FROM β-LEVEL TO BOLD-LEVEL: MODELING NEURAL DYNAMICS FOR VISUAL BRAIN ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing visual brain encoding models typically predict GLM-estimated β responses, which summarize each stimulus-evoked BOLD time course as a single condition-wise amplitude. However, fMRI is acquired as time-resolved BOLD signals, and compressing these signals into a single β amplitude does not retain the temporal information needed to characterize response timing and evolution. We propose **BOLDSyner**, a stimulus-aligned BOLD encoding framework that directly predicts stimulus-aligned BOLD responses by decomposing each condition into a temporal response profile and a signed voxel-wise spatial response pattern, modeling the two factors with asymmetric dual-branch diffusion, and recombining them under direct BOLD-level supervision. Across four NSD subjects, BOLDSyner achieves the state-of-the-art prediction performance at both the β and BOLD levels, increasing β-level correlation from 0.687 to 0.739 and BOLD-level correlation from 0.157 to 0.610. Compared with β-level signals, the predicted BOLD responses recover category-dependent temporal effects observed in measured responses ($r = 0.866$ across 120 subject–ROI–category cells), while preserving the category selectivity and semantic relationships captured by β-level signals. Collectively, these results imply the effectiveness and necessity of mechanistically grounded, time-resolved BOLD signal prediction.

1 INTRODUCTION

Visual brain encoding aims to predict neural responses from visual stimuli and provides a computational framework for studying how visual information is represented in the brain (Naselaris et al., 2011; Kay et al., 2008). Recent advances in deep visual representations and generative approaches have substantially improved neural response prediction (Allen et al., 2022; Yamins et al., 2014; Bao et al., 2025; Mai et al., 2026). However, most visual brain encoding models still predict GLM-estimated β responses or related condition-wise voxel representations for each stimulus (Bao et al., 2025; Mai et al., 2026). These β responses are derived from time-resolved BOLD signals and summarize each stimulus-evoked BOLD time course as a single condition-wise amplitude under a specified design matrix and hemodynamic response model (Friston et al., 1994; 1998; Boynton et al., 1996; Lindquist et al., 2009). While this compact representation is effective for brain encoding and selectivity analyses, it compresses the post-stimulus temporal evolution of the measured response.

This compression becomes limiting when the scientific question concerns not only how strongly a brain region responds, but also when the response emerges, how long it persists, and how it evolves over time. Time-resolved BOLD signals have enabled studies of dynamic response abnormalities in depression (Guo et al., 2015), event segmentation and memory during continuous narratives (Baldassano et al., 2017), and multiscale temporal integration in language comprehension (Lerner et al., 2011). Such questions depend on temporal properties that cannot be directly represented by a single condition-wise β amplitude. In visual encoding, retaining this temporal information further enables us to ask whether different visual categories exhibit distinct stimulus-evoked temporal response patterns within their corresponding category-selective cortical regions.

Temporal variation in BOLD signals has long been studied in neuroimaging through approaches such as HRF estimation, FIR-based response modeling, and hemodynamic matrix factorization (Lindquist et al., 2009; Pedregosa et al., 2015; Hütel et al., 2021). However, stimulus-conditioned prediction of such time-resolved BOLD responses remains largely unexplored in visual brain encoding.

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.

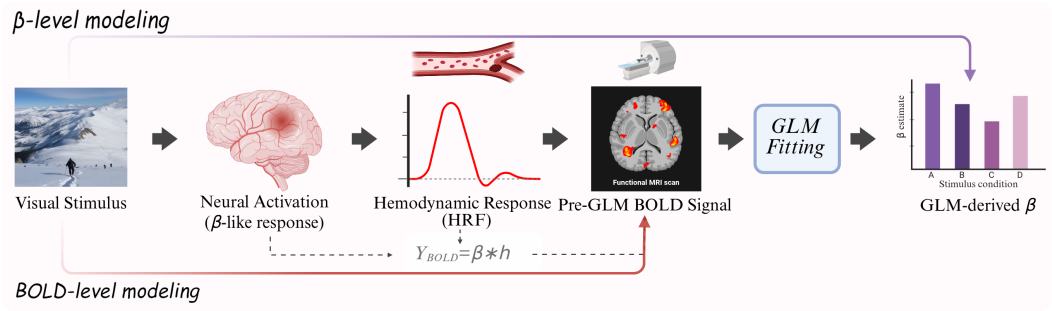

Figure 1: Comparison between conventional β-level modeling and the proposed BOLD-level formulation. Rather than predicting only temporally compressed β-level targets, the BOLD-level formulation models temporal response profiles and voxel-wise spatial response patterns.

As illustrated in Fig. 1, we formulate visual brain encoding directly at the stimulus-aligned BOLD level, predicting the post-stimulus response rather than only its GLM-derived amplitude summary while retaining compatibility with conventional β-level analyses.

On this formulation, we propose **BOLDSyner**, a stimulus-aligned BOLD encoding framework that represents each condition with a temporal response profile and a signed voxel-wise spatial response pattern. This factorization represents post-stimulus temporal evolution together with condition-dependent voxel-wise response patterns. Because the temporal factor is low-dimensional whereas the spatial factor is high-dimensional and exhibits correlations across voxels, BOLDSyner adopts an asymmetric dual-branch diffusion model that directly models the temporal factor and applies covariance-aware perturbations to a compact representation of the spatial factor. The predicted factors are recombined, with the reconstructed BOLD response directly supervised against measured BOLD.

Across four NSD subjects, BOLDSyner achieves the best prediction performance at both the β and BOLD levels, increasing β-level correlation from 0.687 to 0.739 and BOLD-level correlation from 0.157 to 0.610 compared with the state-of-the-art baseline. More importantly, compared with β-level signals, the predicted BOLD responses recover category-dependent temporal effects observed in measured responses ($r = 0.866$ across 120 subject–ROI–category cells), while retaining the category-selective and semantic information accessible through conventional β-level analyses.

The contributions of this paper are summarized as follows:

- We extend visual brain encoding from GLM-derived β responses to stimulus-aligned BOLD responses constructed from pre-GLM fMRI measurements, enabling explicit modeling of post-stimulus temporal structure while retaining compatibility with conventional β-level analyses.
- We propose **BOLDSyner**, which factorizes BOLD responses into temporal response profiles and signed voxel-wise spatial response patterns and models them using an asymmetric dual-branch diffusion framework with branch-specific perturbations and direct BOLD-level supervision.
- We show that BOLDSyner improves both conventional β-level and BOLD-level prediction and recovers category-dependent temporal effects that are not directly represented by a single condition-wise β amplitude, while retaining category-selective and semantic information.

2 Related Work

2.1 Visual Brain Encoding

Visual brain encoding predicts neural responses from visual stimuli and links computational visual representations with human brain activity. Early studies used linear encoding models with hand-crafted visual features to analyze stimulus–response relationships in visual cortex (Naselaris et al., 2011; Kay et al., 2008; Haxby et al., 2001; Stansbury et al., 2013; Huth et al., 2012; Konkle & Caramazza, 2013). Later work improved prediction accuracy by leveraging hierarchical representations from deep visual models and large-scale natural-image datasets such as NSD (Güçlü & van Gerven, 2015; Yamins et al., 2014; Wen et al., 2018; Horikawa & Kamitani, 2017; Radford et al.,

2021; Allen et al., 2022). More recent generative approaches move visual-to-fMRI prediction beyond deterministic regression toward conditional response generation, where diffusion models provide a flexible framework for learning complex conditional distributions (Ho et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022; Ramesh et al., 2022). For example, MindSimulator combines an autoencoder with diffusion-based generation for visual-to-fMRI synthesis (Bao et al., 2025), while SynBrain formulates visual-to-fMRI synthesis through probabilistic representation learning (Mai et al., 2026). These methods demonstrate the potential of generative modeling for neural response synthesis, but they typically predict a single voxel-wise response representation or a latent neural representation for each stimulus. They therefore do not explicitly model the temporally resolved BOLD response profile together with its voxel-wise spatial response pattern.

2.2 Modeling Temporal Dynamics in Neural Responses

BOLD responses are temporally extended measurements shaped by stimulus timing, hemodynamic dynamics, and voxel-wise response amplitudes. Classical GLM analyses model this process by convolving stimulus events with an HRF and estimating condition-specific response amplitudes (Friston et al., 1994; 1998; Boynton et al., 1996). Data-driven HRF estimation, FIR models, low-rank GLM formulations, and hemodynamic matrix factorization further provide flexible ways to characterize temporal response variation and coupled temporal–spatial dependencies in BOLD signals (Glover, 1999; Lindquist et al., 2009; Pedregosa et al., 2013; 2015; Hütel et al., 2021). However, these approaches are primarily designed for response estimation and analysis, and visual encoding pipelines generally use the resulting response estimates as targets for subsequent image-to-neural prediction. BOLDSyner instead incorporates both temporal response dynamics and voxel-wise spatial response patterns into the generative target, predicting the two factors from visual inputs and recombining them into BOLD-level responses.

3 Method

Figure 2 summarizes the pipeline of **BOLDSyner**. Given a visual stimulus, BOLDSyner constructs temporal and voxel-wise spatial response targets from pre-GLM BOLD responses, and then predicts these factors with a dual-branch diffusion model before recombining them into BOLD responses.

3.1 Spatiotemporal BOLD Target Construction

BOLDSyner constructs supervision targets from stimulus-aligned pre-GLM BOLD responses rather than GLM-estimated β maps. Instead of representing each image condition by a single response amplitude at each voxel, we use a temporal response factor U and a signed voxel-wise spatial response factor V to retain both temporal variation and condition-dependent voxel responses.

For each voxel i, we use a $K = 6$ bin FIR representation and fit a rank-one approximation to its condition-by-bin FIR coefficient matrix:

$$\begin{aligned} &Y_i \approx X\beta_i + b_i\mathbf{1}, \\ &\beta_i = [v_{i,1}u_i^\top, \ldots, v_{i,C}u_i^\top]^\top, \quad u_i \in \mathbb{R}^K, \quad v_i \in \mathbb{R}^C. \end{aligned} \tag{1}$$

where Y_i denotes the flattened stimulus-aligned pre-GLM response samples for voxel i, X is the FIR design matrix, C is the number of image conditions, and b_i is an intercept. The vector u_i represents the temporal response profile of voxel i, while v_i contains its condition-wise response loadings. Equivalently, the FIR coefficient matrix is written as

$$B_i = v_i u_i^\top. \tag{2}$$

This approximation separates the FIR coefficients into a temporal profile and condition-wise loadings. Temporal-window selection and higher-rank diagnostics are provided in Appendix A.1.1. We then aggregate the voxel-wise factors into condition-level targets. For image condition c, we define

$$w_{i,c} = |v_{i,c}| + \epsilon, \qquad U_c = \frac{\sum_{i=1}^{N_{\text{voxel}}} w_{i,c}u_i}{\sum_{i=1}^{N_{\text{voxel}}} w_{i,c}}, \qquad V_c = [v_{1,c}, \ldots, v_{N_{\text{voxel}},c}]^\top. \tag{3}$$

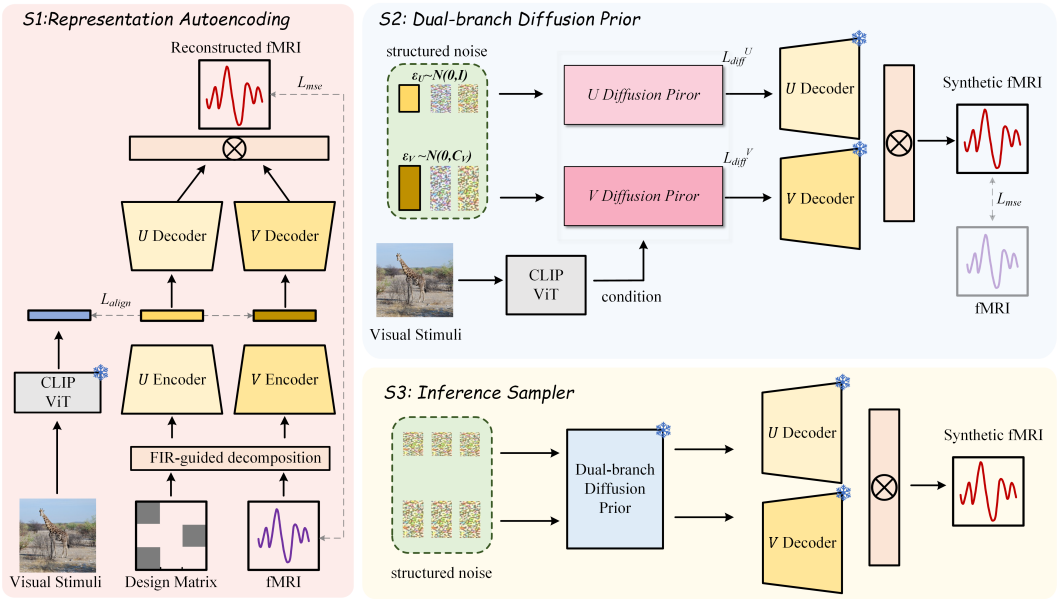

Figure 2: Overview of the proposed **BOLDSyner** framework.

