FUTURE VIDEO GENERATION BETTER ALIGNS WITH THE HUMAN VISUAL CORTEX THAN OBSERVED VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Studying the alignment between the internal representations of vision models and the responses of the visual cortex to the same observed visual stimuli has enabled us to better understand human visual processing. However, studies so far have largely overlooked the fact that the human brain not only processes observed visual stimuli, but also predicts upcoming stimuli based on what has been observed. Accordingly, we hypothesize that internal representations for generating future video frames are better aligned with the predictive nature of human visual processing than representations of the observed video itself. To this end, we compare the alignment between human video-watching fMRI responses in the visual cortex and the internal representations from two types of video diffusion models, an autoregressive (AR) model and its non-AR base model. We first conduct a within-model analysis of the AR video diffusion model and show that the representations for future video generation align better with the visual cortex than the representations of the observed video. We then compare the internal representations of the AR model with those of its non-AR base model and again show that the representations for future video generation align better with the visual cortex than the representations for observed video reconstruction by the base model. Specifically, the alignment of observed video reconstruction is concentrated in lower-order visual cortex, whereas that of future video generation is concentrated in higher-order visual cortex. Finally, we show in a human behavioral experiment that humans prefer videos generated by amplifying the contributions of individual layers that align better with the visual cortex.

1 INTRODUCTION

Studying the representational alignment between vision models and the human brain has enabled us to better understand human visual processing (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Kriegeskorte & Douglas, 2018). For example, by examining which features or layers of the model best explain neural responses, we can infer how visual information may be processed in the human brain (Kriegeskorte et al., 2008; Naselaris et al., 2011; Schrimpf et al., 2018). While these brain–vision model representation alignment studies have provided insights into human visual processing, studies to date have largely focused on representations of the observed visual stimulus (Conwell et al., 2024; Sartzetaki et al., 2025; Höfling et al., 2026). The brain, however, does not only passively respond to the observed stimulus, but also predicts the stimulus that is about to arrive (Rao & Ballard, 1999; Keller & Mrsic-Flogel, 2018; Hogendoorn & Burkitt, 2018). Accordingly, we compare the internal representations of a model that generates future video frames with those of the human visual cortex.

Recently, video generative models based on the diffusion transformer (Peebles & Xie, 2022) have demonstrated the capability to produce photorealistic videos (Brooks et al., 2024; Team Wan et al., 2025). Conventionally, video diffusion models generate video sequences by progressively denoising latent representations across all frames simultaneously. However, such designs disregard the inherent temporal order of causality. To align with the causal order of time in which visual stimuli naturally unfold, recent approaches have integrated the autoregressive (AR) framework with diffusion models to generate the future video conditioned on the observed video (Yin et al., 2024; Huang

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.

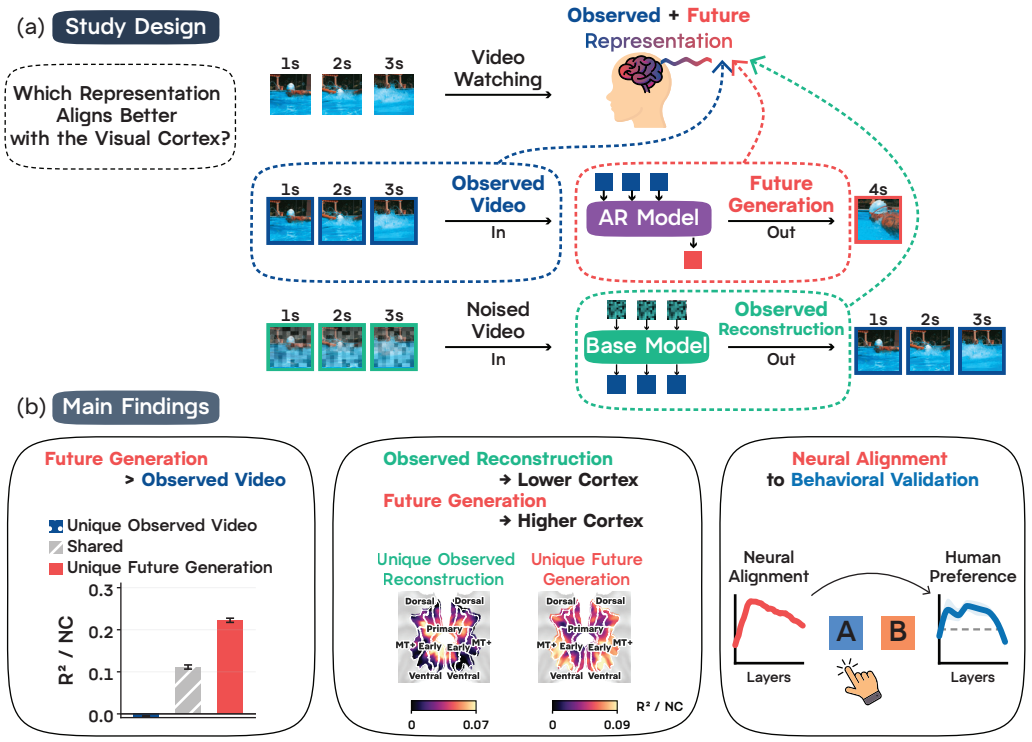

Figure 1: **Study Design and Main Findings.** (a) We map the representations of the observed video, future video generation, and observed video reconstruction onto the visual cortex. (b) Left, future video generation aligns better with the visual cortex than the observed video. Middle, observed video reconstruction aligns better with lower-order visual cortex, whereas future video generation aligns better with higher-order visual cortex. Right, humans prefer videos generated by amplifying the layers that align better with the visual cortex.

et al., 2025; Yuan et al., 2026a). The AR model represents how the visual stimulus unfolds over time to generate a future video that is semantically and physically plausible based on the observed video.

We hypothesize that the internal representations for generating future video frames are better aligned with the predictive nature of human visual processing than the representations of the observed video itself. To this end, we compare the internal representations from two types of video diffusion models, an AR model and its non-AR base model. Using human fMRI responses recorded while subjects watch 3 s naturalistic videos (Lahner et al., 2024), we map the representations of the observed video stimulus, future video generation, and observed video reconstruction (Figure 1a). Moreover, we test whether the alignment of the representation for future video generation is validated at the behavioral level with a human video preference experiment.

Contributions (Figure 1b).

- **We show that future video generation aligns better with the visual cortex than the observed video.** The representation for future video generation aligns better than the representation of the observed video that both the AR model and the human brain receive.
- **We show that the two representations, for future video generation and observed video reconstruction, align with different levels of the visual hierarchy.** The alignment of observed video reconstruction is concentrated in lower-order visual cortex, whereas that of future video generation is concentrated in higher-order visual cortex.
- **We validate the layer profile of neural alignment in human behavior.** Amplifying the contribution of an individual layer during future video generation, we show that humans prefer videos generated by amplifying the layers that align better with the visual cortex.

2 Related Work

2.1 Brain–Vision Model Representation Alignment

The internal representations of vision models turn out to partly resemble the representations of the human visual cortex (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Schrimpf et al., 2018). Presenting the same visual stimulus to a human and to a vision model, and comparing their internal representations, enables us to understand human visual processing. The internal representations of vision models optimized for object recognition align with the representation in the visual cortex (Yamins et al., 2014), and deeper layers align best with higher-order visual cortex (Güçlü & van Gerven, 2015). Large-scale recordings of human cortical responses to natural images and videos (Allen et al., 2022; Lahner et al., 2024) have enabled systematic comparisons across a large pool of vision models. These comparisons show that, for images, the visual training diet exerts a larger and more consistent effect on representation alignment than architecture or task objective (Conwell et al., 2024). For videos, temporal modeling drives representation alignment to lower-order visual cortex while the classification task drives alignment to higher-order visual cortex (Sartzetaki et al., 2025). The comparison has also been extended to the internal representations of an image generative model. The U-Net bottleneck of a latent image diffusion model aligns with the visual cortex best at the noisiest denoising step, and as denoising proceeds the first layer aligns better with lower-order visual cortex while the bottleneck aligns better with higher-order visual cortex (Takagi & Nishimoto, 2023).

2.2 Predictive Nature of Human Visual Processing

The human visual cortex represents not only the response to the observed stimulus but also the future stimulus. According to predictive coding theory, the brain uses its prior experience to build an internal generative model of the external world (Rao & Ballard, 1999; Lee, 2015; Keller & Mrsic-Flogel, 2018). With this internal model, the brain can infer the cause of the observed visual stimuli and predict future stimuli. Indeed, the representation of future visual stimuli has been confirmed in human neuroimaging studies with visual tasks. The human brain represents the next position of a moving object before its input arrives (Hogendoorn & Burkitt, 2018; Blom et al., 2020), and when presented with a partial cue, even lower-order visual cortex represents the full sequence of previously experienced visual stimuli (Ekman et al., 2017). Higher-order visual cortex is thought to represent visual stimuli further into the future than lower-order visual cortex while humans observe naturalistic video (de Vries & Wurm, 2023; Lee et al., 2021).

