NEvo: Neural-Guided Evolutionary Video Synthesis for Dynamic Visual Selectivity

Yingtian Tang[*]
EPFL
yingtian.tang@epfl.ch

Sogand Salehi[*]
EPFL
sogand.salehi@epfl.ch

Ming Zhou[*]
Johns Hopkins University
mzhou52@jhu.edu

Amir Zamir
EPFL
amir.zamir@epfl.ch

Leyla Isik
Johns Hopkins University
lisik@jhu.edu

Martin Schrimpf
EPFL
martin.schrimpf@epfl.ch

Abstract

The human brain processes dynamic visual input through hierarchically organized, functionally specialized regions. While recent *in silico* brain encoding models can synthesize optimal stimuli to probe selectivity in different brain regions, they have been largely limited to static images, leaving dynamic visual processing underexplored. We introduce a novel neural-guided video synthesis framework that generates stimuli optimized for target brain regions across visual cortex. Our method performs evolutionary search over a structured prompt space, guided by a dynamic encoding model that predicts voxel-level responses to video inputs. By maximizing predicted activity for a target ROI, the framework efficiently discovers highly activating dynamic stimuli that consistently surpass handcrafted localizer videos. The synthesized videos recover known selectivities across ventral, dorsal, and lateral pathways, and further reveal systematic differences in sensitivity to temporal dynamics. A searchlight analysis provides new insights into the progression toward increasingly complex social-dynamic features along the lateral stream, further supported by probing with synthesized abstract, non-naturalistic stimuli. Taken together, our framework enables *in silico* exploration of dynamic visual selectivity, with new predictions for *in vivo* experiments. Project website: https://nevo-project.epfl.ch/.

1 Introduction

Human vision is fundamentally dynamic: the visual system continuously integrates changing input to navigate, track objects, guide action, and infer others' intentions [23, 36, 58]. This processing is supported by hierarchically organized, functionally specialized regions across visual cortex [18, 31].

Classical accounts distinguish a ventral stream for object and scene recognition from a dorsal stream for motion and visually guided action [29, 59]. Yet dynamic sensitivity is broadly distributed: MT and V3A encode low- to mid-level motion, the extrastriate body area (EBA) responds to moving bodies and actions, and the superior temporal sulcus (STS) supports biological motion and social perception, together forming a putative lateral stream [10, 40, 56, 66]. Even ventral regions associated with static category selectivity are modulated by temporal structure [25, 37, 71, 81].

Visual cortex has traditionally been characterized through region-wise selectivity, using curated stimuli designed to drive specific regions [41, 67]. More recently, encoding models based on deep neural networks have emerged as strong predictors of brain responses [2, 16, 27, 72, 77, 81], enabling

[*]Equal contribution.

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

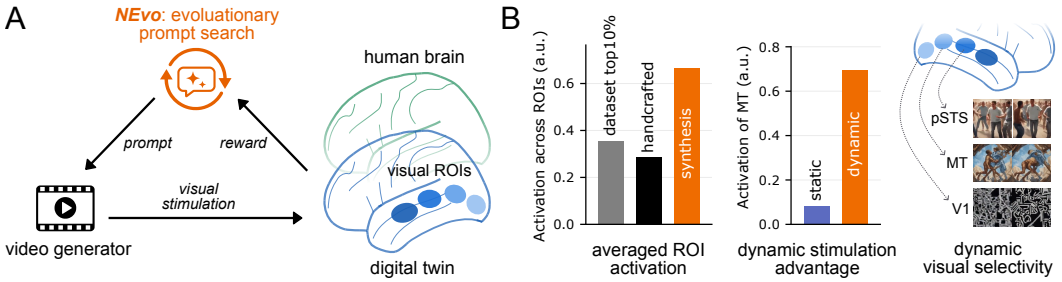

Figure 1: **A. Video synthesis framework.** We propose *NEvo*, a framework that iteratively optimizes video-generation prompts to maximize predicted activation in a target brain region. **B. Overview of results.** By combining dynamic brain encoding with video synthesis, NEvo generates highly activating stimuli, highlights visual dynamics, and enables probing along the lateral visual pathway.

generative synthesis of stimuli that elicit target neural responses [8, 13, 38, 53, 87, 89]. However, these approaches have been confined to static images. How visual cortex represents the temporal structure of natural input therefore remains an open question, particularly in regions selective for motion, interaction, or event dynamics.

We introduce ***NEvo***, a neural-guided evolutionary video synthesis framework for probing dynamic visual selectivity across the human visual cortex. As shown in Fig. 1A, NEvo combines a dynamic brain encoding model with an evolutionary search procedure over a structured prompt space. For a target region of interest (ROI), candidate prompts are rendered into videos by the generation model, scored by the brain encoding model based on predicted ROI activation, and iteratively refined through evolutionary search. To our knowledge, NEvo establishes the first model-based framework for studying visual selectivity under naturalistic dynamic stimulation.

Our contributions are fourfold (Fig. 1B). First, we introduce prompt-based neural-guided optimization over interpretable semantic, motion, and event-level descriptors, enabling targeted video synthesis in high-dimensional spatiotemporal space. Second, we synthesize high-activating stimuli across ventral, dorsal, and lateral visual cortex, exceeding video-dataset retrieval and handcrafted stimuli while recovering known selectivity and revealing regional sensitivity to temporal dynamics. Third, we use synthesis to characterize dynamic selectivity along the visual hierarchy, revealing a motion-to-social progression along the lateral pathway. Lastly, we demonstrate controlled synthesis of non-naturalistic stimuli, highlighting NEvo's potential to generate hypothesis-driven stimuli for future experiments.

2 Related work

2.1 Brain-encoding models

Many studies test whether deep neural networks and the brain exhibit similar internal activation patterns when presented with the same stimuli (e.g., images, videos, audio, or text) [5, 7, 12, 14, 20, 28, 43–45, 62, 63, 73, 78, 79, 83, 84, 90, 91, 93, 94]. In parallel, studies on brain-encoding models map stimulus representations to neural responses across recording modalities (EEG [24], fMRI [72, 81], electrophysiology [76]) and sensory domains (vision [77], language [2], audio [52]). For dynamic vision, recent video fMRI datasets span a broad range of visual dynamics, including naturalistic recordings tens of minutes long [9, 11, 42, 51, 56, 75]; models trained on them accurately, though imperfectly, predict responses to held-out stimuli [16, 81].

2.2 Model-based characterization of visual selectivity

The applications of brain-encoding models are only beginning to emerge. By predicting neural responses to arbitrary stimuli, these models enable systematic probing of functional selectivity. In particular, large-scale dataset retrieval or model-guided synthesis of stimuli that maximally activate specific regions [6, 8, 13, 38, 53, 65, 68, 89, 95] offers a more powerful alternative to traditional small-scale, handcrafted stimulus paradigms [19, 41, 67]. Dataset retrieval uses the encoding model to screen large collections of visual stimuli and select those that maximally activate a target region of

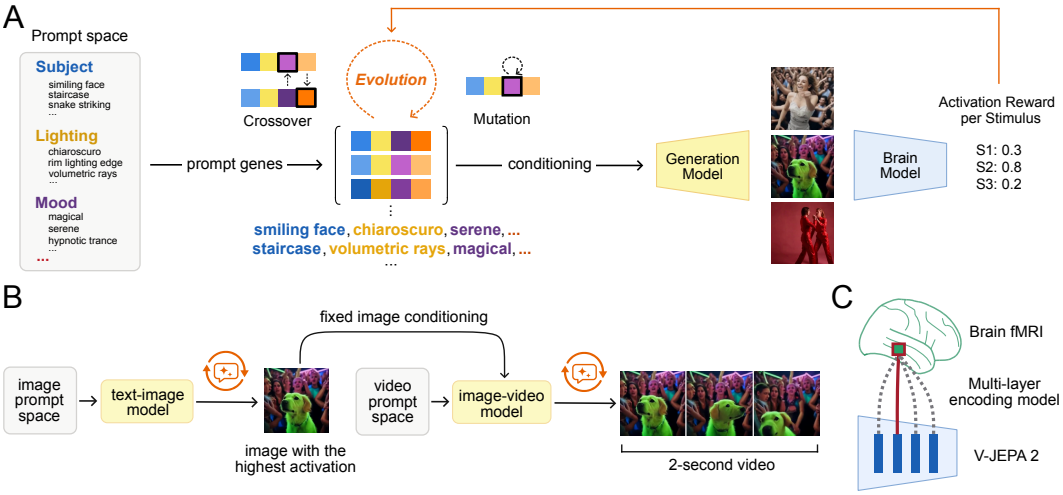

Figure 2: **A. Evolutionary prompting.** Visual attributes are encoded as interpretable genes, decoded into text prompts, and rendered into stimuli by the generation model. Predicted ROI activations from the brain model then guide selection, crossover, and mutation across generations. **B. Two-stage search.** We identify the best-scoring image through image prompt search (Tab. S6), then use it to initialize image-to-video synthesis with a separate video prompt search (Tab. S7). **C. Multi-layer brain encoding.** For each voxel, we select the best predictive layer from multiple model layers.

interest (ROI) [38]. For model-guided synthesis, a common approach is to use gradients from the encoding model to directly optimize the input stimuli, either under statistical constraints [8, 89] or with the aid of generative priors [13, 33, 38, 53, 70].

2.3 Dynamic and social visual processing

Dynamic visual processing is central to natural vision. The dorsal pathway, including MT and V3A, encodes motion, spatial structure, and moving objects or bodies [10, 80, 85], while higher parietal regions integrate this information to support attention and visually guided action [15, 64, 86]. Temporal sensitivity also extends to ventral regions, suggesting that dynamic structure contributes broadly to visual representations [49, 67, 71]. The lateral visual pathway, particularly the superior temporal sulcus (STS), links dynamic visual information to social interaction [3, 39, 55]. Recent work starts to provide evidence for increasing feature complexity along this pathway [56, 66], with posterior STS (pSTS) implicated in biological motion, facial movement, and body action [1, 32, 88], and anterior STS (aSTS) involved in communicative dynamics [50].

3 Method

3.1 *NEvo*: neural-guided evolutionary video synthesis

We introduce *NEvo*, a neural-guided evolutionary framework for synthesizing dynamic visual stimuli that maximally activate a target region of interest (ROI) r under a brain encoding model. NEvo searches over a structured prompt space $\mathcal{P}$, where each $p \in \mathcal{P}$ describes a candidate video in terms of visual content, scene structure, motion, and temporal dynamics. A video generator G maps p to a synthetic video $x = G(p)$, which is evaluated by a dynamic encoding model S_v predicting responses for individual voxels v across visual cortex.