In U_c, voxels with larger-magnitude response loadings for condition c contribute more to the temporal profile. The signed condition-wise loadings are retained in V_c. Collecting all conditions yields $U \in \mathbb{R}^{C\times K}$ and $V \in \mathbb{R}^{C\times N_{\text{voxel}}}$.

These condition-level temporal and spatial factors serve as supervision targets for the representation stage and form the basis of the subsequent diffusion modeling.

3.2 Representation Autoencoding

The BOLD-derived targets differ substantially in dimensionality: U is a low-dimensional temporal response profile, whereas V is a high-dimensional voxel-wise response pattern. We therefore use separate autoencoders for the two factors, mainly to obtain a compact representation of V for subsequent diffusion modeling:

$$Z_U = E_U(U), \quad \hat{U} = D_U(Z_U), \qquad Z_V = E_V(V), \quad \hat{V} = D_V(Z_V). \tag{4}$$

Because U is already low-dimensional, Z_U is used only to reconstruct the temporal factor. In contrast, Z_V is a compact tokenized representation of V and serves as the target of the spatial diffusion branch.

During autoencoder pretraining, we optimize factor reconstruction together with visual–neural alignment and two constraints on the recombined response:

$$\mathcal{L}_{\text{AE}} = \lambda_{\text{rec}} \left(\mathcal{L}_{\text{rec}}^{U} + \mathcal{L}_{\text{rec}}^{V}\right) + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{coef}}\mathcal{L}_{\text{coef}} + \lambda_{\text{resp}}\mathcal{L}_{\text{resp}}^{\text{AE}}, \tag{5}$$

where $\mathcal{L}_{\text{rec}}^{U} = \text{MSE}(\hat{U}, U)$ and $\mathcal{L}_{\text{rec}}^{V} = \text{MSE}(\hat{V}, V)$. The alignment term is applied to Z_V to preserve correspondence between the visual features and the voxel-wise response representation. The coefficient and response terms compare the recombined factors with their targets at the FIR-coefficient and reconstructed-BOLD levels, respectively. Detailed loss formulations are in Appendix A.1.2.

After pretraining, the autoencoders are fixed during diffusion training. The temporal diffusion branch directly models the normalized temporal factor U, whereas the spatial diffusion branch models $Z_V = E_V(V)$. Thus, Z_U is an auxiliary reconstruction code and is not used as a diffusion target.

3.3 Dual-Branch Diffusion Prior

BOLDSyner models the temporal factor U and the spatial representation $Z_V = E_V(V)$ with two conditional diffusion branches. The two variables differ substantially in dimensionality: U is a low-dimensional FIR response profile, whereas Z_V represents a high-dimensional voxel-wise response pattern. We therefore use different perturbations and conditioning for the two branches.

For the temporal branch, we use standard isotropic Gaussian perturbations:

$$r_{U,t_U} = \sqrt{\bar{\alpha}^U_{t_U}}\, U + \sqrt{1-\bar{\alpha}^U_{t_U}}\, \epsilon_U, \qquad \epsilon_U \sim \mathcal{N}(0, I). \tag{6}$$

For the spatial branch, we use correlated Gaussian perturbations:

$$r_{V,t_V} = \sqrt{\bar{\alpha}^V_{t_V}}\, Z_V + \sqrt{1-\bar{\alpha}^V_{t_V}}\, \epsilon_V, \qquad \epsilon_V = C_V^{1/2}\xi_V, \quad \xi_V \sim \mathcal{N}(0, I), \tag{7}$$

so that $\epsilon_V \sim \mathcal{N}(0, C_V)$. The covariance matrix C_V is derived from voxel locations and ROI assignments and mapped to the token representation space. Its construction is detailed in Appendix A.1.3.

The spatial branch predicts its noise term from the noisy spatial representation and visual features. The temporal branch uses the noisy temporal factor, visual features, and a pooled state from the spatial branch. This allows temporal prediction to use the global spatial response associated with the same visual stimulus. The denoising objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t_U,t_V}\left[\lambda_U \left\|\hat{\epsilon}_U - \epsilon_U\right\|_2^2 + \lambda_V \left\|W_V(\hat{\epsilon}_V - \epsilon_V)\right\|_2^2\right], \tag{8}$$

where W_V is the whitening operator associated with C_V. Thus, the temporal branch uses the standard noise-prediction loss, while the spatial branch uses the same covariance matrix in both its forward perturbation and denoising objective. Implementation details are provided in Appendix A.1.4.

After denoising, the spatial representation is decoded to obtain $\hat{V}$, and the temporal prediction gives $\hat{U}$. The two factors are recombined to predict the BOLD response. Branch-wise denoising alone does not ensure that the recombined response matches the measured BOLD signal. We therefore optimize

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{BOLD}}\text{MSE}\left(\mathcal{R}(\hat{U}, \hat{V}), Y\right), \tag{9}$$

where $\mathcal{R}(\cdot,\cdot)$ is the FIR/BOLD reconstruction operator and Y is the measured BOLD response. This final term constrains the recombined temporal and spatial predictions against measured BOLD.

4 Results

We evaluate BOLDSyner on four NSD subjects using the same stimulus split, visual ROI mask, and held-out stimuli across methods (Allen et al., 2022). At the β level, we compare condition-wise voxel responses; at the BOLD level, we compare post-stimulus response windows aligned by stimulus condition and temporal bin.

4.1 Prediction Accuracy at β and BOLD Levels

Table 1 reports three evaluations. The β-level evaluation compares predicted and measured condition-wise voxel responses. For HRF substitution, each predicted voxel-response map is combined with either a mean HRF or an HRF template selected on the validation set. The main BOLD-level evaluation compares predicted and measured response windows using the same temporal window, voxel mask, and evaluation metrics. Linear, MindSimulator (Bao et al., 2025), and SynBrain (Mai et al., 2026) retain their original architectures and training objectives.

BOLDSyner improves BOLD-level prediction without sacrificing β-level accuracy. At the β level, BOLDSyner achieves an MSE of 0.088 and a correlation of 0.739, compared with 0.139 and 0.687 for SynBrain, the best-performing baseline. At the BOLD level, BOLDSyner obtains an MSE of 0.056 and a correlation of 0.610, compared with 0.326 and 0.157 for SynBrain. For BOLDSyner, replacing the predicted temporal response with a fixed HRF reduces the correlation to 0.258–0.260. Thus, the BOLD-level improvement does not come at the cost of condition-wise prediction accuracy and cannot be explained by simply attaching a fixed temporal response to the predicted voxel map.

Figure 3 provides a closer view of the BOLD predictions on Subject 01. Using samples matched by stimulus condition and temporal bin, we compare BOLDSyner with MindSimulator using measured-versus-predicted responses, absolute errors, and mean responses within bins defined by measured BOLD magnitude. BOLDSyner shows closer agreement with measured responses and smaller prediction errors. Results for the remaining subjects are provided in Appendix A.2.1.

Table 1: Average results across four NSD subjects for β-level prediction, fixed-HRF BOLD reconstruction, and BOLD-level prediction. "Val. HRF" denotes the HRF selected on the validation set. Lower MSE and higher correlation indicate better performance; best results shown in bold.

Method	β-level		BOLD-level HRF substitution				Main BOLD-level	
			Mean HRF		Val. HRF			
	MSE↓	Corr↑	MSE↓	Corr↑	MSE↓	Corr↑	MSE↓	Corr↑
Linear	0.400	0.325	0.843	0.025	0.829	0.034	0.752	0.077
MindSimulator (Bao et al., 2025)	0.373	0.366	0.780	0.100	0.749	0.115	0.615	0.131
SynBrain (Mai et al., 2026)	0.139	0.687	0.417	0.141	0.413	0.154	0.326	0.157
BOLDSyner (Ours)	**0.088**	**0.739**	**0.074**	**0.258**	**0.071**	**0.260**	**0.056**	**0.610**

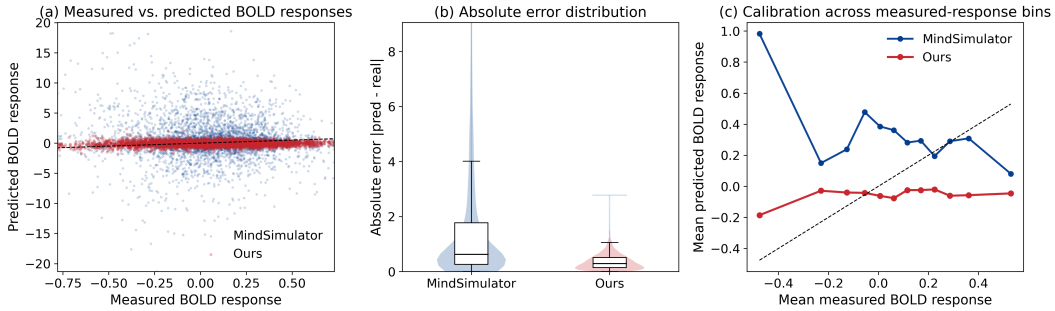

Figure 3: BOLD-level prediction results on Subject 01 using samples aligned by stimulus condition and temporal bin. **(a)** Measured versus predicted responses. **(b)** Absolute prediction errors. **(c)** Mean measured and predicted responses within bins defined by measured BOLD magnitude.

4.2 Temporal Information in Predicted BOLD Responses

A single condition-level β response summarizes response amplitude but does not represent how the response evolves after stimulus onset. We therefore examine whether BOLDSyner preserves this temporal information in the predicted BOLD responses. Across 171 valid subject–ROI pairs from four NSD subjects, the mean correlation between predicted and measured temporal profiles is 0.922 (95% bootstrap CI: 0.902–0.941), with 89.5% of ROI instances achieving $r > 0.8$. Additional ROI-level results are provided in Appendix A.2.2.

We next examine category-dependent temporal differences in Face-, Body-, Place-, and Word-selective ROIs. Figure 4a visualizes the preferred-minus-nonpreferred response contrast; recovery effects are differences between category-specific temporal features (Appendix A.2.2). Measured BOLD responses show clear centroid differences across the predefined ROI–category pairs, with significant effects identified after BH-FDR correction (Fig. 4b).

