 0/1/2; bars show sample SD of the three subject-equal seed means. Ridge fits are deterministic. All model–input combinations are complete; incomplete seeds or folds are not pooled. The y-axis is absolute temporal ROI correlation; finite H extends to 192, with dotted guides to the categorical never endpoint. BOLD was presented silently, so A/T are source-feature controls. TRIBE and RR-stim are history-free references.

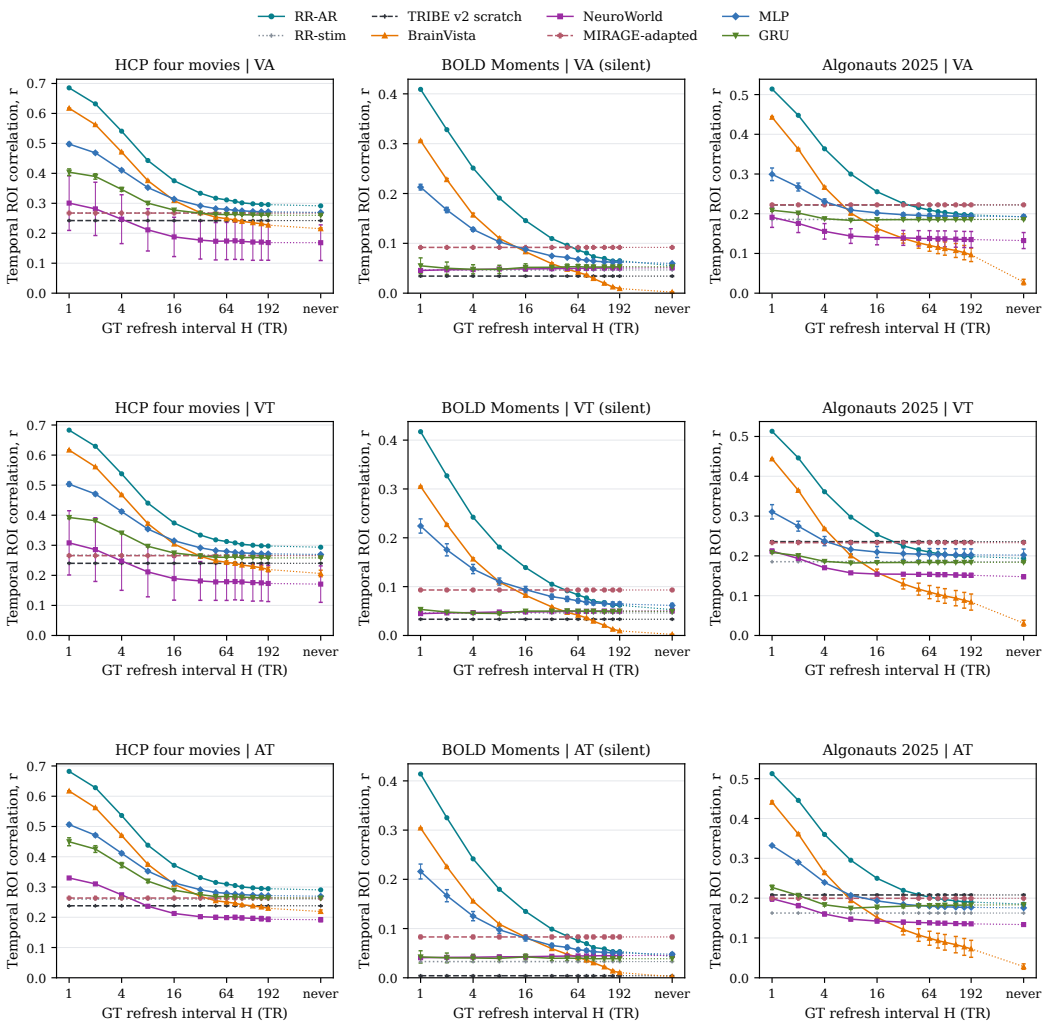

Figure 9: Matched VA, VT and AT controls. All models cover every designated fold and seeds 0/1/2; bars are sample SD of subject-equal seed means. Ridge fits are deterministic. NeuroWorld includes the explicitly documented FP32 recovery fits. BOLD was presented silently.

C.3 Generalization across stimulus encoders

To test whether the benefit of pretrained stimulus representations is specific to V-JEPA, we additionally evaluate DINOv2 (Oquab et al., 2023) and CLIP, using an OpenCLIP checkpoint (Cherti et al., 2023), on HCP four movies. We compare these frozen visual representations with pixels using RR-AR, RR-stim, and MLP, retaining the same subject split and evaluation protocol. Figure 10 shows that all three pretrained encoders improve prediction over pixels, with clearer gains when measured history is sparse or unavailable. In contrast, performance differences among the pretrained encoders are comparatively small, particularly for RR-AR and MLP. Thus, the benefit of richer stimulus representations is not confined to V-JEPA: the main distinction here is between pixels and pretrained features, rather than between individual pretrained encoders.