The ROI-level objective is defined as the average predicted response across voxels in the target ROI:

$$S_r(x) = \frac{1}{|V_r|} \sum_{v \in V_r} S_v(x),$$

Algorithm 1 NEVO: Neural-guided evolutionary video synthesis

Require: Target ROI r, prompt space $\mathcal{P} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, video generator G, encoding model S_v, population size N, generations T, elite fraction P_e, crossover rate P_c, mutation rate P_m

Ensure: Ranked synthetic videos predicted to maximally activate ROI r

Define the ROI score $S_r(x) = |V_r|^{-1} \sum_{v \in V_r} S_v(x)$, where V_r is the voxel set of ROI r
Initialize population $\mathcal{Q}_0 = \{p_i\}_{i=1}^{N}$, with $p_i = (a_{i1}, \ldots, a_{iM})$ and $a_{im} \in \mathcal{A}_m$
for $t = 0, \ldots, T-1$ **do**
    Generate videos $x_i = G(p_i)$ for all $p_i \in \mathcal{Q}_t$
    Evaluate candidates using $s_i = S_r(x_i)$
    Select the top P_e fraction of candidates as elites
    Sample parents from elites and generate offspring via crossover (P_c) and mutation (P_m)
    Form the next population $\mathcal{Q}_{t+1}$ from elites and offspring
end for
return Prompt–video pairs ranked by $S_r(G(p))$

where V_r denotes the set of voxels in ROI r, and $S_v(x)$ is the predicted response of voxel v to video x. The overall synthesis objective is therefore:

$$p^\star = \arg\max_{p \in \mathcal{P}} S_r(G(p)),$$

where $p^\star$ is the prompt whose generated video $x^\star = G(p^\star)$ maximally activates ROI r. As shown in Fig. 2A, the prompt space is structured as a Cartesian product $\mathcal{P} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, where each $\mathcal{A}_m$ is a visual attribute category including appearance (subject, mood, lighting, etc.) and motion (event structure, camera motion, etc.; see Suppl. E). A prompt is represented as a gene sequence $p = (a_1, \ldots, a_M)$, where each $a_m \in \mathcal{A}_m$ selects one option from category m. The selected options are concatenated into a comma-separated text prompt, which is passed to a video diffusion model to generate a two-second video.

We devise an evolutionary search procedure to optimize the synthesis objective (Alg. 1). At each generation, candidate prompts are rendered into videos and scored using S_r. The top $P_e = 0.3$ fraction are selected as parents; offspring are produced via crossover (rate $P_c = 0.5$), where random gene-sequence prefixes are swapped between parents, and per-attribute mutation (rate $P_m = 0.2$), where attributes are independently resampled. The population size is $N = 20$. This iteratively refines the population toward prompts yielding stronger predicted ROI activation. We compare against random search and hill-climbing baselines (Fig. 4E); full hyperparameter details are in Suppl. C.1. As the same encoding model and search procedure apply to any ROI or searchlight location, the framework provides a general approach for probing dynamic visual selectivity across cortex.

3.2 Two-stage video synthesis

To improve search efficiency, we decompose $\mathcal{P}$ into static and dynamic subspaces, $\mathcal{P} = \mathcal{P}_{\text{img}} \times \mathcal{P}_{\text{vid}}$. As shown in Fig. 2B, the first stage searches over $\mathcal{P}_{\text{img}}$ using a fast image generator G_{img}, optimizing static content (subject, scene, texture, lighting, etc.). Each candidate $x_{\text{img}} = G_{\text{img}}(p_{\text{img}})$ is converted into a static video, evaluated using S_r, and the top-scoring candidate is retained as the visual anchor.

The second stage fixes the anchor image and searches over $\mathcal{P}_{\text{vid}}$, which covers dynamic attributes such as motion profile, temporal rhythm, camera motion, event structure, and interaction type. An image-to-video generator G_{vid} maps the anchor and dynamic prompt to a video $x = G_{\text{vid}}(x_{\text{img}}, p_{\text{vid}})$, evaluated by $S_r(x)$. This two-stage strategy is efficient because static content and temporal dynamics can be optimized largely independently, avoiding expensive video evaluations for poor content configurations. We use $T_{\text{img}} = 400$ and $T_{\text{vid}} = 200$ evaluations for the image and video stages, respectively. See Suppl. C.2 for search-space details and Suppl. B.1 for generator details.

3.3 Dynamic brain encoding model

We use V-JEPA 2 [4] as the feature backbone, motivated by prior large-scale evaluations of brain-aligned video models [81]. As shown in Fig. 2C, unlike prior work relying on top-level representations such as final CLIP embeddings [13, 53, 69], we extract multi-block V-JEPA 2 features, average them over space and time, and fit voxel-wise ridge regressions to fMRI responses (Suppl. B.2).

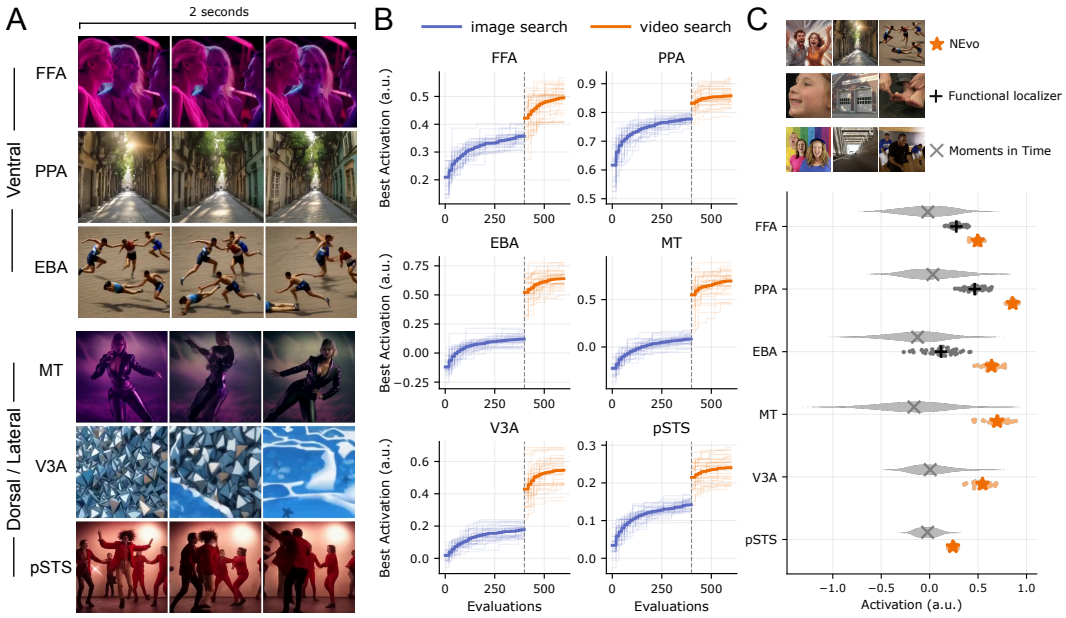

Figure 3: **A. Example synthetic videos from NEvo.** Two-second videos optimized to maximize predicted activation in ventral and dorsal/lateral regions. **B. Two-stage search over iterations.** Optimization trajectories across genetic-search evaluations for each ROI, spanning image and video search phases. Faded lines denote individual seeds; the solid line denotes the mean. **C. Comparison of ROI activations.** Predicted ROI activations are compared across stimuli from NEvo, dynamic functional localizers [67], and the Moments in Time dataset [60].

Each voxel is assigned the block with the best validation performance, allowing cortical locations to align with different model depths. We train the model-to-brain mapping on two fMRI datasets, *BOLDMoments* [51] and a social interaction dataset [56], both projected onto the *fsaverage5* cortical surface. Because the model predicts normalized responses, we report activations in arbitrary units (a.u.). For full framework reproducibility, see Suppl. J.

4 Results

4.1 NEvo synthesizes highly activating videos across visual cortex

We evaluate the two-stage NEvo on representative ROIs spanning ventral, dorsal, and lateral visual streams, including FFA, PPA, EBA, MT, V3A, and pSTS (ROI definition in Suppl. G). All candidates are scored by our dynamic encoding model, using the mean predicted response across target-ROI voxels. To ensure consistent scoring, all videos are stored as files at a fixed quality setting and reloaded before being passed to the encoding model (Suppl. B.2). Fig. 3A shows representative NEvo videos, which recover expected regional selectivities, including face-like content for FFA, scene-like content for PPA, body-related content for EBA, coherent motion for MT/V3A, and interaction-like dynamics for pSTS (more examples in Suppl. A). Fig. 3B shows optimization trajectories across random seeds, with each seed yielding one final optimized video. Mean predicted ROI activation increases across evaluations, with gains in both search stages.

In Fig. 3C, we then compare NEvo stimuli with commonly used dynamic functional localizers and videos from Moments in Time (MiT). Across ROIs, NEvo reaches high predicted activations relative to both reference sets, achieving the top 99.8% of the MiT response and 95.8% of the localizer response on average. These high activations also transfer to held-out brain encoding models (Tab. S1). These results demonstrate that *NEvo* efficiently matches or surpasses the responses elicited by real-video retrieval and dynamic localizers, as evaluated by the brain encoder.

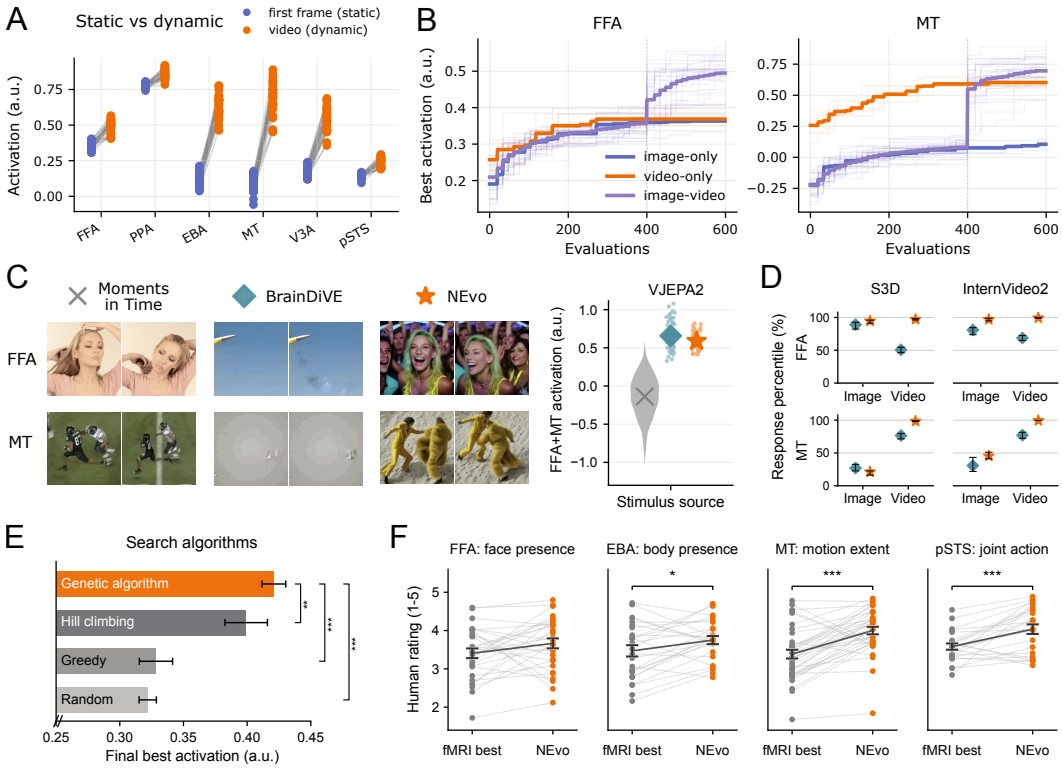

Figure 4: **A. Dynamic stimuli improve activation.** For each ROI, we compare predicted activation between the NEvo-synthesized video and its first frame presented as a static video. **B. A two-stage search.** The two-stage image-to-video search strategy outperforms single-stage search baselines in two representative regions. **C. Ablation of synthesis methods.** NEvo robustly synthesizes high-activating stimuli. While the gradient-based BrainDiVE approach [53] also achieves high predicted activations, its synthesized stimuli exhibit signs of overfitting, as illustrated by the FFA and MT examples (additional samples in Fig. S12; implementation details in Suppl. H). Here we use a VJEPA2-based encoder for both synthesis and evaluation. **D. Held-out validation.** We use the VJEPA2-based encoder for synthesis and S3D- and InternVideo2-based encoders for held-out evaluation. For FFA and MT, we report response percentiles relative to the distribution of responses to Moments in Time videos. Results are shown for both image and video synthesis. **E. Ablation of search methods.** Final best activation of genetic search versus baseline search algorithms (paired Wilcoxon, Holm-corrected; details in Suppl. D and Suppl. C.2). **F. Human ratings.** Participants (n=114) rated NEvo and top-activating videos from fMRI datasets for face presence (FFA), body presence (EBA), motion extent (MT), and joint action (pSTS). More details in Suppl. F.