We then test whether these effects are recovered from predicted BOLD responses. Across 120 subject–ROI–category cells, predicted and measured effects are correlated for centroid ($r = 0.866$), width ($r = 0.732$), and peak latency ($r = 0.709$), with one-sided permutation significance of $p \approx 10^{-4}$ for all three measures (Fig. 4c–e). These results show that predicted BOLD responses preserve category-dependent differences in post-stimulus response timing and temporal profiles that are not represented by a single condition-level β amplitude.

4.3 Category and Semantic Information in Predicted BOLD Responses

We evaluate category selectivity, concept-dependent spatial response patterns, and concept-level representational similarity in predicted BOLD. All BOLDSyner analyses in this section use Subject 01.

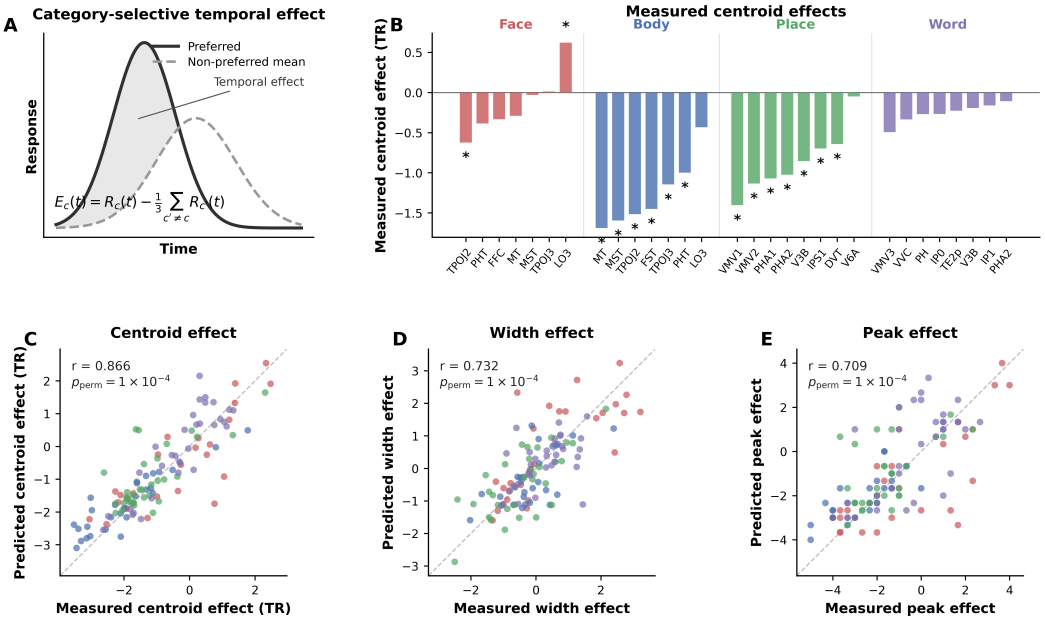

Figure 4: Category-dependent temporal effects in measured and predicted BOLD responses. **(a)** Category-selective response contrast: the preferred-category response profile minus the mean profile of the non-preferred categories. **(b)** Measured centroid effects across predefined Face-, Body-, Place-, and Word-selective ROI–category pairs; stars indicate significant effects after BH-FDR correction. **(c–e)** Predicted versus measured centroid, width, and peak-latency effects across 120 subject–ROI–category cells, computed as differences between category-specific temporal features. Colors denote stimulus categories, and dashed diagonals indicate identity.

Table 2: FLOC-style category evaluation on Places and Bodies. Linear and MindSimulator use their reported results, while BOLDSyner is evaluated from BOLD responses under a fair Top-K protocol.

Top-K images	Places Acc. ↑			Bodies Acc. ↑		
	Linear	MindSimulator	BOLDSyner	Linear	MindSimulator	BOLDSyner
Top 100	0.320	0.526	**0.850**	0.458	**0.917**	0.765
Top 200	0.300	0.456	**0.790**	0.421	**0.842**	0.772
Top 300	0.292	0.413	**0.778**	0.407	**0.814**	0.767
Top 500	0.285	0.373	**0.762**	0.392	0.753	**0.768**
Top 1000	0.279	0.335	**0.743**	0.379	0.642	**0.773**

Category selectivity We conduct a FLOC-style evaluation on Places and Bodies following the Top-K protocol used by MindSimulator. Linear and MindSimulator use their reported FLOC results, while BOLDSyner is evaluated from its predicted BOLD responses under a comparable Top-K protocol. SynBrain is not included because results under this localization protocol are not reported.

As shown in Table 2, BOLDSyner achieves higher Places accuracy than both Linear and MindSimulator across all Top-K settings. For Bodies, it outperforms Linear throughout and exceeds MindSimulator at Top 500 and Top 1000. These results show that predicted BOLD responses retain category-discriminative information used in conventional category-selectivity analyses.

Concept-dependent spatial response patterns We further examine whether semantic concepts remain distinguishable in the predicted spatial responses. We consider five semantic dimensions—animacy, spatial setting, scene naturalness, social content, and textual content—and form concept-specific image sets using text-guided image scoring. For each concept, we compute a one-vs-

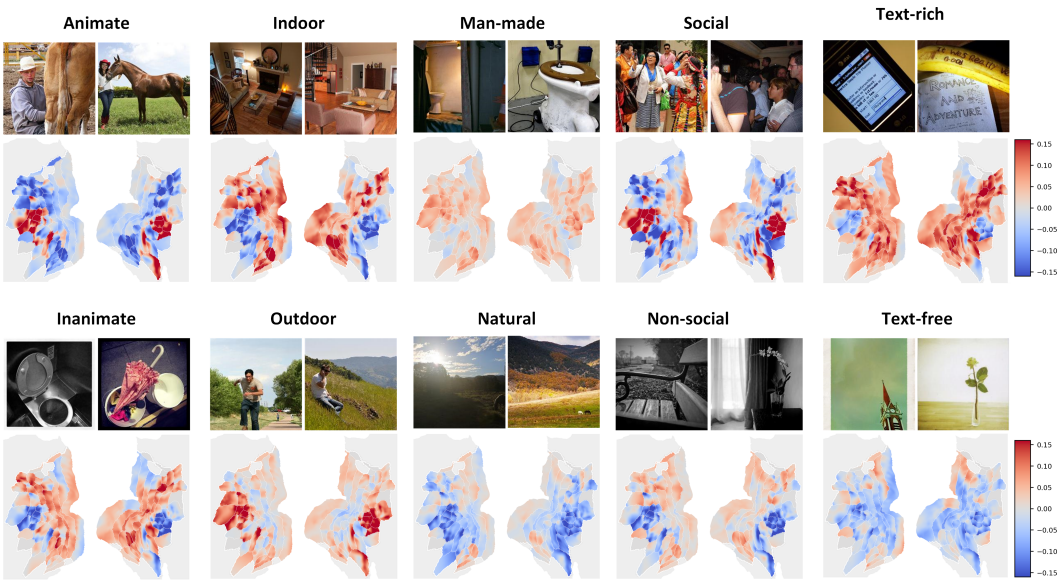

Figure 5: Concept-dependent spatial response patterns in predicted BOLD. Paired concepts are shown with representative images and one-vs-rest response maps on an HCP flat cortical layout.

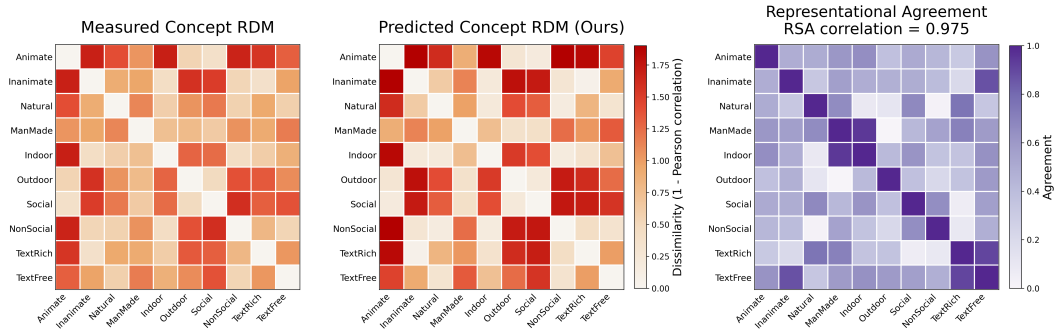

Figure 6: Concept-level representational similarity analysis. RDMs are computed from measured and predicted concept-selective BOLD response patterns, and RSA measures their agreement.

rest response map from the predicted BOLD responses and visualize it on an HCP flat cortical layout (Glasser et al., 2016).

Figure 5 shows distinct spatial response patterns across the evaluated concepts. Qualitatively, some of these differentiated profiles are broadly consistent with known large-scale functional organizations in visual cortex, including posterior temporal responses associated with biological motion and social perception, ventral occipitotemporal responses associated with text-related stimuli, and higher-level visual responses associated with scene and object properties (Allison et al., 2000; Cohen et al., 2000; Dehaene & Cohen, 2011; Epstein & Kanwisher, 1998; Konkle & Oliva, 2012; Vaina et al., 2001). We further average the same concept-selective responses within the 43 HCP-MMP visual ROIs to obtain concept-by-ROI profiles, which are provided in Appendix A.2.3. These results show that concept-related information remains distinguishable in the predicted spatial responses.

Concept-level representational similarity Finally, we test whether predicted BOLD preserves the relationships among semantic concepts observed in measured BOLD. We construct concept-level representational dissimilarity matrices (RDMs) from measured and predicted concept-selective response patterns using one minus Pearson correlation. RSA is then computed between the upper-triangular entries of the two RDMs.

Table 3: Ablation of the dual-branch diffusion design on Subject 01. I denotes isotropic Gaussian perturbation and C_V denotes the covariance-aware perturbation used for the spatial branch. For the single-factor variants, the other factor is provided by the corresponding pretrained autoencoder.

Variant	U perturb.	V perturb.	BOLD Corr ↑	BOLD MSE ↓
AE-only	–	–	0.362	0.087
U-only diffusion	I	–	0.409	**0.051**
V-only diffusion	–	C_V	0.418	0.052
Dual-factor isotropic	I	I	0.262	0.061
Full BOLDSyner	I	C_V	**0.625**	0.052

As shown in Fig. 6, the measured and predicted concept RDMs achieve an RSA correlation of 0.975. This result shows that the predicted BOLD responses preserve the relative relationships among the evaluated semantic concepts. Together, these analyses indicate that predicted BOLD retains category and semantic information alongside the temporal information examined in Sec. 4.2.

4.4 Ablation Study

We evaluate autoencoding alone, single-factor generation, covariance-aware spatial perturbation, and the final BOLD reconstruction loss.

As shown in Table 3, the AE-only model reaches a BOLD correlation of 0.362, showing that the autoencoding stage alone can reconstruct useful response information but is not sufficient for the final prediction task. Adding diffusion to only U or only V increases the correlation to 0.409 and 0.418, respectively, while the full model reaches 0.625. This indicates that both the temporal response profile and the voxel-wise spatial response pattern contribute to BOLD prediction.

Applying isotropic perturbations to both U and V gives a BOLD correlation of 0.262 and an MSE of 0.061. Using isotropic perturbation for U and covariance-aware perturbation for V increases the correlation to 0.625 and reduces the MSE to 0.052.

In a separate paired 300-step diagnostic, removing the final BOLD reconstruction loss increases BOLD MSE from 0.02767 to 3.5710 and decreases correlation from 0.5973 to 0.0591. This shows that directly comparing the recombined prediction with measured BOLD is important for training the two branches to produce a consistent final response. Additional ablations of visual–neural alignment and the autoencoding-stage losses are provided in Appendix A.2.4.

5 Conclusion

We presented BOLDSyner, a visual brain encoding framework that predicts stimulus-aligned BOLD responses while retaining conventional condition-level response prediction. BOLDSyner represents each visual condition with a temporal response profile and a signed voxel-wise spatial response pattern, models the two factors with a dual-branch diffusion prior, and directly constrains their recombined prediction in the measured BOLD space.