2.3 Internal Representation of Diffusion Video Models

A non-AR diffusion video model, which denoises all frames of the video together, learns an internal representation that carries rich visual information (Xiang et al., 2023; Li et al., 2023; Yu et al., 2025; Geyer et al., 2024; Nam et al., 2025). This representation is rich enough to be adapted to vision tasks on the observed video, such as action recognition (Guimarães et al., 2025), depth estimation (King et al., 2026), point tracking (Son et al., 2025; Nam et al., 2025; Yuan et al., 2026b), and physical plausibility estimation (Esmati et al., 2026). In contrast to the non-AR model, an AR diffusion video model sequentially generates the future video based on the observed video, allowing video generation to follow the natural temporal order. The AR model represents how the future visual stimulus unfolds over time to generate a future video that is semantically and physically plausible. Indeed, aligning the representation of an AR diffusion video model to the representation of the future video improves its future video generation quality (Yu et al., 2026).

3 Preliminaries

Human Video fMRI Dataset. We use the BOLD Moments Dataset (Lahner et al., 2024), in which ten subjects watched 1,102 3 s naturalistic videos while whole-brain fMRI was recorded, 1,000 of them for training and 102 for testing. Preprocessing follows the MOSAIC pipeline (Lahner et al., 2025) and leaves one response per video at each cortical vertex (Appendix A.4),

$$Y \in \mathbb{R}^{N \times P}, \quad m \in \{1, \ldots, 10\}, \quad p \in \{1, \ldots, P\}, \quad r \subseteq \{1, \ldots, P\}, \tag{1}$$

where Y is the response of one subject over N videos and P vertices, m indexes subjects, p indexes vertices, r is a region composed of a group of vertices, and $|r|$ is the number of vertices it contains.

Diffusion Video Generative Model. A diffusion video generative model generates a video by iteratively denoising noisy video tokens. A variational autoencoder $\mathcal{E}$ maps a video v to a latent $z_0 = \mathcal{E}(v) \in \mathbb{R}^{C\times T\times H\times W}$, and generation follows a flow-matching path between noise and data (Lipman et al., 2023). A latent on this path at noise level σ interpolates the clean latent z_0 with Gaussian noise,

$$z_\sigma = (1-\sigma)\, z_0 + \sigma\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

We order the schedule $\sigma_0 > \sigma_1 > \cdots > \sigma_{S-1}$, so that step $s = 0$ is the noisiest and step $s = S-1$ the cleanest. The transformer has B layers, numbered $b = 0, \ldots, B-1$. Let $H_{s,b}$ denote the hidden states that layer b produces at noise level σ_s, and let H_s^{in} denote the sequence that enters the first layer.

4 Methods and Experiment Setup

Human Brain Data. Every vertex is labeled with the HCP-MMP atlas (Glasser et al., 2016). Our primary focus is the Visual System, 7,894 vertices grouped into five divisions. The five divisions are Primary Visual, Early Visual, Dorsal Stream, Ventral Stream, and MT+ Complex, and a region r is either one division or the Visual System as a whole. Only vertices with a noise ceiling $\mathrm{NC} > 10\%$ enter the analysis.

Models. The main analysis uses helios-base-v2v (Yuan et al., 2026a) as the AR model, a 40-layer autoregressive video model fine-tuned from the non-AR base model Wan-2.1 T2V-14B (Team Wan et al., 2025). We replicate the analysis on two further AR models, Self-Forcing (Huang et al., 2025) and CausVid (Yin et al., 2024), both 30-layer autoregressive video models fine-tuned from the non-AR base model Wan-2.1 T2V-1.3B.

Video Preparation. Every BOLD Moments Dataset video is 3 s long and square. For helios-base-v2v and Wan-2.1 T2V-14B we resample each clip to 272×272 at 33 fps. For Self-Forcing and CausVid we fit it inside an 832×480 frame, keeping the aspect ratio and padding the sides with black, and sample it at 15 fps.

Autoregressive Model. We use an AR model built on a diffusion video generative model that generates future video frames based on observed video stimuli. The AR model generates the future video by denoising future video tokens. Let z^{obs} denote the clean latent of the observed video and $z^{\text{fut}}_{\sigma_s}$ the latent of the future video at noise level σ_s. The input sequence for every step carries both observed and future tokens at once,

$$H_s^{\text{in}} = \big[\ \underbrace{\text{patch}(z^{\text{obs}})}_{\text{observed video tokens, clean}}\ ;\ \underbrace{\text{patch}(z^{\text{fut}}_{\sigma_s})}_{\text{future video tokens at noise level } \sigma_s}\ \big], \tag{3}$$

with the observed tokens fed in clean at every step and the future tokens carried over from the previous step as they are denoised. Let $c^{\text{obs}} = \text{patch}(z^{\text{obs}})$ denote the observed video embedding, which is the same at every step. The layer activation of the AR model at step s and layer b splits into an observed part and a future part,

$$H^{\text{ar}}_{s,b} = \big[\, H^{\text{obs}}_{s,b}\ ;\ H^{\text{fut}}_{s,b}\,\big]. \tag{4}$$

We extract future tokens $H^{\text{fut}}_{s,b}$, so that each video yields a grid of representations indexed by the denoising step $s = 0, \ldots, S-1$ and the layer $b = 0, \ldots, B-1$.

Future Video Generation with AR Model. For each AR model we feed the 3 s observed video and let the model generate the 1 s future video that follows. Because subjects only observed the 3 s video with no instruction, we leave the text prompt empty. The AR model helios-base-v2v runs 50 denoising steps, and Self-Forcing and CausVid run 4 steps, the default for each model. For each model we extract c^{obs} once, and $H^{\text{fut}}_{s,b}$ at every step and layer.

Base Model. We use a non-AR base model built on a diffusion generative model that reconstructs noised video stimuli. The base model reconstructs the observed video by denoising observed video tokens. We noise z^{obs} by Equation 2 and run one forward pass at each of the noise levels σ_s, so that at every noise level the base model reconstructs the observed video from its noised form. Let $H^{\text{rec}}_{s,b}$ denote the layer activation of the base model at step s and layer b. We extract $H^{\text{rec}}_{s,b}$, which gives a grid over noise levels and layers of the same shape as $H^{\text{fut}}_{s,b}$.

Observed Video Reconstruction with Base Model. For each base model we feed the 3 s observed video noised at the same schedule as its AR model. At each noise level we run one forward pass to denoise the observed video with no text prompt. For each model we extract $H^{\text{rec}}_{s,b}$ at every step and layer.

Vertex-Wise Encoding Model. We map three representations onto the cortical response recorded while subjects observed the videos (Figure 1a). The first is the observed video embedding c^{obs} that the AR model receives (Equation 3). The second is the representation for future video generation $H^{\text{fut}}_{s,b}$ of the AR model (Equation 4). The third is the representation for observed video reconstruction $H^{\text{rec}}_{s,b}$ of the base model. Each of them is flattened and reduced to $d = 6004$ dimensions with a sparse random projection (Appendix A.5), which gives the feature matrices X^{obs}, $X^{\text{fut}}_{s,b}$, and $X^{\text{rec}}_{s,b}$, each in $\mathbb{R}^{N\times d}$ with one row per video. For each representation we fit a vertex-wise encoding model that predicts the response of one cortical vertex (Naselaris et al., 2011). Let X denote any one of these feature matrices. For vertex p we fit ridge regression weights,

$$w_p = \arg\min_{w\in\mathbb{R}^d} \left\|Y_{:,p} - Xw\right\|_2^2 + \lambda_p \|w\|_2^2, \tag{5}$$

with a ridge penalty λ_p chosen for that vertex, and take $\hat{Y}_{:,p} = Xw_p$ as the predicted response of that vertex. The penalty is chosen by nested cross-validation on the training videos, following the stacked-regression scheme of Lin et al. (2024), and for the two grid representations, $H^{\text{fut}}_{s,b}$ and $H^{\text{rec}}_{s,b}$, the best step and layer are selected on the training folds (Appendix A.5). The test videos take no part in fitting or in this selection, and every result we report is measured on test videos.

Encoding Accuracy. Encoding accuracy at a vertex is the coefficient of determination R^2_p of the test-set predictions. Every vertex is assigned a noise ceiling NC_p, the upper bound on how well any model can predict its response given the noise in the fMRI response (Appendix A.4). We report the noise-normalized encoding accuracy

$$\tilde{R}^2_p = R^2_p / \text{NC}_p. \tag{6}$$

The encoding accuracy of a region r is its mean over the vertices of the region, $\tilde{R}^2_r = \frac{1}{|r|}\sum_{p\in r} \tilde{R}^2_p$.