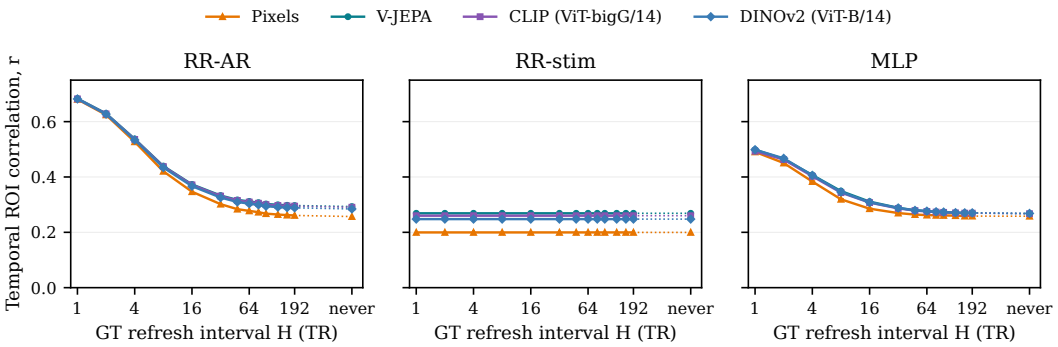

Figure 10: Generalization across visual encoders on HCP four movies. Each panel compares pixels, V-JEPA, CLIP (ViT-bigG/14), and DINOv2 (ViT-B/14) within one predictor. Curves show temporal ROI correlation across 32 held-out subjects and 1,000 parcels. MLP results are averaged over seeds 0/1/2, with bars indicating sample SD across seed-level means; deterministic ridge fits have no seed error bars. The categorical *never* endpoint denotes no further measured-history refresh; dotted connectors are visual guides. Feature dimensions and preprocessing are encoder-specific, so this comparison evaluates representation pipelines rather than a dimension-matched encoder ablation.

D Robustness to Context, Training and Model Capacity

These controls vary the comparison settings without changing the main benchmark. C/K, regularization and solver controls use HCP fold0, BOLD fold0 and Algonauts sub-01; reduced-rank controls use all designated held-out subjects.

D.1 Context length

We compare C=1/4/8 with the benchmark training interval fixed. HCP and BOLD include six neural models at seeds 0/1/2; Algonauts includes TRIBE v2 scratch, BrainVista and NeuroWorld at seed0. Deterministic RR-AR is refitted at each C; RR-stim is unchanged. The benchmark C=8 references are evaluated on the same held-out fold as C=1/4, not pooled over other test folds.

Targets begin at the same absolute index 8. Changing C can change model size and the selected ridge penalty; for history-free predictors it changes the stimulus prefix, not access to measured BOLD. Figure 11 retains model-specific improvements and exceptions to the main ordering. On HCP, RR-AR remains above all six neural references throughout the tested refresh grid at each C. BOLD retains a long-interval MIRAGE crossover; at C=1, some MLP/GRU settings also approach or exceed RR-AR. Algonauts retains the long-interval TRIBE crossover. Thus the broad pattern survives context changes, but not every pairwise ordering is invariant.

D.2 Training interval

In Figure 12, we compare K=8/32 with benchmark K=64 (HCP/BOLD) or 92 (Algonauts), retaining C=8. Fold and seed coverage match the context controls. Training and checkpoint-selection K change together; test stimulus tiling and scored targets remain fixed. Planned window/update counts, rather than total target TRs or compute, are maintained. For TRIBE and MIRAGE, K is the stimulus/target interval, not recurrent BPTT depth. RR's one-step fit is unchanged across K. On HCP, neither shorter interval closes the gap to RR-AR. BOLD retains the MIRAGE crossover at sufficiently long refresh intervals. On Algonauts, the TRIBE crossover remains at K=32 but is absent at K=8; shortening training does not systematically improve the neural references.

D.3 Optimization controls: mini-batch versus closed-form fitting

Could RR's predictive advantage depend on solving its coefficients directly rather than learning them with mini-batches? We fix the linear model, all one-step training rows, C=8, input features, training-only normalization, splits, and the originally validation-selected penalty, then train the coefficients