Across four NSD subjects, BOLDSyner achieves the best average BOLD-level and β-level prediction performance among the evaluated methods. Relative to conventional β-level representations, the predicted BOLD responses recover category-dependent temporal effects observed in measured responses ($r = 0.866$ across 120 subject–ROI–category cells), while retaining category selectivity and semantic relationships accessible through β-level analyses. Overall, these findings demonstrate the substantial value of extending visual brain encoding beyond static condition-wise responses toward time-resolved BOLD prediction.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and code implementation during the preparation of this work. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected and executed by the authors. All experiments, analyses, and scientific conclusions were independently verified by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses only publicly available neuroimaging data and involves no new data collection, human-subject intervention, or deployment. BOLDSyner is developed for research on visual brain encoding and stimulus-evoked BOLD responses. Although improved neural-response prediction may benefit neuroscience research, neural data can contain sensitive information and therefore require appropriate consent, data governance, access control, and safeguards against unauthorized inference.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of data preprocessing, BOLD target construction, model architecture, training objectives, branch-specific diffusion priors, baseline configurations, and evaluation protocols in the main paper and Appendix. Appendix reports implementation details and selected hyperparameter settings, supplemented by the accompanying configuration files. Source code, configuration files, and reproduction scripts are available through the anonymous repository with the submission.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick N. Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

Truett Allison, Aina Puce, and Gregory McCarthy. Social perception from visual cues: role of the STS region. *Trends in cognitive sciences*, 4(7):267–278, 2000.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95 (3):709–721, 2017.

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

Geoffrey M. Boynton, Stephen A. Engel, Gary H. Glover, and David J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13): 4207–4221, 1996.

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

Karl J. Friston, Andrew P. Holmes, Keith J. Worsley, Jean-Baptiste Poline, Christopher D. Frith, and Richard S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994.

Karl J. Friston, Paul Fletcher, Oliver Josephs, Andrew Holmes, Michael D. Rugg, and Robert Turner. Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7(1):30–40, 1998.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Gary H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9 (4):416–429, 1999.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Christine C Guo, Vinh T Nguyen, Matthew P Hyett, Gordon B Parker, and Michael J Breakspear. Out-of-sync: disrupted neural activity in emotional circuitry during film viewing in melancholic depression. *Scientific reports*, 5(1):11605, 2015.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, 2017.

Michael Hütel, Michela Antonelli, Andrew Melbourne, and Sebastien Ourselin. Hemodynamic matrix factorization for functional magnetic resonance imaging. *NeuroImage*, 231:117814, 2021.

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.

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.

Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012.

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

Martin A. Lindquist, Ji Meng Loh, Lauren Y. Atlas, and Tor D. Wager. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage*, 45(1 Suppl):S187–S198, 2009.

Weijian Mai, Jiamin Wu, Yu Zhu, Zhouheng Yao, Dongzhan Zhou, Andrew Luo, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Synbrain: Enhancing visual-to-fmri synthesis via probabilistic representation learning. *Advances in Neural Information Processing Systems*, 38:29253–29284, 2026.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011.

Fabian Pedregosa, Michael Eickenberg, Bertrand Thirion, and Alexandre Gramfort. HRF estimation improves sensitivity of fMRI encoding and decoding models. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pp. 165–169, 2013.

Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Bertrand Thirion, and Alexandre Gramfort. Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, 104:209–220, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Dustin E. Stansbury, Thomas Naselaris, and Jack L. Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013.

Lucia M Vaina, Jeffrey Solomon, Sanjida Chowdhury, Pawan Sinha, and John W Belliveau. Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences*, 98(20):11656–11661, 2001.

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, 2018.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

A Appendix

A.1 Additional Method Details

This appendix provides implementation details omitted from the main text, including spatiotemporal BOLD target construction, representation autoencoding, dual-branch diffusion prior construction, and experimental settings.

An anonymized implementation of BOLDSyner, together with reproduction scripts and additional experimental details, is available at: `https://anonymous.4open.science/r/BOLDSyner_BOLD-Level-Visual-Brain-Encoding-BFE7/`.

A.1.1 Spatiotemporal BOLD Target Construction Details

This section provides implementation details for the BOLD target construction in Sec. 3.1, including response-window extraction, FIR fitting, rank-spectrum analysis, and condition-level aggregation.

Stimulus-aligned BOLD windows For each subject, stimulus-aligned response windows are extracted from run-level pre-GLM BOLD time series. Stimulus events are obtained from the NSD experimental design files, with each image identity treated as one stimulus condition. Voxel time series are restricted to the `nsdgeneral` mask, detrended, and z-score standardized within each run.

The response offset is configured as 4.6 s and rounded to the nearest TR during extraction. With a repetition time of $4/3$ s, this corresponds to an offset of three TRs (approximately 4.0 s). For each valid stimulus event, we extract $K = 6$ consecutive TR bins from this point, corresponding to an 8.0-s response window. The extracted responses are arranged as

$$Y \in \mathbb{R}^{N_{\text{trial}} \times N_{\text{voxel}} \times K} \longrightarrow Y_{\text{FIR}} \in \mathbb{R}^{N_{\text{trial}} K \times N_{\text{voxel}}}, \tag{10}$$

where each row of Y_{FIR} corresponds to an event–bin pair and each column corresponds to a voxel.

Response-window validation We compare the selected response window with an 8.0-s window beginning at stimulus onset. The early window gives an explained-variance value of 0.4790 and a temporal-SNR value of 0.7345, compared with 0.4842 and 0.7501, respectively, for the selected window. The two windows therefore give similar values for both measures.

FIR design and voxel-level fitting Let C denote the number of stimulus conditions. Each condition is expanded with K one-hot FIR basis functions, yielding

$$X_{\text{FIR}} \in \mathbb{R}^{N_{\text{trial}} K \times CK}, \tag{11}$$

with columns ordered by condition and then by FIR bin. The intercept is modeled separately and is not included in the FIR coefficient matrix.

For voxel i, the condition-by-bin FIR coefficient matrix is parameterized as

$$B_i = v_i u_i^\top, \qquad u_i \in \mathbb{R}^K, \quad v_i \in \mathbb{R}^C, \quad B_i \in \mathbb{R}^{C \times K}, \tag{12}$$

where u_i denotes the temporal response profile and v_i contains the condition-wise response loadings. Following the column ordering of X_{FIR}, we define

$$\beta_i = \left[v_{i,1} u_i^\top, \ldots, v_{i,C} u_i^\top \right]^\top \in \mathbb{R}^{CK}. \tag{13}$$

The factors are estimated by

$$\min_{u_i, v_i, b_i} \frac{1}{2} \left\| Y_i - X_{\text{FIR}} \beta_i - b_i \mathbf{1} \right\|_2^2, \tag{14}$$

where $Y_i = Y_{\text{FIR}}[:, i]$ and b_i is the intercept. The objective is optimized independently for each voxel using L-BFGS-B. To resolve the scale ambiguity between u_i and v_i, the first element of u_i is fixed to one, with the response scale absorbed into v_i.

Rank-spectrum analysis To assess the rank-one approximation, we additionally estimate unconstrained voxel-wise FIR coefficient matrices and examine their singular-value spectra. Across 55,723 voxels from four NSD subjects, the first component explains 49.15% of the variance on average, while the first two and three components explain 71.24% and 82.57% cumulatively. Thus, a substantial fraction of the FIR variance is captured by the first few components, while higher-rank components still account for additional variance. We therefore use rank one as a compact approximation in the current model.

Condition-level aggregation After voxel-level fitting, the temporal target for image condition c is obtained by weighting the voxel-wise temporal profiles according to the magnitude of their condition-specific loadings. Specifically,

$$w_{i,c} = |v_{i,c}| + \epsilon, \qquad U_c = \frac{\sum_{i=1}^{N_{\text{voxel}}} w_{i,c} u_i}{\sum_{i=1}^{N_{\text{voxel}}} w_{i,c}}, \qquad U_c \in \mathbb{R}^K, \tag{15}$$

where ϵ is a small constant for numerical stability. The absolute loading magnitude gives greater weight to voxels with larger-magnitude condition-specific loadings while preventing positive and negative loadings from canceling.

The spatial target retains the signed condition-wise loading of each voxel:

$$V_c = [v_{1,c}, v_{2,c}, \ldots, v_{N_{\text{voxel}},c}]^\top, \qquad V_c \in \mathbb{R}^{N_{\text{voxel}}}. \tag{16}$$

Thus, U_c is a condition-level temporal profile obtained from the voxel-wise profiles, whereas V_c contains the signed voxel-wise response loadings.

Collecting all conditions gives

$$U \in \mathbb{R}^{C \times K}, \qquad V \in \mathbb{R}^{C \times N_{\text{voxel}}}. \tag{17}$$

A.1.2 Representation Autoencoding Details

The representation autoencoding stage preserves the BOLD-derived temporal and spatial factors while providing a compact spatial representation for subsequent diffusion modeling. Because the temporal factor U is low-dimensional whereas the spatial factor V contains voxel-wise response loadings, separate autoencoders are used for the two factors.

Given the BOLD-derived factors (U, V), the two branches are defined as

$$Z_U = E_U(U), \quad \hat{U} = D_U(Z_U), \qquad Z_V = E_V(V), \quad \hat{V} = D_V(Z_V). \tag{18}$$

The temporal branch uses an MLP autoencoder, whereas the spatial branch uses a residual autoencoder that maps the voxel-wise spatial factor to a tokenized representation Z_V. The temporal code Z_U is used only for reconstructing U and is not used as a diffusion target. In contrast, Z_V is modeled by the subsequent spatial diffusion branch.

The autoencoding objective is

$$\mathcal{L}_{\text{AE}} = \lambda_{\text{rec}} \left(\mathcal{L}^U_{\text{rec}} + \mathcal{L}^V_{\text{rec}} \right) + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{coef}} \mathcal{L}_{\text{coef}} + \lambda_{\text{resp}} \mathcal{L}^{\text{AE}}_{\text{resp}}. \tag{19}$$

The factor-wise reconstruction losses are

$$\mathcal{L}^U_{\text{rec}} = \text{MSE}(\hat{U}, U), \qquad \mathcal{L}^V_{\text{rec}} = \text{MSE}(\hat{V}, V). \tag{20}$$

The visual–neural alignment loss is applied only to the spatial branch:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{NCE}} \left(g_x(f(x)), g_V(\text{Pool}(Z_V)) \right), \tag{21}$$

where $f(x)$ denotes features extracted by the frozen visual encoder, and $g_x(\cdot)$ and $g_V(\cdot)$ are projection heads. The pooled spatial representation is used only for visual–neural alignment and is not itself used as the spatial diffusion target. No corresponding alignment loss is applied to Z_U.

For the following losses, any normalization applied to the factor inputs is inverted before recomposition. For image condition c, the temporal and spatial factors are recombined into the voxel-by-FIR coefficient matrix

$$\mathcal{B}_c(U_c, V_c) = V_c U_c^\top \in \mathbb{R}^{N_{\text{voxel}} \times K}. \tag{22}$$

The coefficient-space consistency loss is then

$$\mathcal{L}_{\text{coef}} = \text{MSE}\left(\mathcal{B}(\hat{U}, \hat{V}), \mathcal{B}(U, V)\right), \tag{23}$$

where $\mathcal{B}(U, V)$ collects the recomposed FIR coefficients across conditions. This loss constrains the two reconstructed factors jointly after recomposition rather than only in their individual factor spaces.