Variance Partitioning. Two representations can explain the same part of the cortical response. To separate what each explains on its own from what the two explain in common, we fit a joint encoding model that combines representations a and b and measure the unique contribution of each representation. Let $\tilde{R}^2_{\text{stack}}$ be the encoding accuracy of the joint model, and $\tilde{R}^2_a$ and $\tilde{R}^2_b$ the encoding accuracies of a and b alone (Lin et al., 2024),

$$U_a = \tilde{R}^2_{\text{stack}} - \tilde{R}^2_b, \qquad U_b = \tilde{R}^2_{\text{stack}} - \tilde{R}^2_a, \qquad U_{ab} = \tilde{R}^2_a + \tilde{R}^2_b - \tilde{R}^2_{\text{stack}}. \tag{7}$$

U_a and U_b are the unique contributions of the two representations and U_{ab} is their shared contribution. The joint model is explained in Appendix A.6. We partition two pairs. The within-model pair sets the observed video against future video generation, $(X^{\text{obs}}, X^{\text{fut}})$. The cross-model pair sets observed video reconstruction against future video generation, $(X^{\text{rec}}, X^{\text{fut}})$. Here X^{fut} and X^{rec} are the grid cells $X^{\text{fut}}_{s,b}$ and $X^{\text{rec}}_{s,b}$ at the best-aligned step and layer, selected on the training folds for each region and subject (Appendix A.5, Equation 12).

5 Results

5.1 Within-Model Analysis

The AR model receives the 3 s observed video and generates the 1 s future video. We map both the representation of the observed video X^{obs} and the representation for future video generation $X^{\text{fut}}_{s,b}$

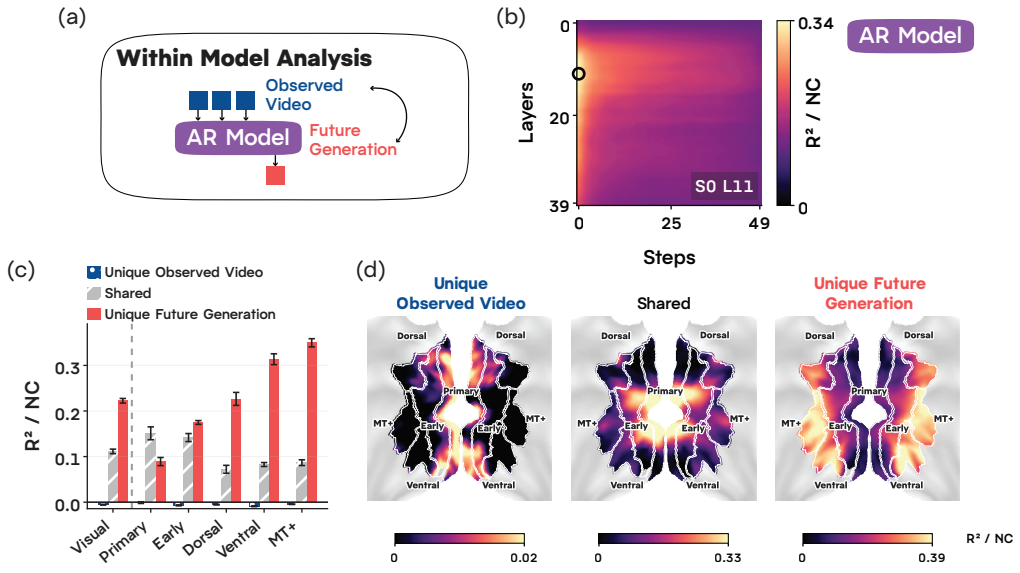

Figure 2: **Within-model analysis.** (a) The AR model receives the 3 s observed video and generates the 1 s future video. (b) Encoding accuracy of the representation for future video generation of the AR model over denoising steps and layers in the Visual System. (c) Unique and shared contributions of the representation for future video generation against the observed video in the Visual System and its five divisions. (d) The same contributions plotted on the cortical surface.

onto the visual cortex, and then use variance partitioning to measure the unique contribution of each representation. We report Helios here and repeat the same analysis on CausVid and Self-Forcing in Appendix A.1.

Future video generation aligns better with the visual cortex than the observed video. The representation for future video generation aligns best with the Visual System at step 0 and layer 11, the cell selected on the training folds (Figure 2b). In the Visual System the unique contribution of future video generation is 0.222, the shared contribution is 0.111, and the unique contribution of the observed video is -0.006 (Figure 2c). The unique contribution of future video generation is above the unique contribution of the observed video in every division ($p = 0.002$, Appendix A.9). Only in Primary Visual does the shared contribution, 0.151, exceed the unique contribution of future video generation, 0.089. In every other division the unique contribution of future video generation exceeds both the unique contribution of the observed video and the shared contribution. The shared contribution and the unique contribution of future video generation are 0.142 and 0.175 in Early Visual, 0.072 and 0.227 in Dorsal Stream, 0.083 and 0.314 in Ventral Stream, and 0.087 and 0.350 in MT+ Complex. On the cortical surface the unique contribution of future video generation grows continuously toward higher-order visual cortex, whereas the shared contribution and the unique contribution of the observed video are concentrated in lower-order visual cortex (Figure 2d).

5.2 Cross-Model Analysis

The base model reconstructs the 3 s observed video from its noised form. We map the representation for observed video reconstruction $X^{\text{rec}}_{s,b}$ onto the visual cortex, and then partition the variance with the representation for future video generation $X^{\text{fut}}_{s,b}$ (Figure 3a). We report Helios here and repeat the same analysis on CausVid and Self-Forcing in Appendix A.1.

Future video generation aligns better with the visual cortex than observed video reconstruction. The representation for observed video reconstruction peaks at step 16 and layer 8, the cell selected on the training folds (Figure 3b). Its best encoding accuracy, 0.301, is below the best encoding accuracy of future video generation, 0.334. In Primary Visual and Early Visual the encoding accuracies of future video generation and observed video reconstruction are similar, 0.240 and 0.240

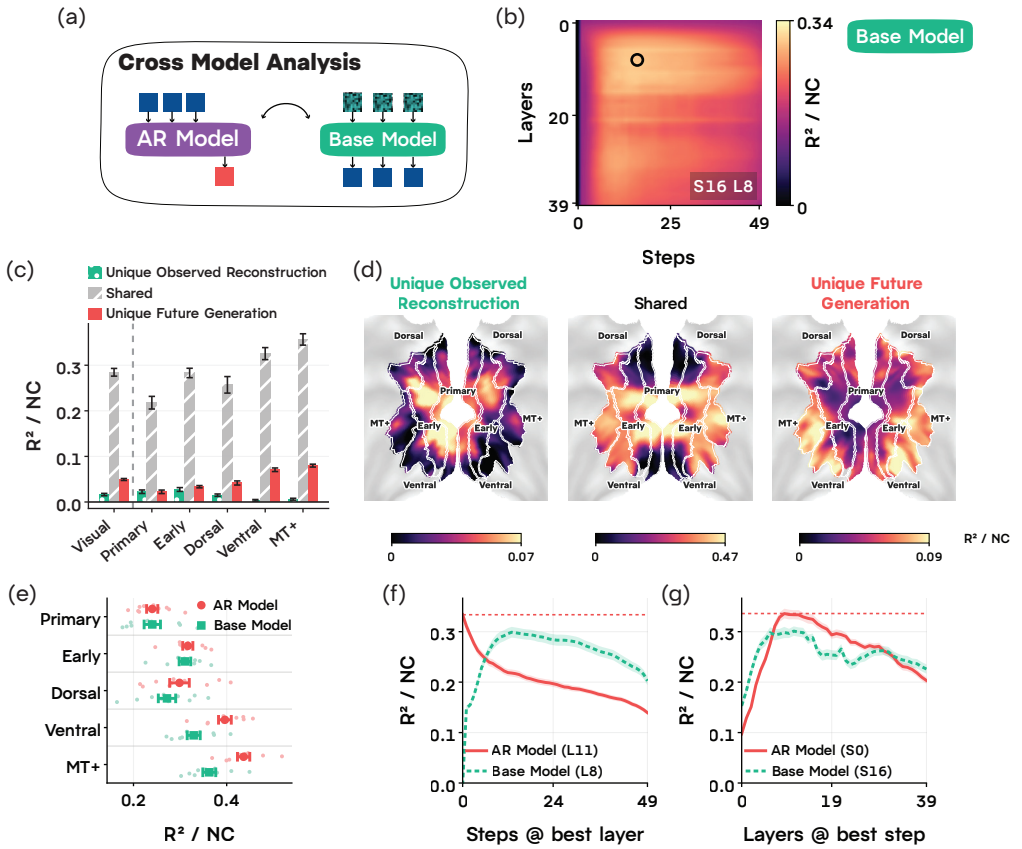

Figure 3: **Cross-model analysis.** (a) The AR model generates the 1 s future video and its base model reconstructs the 3 s noised observed video. (b) Encoding accuracy of the representation for observed video reconstruction of the base model over denoising steps and layers in the Visual System. (c) Unique and shared contributions of future video generation against observed video reconstruction in the Visual System and its five divisions. (d) The same contributions plotted on the cortical surface. (e) Encoding accuracy of the representations for future video generation and observed video reconstruction in the five divisions. (f, g) Encoding accuracy of the two representations along the step axis at each representation's own selected layer and along the layer axis at each representation's own selected step.

in Primary Visual and 0.317 and 0.310 in Early Visual (Figure 3e). However, future video generation aligns better in higher-order visual cortex. In Dorsal Stream future video generation is at 0.299 and observed video reconstruction at 0.272, in Ventral Stream 0.396 and 0.329, and in MT+ Complex 0.436 and 0.363 ($p = 0.002$ in all three).