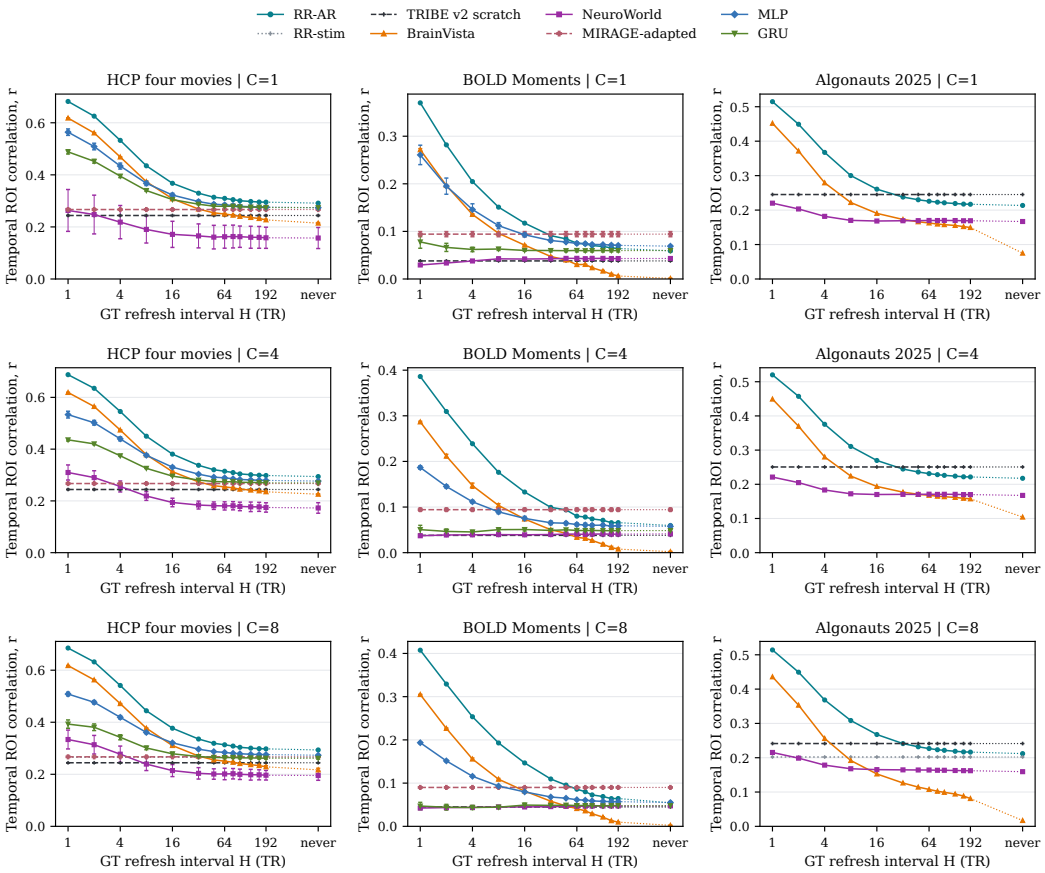

Figure 11: Context-length controls, C=1/4/8 (rows), with the benchmark training interval fixed. Columns show HCP/BOLD fold0 and Algonauts sub-01. HCP/BOLD neural curves use seeds 0/1/2 with sample SD; Algonauts uses seed0 without seed uncertainty. All benchmark reference rows use the same held-out fold as the ablations. Both ridge references are shown; RR-stim does not use BOLD context. All 14 refresh settings are included. Due to the substantially larger data volume, the Algonauts $C = 1$ and $C = 4$ experiments cover only a subset of models.

and unpenalized intercept from zero with Adam. No fitted RR weights or teacher predictions are used in the updates. The experiment covers HCP fold0, BOLD fold0 and Algonauts sub-01, each with seeds 0/1/2.

Matched objective and training protocol. For design matrix Z, targets Y, N training rows and R ROIs, both solvers minimize

$$J(W, b) = \frac{\|ZW + \mathbf{1}b^\top - Y\|_F^2}{NR} + \frac{\lambda}{R}\|W\|_F^2. \tag{3}$$

This is the original one-step ridge objective scaled by $1/R$, not a K-step BPTT loss. Fixed penalties are 0.1, 10 and 1 for HCP, BOLD and Algonauts, respectively. Each epoch shuffles and visits every training row once in batches of 1,024, retaining the final short batch. Adam uses learning rate 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and zero optimizer weight decay; the L2 term is explicit in Eq. 3. A ReduceLROnPlateau schedule monitors only the exact full-training objective (factor 0.2, patience 8, relative improvement threshold 10^{-5}, minimum learning rate 10^{-7}). Training uses FP32 without AMP or TF32. A fresh FP64 Cholesky fit supplies the optimal objective J^* for numerical audits, never optimization targets or initial parameters. Both methods use FP32 inference and the original validation/test evaluator.

The prespecified stopping certificate requires both $(J - J^*)/(J_0 - J^*) \leq 10^{-6}$ and a relative normal-equation residual, including the bias equation, at most 10^{-4}; J_0 is the zero-initialized