We further impose a response-level consistency loss. In the implementation, the FIR design maps the temporal factor to a stimulus-aligned response, while a linear readout of the spatial factor provides its corresponding response amplitude. We denote this reconstruction operation by $\mathcal{R}(\cdot, \cdot)$ and define

$$\mathcal{L}_{\text{resp}}^{\text{AE}} = \text{MSE}\left(\mathcal{R}(\hat{U}, \hat{V}), \mathcal{R}(U, V)\right). \tag{24}$$

Thus, the response loss requires the autoencoded factors to preserve the response implied by the original BOLD-derived factors after recombination. It does not directly use the measured BOLD time series as the target during autoencoder pretraining.

After autoencoder pretraining, the representation modules are fixed during diffusion-prior training. The temporal diffusion branch directly models U, whereas the spatial diffusion branch models the tokenized representation $Z_V = E_V(V)$. Accordingly, neither the auxiliary temporal code Z_U nor the pooled representation $\text{Pool}(Z_V)$ is used as a diffusion target.

A.1.3 Dual-Branch Diffusion Prior Details

This section provides additional details of the asymmetric conditional networks and the spatial covariance used by the dual-branch diffusion prior.

Asymmetric conditional architecture The spatial branch operates on the noisy token representation r_{V,t_V}. After projection and the addition of positional and diffusion-timestep embeddings, the token sequence is processed by a transformer conditioned on the visual representation $f(x)$:

$$h_{V,t_V} = \Phi_V(r_{V,t_V}, t_V, f(x)), \qquad \hat{\epsilon}_V = H_V(h_{V,t_V}, t_V). \tag{25}$$

The temporal branch directly processes the low-dimensional noisy temporal factor r_{U,t_U} with a conditional MLP. Its conditioning combines the projected visual feature with a pooled spatial hidden state:

$$c_U = [f_U(x), \text{Pool}(h_{V,t_V})], \qquad \hat{\epsilon}_U = H_U(r_{U,t_U}, t_U, c_U). \tag{26}$$

Thus, the temporal branch operates directly on the FIR response profile while receiving both visual conditioning and a summary of the spatial branch.

Spatial covariance construction For each voxel retained by the subject-specific `nsdgeneral` mask, we obtain its world coordinate from the reference NIfTI affine and assign its ROI label according to the HCP-MMP1.0 parcellation over the 43 visual ROIs used in our analysis. The voxel-space covariance kernel combines spatial proximity and ROI membership:

$$K_x[p, q] = k(\mathbf{r}_p, \mathbf{r}_q)\, \gamma_{\text{ROI}}(p, q), \qquad \gamma_{\text{ROI}}(p, q) = \begin{cases} \gamma_{\text{same}}, & \rho_p = \rho_q, \\ \gamma_{\text{diff}}, & \rho_p \neq \rho_q, \end{cases} \tag{27}$$

where $\mathbf{r}_p$ and ρ_p denote the coordinate and ROI label of voxel p. Let M denote the mapping between the token representation and voxel-response spaces. The covariance used by the spatial diffusion branch is

$$C_V = M^\top K_x M. \tag{28}$$

Spatially correlated perturbations are sampled as

$$\epsilon_V = C_V^{1/2} \xi_V, \qquad \xi_V \sim \mathcal{N}(0, I), \tag{29}$$

so that $\epsilon_V \sim \mathcal{N}(0, C_V)$. A small diagonal jitter and, when necessary, eigenvalue clipping are used for numerical stability.

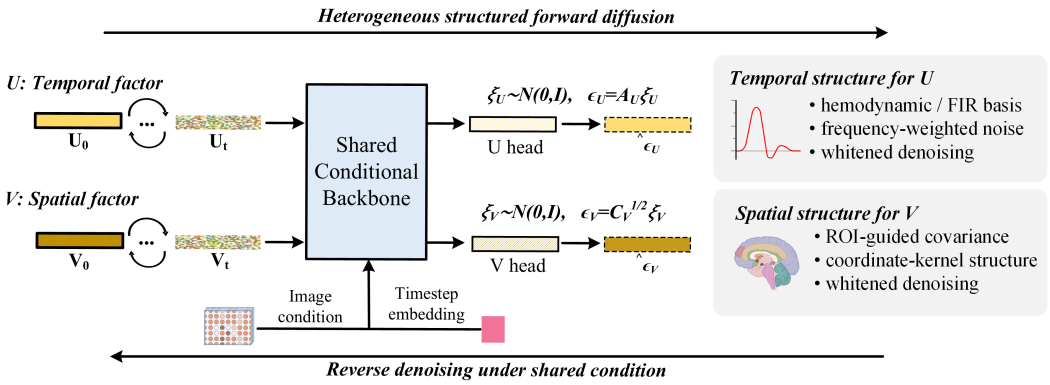

Figure 7: Dual-branch diffusion prior in BOLDSyner. The temporal branch directly models the temporal factor U with isotropic Gaussian perturbations, whereas the spatial branch models the tokenized spatial representation Z_V with covariance-aware perturbations. The temporal branch is additionally conditioned on a pooled spatial hidden representation. The generated factors are decoded, recombined, and jointly constrained in the measured BOLD response space.

A.1.4 Diffusion Objectives and BOLD-level Supervision

The temporal diffusion branch directly models the normalized temporal factor U, whereas the spatial branch models the tokenized representation Z_V. Accordingly, their clean states are $r_U = U$ and $r_V = Z_V$. The temporal branch uses isotropic Gaussian perturbations, $\epsilon_U \sim \mathcal{N}(0, I)$, while the spatial branch uses the covariance-aware perturbations defined in Appendix A.1.3.

The two branches use separate diffusion timesteps t_U and t_V and predict their corresponding noise terms. The denoising objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t_U, t_V}\left[\lambda_U \left\|\hat{\epsilon}_U - \epsilon_U\right\|_2^2 + \lambda_V \left\|W_V\left(\hat{\epsilon}_V - \epsilon_V\right)\right\|_2^2\right], \tag{30}$$

where W_V is the whitening operator associated with C_V. With numerical stabilization,

$$W_V \approx C_V^{-1/2}. \tag{31}$$

Thus, the temporal branch uses the standard isotropic noise-prediction loss, whereas the spatial denoising error is evaluated in the whitened space associated with C_V.

After denoising, the temporal state is inverse-standardized to obtain $\hat{U}$, while the spatial state is decoded by the pretrained spatial decoder and inverse-standardized to obtain $\hat{V}$. The generated factors are recombined through the FIR/BOLD reconstruction operator, and the BOLD-level loss is

$$\mathcal{L}_{\text{BOLD}}^{\text{diff}} = \text{MSE}\left(\mathcal{R}(\hat{U}, \hat{V}), Y\right), \qquad \mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{BOLD}}\mathcal{L}_{\text{BOLD}}^{\text{diff}}, \tag{32}$$

where Y denotes the measured BOLD response.

Unlike the autoencoding-stage response-consistency loss, which compares responses reconstructed from original and autoencoded BOLD-derived factors, this term directly compares the response reconstructed from the visually generated factors with the measured BOLD response.

A.1.5 Experimental Settings and Implementation Details

Dataset and splits We use the Natural Scenes Dataset (NSD) (Allen et al., 2022) and evaluate subj01, subj02, subj05, and subj07. For each subject, approximately 9,000 subject-specific unique images are used for training, while approximately 1,000 shared images are held out for evaluation. A validation subset is separated from the training data for model and checkpoint selection. The validation and test sets are disjoint, and repeated responses to held-out images are averaged before evaluation. All compared methods use the same train/validation/test split and evaluation mask.

Visual conditioning and BOLD targets Visual features are extracted using a pretrained CLIP ViT-L/14 image encoder whose backbone remains frozen throughout training. The BOLD targets are the condition-level temporal factor U and voxel-wise spatial factor V defined in Appendix A.1.1. We use $K = 6$ FIR bins, and the spatial factor is defined over voxels retained by the subject-specific `nsdgeneral` mask. These voxels are further assigned to the 43 predefined visual ROIs according to the HCP-MMP1.0 parcellation for ROI-based analyses and spatial covariance construction.

Training The temporal and spatial autoencoders are pretrained before diffusion training and are subsequently kept fixed. For the conditional diffusion prior, we use AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, batch size 32, and 200 training epochs. The diffusion process uses 1,000 timesteps with a linear noise schedule from $\beta_{\text{start}} = 10^{-4}$ to $\beta_{\text{end}} = 2 \times 10^{-2}$. The checkpoint used for test evaluation is selected on the validation set.

Inference and evaluation At inference, we use a predictor–corrector sampler with 100 sampling steps and generate one response sample for each held-out stimulus. Stochastic evaluations are repeated five times with different random seeds, and the resulting metrics are averaged. The prediction-accuracy results in Table 1 are arithmetic means across the four subjects; Table 4 reports subject-wise results.

Computational environment All models are implemented in PyTorch and trained on Linux with CUDA-enabled NVIDIA GPUs. The main experiments use a single NVIDIA RTX 4090 GPU with 24 GB memory per subject. A complete single-subject run, including autoencoder pretraining, diffusion training, and test inference, requires approximately 40 GPU hours.

A.2 Additional Results and Analysis Details

A.2.1 Additional Prediction Accuracy Results

This section provides the subject-wise results and additional BOLD-level diagnostics corresponding to Sec. 4.1.

Evaluation protocol The evaluated methods use different native neural-response targets. MindSimulator and SynBrain predict voxel-wise β-level responses rather than stimulus-aligned pre-GLM BOLD responses. We therefore retain their original architectures, training objectives, and prediction targets, and evaluate their predicted outputs under the same subject-specific split, voxel mask, stimulus alignment, and metrics used for BOLDSyner.

We report three evaluations. The β-level evaluation compares condition-wise voxel responses. For BOLDSyner, the β-level prediction is given by the scale-restored predicted spatial factor $\hat{V}$, which is evaluated as a condition-wise voxel-response map against the same GLM-estimated β target used for the baseline methods. The HRF-substitution evaluation combines each predicted voxel-response map with either a mean HRF or an HRF template selected on the validation set. The main BOLD-level evaluation compares predicted and measured responses over the same stimulus-aligned temporal window. Thus, the baseline methods are evaluated from their original predicted response representations rather than retrained as BOLD-prediction models.

Subject-wise results Table 4 gives the subject-wise results corresponding to the four-subject averages in Table 1. In the main BOLD-level evaluation, BOLDSyner achieves the lowest MSE and highest correlation for all four subjects. At the β level, it achieves the highest correlation for all four subjects and the lowest or tied-lowest MSE. For BOLDSyner, replacing the predicted temporal response with a fixed HRF consistently gives substantially lower BOLD correlations than the main BOLD prediction, consistent with the averaged results in Sec. 4.1.

Additional BOLD-level diagnostics Figures 8, 9, and 10 provide additional subject-wise examples corresponding to Fig. 3. For Subjects 02 and 07, we show measured versus predicted responses, absolute prediction errors, and mean responses within bins defined by measured BOLD magnitude. All comparisons use matched samples aligned by stimulus condition and temporal bin. BOLDSyner shows closer agreement with measured responses and lower prediction errors than MindSimulator, consistent with the quantitative results in Table 4.

Table 4: Subject-wise results corresponding to Table 1. We report β-level prediction, BOLD reconstruction using fixed HRFs, and the main BOLD-level evaluation. "Val. HRF" denotes the HRF template selected on the validation set. Lower MSE and higher correlation indicate better performance; the best result within each subject is shown in bold.