Observed video reconstruction aligns with lower-order visual cortex, and future video generation aligns with higher-order visual cortex. In Primary Visual the two unique contributions are 0.023 for observed video reconstruction and 0.022 for future video generation, in Early Visual 0.027 and 0.034, in Dorsal Stream 0.015 and 0.042, in Ventral Stream 0.004 and 0.071, and in MT+ Complex 0.006 and 0.080 (Figure 3c). On the cortical surface the unique contribution of future video generation grows continuously toward higher-order visual cortex, whereas the unique contribution of observed video reconstruction is concentrated in lower-order visual cortex (Figure 3d).

Alignment of observed video reconstruction peaks at a less noisy step than future video generation. The alignment profiles of observed video reconstruction and future video generation differ along the step axis (Figure 3f). The alignment of future video generation is highest at the noisiest step and decreases monotonically as denoising proceeds, whereas the alignment of observed video

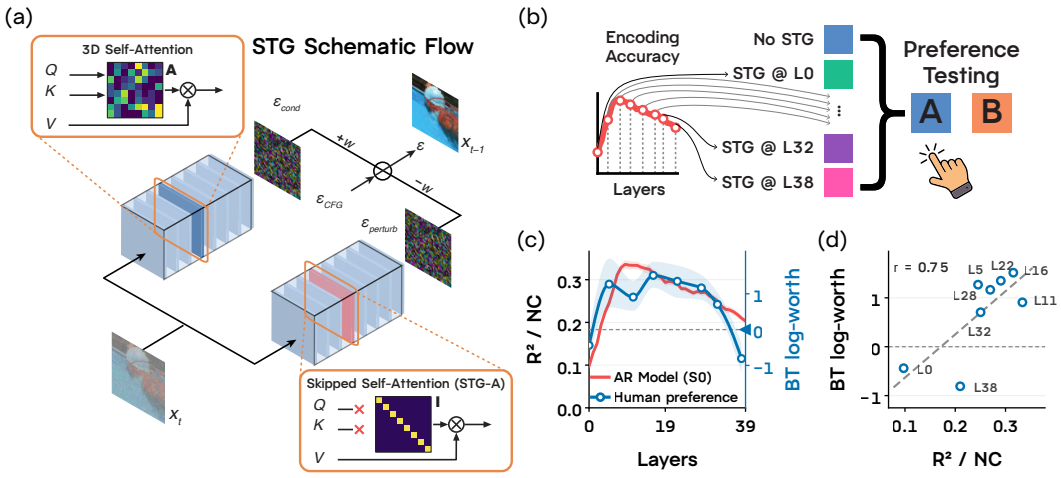

Figure 4: **Layer-wise guidance and human preference.** (a) STG skips the self-attention of one layer and pushes the sample away from the perturbed prediction (Equation 15). The schematic follows the layout of Ahn et al. (2024). (b) Pairwise preference test among the nine arms, the unguided arm (No STG) and eight STG-guided arms. (c) Encoding accuracy of the AR model at step 0 in the Visual System and human preference as Bradley–Terry log-worth along the layer axis. (d) Encoding accuracy against human preference, with the Pearson r.

reconstruction peaks early in the noise schedule and declines after it. Along the layer axis the alignment of both representations follows an inverted U with the maximum at an early-to-middle layer (Figure 3g).

5.3 Neural Alignment to Behavioral Validation

We finally ask whether the neural alignment of future video generation extends to human behavior. Spatiotemporal skip guidance (STG; Hyung et al., 2025) amplifies the contribution of a layer during video generation (Figure 4a, Appendix A.7). We test whether the neural alignment of future video generation is validated by human preference for videos generated under layer-wise amplification with STG. We apply STG to one layer $b \in \{0, 5, 11, 16, 22, 28, 32, 38\}$ at a time while the AR model generates a 5 s future video conditioned on the observed 3 s video. We then run A/B tests on the combined 8 s videos with five human observers, 720 preference trials in total (Figure 4b, Appendix A.8).

Humans prefer videos generated by amplifying well-aligned layers. The neural alignment of future video generation and human preference along the layer axis both follow an inverted U (Figure 4c). Among the eight guided layers, encoding accuracy at step 0 is highest at layer 11 and preference at layer 16. Over the eight layers the two measures correlate at $r = 0.75$ ($p = 0.029$, Figure 4d).

6 Discussion

Future video generation aligns better than the observed video. The representation for future video generation aligns better with the visual cortex than the observed video itself. Moreover, the observed tokens themselves add no unique contribution beyond future video generation in any division (Appendix A.2), which further confirms that future video generation is the key to brain alignment. This may be indirect evidence that the visual cortex performs future-oriented computation, especially in higher-order visual cortex, where human experiments have also shown predictive signals (Lee et al., 2021; de Vries & Wurm, 2023; Tarder-Stoll et al., 2024).

Why the noisiest step and an early-to-middle layer. The representation for future video generation aligns best at the noisiest step and an early-to-middle layer. As denoising goes on, the future

video settles into one of the many futures that could follow. Interestingly, this uncommitted state aligned best with the visual cortex. This finding is in line with Bayesian theories of neural coding, where the activity of a neural population encodes the whole range of possible stimuli rather than a single-valued prediction (Ma et al., 2006; van Bergen & Jehee, 2019). On the other hand, the representation for observed video reconstruction aligns best at a less noisy step. Moreover, the unique contribution of observed video reconstruction is concentrated in lower-order visual cortex, which processes simple visual features such as edges, contours, and textures (Rolls, 2024). As such fine details are refined in the later stages of the denoising schedule (Wang & Vastola, 2023), these results suggest that, compared with future video generation, observed video reconstruction focuses more on refining such fine details than on grasping the global semantics of the video. Along the layer axis, prior work reports that the visual representation is rich in the intermediate layers of diffusion transformers (Nam et al., 2025; Yuan et al., 2026b; Yu et al., 2025) and poor in the first and last layers (Yuan et al., 2026b), which agrees with our finding.

Comparison with a wide range of vision models. Previous studies compared the internal representations of vision models of various architectures, training datasets, and objectives to understand human visual processing (Conwell et al., 2024; Sartzetaki et al., 2025). We also compared the internal representations of a wide range of vision models with the representation for future video generation (Appendix A.3). The internal representations of self-supervised video models aligned best with the visual cortex, as reported before (Höfling et al., 2026; Tang et al., 2025). However, the gap in brain alignment between the representation for future video generation and the representations of self-supervised video models narrowed toward higher-order visual cortex. This may be further indirect evidence that higher-order visual cortex implements a mechanism that predicts future visual stimuli.

From neural alignment to human preference. STG guides the video generation path away from a perturbed weak model in which the self-attention of one layer is skipped (Hyung et al., 2025). Applying STG to a layer amplifies the representation of that layer. In our experiment, humans preferred the future videos generated by amplifying the layers that aligned better with the visual cortex. A layer that aligns better with the visual cortex may hold a richer visual representation, and amplifying it may make the generated video more natural to a human. This is in line with previous studies showing that, for diffusion-based generative models, aligning the internal representation to a richer visual representation improves the quality of the generated visual contents (Yu et al., 2025; Zhang et al., 2025).

Limitations. First, our analysis framework, which feeds a 3 s observed video and generates the 1 s future video that follows, limits understanding of how the brain represents the past and the future over a longer context and a longer horizon (Hasson et al., 2008; Lee et al., 2021; Brunec & Momennejad, 2022; Tarder-Stoll et al., 2024). Second, the cortical response evolves over time while the subject watches the video (de Vries & Wurm, 2023), but we map one response per video onto the visual cortex and do not compare the time course of the response with the internal representation of the AR model. Finally, our behavioral experiment shows that the layer profile of neural alignment is reflected in human preference, but it does not provide a causal account of how neural alignment gives rise to preference.