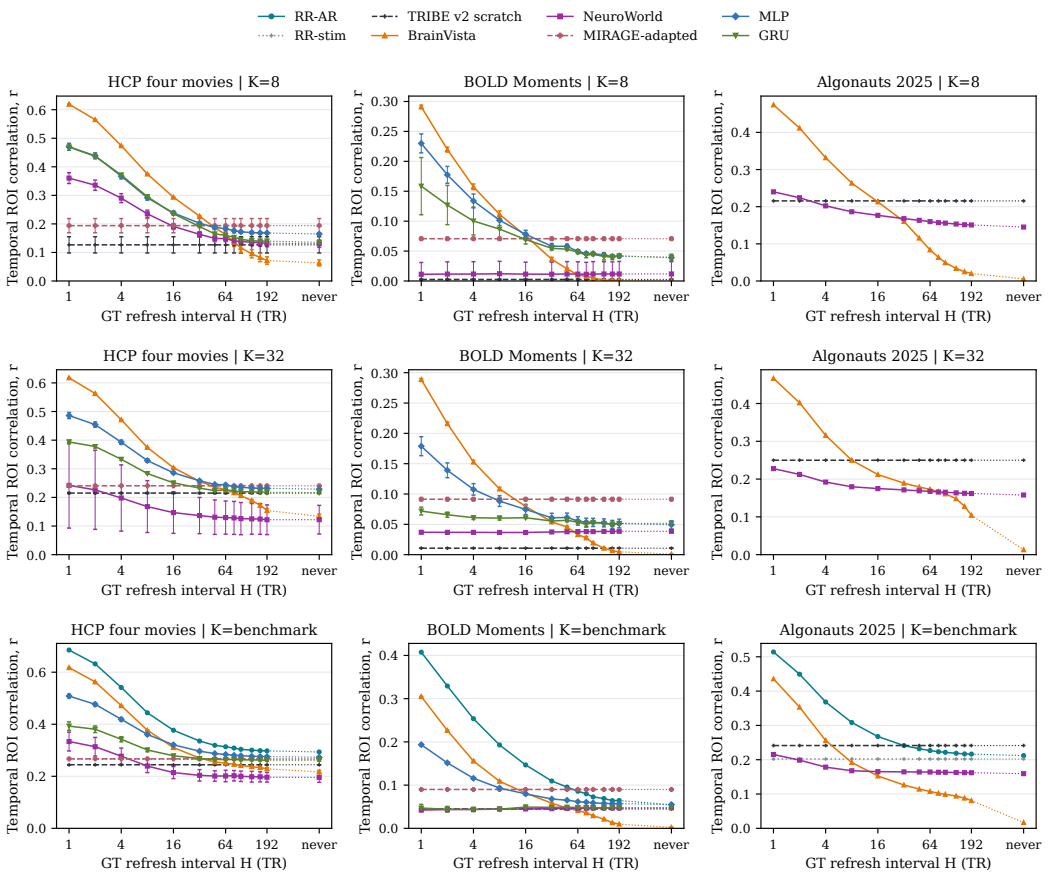

Figure 12: Training-interval controls, K=8/32/benchmark (rows), C=8 throughout. Benchmark K is 64 for HCP/BOLD and 92 for Algonauts. Both ridge references are repeated unchanged. All rows use the same held-out fold; model and seed coverage match Figure 11. K changes both training and checkpoint-selection intervals, not just gradient depth. Due to the substantially larger data volume, the Algonauts $K = 8$ and $K = 32$ experiments cover only a subset of models.

objective. Otherwise training stops at 300 epochs. The terminal iterate is evaluated at every standard and extended H, without validation or test checkpoint selection. The full-training objective and convergence audit are not computed from held-out targets.

Result and conclusion boundary. All nine runs reach the 300-epoch ceiling without satisfying both numerical criteria. Nevertheless, their predictive performance nearly coincides with the closed-form reference (Figure 13). The maximum absolute correlation discrepancy over all three seeds and all 14 H conditions is 3.44×10^{-5} on HCP, 1.96×10^{-4} on BOLD, and 2.51×10^{-5} on Algonauts. The largest corresponding scalar-mean regularized objective gaps are 2.40×10^{-6}, 3.85×10^{-5} and 1.10×10^{-6}. Fresh closed-form correlations agree with the original RR at standard H to within 5×10^{-8}. Thus mini-batch optimization from zero reproduces the linear reference's predictive performance in these tested folds: closed-form solving is not necessary for that performance. This is not certified numerical convergence, a formal equivalence test, or evidence that nonlinear optimization is adequate.

D.4 Regularization strength

Figure 14 displays the results across different regularization strength. The fixed diagnostic splits are HCP fold0, BOLD fold0 and Algonauts sub-01. We evaluate every prespecified $\lambda \in \{0, 0.01, 0.1, 1, 10, 100\}$ on the same design. All endpoints are reported at the standard H