Subject	Method	β-level		BOLD-level HRF substitution				Main BOLD-level	
		MSE ↓	Corr ↑	Mean HRF MSE ↓	Mean HRF Corr ↑	Val. HRF MSE ↓	Val. HRF Corr ↑	MSE ↓	Corr ↑
01	Linear	0.406	0.325	0.917	0.014	0.877	0.016	0.823	0.032
	MindSimulator	0.372	0.386	0.817	0.072	0.751	0.111	0.744	0.113
	SynBrain	0.079	0.715	0.224	0.138	0.216	0.157	0.203	0.172
	BOLDSyner (Ours)	**0.024**	**0.719**	**0.056**	**0.268**	**0.056**	**0.269**	**0.052**	**0.625**
02	Linear	0.391	0.358	0.799	0.045	0.780	0.054	0.667	0.147
	MindSimulator	0.375	0.386	0.790	0.095	0.748	0.101	0.528	0.115
	SynBrain	**0.134**	0.667	0.576	0.146	0.575	0.158	0.450	0.132
	BOLDSyner (Ours)	**0.134**	**0.754**	**0.067**	**0.303**	**0.066**	**0.304**	**0.046**	**0.597**
05	Linear	0.407	0.353	0.779	0.026	0.779	0.041	0.789	0.075
	MindSimulator	0.371	0.387	0.801	0.126	0.786	0.134	0.581	0.165
	SynBrain	0.190	0.739	0.490	0.134	0.484	0.147	0.324	0.157
	BOLDSyner (Ours)	**0.064**	**0.750**	**0.079**	**0.274**	**0.071**	**0.276**	**0.043**	**0.644**
07	Linear	0.396	0.265	0.878	0.016	0.878	0.023	0.730	0.054
	MindSimulator	0.373	0.303	0.712	0.107	0.709	0.112	0.608	0.130
	SynBrain	0.151	0.626	0.377	0.144	0.376	0.152	0.325	0.165
	BOLDSyner (Ours)	**0.128**	**0.732**	**0.093**	**0.188**	**0.089**	**0.189**	**0.084**	**0.575**

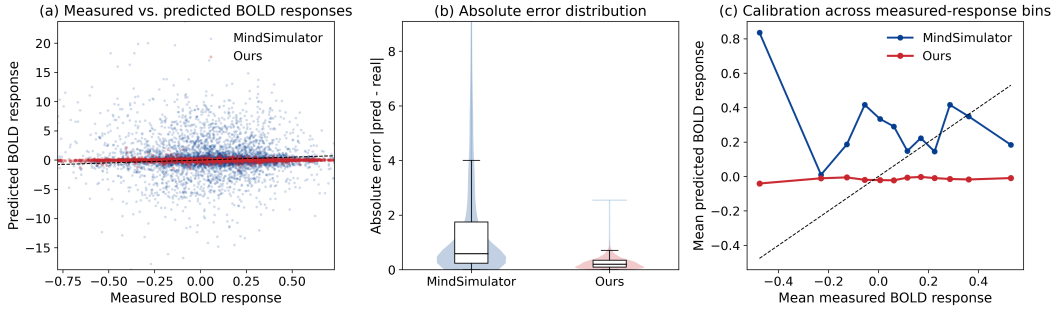

Figure 8: Additional BOLD-level prediction results for Subject 02. Panels show measured versus predicted responses, absolute prediction errors, and mean measured and predicted responses within bins defined by measured BOLD magnitude. Samples are aligned by stimulus condition and temporal bin.

For Subject 05, the sampled responses span a narrower amplitude range, so we additionally show error-based views: the absolute-error distribution, mean absolute error across measured-response bins, and cumulative absolute-error distribution. The evaluation itself is unchanged. These views likewise show lower BOLD-level prediction errors for BOLDSyner than for MindSimulator.

A.2.2 Additional Temporal Analyses

This section provides additional analyses corresponding to Sec. 4.2. We first show hierarchy-level BOLD response profiles for all four subjects, then report temporal results at the individual-ROI level, and finally provide additional analyses of category-dependent temporal effects.

Hierarchy-level temporal profiles As a supplementary visualization, we group the 43 visual ROIs into low-, mid-, and high-level regions using the fixed hierarchy definition in Table 5. For each subject, measured and predicted BOLD responses are averaged within each hierarchy group across the post-stimulus FIR bins. Figures 11, 12, 13, and 14 show the results for Subjects 01, 02, 05, and 07, respectively.

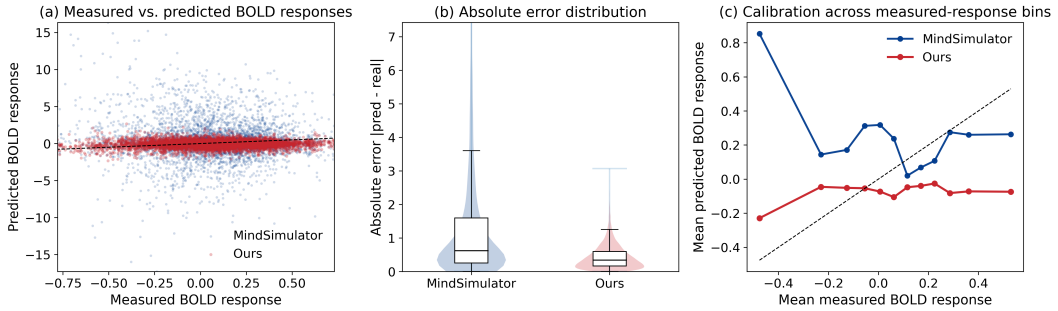

Figure 9: Additional BOLD-level prediction results for Subject 07. Panels show measured versus predicted responses, absolute prediction errors, and mean measured and predicted responses within bins defined by measured BOLD magnitude. Samples are aligned by stimulus condition and temporal bin.

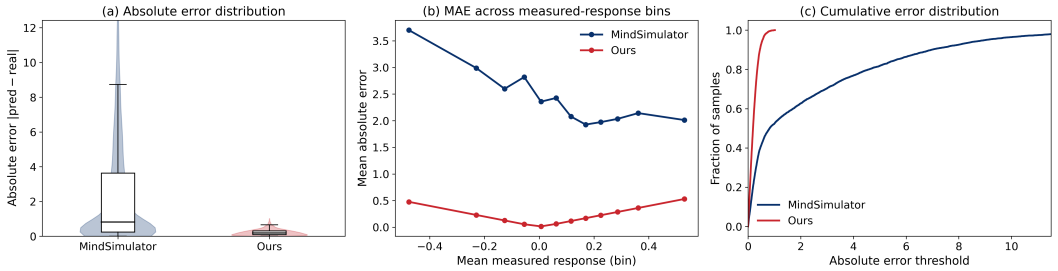

Figure 10: Additional BOLD-level prediction results for Subject 05. Panels show the absolute prediction-error distribution, mean absolute error across measured-response bins, and cumulative absolute-error distribution. The quantitative evaluation is the same as for the other subjects.

Across subjects, the predicted profiles follow the main temporal changes in the measured responses, including the rise, peak or plateau, and subsequent decline. These plots provide a summary after averaging across broad ROI groups. The individual-ROI analysis below evaluates temporal agreement without this hierarchy-level averaging.

Individual-ROI temporal profiles To evaluate temporal agreement more directly, we compare measured and predicted temporal profiles at the individual-ROI level using the same FIR setting ($K = 6$). Across the four NSD subjects, 171 valid subject–ROI pairs are included. Of the 172 possible subject–ROI pairs, Subject 01's 7PL contains no voxels within the `nsdgeneral` analysis mask, leaving 171 valid pairs. The mean ROI-wise temporal-profile correlation is 0.922, with a 95% bootstrap confidence interval of 0.902–0.941. In addition, 89.5% of ROI instances achieve $r > 0.8$, and 79.5% achieve $r > 0.9$.

We also compare temporal centroid and response width derived from the measured and predicted profiles. Temporal centroid shows a predicted–measured correlation of $r = 0.744$ with a mean absolute error of 0.160, while response width shows a correlation of $r = 0.589$. These results show that the agreement between measured and predicted temporal responses is also present at the individual-ROI level rather than arising only from averaging across broad ROI groups.

Category-dependent temporal effects We further examine temporal differences within category-selective visual regions. The 30 ROI–category pairs are fixed from the most strongly activated HCP-MMP ROIs in the existing spatial category-selectivity analysis: seven pairs each for Face and Body, and eight each for Place and Word. The selection criterion is spatial category selectivity; the same pairing is used for all four subjects in the temporal analysis. For each subject, ROI, and

Table 5: Hierarchy-based grouping of the 43 visual ROIs used for the supplementary temporal-profile visualization.

Level	Interpretation	ROIs
Low-level	Early visual areas	V1, V2, V3, V4
Mid-level	Intermediate extrastriate, dorsal/lateral, and motion-related visual areas	V3A, V3B, V3CD, V4t, V6, V6A, V7, V8, LO1, LO2, LO3, MT, MST, FST, IPS1, MIP, PreS
High-level	Higher-order ventral, temporal, parietal, and visual association areas	FFC, VVC, VMV1, VMV2, VMV3, PHA1, PHA2, PHA3, PIT, PH, PHT, TE1p, TE2p, TF, TPOJ2, TPOJ3, PGp, PGi, PGs, IP0, IP1, DVT

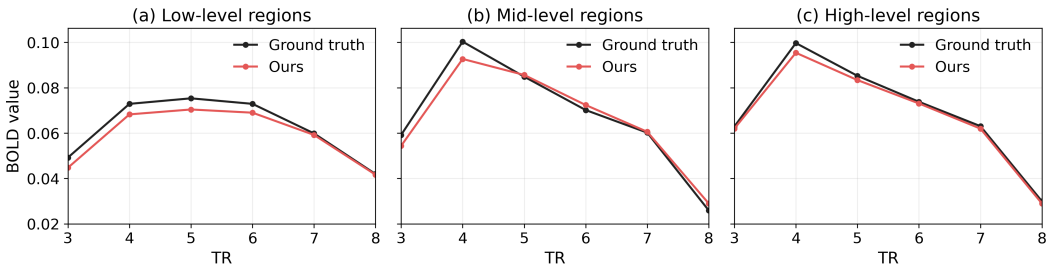

Figure 11: Hierarchy-level BOLD response profiles for Subject 01. Curves show group-averaged measured and predicted responses across post-stimulus FIR bins for low-, mid-, and high-level visual ROI groups.

preferred category c, we visualize the category-dependent response contrast as

$$E_{r,c}(t) = P_{r,c}(t) - \frac{1}{|\mathcal{C}| - 1} \sum_{c' \neq c} P_{r,c'}(t), \tag{33}$$

where $P_{r,c}(t)$ denotes the temporal response profile for category c in ROI r, and $\mathcal{C} = \{\text{Face}, \text{Body}, \text{Place}, \text{Word}\}$. Each profile $P_{r,c}$ is averaged over images of category c after averaging repeated presentations of each image. For a profile $P = (P_1, \ldots, P_K)$, let $t_k = k + 2$ denote the sampled time in TR units and set $w_k = \max(P_k, 0)$. If $\sum_k w_k \leq 10^{-12}$, we instead use $w_k = \max(P_k - \min_j P_j, 0)$; if the total weight remains at most 10^{-12}, we use uniform weights. The temporal features are

$$\begin{aligned} \mu(P) &= \frac{\sum_k t_k w_k}{\sum_k w_k}, \\ \omega(P) &= \frac{\sum_k (t_k - \mu(P))^2 w_k}{\sum_k w_k}, \\ \tau(P) &= t_{\arg\max_k P_k}. \end{aligned} \tag{34}$$

Here, μ is the temporal centroid, ω is response width defined as the weighted temporal variance, and τ is peak latency, with ties resolved by the earliest bin. Centroid and peak latency are measured in TR units, and width in TR^2. For the recovery correlations, the effect for each feature $m \in \{\mu, \omega, \tau\}$ is computed from the category-specific profiles as

$$\Delta^m_{r,c} = m(P_{r,c}) - \frac{1}{|\mathcal{C}| - 1} \sum_{c' \neq c} m(P_{r,c'}). \tag{35}$$

Among the 30 predefined ROI–category pairs, 15 measured temporal-centroid effects are significant at $q < 0.05$ after BH-FDR correction across these 30 pairs, using two-sided image-level permutation p-values from 5,000 permutations. Most of these significant effects occur in Body-related ROIs (6/7)

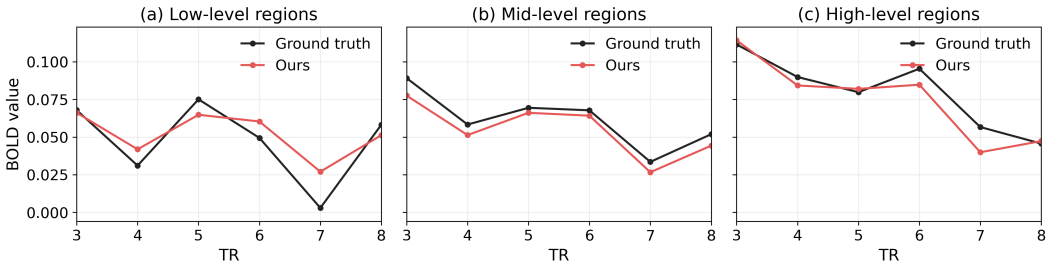

Figure 12: Hierarchy-level BOLD response profiles for Subject 02. Each panel compares group-averaged measured and predicted responses across post-stimulus FIR bins for low-, mid-, and high-level visual ROI groups.