7 Conclusion

In this study, we examined the representational alignment of an AR and a non-AR video generative model with the visual cortex. We tested which of the three representations, the observed video, future video generation, and observed video reconstruction, aligns best with the visual cortex. With vertex-wise encoding and variance partitioning, we showed that future video generation aligns better than both the observed video and its reconstruction. More specifically, the alignment advantage of the representation for future video generation grows along the visual hierarchy. In contrast, the unique contributions of the observed video and its reconstruction are concentrated in lower-order visual cortex. Finally, by amplifying one layer during future video generation, we showed that humans preferred the videos generated from the layers that align well with the visual cortex.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the research methodology and experiments, to implement methods, to assist with translation, and to interpret results. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide ingredients for or write mathematical proofs, to propose or refine hypotheses, to clean or reformat datasets, or to support qualitative or thematic data analysis. Additionally, we used generative AI tools to create and edit software code, to draft parts of the paper, to summarize existing literature, to identify relevant literature and search for information, for brainstorming, to edit the paper to improve readability, and to format references. We have reviewed all AI-assisted work. LLM-generated code was verified and tested by the authors, and all AI-assisted text, translations, and references were checked and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The fMRI data are from the publicly released BOLD Moments Dataset and are used in their anonymized form. The human preference experiment was a video-watching task with no risk to the observers, who were adult volunteers and gave informed consent, and no identifying information about them was collected or reported.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jung, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.48550/arXiv.2403.17377.

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 Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Tessel Blom, Daniel Feuerriegel, Philippa Johnson, Stefan Bode, and Hinze Hogendoorn. Predictions drive neural representations of visual events ahead of incoming sensory information. *Proceedings of the National Academy of Sciences*, 117(13):7510–7515, 2020. doi: 10.1073/pnas.1917777117.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Iva K. Brunec and Ida Momennejad. Predictive representations in hippocampal and prefrontal hierarchies. *Journal of Neuroscience*, 42(2):299–312, 2022. doi: 10.1523/JNEUROSCI.1327-21.2021.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y.

Ingmar E. J. de Vries and Moritz F. Wurm. Predictive neural representations of naturalistic dynamic input. *Nature Communications*, 14:3858, 2023. doi: 10.1038/s41467-023-39355-y.

Matthias Ekman, Peter Kok, and Floris P. de Lange. Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, 8:15276, 2017. doi: 10.1038/ncomms15276.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. arXiv preprint arXiv:2606.05328, 2026.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16 (1):111–116, 2019. doi: 10.1038/s41592-018-0235-4.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2307.10373.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

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. doi: 10.1523/JNEUROSCI.5023-14.2015.

Rogério Guimarães, Frank Xiao, Pietro Perona, and Markus Marks. Diffusion-based action recognition generalizes to untrained domains. *arXiv preprint arXiv:2509.08908*, 2025.

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. doi: 10.1523/JNEUROSCI.5487-07.2008.

Larissa Höfling, Matthias Tangemann, Lotta Piefke, Susanne Keller, Katrin Franke, and Matthias Bethge. Only brains align with brains: Cross-region alignment patterns expose limits of normative models. In *International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2604.21780.

Hinze Hogendoorn and Anthony N. Burkitt. Predictive coding of visual object position ahead of moving objects revealed by time-resolved EEG decoding. *NeuroImage*, 171:55–61, 2018.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009, 2025.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.48550/arXiv.2411.18664.

Georg B. Keller and Thomas D. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018. doi: 10.1016/j.neuron.2018.10.003.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014. doi: 10.1371/journal.pcbi.1003915.

Michael King, Aravindh Mahendran, Matthew Koichi Grimes, Fedor Kitashov, Adham Elarabawy, Pedro Velez, Maks Ovsjanikov, and Viorica Pătrăucean. Gen4u: Unifying video generation and understanding via diffusion. *arXiv preprint arXiv:2607.06856*, 2026.

Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21:1148–1160, 2018. doi: 10.1038/s41593-018-0210-5.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15(1):6241, 2024. doi: 10.1038/s41467-024-50310-3.

Benjamin Lahner, Mayukh Deb, N. Apurva Ratan Murty, and Aude Oliva. MOSAIC: A scalable framework for fMRI dataset aggregation and modeling of human vision. bioRxiv preprint, 2025.

Caroline S. Lee, Mariam Aly, and Christopher Baldassano. Anticipation of temporally structured events in the brain. *eLife*, 10:e64972, 2021. doi: 10.7554/eLife.64972.

Tai Sing Lee. The visual system's internal model of the world. *Proceedings of the IEEE*, 103(8): 1359–1378, 2015. doi: 10.1109/JPROC.2015.2434601.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 287–296, 2006. doi: 10.1145/1150402.1150436.

Ruogu Lin, Thomas Naselaris, Kendrick Kay, and Leila Wehbe. Stacked regressions and structured variance partitioning for interpretable brain maps. *NeuroImage*, 298:120772, 2024. doi: 10.1016/j.neuroimage.2024.120772.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2210.02747.

Wei Ji Ma, Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.48550/arXiv.2506.17220.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11:e77599, 2022. doi: 10.7554/eLife.77599.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

Edmund T. Rolls. Two what, two where, visual cortical streams in humans. *Neuroscience & Biobehavioral Reviews*, 160:105650, 2024. doi: 10.1016/j.neubiorev.2024.105650.

Christina Sartzetaki, Gemma Roig, Cees G. M. Snoek, and Iris I. A. Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.1101/2024.12.05.626975.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which artificial neural network for object recognition is most brain-like? bioRxiv preprint, 2018.

Soowon Son, Honggyu An, Chaehyun Kim, Hyunah Ko, Jisu Nam, Dahyun Chung, Siyoon Jin, Jung Yi, Jaewon Min, Junhwa Hur, and Seungryong Kim. Repurposing video diffusion transformers for robust point tracking. *arXiv preprint arXiv:2512.20606*, 2025.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Many-two-one: Diverse representations across visual pathways emerge from a single objective. *bioRxiv*, 2025. doi: 10.1101/2025.07.22.664908.

Hannah Tarder-Stoll, Christopher Baldassano, and Mariam Aly. The brain hierarchically represents the past and future during multistep anticipation. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-53293-3.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Shiwei Zhang, Siyang Sun, Wei Lin, Wenyuan Yu, Yingya Zhang, Yun Zheng, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

Ruben S. van Bergen and Janneke F. M. Jehee. Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *Journal of Neuroscience*, 39(41):8164–8176, 2019.

Matteo Visconti, Fatma Deniz, Tom Dupré la Tour, and Jack L. Gallant. Encoding models in functional magnetic resonance imaging: The voxelwise encoding model framework. PsyArXiv preprint (v4), 2026. URL https://osf.io/preprints/psyarxiv/nt2jq_v4.

Binxu Wang and John J. Vastola. Diffusion models generate images like painters: An analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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. doi: 10.1073/pnas.1403112111.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. arXiv preprint arXiv:2412.07772, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Yonghao Yu, Lang Huang, Runyi Li, Zerun Wang, and Toshihiko Yamasaki. Video-mirai: Autoregressive video diffusion models need foresight. *arXiv preprint arXiv:2606.03971*, 2026.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. arXiv preprint arXiv:2603.04379, 2026a.

Tianyu Yuan, Yuanbo Yang, Lin-Zhuo Chen, Yao Yao, and Zhuzhong Qian. Denoise to track: Harnessing video diffusion priors for robust correspondence. *arXiv preprint arXiv:2512.04619*, 2026b. doi: 10.48550/arXiv.2512.04619.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. doi: 10.48550/arXiv.2505.23656.

A Appendix

A.1 Replication on Other Autoregressive Video Models

We repeat the within-model and cross-model analyses on two other AR video models, CausVid (Yin et al., 2024) and Self-Forcing (Huang et al., 2025). Both are fine-tuned from the non-AR base model Wan-2.1 T2V-1.3B and generate the future video in four denoising steps. We extract the representations for future video generation and for observed video reconstruction from all four steps and all 30 layers, and run the same encoding and variance partitioning as in the main analysis (Figures 5–8).