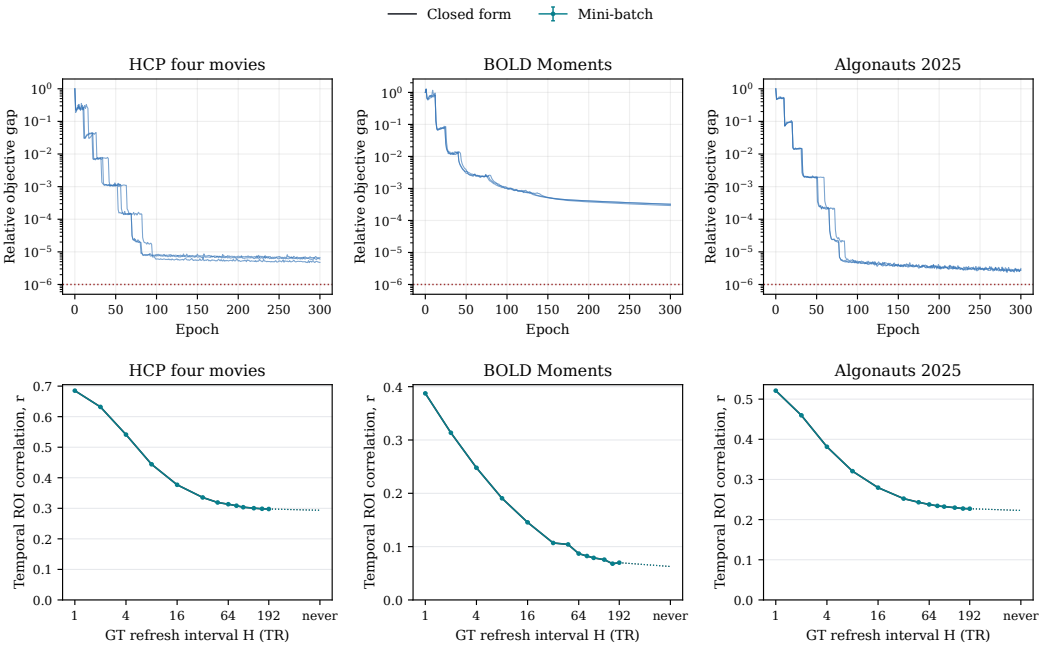

Figure 13: Same-objective mini-batch optimization control, HCP/BOLD fold0 and Algonauts sub-01, seeds 0/1/2. Top: $(J - J^*)/(J_0 - J^*)$ during zero-initialized Adam training; thin lines are individual seeds and the thick line is their mean. The dotted line is the prespecified objective-gap threshold, not a performance-equivalence margin. All nine runs reach 300 epochs without satisfying both numerical convergence criteria. Bottom: absolute temporal ROI correlation of the terminal mini-batch models (mean $\pm$ sample SD across three seeds) and the fresh closed-form reference. The near-overlap reflects predictive agreement, not certified convergence. Dotted connectors to never are visual guides to a categorical finite-record condition. No validation/test checkpoint selection or RR initialization is used.

values and at 128,160,192 and never. The results show that the performance is sensitive to the regularization strength.

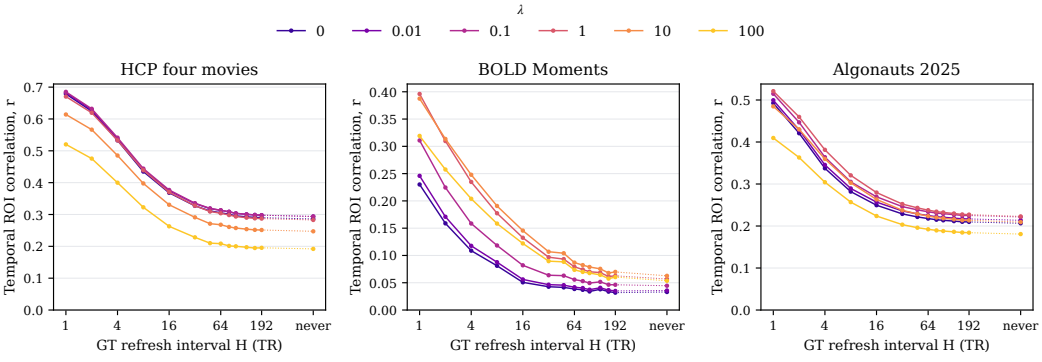

Figure 14: Fixed-grid regularization controls, seed0, one declared fold per dataset. Every tested penalty is shown through H=192 and never, the finite-record no-refresh endpoint.

D.5 REDUCED-RANK CONTROLS

All fixed ranks are reported. Coefficient counts include the intercept but exclude frozen stimulus encoders; they count stored factors, not intrinsic degrees of freedom. These are deterministic seed0 fits.

The main benchmark uses dense RR-AR. Here we constrain the rank of its joint history–stimulus coefficient matrix under the same ridge objective. Figure 15 shows that substantial compression preserves much of its long-horizon accuracy. The estimator is described below, and parameter accounting is detailed in Appendix B.5.

Estimator and implementation. We minimize the objective in Eq. 2 subject to $\mathrm{rank}(W) \leq r$. If V_r contains the leading orthonormal eigenvectors of $Q^\top W_\lambda$, the reduced-rank solution is

$$W_{\lambda,r} = (W_\lambda V_r) V_r^\top. \tag{4}$$

Completing the square in the ridge objective gives a rank-r approximation in the $(G + \lambda I)$ metric, rather than ordinary singular-value truncation of the coefficient matrix. The zero-penalty case uses the numerical minimum-norm convention. The factorized map stores $(d + R)r + R$ coefficients including the unpenalized intercept, versus $dR + R$ for the dense map. Regularization is selected separately on validation data for each prespecified rank. Fitting uses FP64 and prediction FP32, with all designated folds included. Both ridge regression and its reduced-rank form are established estimators; we use them as controls on predictive model complexity.