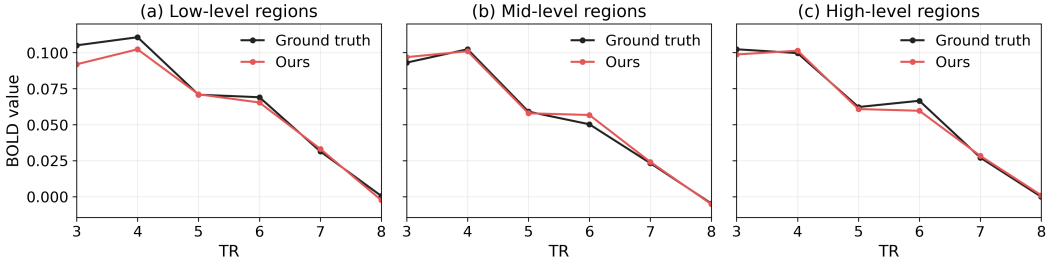

Figure 13: Hierarchy-level BOLD response profiles for Subject 05. Each panel compares group-averaged measured and predicted responses across post-stimulus FIR bins for low-, mid-, and high-level visual ROI groups.

and Place-related ROIs ($7/8$). Across all 120 subject–ROI–category cells, predicted and measured temporal-centroid effects have a correlation of $r = 0.866$ and a direction agreement of 90.8%. A one-sided category-label permutation test with 10,000 permutations gives $p \approx 10^{-4}$. For each subject–ROI combination, we permute the predicted category labels across the four categories and recompute $\Delta^m_{r,c}$, while keeping the measured effects fixed. The null statistic is the Pearson correlation across the 120 cells, and the one-sided p-value is $(1 + \#\{r_{\text{perm}} \geq r_{\text{obs}}\})/10{,}001$.

The same analysis gives a predicted–measured correlation of $r = 0.732$ for response-width effects, with 82.5% direction agreement, and $r = 0.709$ for peak-latency effects, with 81.7% direction agreement. The corresponding one-sided category-label permutation tests also give $p \approx 10^{-4}$ for both measures. These results show that the agreement between measured and predicted category-dependent temporal effects is not limited to the centroid measure.

Figure 15 further illustrates the Body–MT temporal effect. The solid and dashed curves in Fig. 15a show the mean measured and predicted Body-preference temporal effects across the four subjects, respectively. The shaded regions indicate the corresponding mean $\pm$ SEM across subjects. The two curves show similar temporal changes across the post-stimulus window. Figure 15b places the four subject-specific Body–MT effects within the population-level centroid comparison, linking this ROI–category example to the overall predicted–measured correspondence.

A.2.3 Additional Category and Semantic Analyses

This section provides additional analyses corresponding to Sec. 4.3. All BOLDSyner analyses in this subsection use Subject 01, including the four-class FLOC evaluation, concept maps, ROI profiles, and extended RSA. We first extend the category-selectivity analysis to Face, Body, Place, and Word, then report ROI-level concept-response profiles and additional concept maps, and finally evaluate representational similarity over a broader 24-concept set.

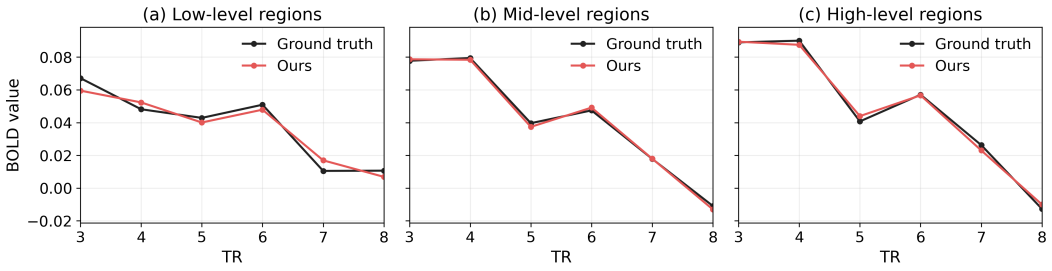

Figure 14: Hierarchy-level BOLD response profiles for Subject 07. Each panel compares group-averaged measured and predicted responses across post-stimulus FIR bins for low-, mid-, and high-level visual ROI groups.

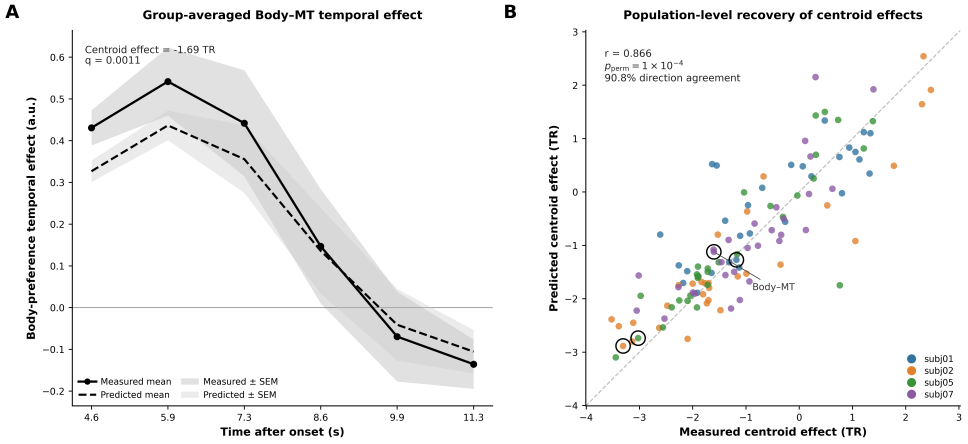

Figure 15: Body–MT temporal effect and population-level centroid recovery. **(a)** Body-preference temporal effects averaged across four subjects. The solid line shows the mean measured effect and the dashed line shows the mean predicted effect. Shaded regions indicate mean ± SEM across the four subjects. The temporal effect is defined as the preferred-category response minus the mean response of the non-preferred categories. **(b)** Predicted versus measured temporal-centroid effects across 120 subject–ROI–category cells. The highlighted points denote the four subject-specific Body–MT cases. Across all cells, predicted and measured centroid effects are correlated at $r = 0.866$, with 90.8% direction agreement and $p_{\text{perm}} = 10^{-4}$.

Four-class category analysis We first extend the FLOC-style analysis to four visual categories: Face, Body, Place, and Word. For each category, we compute a one-vs-rest response map from the predicted BOLD responses. Figure 16 shows the resulting spatial response patterns on the HCP-MMP flat cortical layout.

Table 6 reports one-vs-rest specificity and accuracy across different Top-K image sets using the same image-selection procedure as in the main-text FLOC-style evaluation. Across Face, Body, Place, and Word, the predicted BOLD responses retain category-discriminative information over a range of Top-K settings, although the absolute values differ across categories.

ROI-level concept response profiles To complement the concept-dependent cortical maps in Sec. 4.3, we average the same one-vs-rest responses within the 43 HCP-MMP visual ROIs. This yields a concept-by-ROI profile for each semantic concept. Figure 17 shows the ten concept poles used in the main-text analysis, and Fig. 18 extends the visualization to a broader set of image-defined concepts.

The ROI-level profiles differ across concepts, consistent with the concept-dependent spatial response patterns shown in the main text. These plots provide a parcel-level view of the same predicted responses without making additional functional-localization claims.

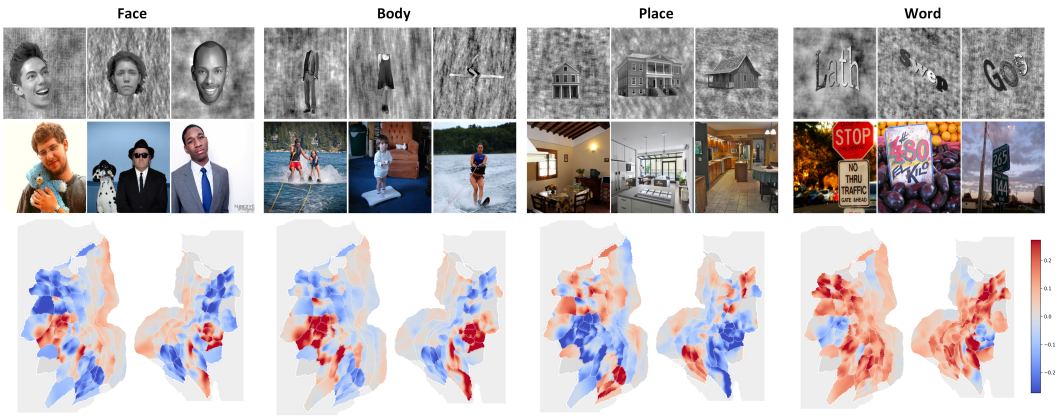

Figure 16: Four-class category-selective response patterns in predicted BOLD. For each category, we show NSD fLoc exemplars, selected natural images, and the corresponding one-vs-rest response map on an HCP-MMP flat cortical layout.

Table 6: Four-class FLOC-style evaluation using predicted BOLD responses. For each category and Top-K image set, we report one-vs-rest specificity (Spec.) and accuracy (Acc.). Higher values indicate better performance.

Top-K images	Face		Body		Place		Word	
	Spec.	Acc.	Spec.	Acc.	Spec.	Acc.	Spec.	Acc.
Top 100	0.867	0.778	0.947	0.765	0.921	0.850	0.586	0.654
Top 200	0.850	0.775	0.955	0.772	0.919	0.790	0.568	0.625
Top 300	0.847	0.743	0.932	0.767	0.893	0.778	0.594	0.617
Top 500	0.866	0.760	0.924	0.768	0.870	0.762	0.588	0.604
Top 1000	0.871	0.752	0.907	0.773	0.881	0.743	0.567	0.577

Additional concept maps and extended RSA We further examine 24 image-defined concepts spanning human-related, animal-related, scene-related, place-related, object/content-related, and context-related categories. Figure 19 shows the corresponding one-vs-rest response maps derived from predicted BOLD responses. These maps are used to visualize differences among predicted concept responses rather than to assign specific functions to individual cortical regions.

We finally repeat the representational similarity analysis using the same 24 concepts. Measured and predicted RDMs are constructed from the one-vs-rest concept-response patterns using one minus Pearson correlation as the dissimilarity measure. RSA is computed from the off-diagonal upper-triangular entries of the two RDMs. As shown in Fig. 20, the measured and predicted RDMs have an RSA correlation of $r = 0.742$. Thus, the predicted responses retain a positive correspondence with measured BOLD in the relative relationships among concepts when the analysis is extended to this broader concept set.