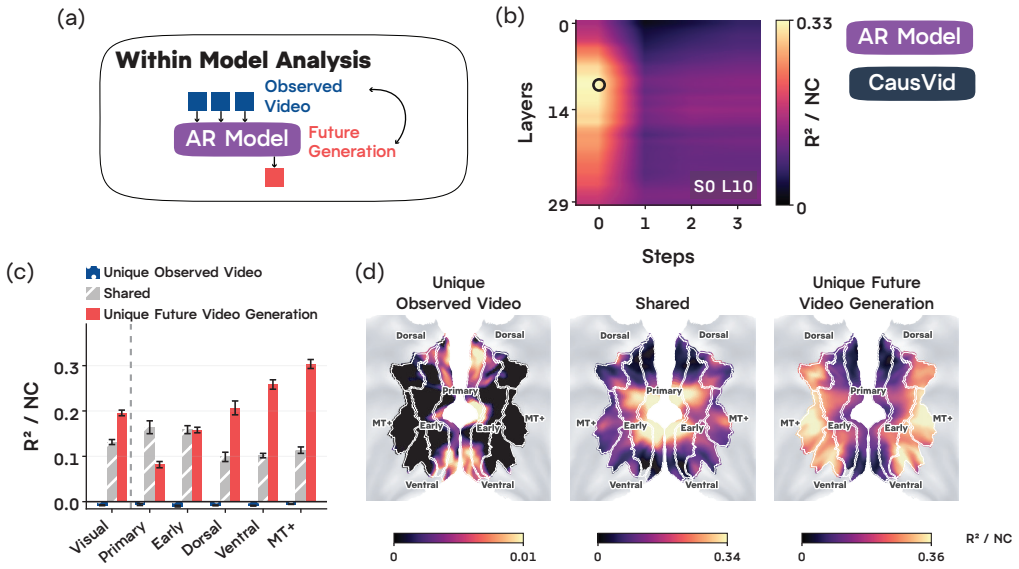

Figure 5: **Within-model analysis of CausVid.**

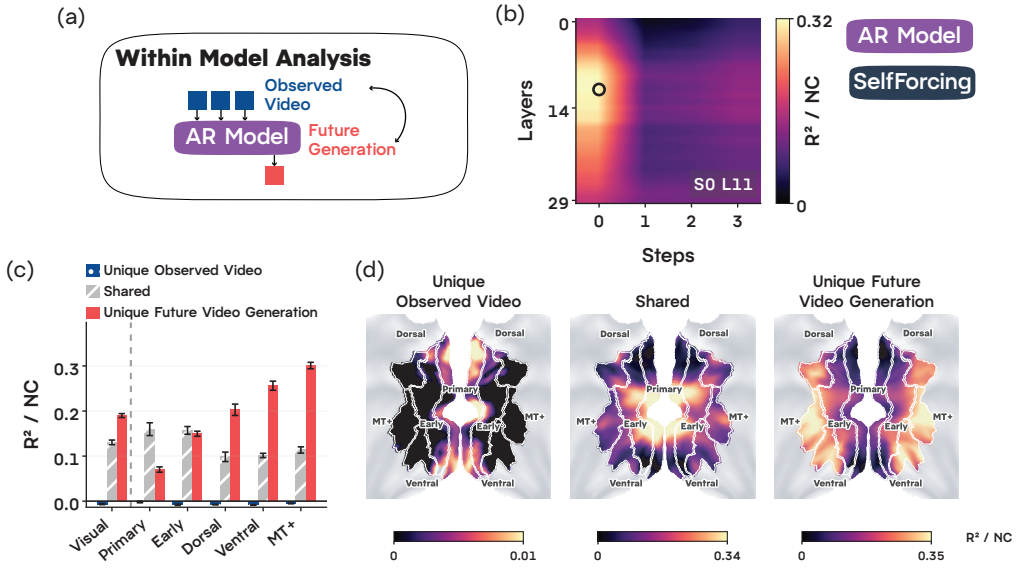

Figure 6: **Within-model analysis of Self-Forcing.**

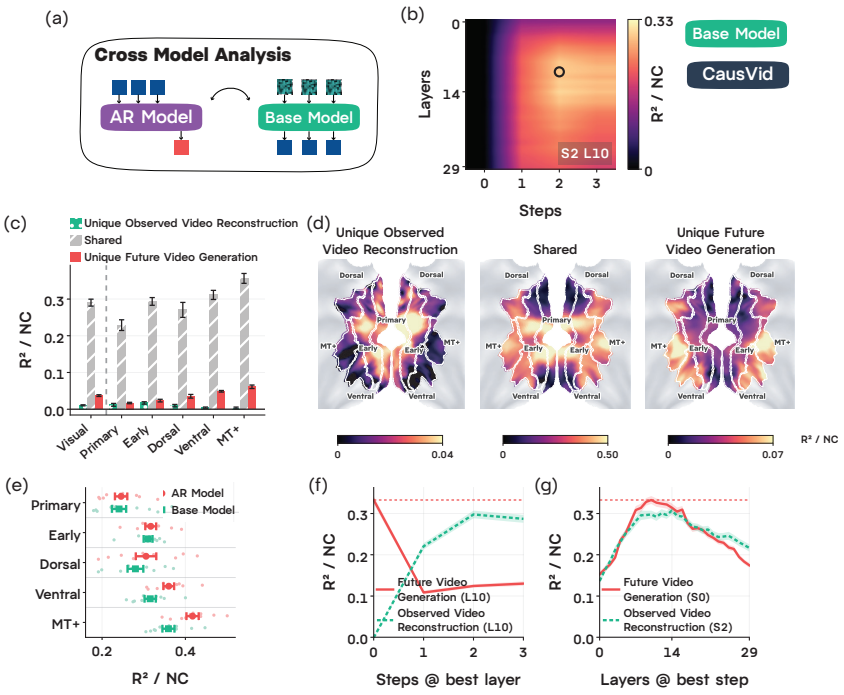

Figure 7: **Cross-model analysis of CausVid.**

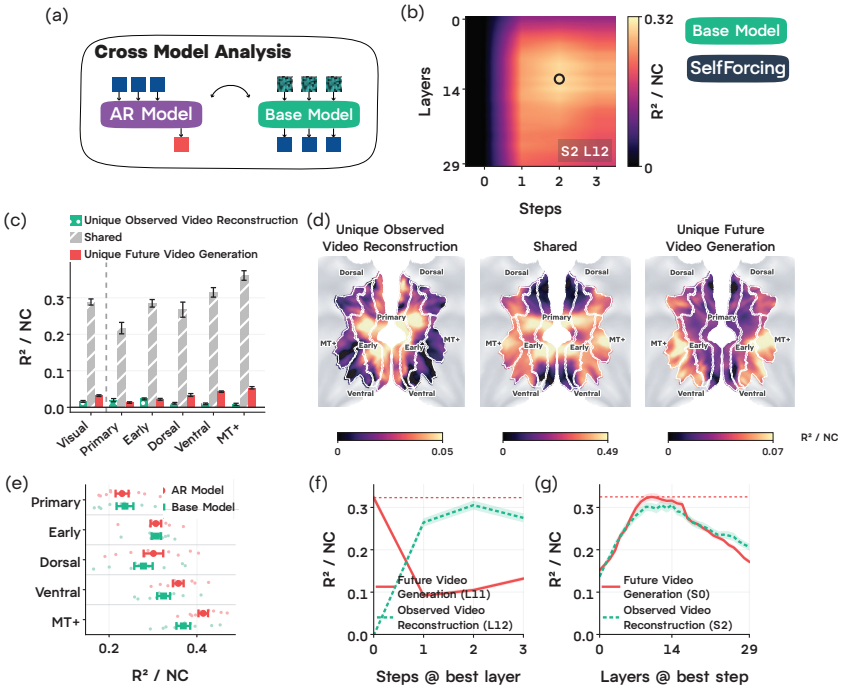

Figure 8: **Cross-model analysis of Self-Forcing.**

A.2 Ablation Study

Token Readout. The activation of the AR model while it generates the future video has an observed part and a future part, $H^{\text{ar}}_{s,b} = [\, H^{\text{obs}}_{s,b} \;;\; H^{\text{fut}}_{s,b} \,]$ (Equation 4). In the main analysis we read out

the future part $H^{\text{fut}}_{s,b}$. For the ablation we also read out the observed part $H^{\text{obs}}_{s,b}$ and the concatenation $[\, H^{\text{obs}}_{s,b} \;;\; H^{\text{fut}}_{s,b} \,]$, and apply the same sparse random projection and vertex-wise ridge regression as in the main analysis.

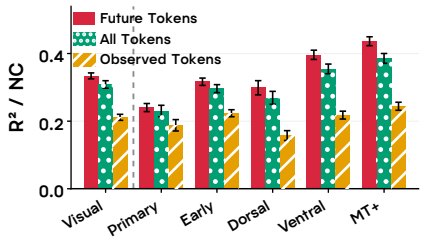

Figure 9: **Encoding accuracy of future video generation under different token readouts.**

Future Tokens Align Best with the Visual Cortex. Reading out only the future tokens aligns best with the visual cortex (Figure 9). In the Visual System the encoding accuracy of the future tokens reaches $\tilde{R}^2 = 0.334$ against 0.211 for the observed tokens, with the concatenated tokens between them at 0.309.

Variance Partitioning against the Observed Tokens. The observed tokens $H^{\text{obs}}_{s,b}$ also form a grid over denoising steps and layers, so we partition the variance between $H^{\text{obs}}_{s,b}$ and $H^{\text{fut}}_{s,b}$. The cell of each representation is selected on the training folds for each region and subject.