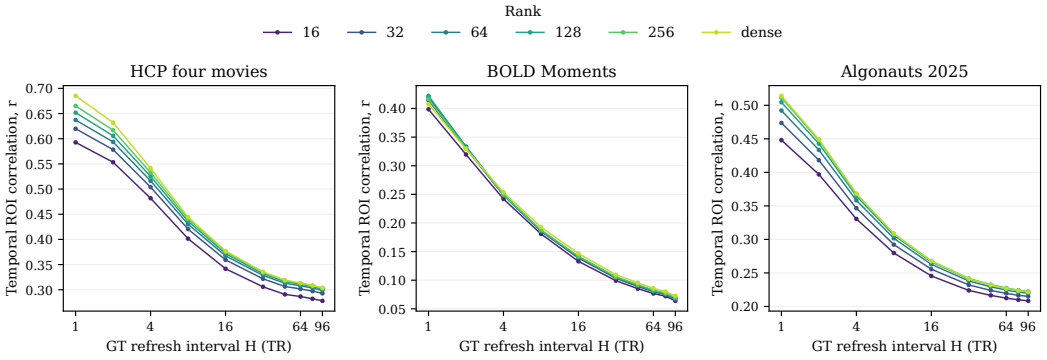

Figure 15: Reduced-rank controls for the joint RR-AR predictor. All prespecified ranks and the matched dense fit use video, audio, and text features, with C=8 and separately validation-selected ridge penalties. Curves show absolute temporal ROI correlation over all designated folds at H=1–96. Fits are deterministic; seed0 does not denote independent replication. Parameter accounting follows Appendix B.5.

E Regional Predictability and Cross-Subject Analyses

This section supports the main-text regional analysis with two complementary questions: whether predicted trajectories preserve large-scale functional organization, and how regional prediction accuracy relates to response consistency across training subjects. Unless stated otherwise, regional analyses include every designated fold, all 1,000 cortical targets in each dataset, and seeds 0/1/2.

For each parcel, we first average its temporal prediction accuracy across valid clips from the same held-out subject and then average equally across held-out subjects. When comparing two models, we use exactly the same subjects, records, parcels, prediction horizons, and valid time points. All 1,000 parcels are included.

E.1 Preservation of Functional Organization

Temporal correlation evaluates each parcel separately. We additionally ask whether the joint predicted trajectory preserves the correlation structure among parcels. For each continuous held-out record and prediction at $H = 32$, we compute the parcel-by-parcel Pearson functional-connectivity (FC) matrix for both the prediction and its matched ground truth. Off-diagonal correlations are averaged in Fisher-z space within subject.

For the mean-FC display in Figure 16, these matrices are averaged equally over subjects and, where applicable, seeds before applying the inverse Fisher transform. In the HCP row, lower triangles show

group-average FC in Pearson-r units, whereas upper triangles show intersubject FC variability. For each edge, variability is computed as the sample standard deviation across subjects of the subject-level Fisher-z FC values, after averaging across seeds within subject for stochastic models. Deterministic ridge models use their single fit. We report this variability only for HCP; BOLD Moments and Algonauts have smaller subject samples. Parcels are ordered by the seven Yeo networks.

The scalar FC similarity in Table 7 is computed separately for every subject and seed as the Pearson correlation between the off-diagonal upper-triangular entries of the predicted and ground-truth FC matrices; seed-specific values are averaged within subject, and the table reports the mean and sample SD across subjects. This subject-first statistic is intentionally not the correlation between two group-average matrices.