A.2.4 Additional Ablation Analyses

Section 4.4 evaluates the main components of BOLDSyner. Here, we provide additional details for the final BOLD-loss diagnostic and ablations of auxiliary representation-stage terms. Unless otherwise specified, the experiments are conducted on Subject 01.

Final BOLD reconstruction loss To isolate the effect of the final BOLD reconstruction loss, we conduct paired 300-step runs that differ only in whether this loss is enabled. Removing the loss increases BOLD MSE from 0.02767 to 3.5710 and decreases correlation from 0.5973 to 0.0591. This confirms that the final BOLD loss is important for constraining the temporal and spatial predictions

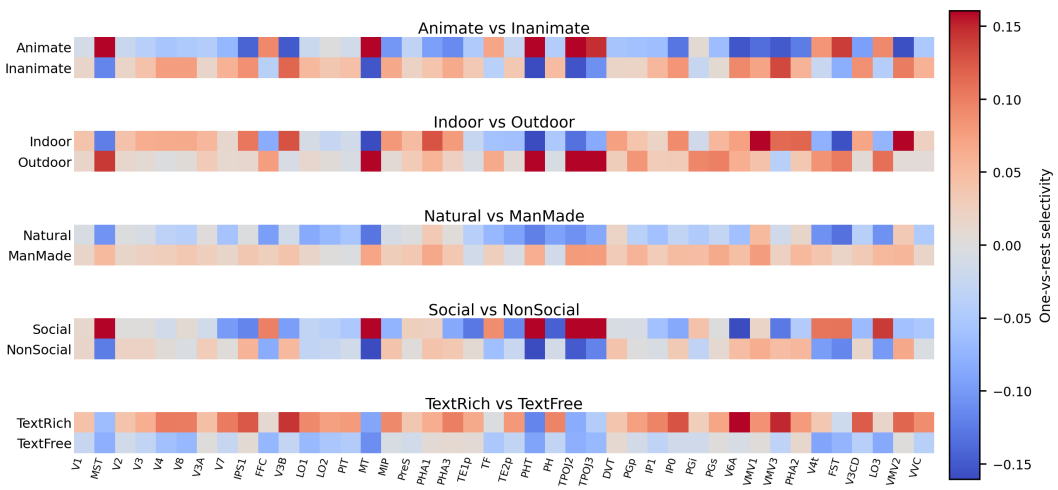

Figure 17: ROI-level response profiles for the semantic concepts used in the main text. Rows correspond to semantic concepts and columns to the 43 HCP-MMP visual ROIs. Values are one-vs-rest response scores computed from predicted BOLD responses and averaged within each ROI. All rows use the same color scale.

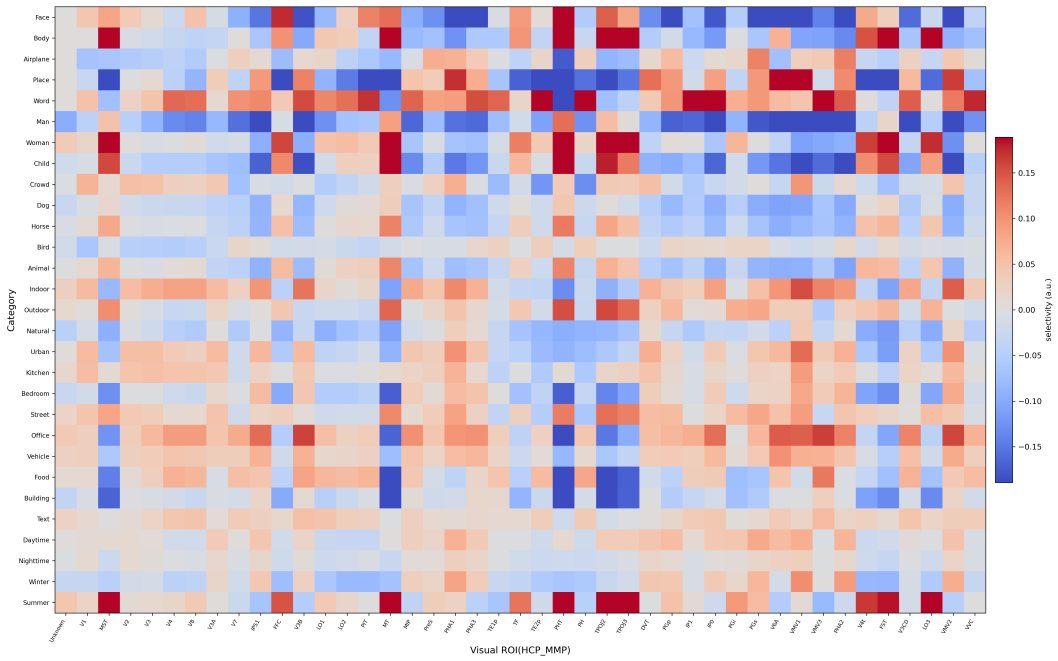

Figure 18: Extended ROI-level response profiles for a broader set of image-defined semantic concepts. Rows correspond to semantic concepts and columns to the 43 HCP-MMP visual ROIs. Values are one-vs-rest response scores computed from predicted BOLD responses and averaged within each ROI. All rows use the same color scale.

after recombination. Because this is a short-run diagnostic, its absolute values are not directly compared with the fully trained results in Table 3.

Visual–neural alignment The alignment loss in the spatial autoencoding branch provides an additional constraint between the visual features and the spatial neural representation. For each ROI, we average the target and reconstructed spatial factors V across its voxels and compute their Pearson correlation across conditions; the ROI correlations are then averaged within each hierarchy

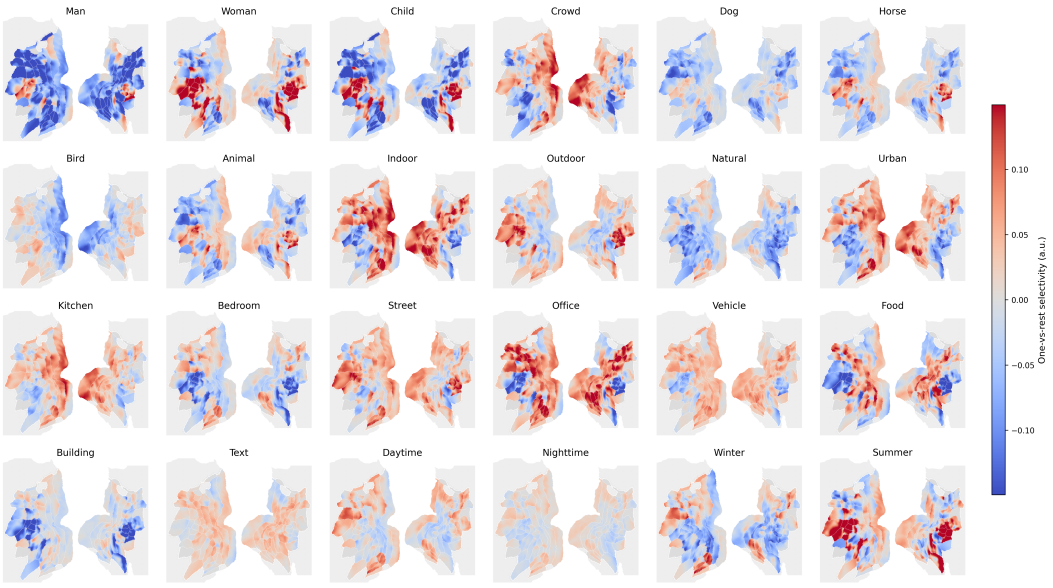

Figure 19: Additional one-vs-rest response maps for 24 image-defined semantic concepts. Each panel shows the response pattern derived from predicted BOLD responses on the HCP-MMP flat cortical layout. All maps use the same color scale.

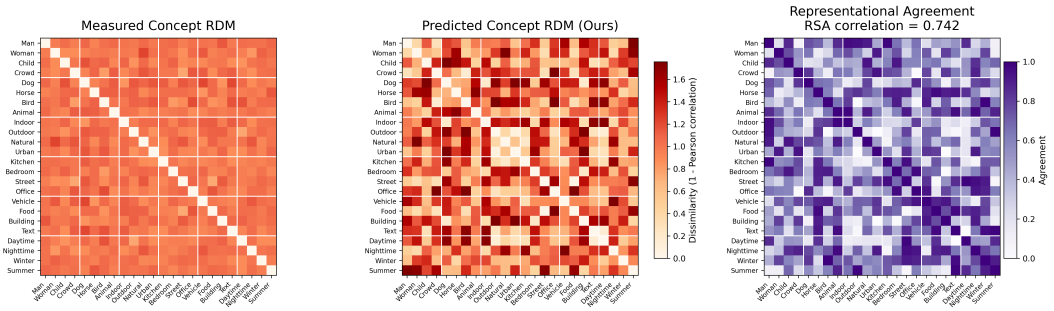

Figure 20: Representational similarity analysis over 24 image-defined semantic concepts. Measured and predicted RDMs are constructed from one-vs-rest concept-response patterns using one minus Pearson correlation. RSA is computed from the off-diagonal upper-triangular entries of the two RDMs and reaches $r = 0.742$.

group. Removing this term decreases these mean V reconstruction correlations in low-, mid-, and high-level ROIs from 0.9982/0.9857/0.9745 to 0.9853/0.9685/0.9547. The consistent but relatively small changes support using visual–neural alignment as an auxiliary representation constraint.

Temporal autoencoding pathway The temporal latent Z_U is used only to reconstruct the temporal factor U during autoencoding and is not a diffusion target. The temporal-pathway and recombination-loss ablations are separate paired diagnostics, each evaluated against its own reference run. In the following AE diagnostics, Oracle-U fixes the temporal factor to its target value and compares responses reconstructed using $\hat{V}$ with those reconstructed using the target V, isolating the effect of spatial reconstruction. Both responses use the same FIR-mapped target U, with spatial amplitudes given by the voxel means of the respective inverse-standardized spatial factors. Removing Z_U changes V reconstruction correlation from 0.5051 to 0.4958 and Oracle-U BOLD MSE/correlation from 0.0077/0.9810 to 0.0107/0.9762. These small differences indicate that Z_U mainly supports temporal reconstruction rather than providing a major independent contribution to the final prediction.

AE-stage recombination constraints We further remove both the coefficient-space consistency loss $\mathcal{L}_{\mathrm{coef}}$ and the response-consistency loss $\mathcal{L}_{\mathrm{resp}}^{\mathrm{AE}}$. Without these terms, Oracle-U BOLD MSE increases from 0.0095 to 0.0239 and correlation decreases from 0.9790 to 0.9225, while V reconstruction correlation remains nearly unchanged (0.4129 vs. 0.4133). These results indicate that the two losses mainly improve consistency after the reconstructed temporal and spatial factors are recombined, rather than substantially changing the individual spatial reconstruction.

A.3 LIMITATIONS

BOLDSyner has several limitations. First, the current temporal–spatial target uses a rank-one approximation. Although rank one provides a compact factorization, higher-rank components still explain additional FIR variance, with cumulative explained variance increasing from 49.15% at rank one to 71.24% and 82.57% at ranks two and three.

Second, the temporal factor is defined over a fixed six-bin post-stimulus window. The current analysis therefore focuses on short-range stimulus-evoked responses and does not cover longer or continuous BOLD dynamics.

Third, our experiments use four NSD subjects with high-quality 7T fMRI and repeated stimulus presentations. Generalization to lower-SNR recordings, fewer repetitions, and different acquisition settings remains to be evaluated.

Finally, the current model is trained separately for each subject, and our analyses focus on response prediction, temporal responses, category selectivity, and semantic relationships. Extending BOLDSyner to cross-subject modeling and broader neuroscience applications remains an important direction for future work.