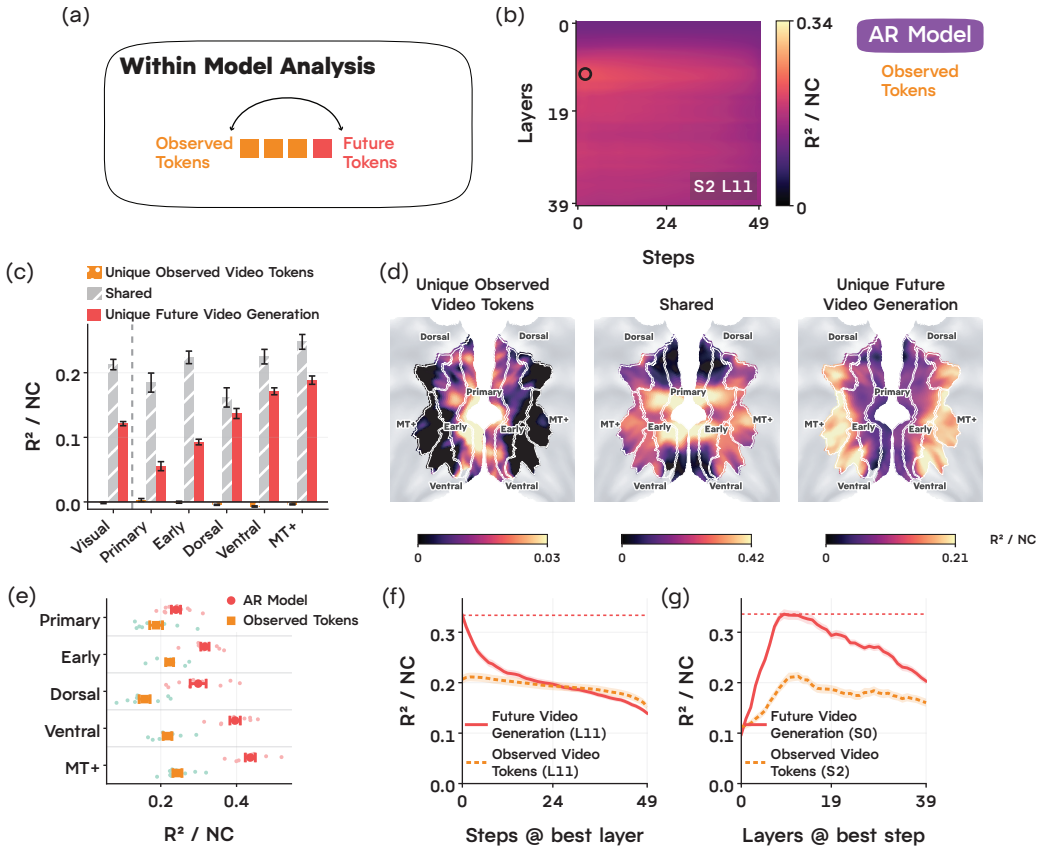

Figure 10: **Within-model analysis against the observed tokens.**

Future Tokens Keep Their Unique Contribution against the Observed Tokens. The observed tokens align best with the Visual System at step 2 and layer 11 (Figure 10b). Again, future video generation aligns better than the observed tokens. In the Visual System the shared contribution is 0.213, the unique contribution of future video generation is 0.121, and the unique contribution of the observed tokens is -0.001 (Figure 10c). The unique contribution of future video generation is above the unique contribution of the observed tokens in every division ($p = 0.002$), and it grows along the hierarchy, from 0.055 in Primary Visual and 0.093 in Early Visual to 0.137 in Dorsal Stream, 0.171 in Ventral Stream, and 0.189 in MT+ Complex. On the cortical surface the unique contribution of future video generation is again concentrated in higher-order visual cortex (Figure 10d).

A.3 Comparison with a Wide Range of Vision Models

We compared the internal representations of a wide range of vision models with the representation for future video generation of the AR model. We included both video and image vision models, 80 models in total.

Vision Model Activation Extraction. A video vision model that takes n frames spaced δ frames apart sees $n\delta$ of the T frames of a video in one pass, so we run $K = \lceil T/(n\delta) \rceil$ passes $v^{(1)}, \ldots, v^{(K)}$ to cover the whole video. An image vision model has no temporal window, so a pass is a single frame and $K = T$. Let h_ℓ denote the activation at layer ℓ of a vision model, where $\ell = 1, \ldots, L$ indexes the composite stages of the backbone. The representation of the vision model X_ℓ^{vis} holds one row per video, the layer activation flattened over tokens and channels and averaged over passes,

$$\left[X_\ell^{\text{vis}}\right]_v \;=\; \frac{1}{K}\sum_{k=1}^{K} \operatorname{vec} h_\ell\big(v^{(k)}\big). \tag{8}$$

Vision models have no denoising axis, so their grid reduces to the layer axis alone. X_ℓ^{vis} goes through the same sparse random projection and the same nested cross-validation (Appendix A.5). Figure 11 shows the encoding accuracy of every model in the Visual System, and Figure 12 the best model of each group in each division.

A.4 fMRI Dataset and Preprocessing

Stimuli and Acquisition. The BOLD Moments Dataset records whole-brain fMRI responses of ten subjects to 1,102 naturalistic videos of 3 s (Lahner et al., 2024). Each video was shown at a visual angle of 5° while the subject held central fixation, and each trial was the 3 s video followed by a 1 s inter-trial interval. The 1,000 training videos were presented three times each and the 102 test videos ten times each. Data were acquired on a 3 T Siemens Trio scanner with a 32-channel head coil. Functional images were T2*-weighted gradient-echo echo-planar images at 2.5 mm isotropic resolution with a repetition time of 1.75 s, an echo time of 30 ms, a flip angle of 71°, 54 slices, and a multi-band acceleration factor of 2. A T1-weighted anatomical image at 1 mm isotropic resolution was acquired as the anatomical reference, and dual-echo field maps were acquired at the beginning of every session to correct the spatial distortion of the functional images.

Preprocessing. We use the release of the BOLD Moments Dataset under the MOSAIC framework, which aggregates vision fMRI datasets under a shared preprocessing pipeline (Lahner et al., 2025). fMRIPrep preprocesses each subject and registers the data to the fsLR32k cortical surface (Esteban et al., 2019), and GLMsingle estimates one beta per presentation from the preprocessed time series (Prince et al., 2022). The betas of each session are normalized by the mean and standard deviation of the training videos presented in that session. Averaging the normalized betas over the repetitions of a video leaves one response per video at each vertex.

Noise Ceiling. The noise ceiling is the upper bound on how well any model can predict the response of a vertex. In the MOSAIC pipeline the noise ceiling is calculated from the ten presentations of each test video. The variance across the ten presentations of the same video is the noise variance σ^2_{noise}, and the remaining variance of the response is the signal variance σ^2_{signal}. The response we model is the average over $n = 10$ presentations, which reduces the noise variance to $\sigma^2_{\text{noise}}/n$. The