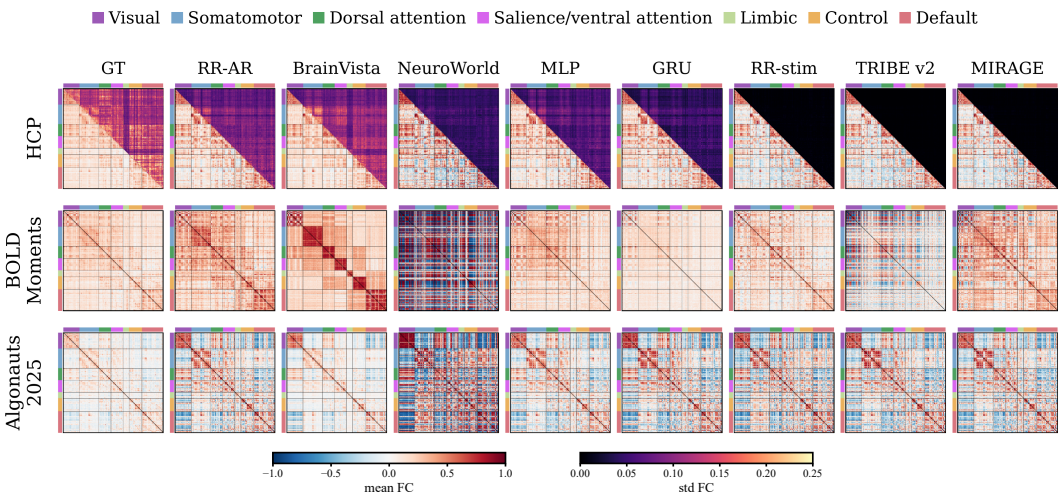

Figure 16: Ground-truth and predicted functional connectivity at $H = 32$ across three datasets, with parcels ordered by Yeo-7 network. In the HCP row, lower triangles show group-average FC in Pearson-r units (left colorbar), and upper triangles show intersubject variability, quantified as the sample SD across subjects of subject-level Fisher-z FC values after averaging across seeds within subject (right colorbar, range 0–0.25). BOLD Moments and Algonauts show mean FC in both triangles; intersubject variability is reported only for HCP.

Figure 16 shows that the predictions recover broad within- and between-network organization rather than merely matching each parcel's marginal time course. The amount of recovered structure nevertheless varies across models and datasets. RR-AR is among the strongest FC predictors in all three datasets, with subject-level similarities of 0.730 ± 0.042, 0.626 ± 0.045, and 0.691 ± 0.008 for HCP, BOLD Moments, and Algonauts, respectively (Table 7). Relative to its stimulus-only counterpart, RR-AR improves FC similarity by 0.142 on HCP, 0.188 on BOLD Moments, and 0.094 on Algonauts. The advantage is not exclusive: BrainVista is close to RR-AR on HCP and Algonauts, and MLP is close on Algonauts. Conversely, methods with similar temporal-accuracy curves can differ visibly in their recovered FC patterns. On HCP, history-conditioned predictors produce appreciable intersubject FC variability, whereas stimulus-only predictors show little to no variability across subjects receiving the same stimulus inputs (Figure 16, upper triangles). FC preservation therefore supplies a complementary systems-level criterion, rather than a restatement of parcelwise prediction accuracy or evidence that the learned predictor identifies causal connectivity.

E.2 Training-subject response consistency and held-out regional prediction

Figure 5 in Section 4.3 relates training-subject response consistency to held-out parcel accuracy and to the gain from measured history at $H = 32$. Here we specify the alignment and aggregation, and use $H = 96$ as a longer-horizon sensitivity analysis.

Alignment without held-out leakage. For each evaluation fold, response consistency is computed only from its training subjects. HCP uses 36 aligned timeline groups from 107 training subjects and

Table 7: Similarity between predicted and ground-truth functional connectivity at $H = 32$, measured by Pearson correlation between their off-diagonal upper-triangular entries. Values report the mean $\pm$ sample SD across held-out subjects after averaging the three seed-specific correlations within each subject.

Dataset	RR-AR	RR-stim	TRIBE v2	BrainVista	NeuroWorld	MIRAGE	MLP	GRU
HCP four movies	$\mathbf{0.73 \pm 0.04}$	0.59 ± 0.05	0.56 ± 0.05	0.72 ± 0.03	0.48 ± 0.04	0.58 ± 0.05	0.65 ± 0.04	0.63 ± 0.05
BOLD Moments	$\mathbf{0.63 \pm 0.05}$	0.44 ± 0.04	0.18 ± 0.03	0.47 ± 0.04	0.19 ± 0.03	0.39 ± 0.03	0.38 ± 0.03	0.35 ± 0.05
Algonauts 2025	0.69 ± 0.01	0.60 ± 0.01	0.63 ± 0.01	$\mathbf{0.70 \pm 0.01}$	0.47 ± 0.01	0.64 ± 0.01	0.69 ± 0.01	0.64 ± 0.01

evaluates 32 held-out subjects. Each of five BOLD Moments folds uses eight training subjects and evaluates two held-out subjects. Each of four Algonauts LOSO folds uses three training subjects across 297 aligned groups and evaluates the remaining subject. HCP and Algonauts groups match stimulus identity and time coordinate.

BOLD Moments requires event rather than run alignment because presentation order varies across subjects. Each event is sampled at onset plus 1.5 seconds on the response grid that has already been shifted by three TRs. Linear interpolation uses the two neighboring samples only when both lie within the event and valid analysis interval; no additional HRF shift is applied. Repeated presentations are averaged within training subject, after which shared video identities are compared in a common order. The BOLD Moments quantity is therefore event-pattern correspondence, not temporal ISC.