4.2 Dynamic structure enhances regional activation beyond static appearance

We test whether optimized videos drive ROIs beyond static appearance alone. We track predicted activation across the two-stage search, and compare each optimized video with its first-frame static control. On average, the video phase further improves activation after image search (Fig. S6; mean improvement: 0.27 ± 0.03, 95% CI; same CI convention below), though with higher computational cost. Across individual ROIs, optimized videos outperform first-frame controls (Fig. 4A), indicating broad sensitivity to dynamic information, with the strongest effects in motion-related regions (e.g., MT; 0.61 ± 0.05) and measurable effects in ventral regions (e.g., FFA; 0.14 ± 0.02).

In Fig. 4B, we evaluate the two-stage image-to-video search strategy using five seeds for FFA and MT. The two-stage approach consistently outperforms both video-only and image-only search under matched conditions, demonstrating the benefit of spatial and temporal domain decomposition. Meanwhile, the two-stage search is substantially faster than the video-only search (Fig. S6).

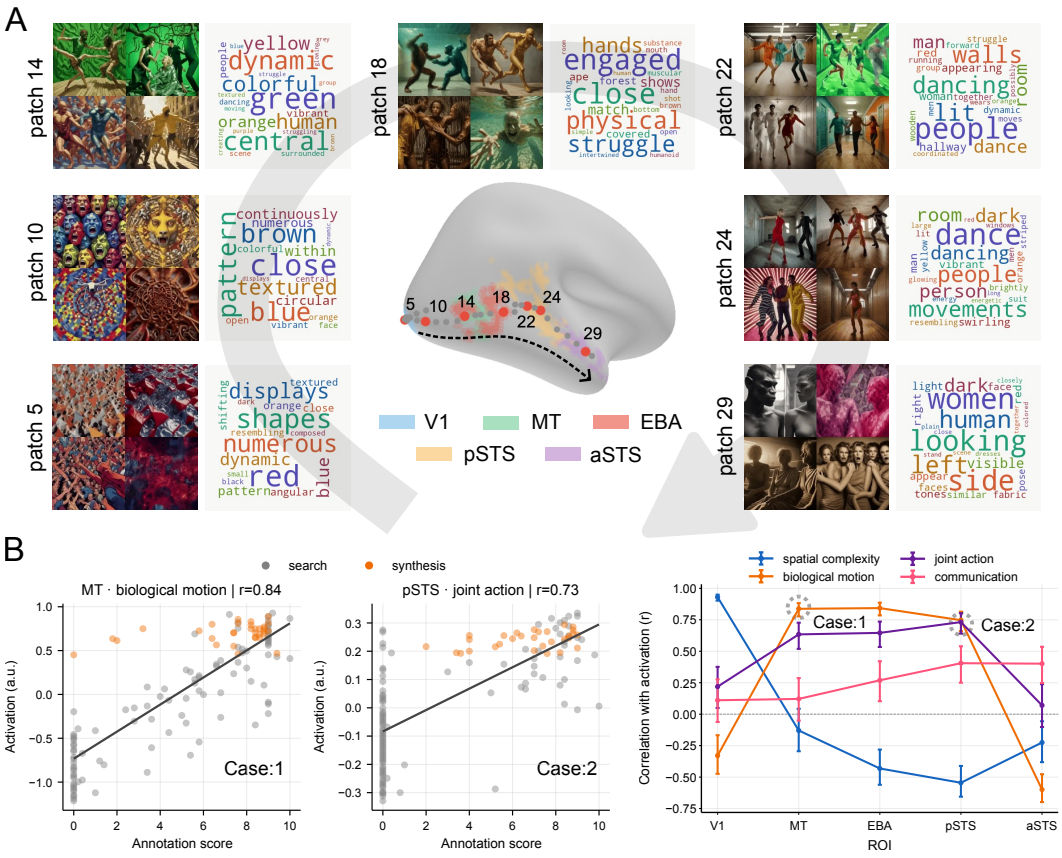

Figure 5: **A. Searchlight synthesis.** Top-synthesis stimuli are shown alongside word clouds from Gemini-annotated descriptions for example searchlight patch along the lateral stream trajectory (V1 to aSTS). Patch indices are marked on the cortical surface. Left-hemisphere results and dataset-search comparisons are shown in Figs. S9 and S10. **B. Visual property–activation correlations.** Left: Representative correlations between annotated visual-property scores and predicted activations for two example cases: biological motion in MT and joint action in pSTS. Orange points indicate synthesized stimuli and gray points indicate search stimuli from MiT; additional correlations in Fig. S11. Right: Correlation profiles across lateral-stream ROIs for visual properties; Error bars denote 95% CIs derived via Fisher z-transformation; full results in Tab. S12.

4.3 Ablating components of the synthesis framework

We next ablate key components of NEvo to assess design choices (Fig. 4C–E). In Fig. 4C, we compare NEvo with BrainDiVE, a canonical gradient-based synthesis method, for FFA and MT. First, we use the VJEPA2-based encoder for both synthesis and evaluation. In the dynamic setting, BrainDiVE achieves high activation but fails to generate meaningful stimuli (additional examples in Fig. S12). In contrast, NEvo reliably produces high-activating and meaningful stimuli.

Second, we use the VJEPA2-based encoder for synthesis and S3D- and InternVideo2-based encoders for held-out evaluation (Fig. 4D). NEvo achieved higher response percentiles (relative to 1,000 Moments in Time videos) for both images and videos, with the largest advantage in video synthesis (97.3–99.5% for NEvo versus 50.7–77.1% for BrainDiVE), indicating strong generalization beyond the guiding brain encoder. NEvo also mostly outperformed BrainDiVE in image synthesis, the setting for which BrainDiVE was originally proposed. Interestingly, both methods show substantial overfitting when synthesizing images for MT, possibly reflecting MT's intrinsic preference for dynamic information. Extending the generalization evaluation for NEvo across ROIs (Table S1), its median predicted responses ranked between the 78.7th and 99.2nd percentiles, exceeding the 95th



Figure 6: **Controlled synthesis from a non-naturalistic anchor.** Starting from a fixed first frame of two stacked plasticine discs, we run video-stage NEvo to maximize predicted pSTS or MT activation. For each ROI, examples show the highest and lowest activation boosts across seeds, measured relative to the static first-frame baseline. pSTS optimization introduces face-like features and coordinated interactions, whereas MT optimization yields motion-rich dynamics with less biological or social content, consistent with Fig. 5. Additional examples and controls are shown in Fig. S14.

percentile for EBA, pSTS, and FFA in both held-out encoders. Overall, these results indicate greater robustness of evolutionary prompting than gradient-based guidance for video synthesis.

Fig. 4E compares search algorithms using activation across FFA and MT. Genetic search significantly outperforms random search, greedy search, and hill climbing, supporting its effectiveness (Suppl. D).

Finally, though neuroimaging validation is beyond the scope of this work, we conducted a human behavioral experiment to assess whether NEvo videos exhibit the region-specific visual features they were synthesized to maximize (Fig. 4F; Suppl. F). We recruited 133 online participants via Prolific, retaining 114 after quality checks. For each ROI, participants rated synthesized NEvo videos and top-activating naturalistic videos drawn from two existing fMRI datasets [51, 56] on the corresponding target feature: motion extent (MT), body presence (EBA), face presence (FFA), or joint action (pSTS). Videos were intermixed and presented in randomized order. Two-sided paired t-tests on participant-level mean ratings showed that NEvo videos received significantly higher ratings than top-activating naturalistic videos for EBA ($p = 0.035$), MT ($p < 0.001$), and pSTS ($p < 0.001$), with a non-significant trend for FFA ($p = 0.086$).

4.4 Synthesize videos characterize selectivity across lateral visual stream

To use NEvo for dynamic visual selectivity discovery, we apply it along a searchlight trajectory across the lateral visual stream, a recently proposed pathway for social processing that remains only partially characterized [56, 66]. We define this trajectory by connecting landmark ROIs from V1 to aSTS along cortical-surface geodesics (shortest paths on the cortical surface; Suppl. G), spanning V1, MT, EBA, pSTS, and aSTS. For each searchlight patch along this trajectory, we use NEvo to synthesize high-activating videos. We then summarize their dominant content using Gemini-2.5-Flash [30] to generate one-sentence video descriptions and aggregate these descriptions into word clouds.

The synthesized stimuli reveal a systematic progression of feature selectivity along the lateral stream (Fig. 5A). Early visual regions prefer repeated, colorful, high-contrast textures, with descriptions dominated by low-level terms such as “textured,” “numerous,” and “red” (patches 5 and 10). Mid-level lateral regions show increasing motion and body content (patch 14), consistent with MT and EBA selectivity. Farther along the pathway, stimuli become dominated by physical interactions and synchronized, coordinated body movements, such as “struggling” and “dancing” (patches 18, 22, and 24). In aSTS, synthesized stimuli shift toward face-to-face social contact (patch 29). Together, these results show a progressive gradient from low-level spatial structure to body motion and high-level social interaction along the lateral visual stream.

4.5 Synthesized videos enhance activation through improved visual properties

We further characterize the functional relevance of the synthesized stimuli. We again use Gemini-2.5-Flash to annotate low-level structure, motion, body, scene, and social-interaction properties for both

synthesized videos and naturalistic videos sampled uniformly across each ROI's predicted activation range (Suppl. E). We compute Pearson correlations between each property and model-predicted activation, and overlay synthesized stimuli on the naturalistic video distributions (Fig. 5B left). Representative ROIs show strong property–activation correlations: MT with biological motion ($r = 0.84$, $p < 0.001$), pSTS with joint action ($r = 0.73$, $p < 0.001$). Across ROIs, synthesized stimuli consistently occupy the high-property, high-activation extreme, suggesting that *NEvo* converges on functionally characteristic stimuli rather than incidental outliers. Along the lateral pathway, property–activation correlations reveal a gradual transition toward increasingly complex social features, peaking in pSTS, shifting toward less motion-focused communicative features in aSTS (Fig. 5B, right).