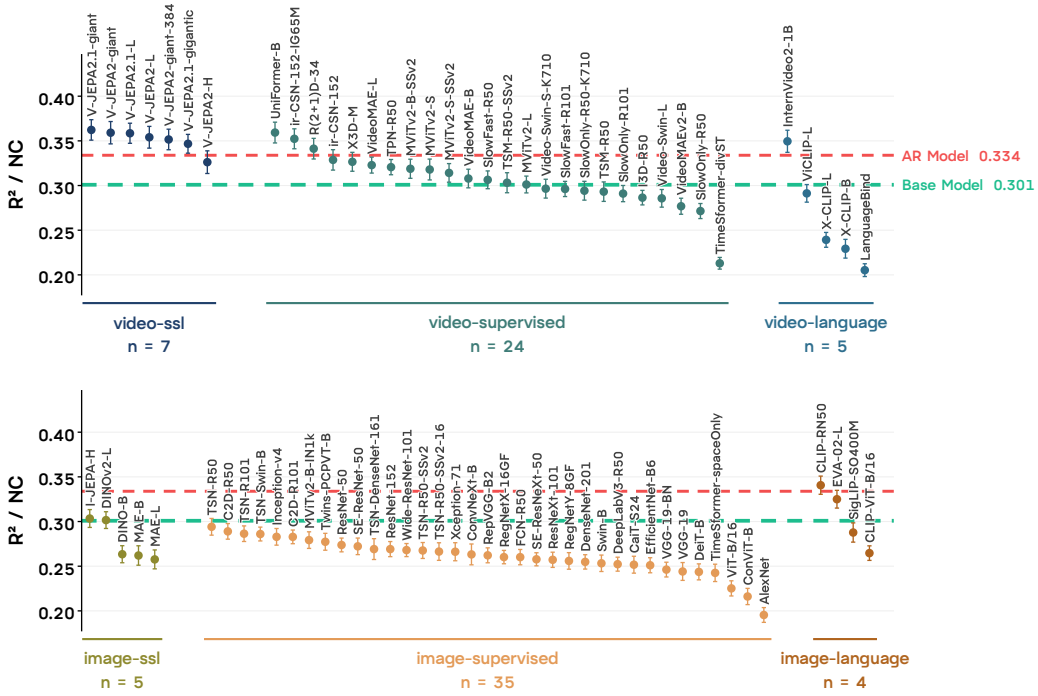

Figure 11: **Encoding accuracy of vision models in the Visual System.**

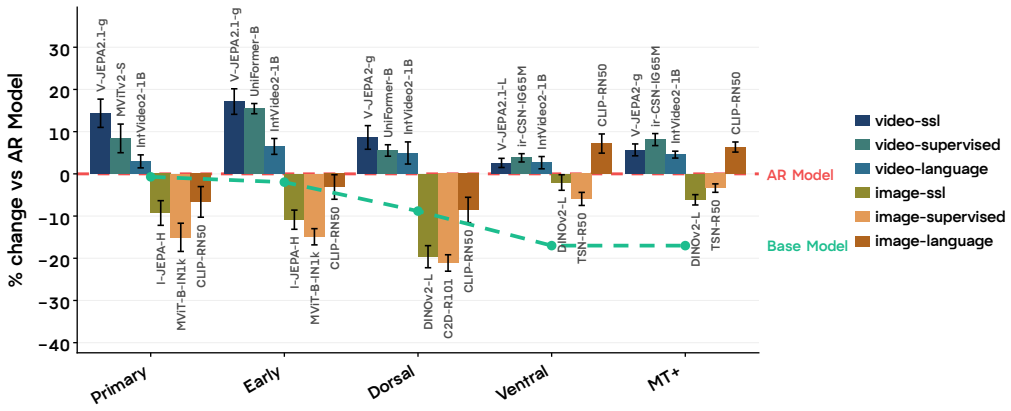

Figure 12: **Best vision model of each group in each division.**

noise ceiling is the fraction of the variance of the averaged response that is signal,

$$\mathrm{NC} = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{signal}} + \sigma^2_{\text{noise}}/n}. \tag{9}$$

Atlas. The HCP-MMP atlas parcellates each hemisphere into 180 areas grouped into 22 divisions by function and anatomy, and those 22 divisions into six major systems (Glasser et al., 2016). One of the six is the Visual System, which comprises the five divisions. Primary Visual is V1, 1,618 vertices. Early Visual is V2, V3, and V4, 2,565 vertices. The Dorsal Stream is V6, V3A, V7, IPS1, V3B, and V6A, 1,116 vertices. The Ventral Stream is V8, FFC, PIT, VMV1, VMV2, VMV3, and

VVC, 1,250 vertices. The MT+ Complex is MST, LO1, LO2, MT, PH, V4t, FST, V3CD, and LO3, 1,345 vertices.

A.5 Feature Construction and Encoding Model

Sparse Random Projection. Every representation is flattened over tokens and channels and reduced to $d = 6004$ dimensions with a sparse random projection Φ (Li et al., 2006),

$$X^{\text{obs}} = \Phi\big(c^{\text{obs}}\big), \qquad X^{\text{fut}}_{s,b} = \Phi\big(H^{\text{fut}}_{s,b}\big), \qquad X^{\text{rec}}_{s,b} = \Phi\big(H^{\text{rec}}_{s,b}\big), \tag{10}$$

each lying in $\mathbb{R}^{N\times d}$. The number of projections follows the Johnson–Lindenstrauss lemma, and $\epsilon = 0.1$ for the $N = 1{,}102$ videos gives $d = 6004$.

Ridge Regression. We follow the ridge regression and parameter selection of Lin et al. (2024). The encoding model of Equation 5 is solved in closed form,

$$\hat{B} = (X^\top X + \lambda I)^{-1} X^\top Y, \qquad \hat{Y} = X\hat{B}. \tag{11}$$

The penalty λ is chosen for each vertex from 16 values spaced by factors of ten between 10^{-6} and 10^{9}. The outer loop divides the 1,000 training videos into five folds and predicts each fold from a model fit on the remaining four, which yields an out-of-fold prediction for every training video. The inner loop divides those four folds again into five and chooses λ at each vertex by the inner cross-validation error. The model is then refit on all training videos with the chosen λ.

Cell Selection. Let $g = (s, b)$ denote a grid cell and let X_g be its feature matrix. Let $R^{2,\text{oof}}_{m,p}(X)$ denote the out-of-fold R^2 at vertex p in subject m. For a region r in subject m we select the cell $g^\star_{r,m}$ that maximizes that subject's own out-of-fold score, averaged over the vertices of the region,

$$g^\star_{r,m} = \arg\max_{g} \frac{1}{|r|} \sum_{p\in r} R^{2,\text{oof}}_{m,p}\big(X_g\big), \tag{12}$$

and use it for every vertex of r in that subject. Both the ridge penalty and $g^\star_{r,m}$ come from the training folds, so the test videos take no part in model selection. The encoding accuracy reported at a cell is the mean over subjects of each subject's test score at the cell selected on their own training set.

A.6 Variance Partitioning

We use variance partitioning to separate what each feature space explains uniquely from what they explain in common (Visconti et al., 2026; Lin et al., 2024). For a pair a and b their predictions are combined at each vertex into a joint model, with convex weights fit to the training responses,

$$\hat{Y}^{\text{stack}} = w_a\hat{Y}^a + w_b\hat{Y}^b, \qquad (w_a, w_b) = \arg\min_{w_a+w_b=1,\ w\geq 0} \big\|Y_{\text{train}} - (w_a\hat{Y}^a_{\text{oof}} + w_b\hat{Y}^b_{\text{oof}})\big\|^2, \tag{13}$$

which has a closed-form solution at each vertex. Let $\tilde{R}^2_a$, $\tilde{R}^2_b$, and $\tilde{R}^2_{\text{stack}}$ denote the normalized test scores of the two feature spaces and of their combination. The unique and shared contributions are

$$U_a = \tilde{R}^2_{\text{stack}} - \tilde{R}^2_b, \qquad U_b = \tilde{R}^2_{\text{stack}} - \tilde{R}^2_a, \qquad U_{ab} = \tilde{R}^2_a + \tilde{R}^2_b - \tilde{R}^2_{\text{stack}}. \tag{14}$$

A.7 Spatiotemporal Skip Guidance

Spatiotemporal skip guidance (STG) is a sampling-time guidance for video diffusion models (Hyung et al., 2025). In its attention variant (STG-Attention) it forms a perturbed weak model by skipping the self-attention of one layer b and pushes the generation path away from that weak model,

$$\epsilon = \epsilon_{\text{CFG}} + w\big(\epsilon_{\text{cond}} - \epsilon_{\text{perturb}}\big), \qquad \epsilon_{\text{CFG}} = \epsilon_{\text{uncond}} + s\big(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}\big). \tag{15}$$

We set $w = 1.5$ and the classifier-free guidance scale $s = 5.0$. Here ϵ_{cond} and ϵ_{uncond} are the predictions of the intact model under the conditional and the unconditional prompt, and $\epsilon_{\text{perturb}}$ is the prediction under the conditional prompt with the self-attention of layer b skipped. In the skipped layer the attention scores are replaced by the identity.

We apply STG while the AR model generates the 5 s video that follows the 3 s observed video. The unguided arm uses classifier-free guidance alone, and each of the other eight arms adds the STG term for one layer $b \in \{0, 5, 11, 16, 22, 28, 32, 38\}$.

A.8 Human Preference Evaluation

Stimuli. For each of the 102 test videos the AR model generates the 5 s continuation under nine arms, eight with STG applied at one layer each and one baseline that uses classifier-free guidance alone. Each video is 8 s long, the 3 s original video followed by the 5 s continuation, rendered at 272×272 and 33 fps without sound.

Trial Design. In each trial an observer watches two videos side by side and chooses the one they prefer. The nine arms give $\binom{9}{2} = 36$ pairs of arms, and an observer sees every pair four times, each time with a different source video, so an observer completes 144 trials. The left and right positions are balanced within each pair. Five observers completed the test, 720 trials in total.

Bradley–Terry Model. To rank the arms from the pooled choices we fit a Bradley–Terry model (Bradley & Terry, 1952), in which arm i is chosen over arm j with probability

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j}, \tag{16}$$

where π_i is the worth of arm i. We report the log-worth $\theta_i = \log \pi_i$ with the unguided arm anchored at $\theta = 0$. A pseudo-count of 0.1 wins and 0.1 losses between every pair of arms keeps the estimate finite.

A.9 Statistical Testing

Paired comparisons across the ten subjects use an exact two-sided sign-flip test, whose smallest attainable p is 0.002. The correlation between the layer profiles of encoding accuracy and of preference is the Pearson r over the eight guided layers, with a two-sided p from an exact permutation test. Intervals on the log-worth are 95% intervals from 2,000 bootstrap resamples of observers.