Estimation and aggregation. Within every aligned group and parcel, each training subject's response is centered and standardized over matched samples and correlated with the mean standardized response of the remaining training subjects. Valid correlations are averaged in Fisher-z space over subjects and then equally over groups, followed by the inverse transform. Constant responses and constant peer means are omitted. Each held-out subject's parcel accuracy is paired with the response-consistency estimate from that subject's training fold. Parcel values are averaged equally over held-out subjects, and descriptive Spearman associations are computed across all 1,000 parcels.

Table 8: Across-parcel Spearman associations between training-subject response consistency and held-out prediction at $H = 96$. History gain is RR-AR minus RR-stim. These longer-horizon summaries are descriptive spatial associations, not tests over independent parcels.

Dataset	Subjects	Parcels	RR-AR	RR-stim	History gain
HCP four movies	32	1,000	0.986	0.989	-0.498
BOLD Moments	10	1,000	0.809	0.871	-0.381
Algonauts 2025	4	1,000	0.969	0.962	-0.170

The main-text $H = 32$ result and the $H = 96$ sensitivity analysis give the same two-part picture. First, parcels with more consistent training-subject responses are easier to predict with both RR-AR and RR-stim; the relationship therefore characterizes general regional predictability rather than uniquely explaining the linear–neural model gap. Second, measured history contributes a different spatial profile. At $H = 96$, consistency is negatively associated with RR-AR minus RR-stim in every dataset, most strongly on HCP and least strongly on Algonauts. The association is not a universal inverse law: it varies across subjects and within networks, and Control or Default need not show the largest gain in every dataset.

Response consistency measures correspondence, not activation magnitude, raw intersubject variance, or temporal SNR. Prepared-signal variance is not comparable across datasets, particularly because Algonauts applies upstream whole-run standardization. Together, the main-text and longer-horizon analyses support a limited conclusion: shared responses are generally easier to predict, whereas measured history contributes a complementary regional profile. They do not identify a biological mechanism or uniquely explain the performance gap between linear and neural predictors.

F Simulation Design for Separating Neural and Hemodynamic Time Scales

Empirical BOLD prediction combines two distinct temporal processes: the evolution of the underlying neural state and its transformation through the hemodynamic observation process. To separate these factors, we construct a paired simulation in which identical neural trajectories are observed through hemodynamic models with different time scales. The purpose of this simulation is to test sensitivity to the observation process under controlled conditions, rather than to reproduce the full biological complexity of measured fMRI.

Neural and stimulus dynamics. The simulator contains 11 audiovisual ROIs with fixed linear neural dynamics driven by two stimulus channels representing visual and auditory input. Each five-minute sequence begins with 20 seconds of fixation, followed by one of three stimulus schedules: sustained stimulation for 280 seconds, 5 seconds on/25 seconds off, or 20 seconds on/20 seconds off. Fifty simulated subjects are divided into fixed training, validation, and test sets of 35, 5, and 10 subjects, respectively.

Paired hemodynamic intervention. For every subject and stimulus schedule, the same latent neural trajectory is passed through fast, medium, and slow Balloon–Windkessel observation models, with nominal HRF times to peak of 2, 6, and 10 seconds, respectively. Within each condition, all ROIs share a common nominal HRF. The comparison therefore holds the stimuli, latent neural trajectories, and subject assignments fixed while varying the hemodynamic observation time scale. The neural state before hemodynamic transformation is sampled on the same 1-s grid and retained as a matched no-HRF reference. This reference is a simulated latent state; it should not be interpreted as an electrophysiological measurement, a deconvolved BOLD signal, or an estimate of latent neural activity from the empirical datasets.

Neural simulation, BOLD sampling, and refresh intervals. Neural dynamics are simulated with a time step of $\Delta t = 5$ ms. The resulting neural trajectories drive the Balloon–Windkessel observation model, and its BOLD output is sampled at $\mathrm{TR} = 1$ s for model fitting and evaluation. The matched no-HRF reference is sampled from the neural trajectory on the same 1-s grid. Thus, 5 ms specifies the neural simulation time step, while prediction intervals are defined on the 1-s sampling grid. RR-AR receives the two-dimensional synthetic stimulus drive directly, without a pretrained media encoder, and an observed context of $C = 8$ samples (8 s). Refresh intervals $H \in \{1, 2, 4, 8, 16, 32, 48, 64, 80, 96, 128, 192\}$ therefore correspond to 1–192 s; in particular, $H = 32$ corresponds to 32 s. A rollout length of $K = 64$ samples corresponds to 64 s. The train/validation/test split and scoring times are matched across observation conditions. RR-AR fitting is deterministic.

Interpretation and limitations. This paired design isolates how the observation time scale changes apparent predictive persistence while holding the simulated neural trajectory fixed. It does not establish that hemodynamics explains the empirical performance gap between linear and neural predictors. In particular, the latent dynamics in this intervention are linear, the system contains only 11 ROIs, and the specified HRFs span a controlled family rather than the full variability of human hemodynamics. Conclusions from this simulation should therefore be interpreted as a sensitivity analysis of the measurement process, not as a mechanistic model of the empirical datasets.