As a controlled application of *NEvo*, we fix the static appearance to a first frame containing two abstract objects and optimize only the video dynamics for different ROIs across multiple random seeds (Suppl. I). As shown in Fig. 6, pSTS optimization introduces face-like features and coordinated interactions, consistent with social-dynamic selectivity, whereas MT optimization yields translation- and oscillation-rich motion with fewer biological or social attributes. These results show that NEvo can dissociate region-specific dynamic selectivity from abstract anchors, providing a controlled paradigm for generating targeted, hypothesis-driven stimuli.

5 Discussion

We introduced *NEvo*, a neural-guided evolutionary framework for synthesizing dynamic stimuli optimized for target regions across human visual cortex. Synthesized videos recover expected selectivities and elicit strong predicted responses, exceeding established localizers even in well-characterized regions such as FFA and MT, and generate testable hypotheses for less-characterized regions (pSTS and aSTS) along the lateral visual stream.

NEvo formulates video synthesis as black-box optimization over interpretable semantic, motion, and event-level descriptors, moving model-based stimulus design beyond static images into the substantially larger search space of video. Compared with direct gradient-based optimization, evolutionary prompting constrains synthesis toward semantically meaningful and temporally coherent videos, while remaining flexible enough to target different cortical regions. It also demonstrates strong robustness across different encoding models.

Our results show that temporal structure broadly shapes visual selectivity: optimized videos outperform first-frame controls in motion-sensitive regions such as MT, and even in ventral regions such as PPA. The two-stage strategy exploits this by first identifying strong appearance anchors, and then optimizing anchor-specific motion dynamics, avoiding many uninformative video generations.

Beyond predefined ROIs, *NEvo* enables model-based discovery of dynamic selectivity across the cortical surface. Along the lateral visual stream, searchlight synthesis and activation-property analysis reveal a smooth progression toward increasingly complex socially relevant features: from motion-rich, body-involved physical movement, to synchronized, coordinated joint interaction, to face-centered, motion-suppressed communicative content. Controlled synthesis dissociates these region-specific preferences even with abstract objects. Together, these findings provide convergent evidence for functional gradients and reveal finer-grained structure than prior ROI-based observations [17, 56, 66].

Limitations. *NEvo* depends on the accuracy and coverage of the underlying encoding model, so synthesized stimuli may reflect both neural selectivity and model bias; future validation with human neuroimaging experiments is therefore necessary. The current model is vision-only, limiting characterization of regions such as aSTS, where communication often relies on speech and audiovisual cues. Finally, the structured prompt space improves interpretability but may omit relevant visual dimensions not captured by the schema.

Broader impact. *NEvo* may benefit neuroscience and clinical research by enabling scalable, controlled design of dynamic stimuli and more reproducible experiments. Potential risks include amplifying biases from framework components, including video generators and encoding models.

Future work should incorporate multimodal encoding models, richer video generators, and search spaces suited to longer event structures. Critically, closed-loop neuroimaging should test whether *NEvo*-synthesized videos elicit the predicted human brain responses, potentially extending neural-guided synthesis from mapping healthy brain function to designing targeted clinical interventions.

Acknowledgments and Disclosure of Funding

This work was supported by the Swiss National Science Foundation (SNSF) under Grant No. 10.003.772. This work was also supported under project ID 43 as part of the Swiss AI Initiative, through a grant from the ETH Domain, with computational resources provided by the Swiss National Supercomputing Centre (CSCS) on the Alps infrastructure. This work has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI). This work was also supported by the National Institutes of Health (NIH) under Grant No. R01MH132826.

For conflicts of interest, M.S. declares Read/Write Labs and Neurosoft Bioelectronics.

This work used OpenAI Codex to refine the code and improve the writing.

References

[1] Ralph Adolphs. Cognitive neuroscience of human social behaviour. *Nature reviews neuroscience*, 4(3):165–178, 2003.

[2] Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350, 2025.

[3] 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.

[4] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[5] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 25164–25178. Curran Associates, Inc., 2021. URL `https://papers.nips.cc/paper/2021/hash/d384dec9f5f7a64a36b5c8f03b8a6d92-Abstract.html`.

[6] Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. Mindsimulator: Exploring brain concept localization via synthetic fmri. *arXiv preprint arXiv:2503.02351*, 2025.

[7] Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

[8] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, May 2019. doi: 10.1126/science.aav9436. URL `https://www.science.org/doi/10.1126/science.aav9436`. Publisher: American Association for the Advancement of Science.

[9] Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9:91, March 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01173-0. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8938409/`.

[10] Richard T. Born and David C. Bradley. STRUCTURE AND FUNCTION OF VISUAL AREA MT. *Annual Review of Neuroscience*, 28(1):157–189, July 2005. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.26.041002.131052. URL `https://www.annualreviews.org/doi/10.1146/annurev.neuro.26.041002.131052`.

[11] Julie A Boyle, Basile Pinsard, and Lune Pierre Bellec. Cneuromod data collection complete: 200h of individual fmri across diverse naturalistic and controlled tasks to build neuroai models. In *Conference on Cognitive Computational Neuroscience (CCN 2025)*, 2025.

[12] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, February 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1. Number: 1 Publisher: Nature Publishing Group.

[13] Diego García Cerdas, Christina Sartzetaki, Magnus Petersen, Gemma Roig, Pascal Mettes, and Iris Groen. Brainactiv: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation. *bioRxiv*, pages 2024–10, 2024.

[14] Minkyu Choi, Kuan Han, Xiaokai Wang, Yizhen Zhang, and Zhongming Liu. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *Advances in Neural Information Processing Systems*, 36:50408–50428, 2023.

[15] Kristl G. Claeys, Delwin T. Lindsey, Erik De Schutter, and Guy A. Orban. A higher order motion region in human inferior parietal lobule: evidence from fMRI. *Neuron*, 40(3):631–642, October 2003. ISSN 0896-6273. doi: 10.1016/s0896-6273(03)00590-7.

[16] Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction. *arXiv preprint arXiv:2507.22229*, 2025.

[17] Ben Deen, Kami Koldewyn, Nancy Kanwisher, and Rebecca Saxe. Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral cortex*, 25(11): 4596–4609, 2015.

[18] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434, February 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.01.010. URL https://www.cell.com/neuron/abstract/S0896-6273(12)00092-X. Publisher: Elsevier.

[19] Russell Epstein, Alison Harris, Damian Stanley, and Nancy Kanwisher. The parahippocampal place area: recognition, navigation, or encoding? *Neuron*, 23(1):115–125, 1999.

[20] Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex. *bioRxiv*, pages 2023–12, 2023.

[21] Bruce Fischl. FreeSurfer. 62(2):774–781, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.01.021. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811912000389.

[22] Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 2011.

[23] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, November 2014. ISBN 978-1-317-57938-0. Google-Books-ID: 8BSLBQAAQBAJ.

[24] Alessandro T Gifford, Domenic Bersch, Daniel Janini, Gemma Roig, and Radoslaw M Cichy. Catalyzing in silico neuroscience with a toolkit of accurate encoding models of the brain. 2025.

[25] Sharon Gilaie-Dotan, Ayse P Saygin, Lauren J Lorenzi, Ryan Egan, Geraint Rees, and Marlene Behrmann. The role of human ventral visual cortex in motion perception. *Brain*, 136(9): 2784–2798, 2013.

[26] 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, August 2016. ISSN 1476-4687. doi: 10.1038/nature18933. URL https://www.nature.com/articles/nature18933. Publisher: Nature Publishing Group.

[27] Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *arXiv preprint arXiv:2411.05712*, 2024.

[28] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25 (3):369–380, March 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4. Number: 3 Publisher: Nature Publishing Group.

[29] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992. ISSN 0166-2236. doi: 10.1016/0166-2236(92)90344-8.

[30] Google. Gemini. https://gemini.google.com/, 2026. Large language model; accessed 2026-05-06.

[31] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27 (1):649–677, 2004.

[32] Emily D Grossman and Randolph Blake. Brain areas active during visual perception of biological motion. *Neuron*, 35(6):1167–1175, 2002.

[33] Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R Sabuncu, and Amy Kuceyeski. Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

[34] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

[35] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

[36] Uri Hasson, Janice Chen, and Christopher J. Honey. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313, June 2015. ISSN 1364-6613. doi: 10.1016/j.tics.2015.04.006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4457571/.

[37] James V Haxby, M Ida Gobbini, and Samuel A Nastase. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *Neuroimage*, 216:116561, 2020.

[38] Ethan Hwang, Hossein Adeli, Wenxuan Guo, Andrew Luo, and Nikolaus Kriegeskorte. In silico mapping of visual categorical selectivity across the whole brain. *arXiv preprint arXiv:2510.21142*, 2025.

[39] Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152, October 2017. doi: 10.1073/pnas.1714471114.

[40] Jan Jastorff, Ivo D. Popivanov, Rufin Vogels, Wim Vanduffel, and Guy A. Orban. Integration of shape and motion cues in biological motion processing in the monkey STS. *NeuroImage*, 60(2):911–921, April 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.12.087. URL https://www.sciencedirect.com/science/article/pii/S1053811912000286.

[41] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997.

[42] Umit Keles, Julien Dubois, Kevin J. M. Le, J. Michael Tyszka, David A. Kahn, Chrystal M. Reed, Jeffrey M. Chung, Adam N. Mamelak, Ralph Adolphs, and Ueli Rutishauser. Multimodal single-neuron, intracranial EEG, and fMRI brain responses during movie watching in human patients. *Scientific Data*, 11(1):214, February 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03029-1. URL https://www.nature.com/articles/s41597-024-03029-1. Publisher: Nature Publishing Group.

[43] Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018. ISSN 08966273. doi: 10.1016/j.neuron.2018.03.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627318302502.

[44] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11): e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915.

[45] Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, October 2019. doi: 10.1073/pnas.1905544116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1905544116. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[46] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[47] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[48] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, March 2006. doi: 10.1073/pnas.0600244103.

[49] Emel Küçük, Matthew Foxwell, Daniel Kaiser, and David Pitcher. Moving and static faces, bodies, objects and scenes are differentially represented across the three visual pathways. preprint, Neuroscience, December 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.11.30.518408.

[50] Juha M Lahnakoski, Enrico Glerean, Juha Salmi, Iiro P Jääskeläinen, Mikko Sams, Riitta Hari, and Lauri Nummenmaa. Naturalistic fmri mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in human neuroscience*, 6:233, 2012.

[51] 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, July 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-50310-3. URL https://www.nature.com/articles/s41467-024-50310-3. Publisher: Nature Publishing Group.

[52] Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H. Carney, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, December 2023. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-023-01468-4. URL https://www.nature.com/articles/s41593-023-01468-4.

[53] Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36:75740–75781, 2023.

[54] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[55] Emalie McMahon and Leyla Isik. Seeing social interactions. *Trends in Cognitive Sciences*, 27(12):1165–1179, December 2023. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2023.09.001. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(23)00248-6. Publisher: Elsevier.

[56] Emalie McMahon, Michael F. Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, December 2023. ISSN 0960-9822. doi: 10.1016/j.cub.2023.10.015. URL https://www.sciencedirect.com/science/article/pii/S0960982223013738.

[57] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886. IEEE, 2020.

[58] David Milner and Mel Goodale. *The visual brain in action*, volume 27. Oup Oxford, 2006.

[59] Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, January 1983. ISSN 0166-2236. doi: 10.1016/0166-2236(83)90190-X. URL https://www.sciencedirect.com/science/article/pii/016622368390190X.

[60] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[61] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. Moments in Time Dataset: one million videos for event understanding, February 2019. URL http://arxiv.org/abs/1801.03150. arXiv:1801.03150 [cs].

[62] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-Driven Convolutional Recurrent Models of the Visual System. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6be93f7a96fed60c477d30ae1de032fd-Abstract.html.

[63] Paula Olszewski-Kubilius, Rena F. Subotnik, Lauren Cassani Davis, and Frank C. Worrell. Benchmarking Psychosocial Skills Important for Talent Development. *New Directions for Child and Adolescent Development*, 2019(168):161–176, 2019. ISSN 1534-8687. doi: 10.1002/cad.20318. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cad.20318. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cad.20318.

[64] Guy A. Orban, Alessia Sepe, and Luca Bonini. Parietal maps of visual signals for bodily action planning. *Brain Structure and Function*, 226(9):2967–2988, December 2021. ISSN 1863-2661. doi: 10.1007/s00429-021-02378-6. URL https://doi.org/10.1007/s00429-021-02378-6.

[65] Pawel Pierzchlewicz, Konstantin Willeke, Arne Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images. *Advances in Neural Information Processing Systems*, 36:32574–32601, 2023.

[66] David Pitcher and Leslie G. Ungerleider. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, 25(2):100–110, February 2021. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2020.11.006. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(20)30278-3. Publisher: Elsevier.

[67] David Pitcher, Daniel D. Dilks, Rebecca R. Saxe, Christina Triantafyllou, and Nancy Kanwisher. Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, 56(4):2356–2363, June 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.03.067.

[68] Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10, May 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.04.005. URL https://www.sciencedirect.com/science/article/pii/S0092867419303915.

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[70] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[71] Sophia Robert, Leslie G. Ungerleider, and Maryam Vaziri-Pashkam. Disentangling Object Category Representations Driven by Dynamic and Static Visual Input. *The Journal of Neuroscience*, 43(4):621–634, January 2023. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0371-22.2022. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0371-22.2022.

[72] Christina Sartzetaki, Gemma Roig, Cees GM Snoek, and Iris Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[73] Christina Sartzetaki, Anne W Zonneveld, Pablo Oyarzo, Alessandro T Gifford, Radoslaw M Cichy, Pascal Mettes, and Iris IA Groen. The human brain as a dynamic mixture of expert models in video understanding. *bioRxiv*, pages 2025–10, 2025.

[74] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023. URL https://arxiv.org/abs/2311.17042.

[75] Clara Sava-Segal, Chandler Richards, Megan Leung, and Emily S Finn. Individual differences in neural event segmentation of continuous experiences. *Cerebral Cortex*, 33(13):8164–8178, July 2023. ISSN 1047-3211. doi: 10.1093/cercor/bhad106. URL https://doi.org/10.1093/cercor/bhad106.

[76] 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? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

[77] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3):413–423, November 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.040. URL https://www.cell.com/neuron/abstract/S0896-6273(20)30605-X. Publisher: Elsevier.

[78] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. doi: 10.1073/pnas.2105646118. URL

https://www.pnas.org/doi/abs/10.1073/pnas.2105646118. Publisher: Proceedings of the National Academy of Sciences.

[79] Martin Schrimpf, Paul McGrath, Eshed Margalit, and James J. DiCarlo. Do Topographic Deep ANN Models of the Primate Ventral Stream Predict the Perceptual Effects of Direct IT Cortical Interventions? preprint, Neuroscience, January 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.01.09.572970.

[80] Eero P. Simoncelli and David J. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, March 1998. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00183-1. URL https://www.sciencedirect.com/science/article/pii/S0042698997001831.

[81] Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Diverse perceptual representations across visual pathways emerge from a single objective. *bioRxiv*, pages 2025–07, 2025.

[82] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[83] Jessica AF Thompson, Hannah Sheahan, Tsvetomira Dumbalska, Julian D Sandbrink, Manuela Piazza, and Christopher Summerfield. Zero-shot counting with a dual-stream neural network model. *Neuron*, 112(24):4147–4158, 2024.

[84] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

[85] Roger BH Tootell, Janine D Mendola, Nouchine K Hadjikhani, Patrick J Ledden, Arthur K Liu, John B Reppas, Martin I Sereno, and Anders M Dale. Functional analysis of v3a and related areas in human visual cortex. *Journal of Neuroscience*, 17(18):7060–7078, 1997.

[86] Annalisa Tosoni, Gaspare Galati, Gian Luca Romani, and Maurizio Corbetta. Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience*, 11(12):1446–1453, December 2008. ISSN 1546-1726. doi: 10.1038/nn.2221. URL https://www.nature.com/articles/nn.2221. Number: 12 Publisher: Nature Publishing Group.

[87] Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, March 2024. ISSN 2397-3374. doi: 10.1038/s41562-023-01783-7. URL https://www.nature.com/articles/s41562-023-01783-7. Publisher: Nature Publishing Group.

[88] Jon Walbrin, Paul Downing, and Kami Koldewyn. Neural responses to visually observed social interactions. *Neuropsychologia*, 112:31–39, 2018.

[89] Edgar Y. Walker, Fabian H. Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G. Fahey, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, December 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0517-x. URL https://www.nature.com/articles/s41593-019-0517-x. Publisher: Nature Publishing Group.

[90] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in neural information processing systems*, 32, 2019.

[91] Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058):470–477, 2025.

[92] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer, 2024.

[93] Konstantin F Willeke, Polina Turishcheva, Alex Gilbert, Goirik Chakrabarty, Hasan A Bedel, Paul G Fahey, Yongrong Qiu, Marissa A Weis, Michaela Vystrčilová, Taliah Muhammad, et al. Omnimouse: Scaling properties of multi-modal, multi-task brain models on 150b neural tokens. *arXiv preprint arXiv:2604.18827*, 2026.

[94] 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, June 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1403112111. URL https://pnas.org/doi/full/10.1073/pnas.1403112111.

[95] Jacob Yeung, Andrew F Luo, Gabriel Sarch, Margaret M Henderson, Deva Ramanan, and Michael J Tarr. Neural representations of dynamic visual stimuli. *arXiv preprint arXiv:2406.02659*, 1(2):3, 2024.

A Additional visualizations

We provide additional qualitative examples of synthetic videos generated by *NEvo*. Figs. S1 and S2 show additional two-second videos optimized for different target ROIs. Fig. S3 shows examples generated with improved prompt-quality constraints and higher-resolution rendering.

Fig. S4 provides additional stimuli from the lateral-stream searchlight analysis. These examples illustrate how NEvo synthesizes different feature combinations along the cortical hierarchy, ranging from high-frequency textures near early visual cortex to body motion, dyadic interaction, and crowded social scenes in higher lateral temporal regions. Notably, several synthesized stimuli combine visual features that are uncommon in natural video datasets, such as abstract social configurations without clear identity cues or dense pattern fields without natural objects.

In Fig. S12, we show additional samples from applying BrainDiVE to dynamic stimulus synthesis. Many samples contain artifacts introduced by gradient-based modification, and none reflects the target ROI selectivity.

B Concrete model choices in *NEvo*

B.1 Generation models

For the image search stage, we use SDXL-Turbo [74], a distilled single-step variant of Stable Diffusion XL. Images are generated at 512×512 resolution in a single inference step with guidance scale 0.0.

For the video search stage, we use LTX-Video 2.3 [34, 35], a 22B distilled diffusion transformer with a spatial upsampler and Gemma-3 12B [82] as the text encoder. Videos are generated at 256×256 resolution, 24 fps, and 2 seconds duration. For image-to-video generation, the anchor image from the image search stage is provided as a conditioning input at frame index 0 with strength 1.0. For text-to-video generation, no image conditioning is used. All videos are generated with fp8 quantization.

In general, higher generation resolutions yield higher predicted brain responses (Fig. S13), even though the brain encoding model normalizes all inputs to the same quality (Suppl. B.2). The settings above were chosen to balance performance with computational cost.

B.2 Brain-encoding models

To build the VJEPA2 encoding model, we use all videos and corresponding fMRI response betas from *BOLDMoments* [51] and a social interaction dataset [56]. Each video is associated with a single beta volume; therefore, we predict one response vector per video rather than a time-resolved response. We project the volumetric data onto the *fsaverage* surface and normalize responses per voxel within each dataset.

Videos are input to VJEPA2 at 12 fps. We extract features from `backbone.blocks.`i`.attn`, with $i \in \{13, 15, 17, 19, 21, 23\}$. Each feature tensor has dimensions time $\times$ height $\times$ width $\times$ channels. For each layer, we average features over time, height, and width, and independently regress the resulting feature vector onto each voxel using ridge regression. The ridge penalty is selected by cross-validation over 17 logarithmically spaced values between 10^{-8} and 10^{8}. We split the data into 90% training and 10% validation sets, and assign each voxel to the layer with the best validation performance.

We additionally construct held-out encoding models using S3D pretrained on HowTo100M with MIL-NCE [57] and the visual backbone of InternVideo2-Stage2-1B [92]. As with the VJEPA2 encoding model, we extract layer-wise visual features, pool them over space and time, and fit linear readouts to predict surface-based fMRI responses, producing one response vector per video. Other training settings are the same as those of the VJEPA2 encoder.

In Fig. S5, we additionally compare to a purely static encoding model. The CLIP-based model is constructed by extracting features from individual video frames using the last six layers of the CLIP vision encoder, concatenating them into a single feature representation, and applying the same averaging and regression procedure. For the "last-layer" variants, we use only the final layer rather

than the multi-layer feature set described above. For this particular figure, to assess out-of-distribution generalization in this analysis, models are trained on the social interaction dataset and tested on *BOLDMoments*. The final ceiled predictivity is measured as the Pearson correlation between predicted and measured responses, normalized by the internal consistency of the brain signal [76].

To ensure that synthetic-video scoring is not confounded by generation quality, particularly resolution, we standardize all videos before final activation prediction. Each generated video is exported with FFmpeg at 256×256 resolution, 24 fps, 2 s duration, CRF 18, medium preset, and `yuv420p` pixel format, then reloaded as tensors for encoding-model prediction. Image inputs are converted into 2-second static videos using the same format.

C Details of the *NEvo* search procedure.

C.1 Hyperparameter summary

We summarize the hyperparameters used in Sec. 3.1 and Sec. 3.2. Unless otherwise stated, we use a population size of $N = 20$, an elite fraction of $P_e = 0.3$, a single-point crossover rate of $P_c = 0.5$, and a per-attribute mutation rate of $P_m = 0.2$. These correspond directly to the inputs of Alg. 1.

The population size N controls the number of candidate prompts evaluated per generation. The elite fraction P_e determines the proportion of top-scoring candidates retained as parents, enforcing selection pressure toward high-activation stimuli. The crossover rate P_c specifies the probability of recombining two parent prompts via single-point crossover, exchanging subsets of attributes to form offspring. The mutation rate P_m is applied independently to each attribute, randomly resampling it from its domain, enabling exploration of the prompt space.

For the two-stage search, we use $T_{\text{img}} = 400$ evaluations for the image stage and $T_{\text{vid}} = 200$ evaluations for the video stage. The image stage optimizes static content attributes to identify promising configurations, which initialize the video stage where motion and temporal attributes are further refined. This decomposition exploits the approximate factorization of content and dynamics, reducing the number of expensive video evaluations. We chose these evaluation budgets because performance generally plateaued beyond these points.

Hyperparameter sensitivity is evaluated in Fig. S8, where we vary mutation rate, crossover rate, elite fraction, and population size while keeping the remaining parameters fixed at their default values. Additional search results for more ROIs are shown in Fig. S7.

C.2 Design of the search space

The prompt search space $\mathcal{P} = \mathcal{P}_{\text{img}} \times \mathcal{P}_{\text{vid}}$ is organized into categories, where each category represents one dimension of the visual or temporal content of a stimulus. Options within each category span the relevant range of values for that dimension.

For the image space $\mathcal{P}_{\text{img}}$, categories are grouped by the level of the visual hierarchy they target. Low-level categories such as SpatialFrequency, EdgeDensity, and Pattern cover basic image statistics relevant for early visual cortex. Mid-level categories such as Texture and Surface cover material properties. Higher-level categories cover object shape (ObjectForm), body configuration (BodyContent, BodyAction), face identity and gaze (FaceContent, FaceViewpoint, FaceOcclusion), and social interaction (SocialDynamics, SocialCue, IntentionSignal). Scene-level categories such as SceneCategory and SceneGeometry target scene-selective regions.

The image space also includes motion-related categories such as ScenePhysics and MotionSourceType. Static images cannot depict motion directly, but they can imply it through cues like frozen mid-air poses or dynamic scene composition. These categories allow the image search to find content that implies motion, serving as a useful initialization for regions that care about dynamic features. The video search then builds on this by animating the anchor image with explicit temporal dynamics.

Not all categories are directly motivated by known functional selectivity. Categories like Mood, Style, and RenderType serve as quality anchors that help the generative model produce coherent images, and as exploratory dimensions that give the search freedom to discover effective combinations beyond our prior expectations.

The video motion space $\mathcal{P}_{\text{vid}}$ covers temporal dimensions only, since the anchor image already provides visual content in I2V mode. Categories cover motion source and scale (MotionSourceType, MotionScale), speed and rhythm (MotionStrength, TemporalRhythm), the primary action (PrimaryAction), and how agents interact over time (Interaction, SocialContingency, EventStructure).

The image and video motion search spaces are summarized in Tables S6 and S7, with the complete listings provided in Tables S8 and S9.

In Figs. 3–4, we constrain the search space to categories relevant to each target ROI. Categories broadly aligned with the known selectivity of the ROI are kept active, while all other categories are fixed to their empty option. This reduces the effective search dimensionality and focuses the search budget on informative attributes. For the lateral-stream analysis in Fig. 5, we use the full unconstrained search space at all searchlight locations, allowing *NEvo* to discover effective feature combinations without imposing prior assumptions about regional preferences.

Constrained versus unconstrained search space. To test whether constraining the search space affects the final result, we reran *NEvo* for all ROIs in Fig. 3 using both the constrained and the full unconstrained search space (Table S2). The unconstrained search yields lower activation for FFA and slightly higher activation for MT, with no significant difference for EBA and pSTS. Overall, these results continue to support the strength of our approach.

Robustness to search-space design. To test whether the synthesized content depends on our specific choice of categories and options, we constructed an alternative search space by prompting a large language model (Claude) to make the original space denser and more expressive, while including diverse options. We then ran the image stage of *NEvo* on both spaces for four ROIs (PPA, FFA, V3, EBA).

To compare the resulting images quantitatively, we embedded all synthesized images in CLIP space and trained a classifier to predict the target ROI from the image embedding. The classifier was trained on images from one search space and tested on images from the other, in both directions (Table S3; chance level 0.25). Mean accuracy across ROIs is 0.808, with perfect transfer for PPA and the lowest accuracy for V3 (0.645). Misclassifications for V3 coincide with the occasional presence of body parts in the synthesized images, and those for EBA with differences in whether full bodies or body parts are depicted.

D Alternative search algorithms

To test whether the evolutionary strategy of *NEvo* is needed, we compare it against three alternative black-box optimizers over the same prompt space $\mathcal{P} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$. All methods maximize the same ROI score $S_r(G(p))$ under a shared budget of 200 evaluations per run.

Random search. Prompts are sampled uniformly from $\mathcal{P}$, and the highest-scoring one is returned. This serves as a null baseline that uses no information from previous evaluations.

Greedy coordinate search. Starting from the best of 8 random prompts, the search optimizes one attribute at a time: every option in $\mathcal{A}_m$ is evaluated with all other attributes held fixed, and the best option is kept. Attributes are visited in random order, for up to three passes over all attributes.

Hill climbing. A $(1 + \lambda)$ local search with $\lambda = 8$, equivalent to *NEvo* with a population size of 1 and no crossover. At each step, 8 neighbours of the current prompt are generated by resampling k randomly chosen attributes, with k decreasing from 3 to 1 over the run. The current prompt is replaced by the best neighbour only if that neighbour scores higher. After 8 consecutive steps without improvement, the search restarts from a new random prompt, while the best prompt found so far is retained.

Evaluation. Each method is run with 21 random seeds on each of 5 ROIs, using the full unconstrained search space (Suppl. C.2). The results are shown in Fig. 4E. *NEvo* reaches the highest final activation and significantly outperforms all three baselines. Hill climbing ranks second, followed by greedy coordinate search and random search.

E Automated visual-content annotation for synthesis analysis

To characterize the visual content of both the synthesized and naturalistic videos, we conducted an automated annotation analysis using a video understanding model (Gemini-2.5-Flash). For each video, the model was prompted to return a brief one-sentence summary of the visible content and to rate a set of perceptual and social properties on a standardized numerical scale.

Annotation prompt. The model was given the following structured instruction:

> *You are a video annotation assistant. Your task is to rate a short video clip on a set of perceptual and social properties. Return strict JSON only — no markdown fences, no prose outside the JSON object. The JSON must contain exactly two keys:* `summary` *(a ≤35-word description of the visible content) and* `properties` *(a JSON object with each property name as a key, each value containing a* `score` *and a ≤20-word* `evidence` *string grounded in specific visible cues). All properties use a 0–10 integer scale (0 = absent/minimal, 10 = extreme/maximal), except* `person_count_estimate` *which uses values 0–5 only. Social properties are set to 0 when fewer than two agents are visible. Each score should be based solely on visible content in the clip, rating the full clip rather than a single frame.*

Annotation properties. The full set of annotated visual properties is listed in Table S10, organized by functional category. Properties span low-level visual features (e.g., spatial complexity, texture), scene-level content (e.g., spatial expanse, navigation), motion (e.g., biological motion, camera motion), action (e.g., goal-directed action), human/body content (e.g., face and body presence), and social interaction (e.g., joint action, communication).

Visual property–activation correlation. For each target ROI, we sampled a set of naturalistic videos ($n = 100$) drawn uniformly across the full range of model-predicted activation (from lowest to highest) from the Moments in Time [61] training set (after excluding 11,971 videos with no visual content, e.g., entirely black frames; $n = 715{,}334$ usable videos), and synthesized a matched set of videos ($n = 32$) using NEvo. This procedure was applied to eight ROIs (V1, V3A, FFA, PPA, MT, EBA, pSTS, and aSTS), yielding a total of 1,052 test videos. All videos were annotated using the prompt described above. For each ROI, we computed the Pearson correlation between each annotated visual property score and the corresponding model-predicted activation, and report results for the representative properties (Fig. 5B) and top-correlated properties (Fig. S11) per ROI. Full results across all properties and ROIs are provided in Table S12.

Reliability of annotation. To assess the stability of Gemini-2.5-Flash annotations, each of the 1,052 test videos was annotated five times using the same prompt, with a temperature of 0.2 to promote deterministic outputs. Test-retest reliability was evaluated using the intraclass correlation coefficient ICC(C,1) (single-measure consistency formulation[47, 54]), quantifying annotation consistency across five repeated runs on the same videos. Results are reported in Table S11. Final annotation scores were averaged across the five repeated runs before computing property–activation correlations.

Word cloud generation. For each searchlight cluster, we annotated $n = 32$ synthesized videos using the same Gemini prompt to obtain one-sentence descriptions. Descriptions were tokenized, and standard English stopwords as well as domain-specific terms related to video generation style and generic scene descriptors were removed before computing word frequencies.

F Human behavioral annotation experiment

To verify that NEvo-synthesized videos exhibit prominent region-specific visual features, we conducted an online behavioral annotation experiment comparing NEvo videos against top-activating naturalistic videos from two existing fMRI datasets [51, 56].

Stimuli. For each target ROI (MT, EBA, FFA, pSTS), we selected 25 NEvo-synthesized videos and 25 top-activating naturalistic videos as controls. The naturalistic videos were drawn from two

existing fMRI datasets in proportion to their dataset sizes: 20 from BOLD Moments Dataset[51] and 5 from the social interaction dataset [56].

Task and rating scale. Participants rated each video on a 5-point Likert scale for the property corresponding to the target ROI:

- MT — motion extent: "How much overall visual motion is present across the video frame?" (1 = Completely static; 5 = Continuous fast motion)
- EBA — body presence: "To what extent are bodies or body parts visible and prominent in the video?" (1 = No body; 5 = Body/bodies/body parts fill the frame)
- FFA — face presence: "To what extent are faces visible and prominent in the video?" (1 = No face; 5 = Face(s) fill the frame)
- pSTS — joint action: "To what extent are the individuals engaged in a joint action together?" (1 = Entirely independent; 5 = Fully joint action)

Synthesized and naturalistic videos were intermixed and presented in randomized order within each ROI condition.

Participants and quality control. We recruited 133 participants online via Prolific. Participants were excluded if they showed inconsistent responses across repeated catch trials, responded too rapidly, or gave near-identical ratings across all trials. After these quality checks, 114 participants were retained for analysis. For each ROI, we computed participant-level mean ratings separately for NEvo and naturalistic videos, and applied two-sided paired t-tests to assess whether ratings differed between the two conditions.

G ROI construction

Functional ROIs. ROIs for EBA, FFA, PPA, MT, pSTS, and aSTS were defined in group-level *fsaverage5* surface space using individual subject localizer data from the social interaction dataset [56]. For each subject, binary ROI masks in T1w volumetric space were projected onto the *fsnative* cortical surface using `mri_vol2surf` with nearest-neighbor interpolation, and subsequently resampled to *fsaverage5* using `mri_surf2surf`. Both steps were performed using FreeSurfer 8.1.0 [21] separately for each hemisphere. Group-level masks were then constructed by majority voting: a vertex was included if at least one subject had it labeled as part of that ROI, maximizing spatial coverage while retaining functionally relevant vertices. Unless otherwise specified, ROI activation is computed across both hemispheres.

Early visual cortex ROIs. V1 and V3A are not covered by the functional localizers in the social interaction dataset. We therefore defined them using the corresponding labels from the multi-modal parcellation atlas of Glasser et al. [26], resampled to the *fsaverage5* surface.

Searchlight patches. To characterize functional selectivity along the lateral visual stream beyond predefined ROIs, we constructed a sequence of searchlight patches tracing a continuous trajectory through landmark regions on the *fsaverage5* cortical surface. For each hemisphere, we specified an ordered list of landmark ROIs (V1, MT, EBA, pSTS, and aSTS) and connected consecutive ROIs along surface geodesics, yielding a one-dimensional path through the lateral stream. Along this path, we placed searchlight centers [48] at a fixed stride of 5 mm, and defined each patch as the set of cortical vertices within a 10 mm geodesic radius of its center. This procedure yields 28 patches in the left hemisphere and 30 patches in the right hemisphere. Patches are indexed sequentially from posterior to anterior within each hemisphere; indices reported in Fig. 5A and Fig. S9 refer to this ordering.

H Comparison with BrainDiVE

Sensitivity to guidance strength in video synthesis. We swept the BrainDiVE guidance scale and measured response changes in the guiding VJEPA2-based encoder and the independent S3D-based

encoder. We summarize transfer using the descriptive ratio

$$T = \frac{\Delta r_{\mathrm{S3D}}}{\Delta r_{\mathrm{VJEPA2}}}.$$

Increasing guidance strength consistently increased the predicted response in the guiding encoder, but improvements in S3D peaked at a scale of 0.4 and subsequently declined (Table S4). We therefore selected 0.4 for BrainDiVE for the video comparison in Table S1.

Sensitivity to the image generator. We applied gradient-based guidance to both Stable Diffusion v1.5 and FLUX.2-klein-4B. Across the tested scales, transfer ratios were lower for FLUX.2-klein-4B (Table S5). In particular, stronger guidance continued to increase VJEPA2-predicted responses while S3D response gains declined, reaching approximately zero for FLUX.2-klein-4B at a scale of 0.8. This divergence is consistent with increasing optimization of encoder-specific features and highlights the sensitivity of the tested gradient-based pipeline to both guidance strength and generator choice.

I Controlled synthesis in non-naturalistic scenario

Synthesis methods. For the controlled synthesis, we use Gemini [30] to generate initial images depicting two abstract objects with matched backgrounds and color schemes, and then apply the second-stage image-conditioned video search to animate them. The search pipeline follows Suppl. C with two modifications. First, we use HunyuanVideo [46] as the generation model. Second, we add prompt constraints to ensure that the generated videos contain only the two objects, without extra objects or deformation into human-like figures.

Controlled synthesis results. Fig. S14 shows controlled-synthesis examples for pSTS and MT across multiple non-naturalistic image anchors, including plasticine stacks, deformable blobs, and rubber ducks. For each anchor, we run multiple seeds ($n = 384$) and display videos from the top and bottom 3% of predicted activation boost relative to the static anchor. Despite prompt constraints, pSTS strongly favors face-like and interactive content, whereas MT shows weaker sensitivity to such content and is driven more by strong motion.

J Reproducibility

For each two-stage *NEvo* run targeting a given ROI, we fix the random seed for both NumPy and PyTorch. Because text- and image-conditioned generation can remain partially stochastic even under fixed seeds, exact stimulus-level reproducibility is not always guaranteed: repeated runs may produce semantically similar but visually non-identical images or videos. Nevertheless, final ROI scores are expected to be stable across runs.

For each ROI, we run 32 independent seeds and aggregate the highest-scoring stimuli for analysis. A single seed is run on one A100 GPU with 80 GB memory, 16 CPU cores, and at most 64 GB CPU memory. Under the default hyperparameters in Suppl. C.1, one full two-stage run takes approximately 1200 s.

Our supplementary files include the prompt configurations, random seeds, and algorithm code used for all reported analyses.

Table S1: Held-out response percentiles (%) relative to 1,000 Moments in Time videos ($n = 32$ per ROI). Brackets indicate bootstrap 95% CIs (10,000 resamples).

Held-out encoder	EBA	pSTS	FFA	MT	PPA	V3A
S3D	98.7 [95.8–99.7]	97.1 [92.1–98.7]	96.3 [93.4–97.7]	95.0 [80.1–97.5]	90.1 [86.5–93.0]	78.7 [54.8–92.3]
InternVideo2	99.2 [96.2–100.0]	98.2 [93.1–99.3]	99.1 [97.4–99.2]	95.4 [83.6–97.2]	87.6 [84.3–90.2]	78.8 [54.5–85.5]

Table S2: Final best activation of *NEvo* with constrained and unconstrained search spaces. n: number of runs; $\Delta\%$: relative change from constrained to unconstrained, with 95% confidence interval.

ROI	n	Constrained	Unconstrained	$\Delta\%$	95% CI	p
FFA	25	0.4720	0.4007	-15.1	$[-21.2, -9.0]$	0.0002
EBA	20	0.5964	0.5594	-6.2	$[-16.0, +4.0]$	0.2943
pSTS	20	0.2452	0.2404	-2.0	$[-9.6, +5.2]$	0.8408
MT	19	0.6762	0.7067	$+4.5$	$[-0.1, +9.1]$	0.0494

Table S3: Cross-search-space classifier accuracy. A classifier trained on CLIP embeddings of images synthesized in one search space is tested on images from the other. Chance level is 0.25.

ROI	Original → alternative	Alternative → original	Mean
PPA	1.000	1.000	1.000
FFA	0.825	1.000	0.912
V3	0.696	0.594	0.645
EBA	0.442	1.000	0.721
Overall	0.714	0.902	0.808

Table S4: Guidance-scale sweep for the BrainDiVE video adaptation. Δr denotes the change in predicted FFA response. The selected scale, 0.4, maximizes the S3D response gain.

Guidance scale	$\Delta r_{\mathrm{VJEPA2}}$	Δr_{S3D}	Transfer ratio T
0.05	+0.649	+0.126	0.19
0.10	+1.024	+0.343	0.34
0.20	+1.422	+0.524	0.37
0.40	+1.694	**+0.588**	0.35
0.80	+1.907	+0.528	0.28
1.60	+1.993	+0.184	0.09

Table S5: Guidance-scale sweeps for gradient-based image synthesis with two diffusion generators. Transfer ratios are computed as $\Delta r_{\mathrm{S3D}}/\Delta r_{\mathrm{VJEPA2}}$.

Generator	Guidance scale	$\Delta r_{\mathrm{VJEPA2}}$	Δr_{S3D}	Transfer ratio T
Stable Diffusion v1.5	0.05	+0.410	+0.485	1.18
	0.10	+0.623	+0.579	0.93
	0.20	+0.733	+0.699	0.95
	0.40	+0.809	+0.497	0.61
	0.80	+0.880	+0.184	0.21
FLUX.2-klein-4B	0.05	+0.820	+0.620	0.76
	0.10	+1.028	+0.552	0.54
	0.20	+1.262	+0.279	0.22
	0.40	+1.341	+0.091	0.07
	0.80	+1.373	-0.009	-0.01

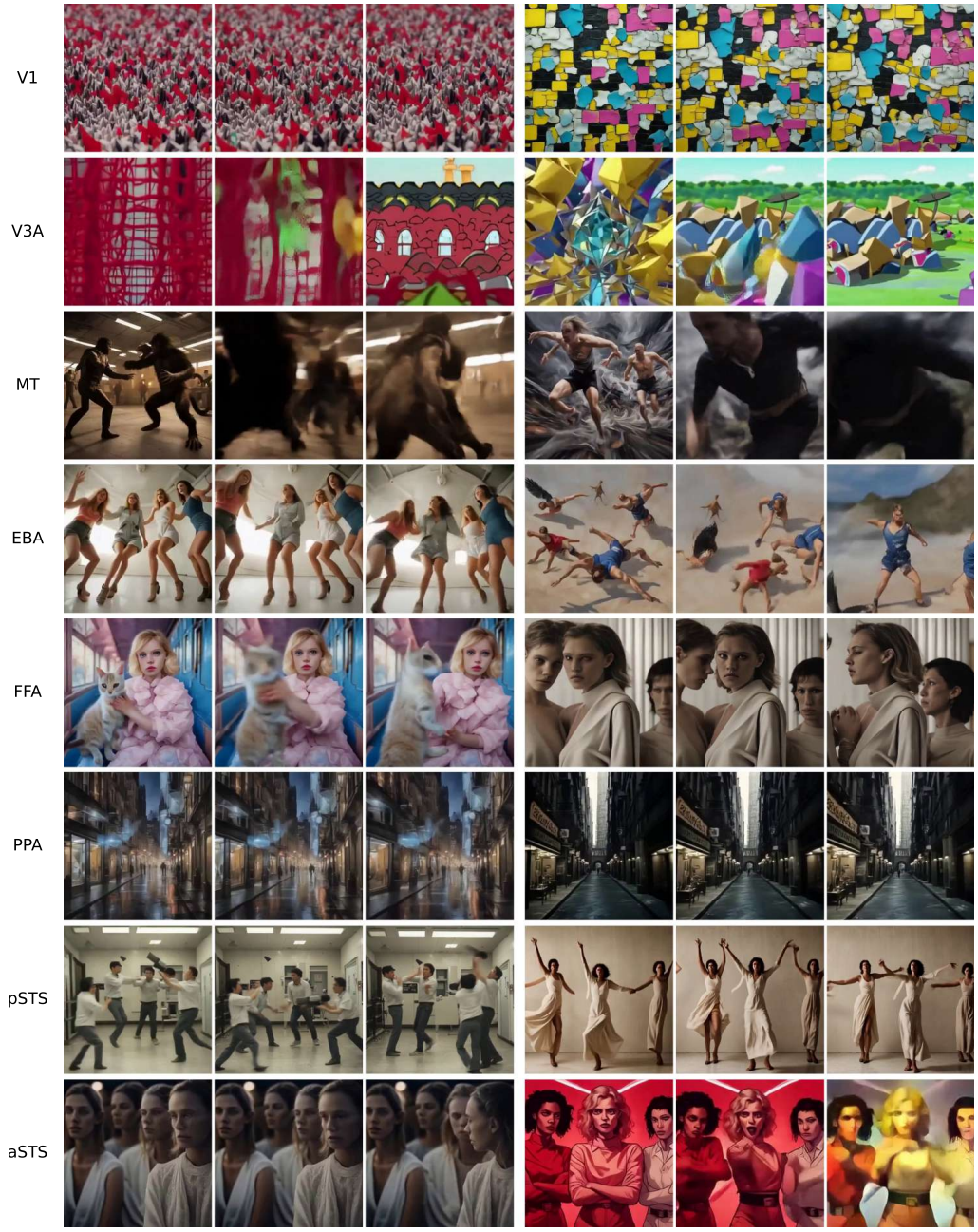

Figure S1: Additional examples of synthetic videos from NEvo. Two-second videos optimized to maximize predicted activation in different ROIs.

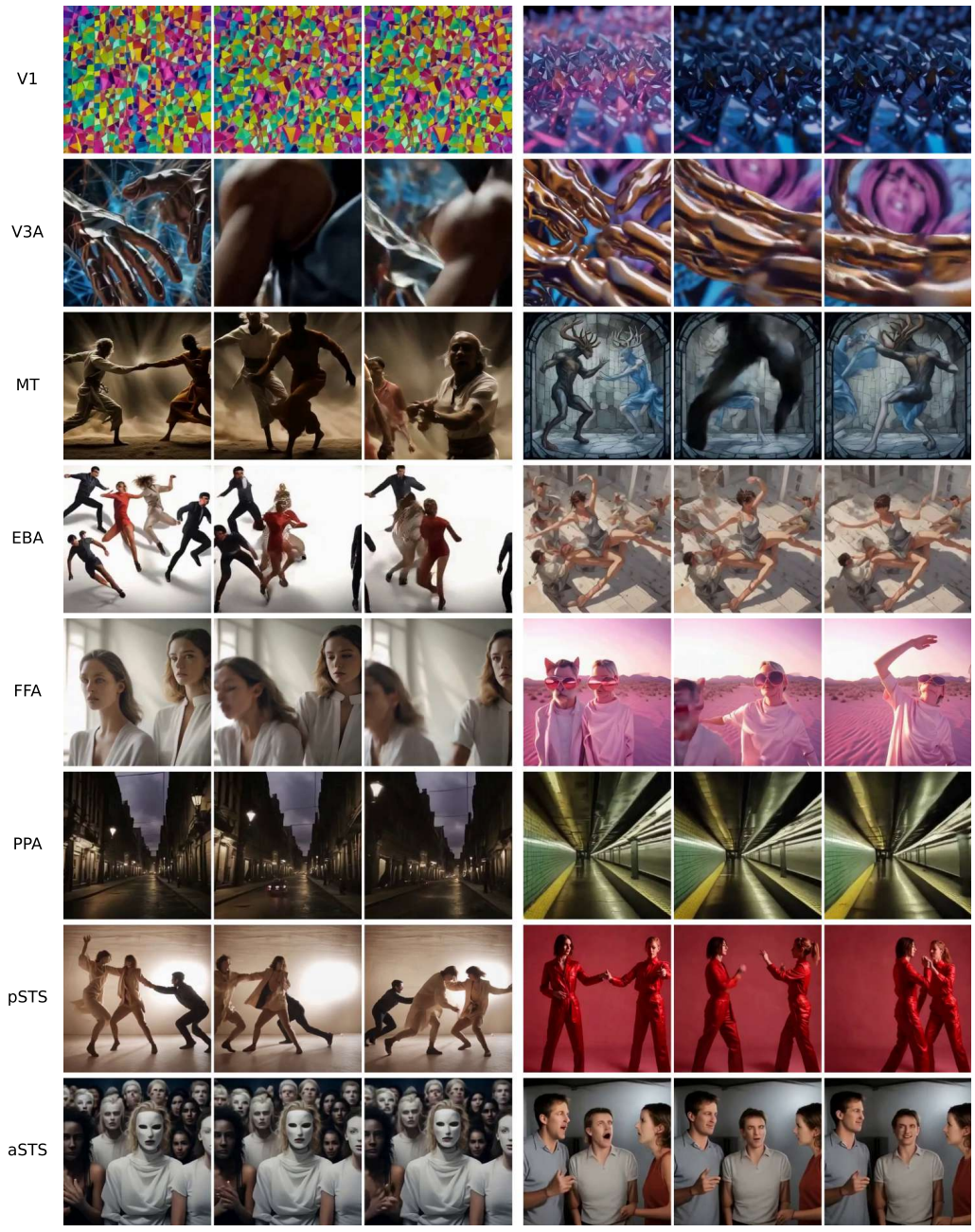

Figure S2: Additional examples of synthetic videos generated by NEvo, continuing Fig. S1.

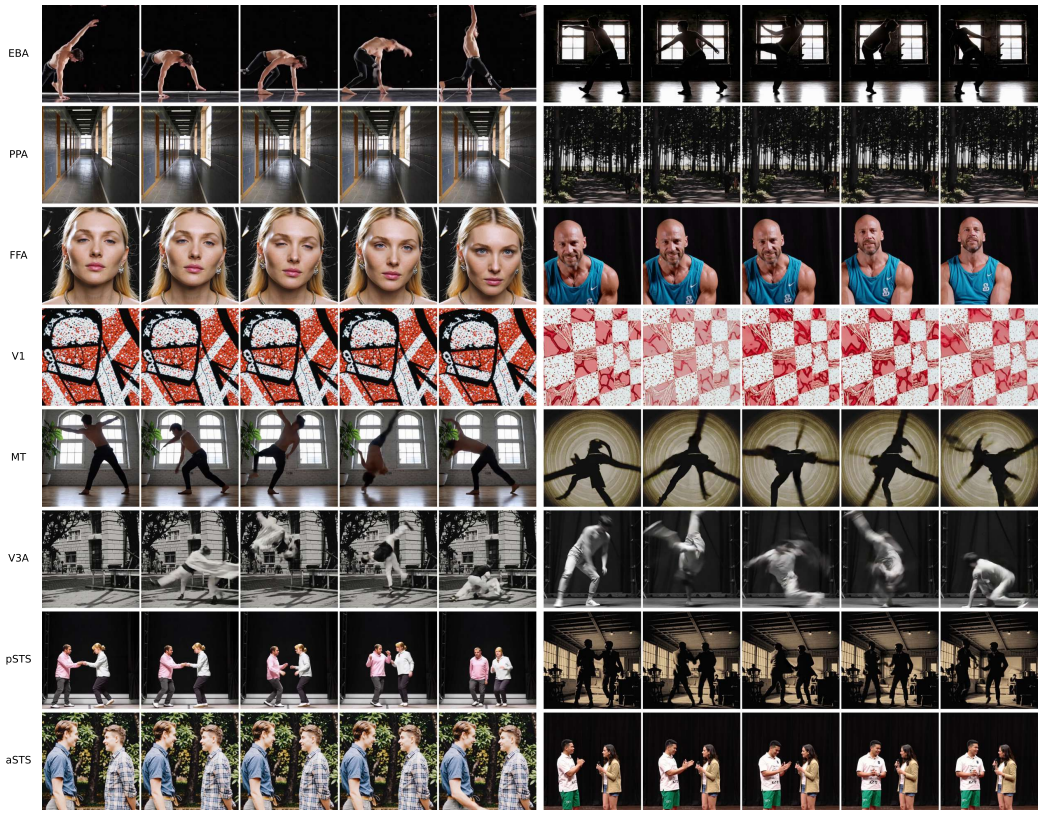

Figure S3: Additional examples of synthetic videos generated by NEvo with prompts reinforced for high quality. For better presentation, we show five sampled frames from each video.

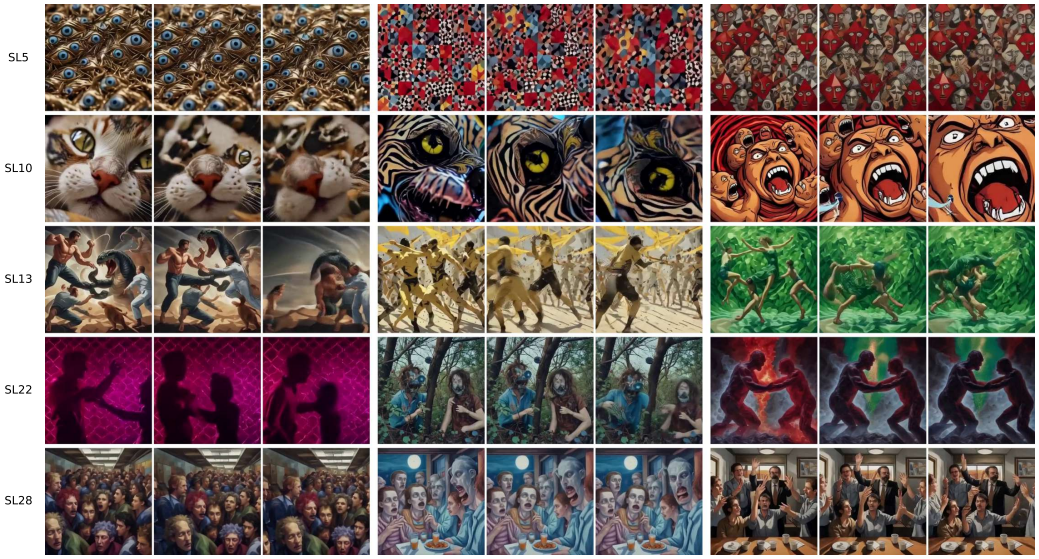

Figure S4: Synthetic stimuli from *NEvo* along the lateral visual stream, highlighting combinations of visual features that are unlikely to appear in natural video datasets. Each row corresponds to a searchlight location (see Fig. 5A), with three independent seeds shown per location. SL5 (SL stands for search light), near V1, produces dense repeating textures and pattern fields with high spatial frequency content. SL10, situated between V1 and EBA, generates high-contrast animal textures and close-up predator faces, reflecting a transition from low-level edge and texture processing toward body and object selectivity. SL13, near EBA and MT, consistently produces scenes of multiple figures engaged in vigorous physical interaction such as fighting and dancing with circular movements. SL22, overlapping pSTS, favors silhouetted dyadic interactions in dramatic lighting, emphasizing the social configuration over identity or texture. SL28, near aSTS, generates crowded social scenes with multiple expressive faces and group interactions. Across locations, the synthesized stimuli reflect known functional preferences of each region while combining features, such as abstract social configurations without identifiable faces, or pattern fields without natural objects, that are rare or absent in naturalistic video datasets.

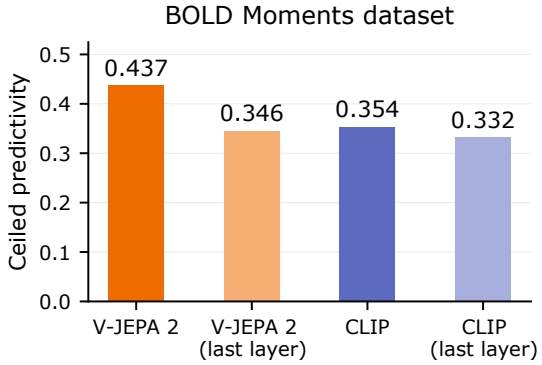

Figure S5: **Ablation of encoding models.** Our dynamic multi-layer encoding outperforms the commonly used CLIP last-layer baseline [13, 53] and its own last-layer control (Suppl. B.2).

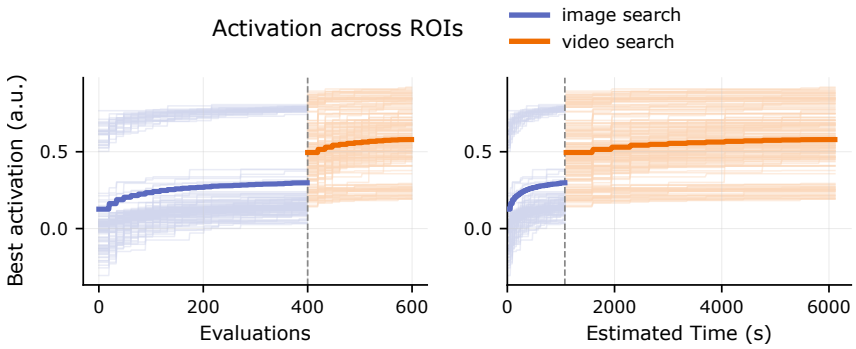

Figure S6: **A. Two-stage search over time.** Optimization trajectories across search evaluations and wall-clock time, aggregated across ROIs.

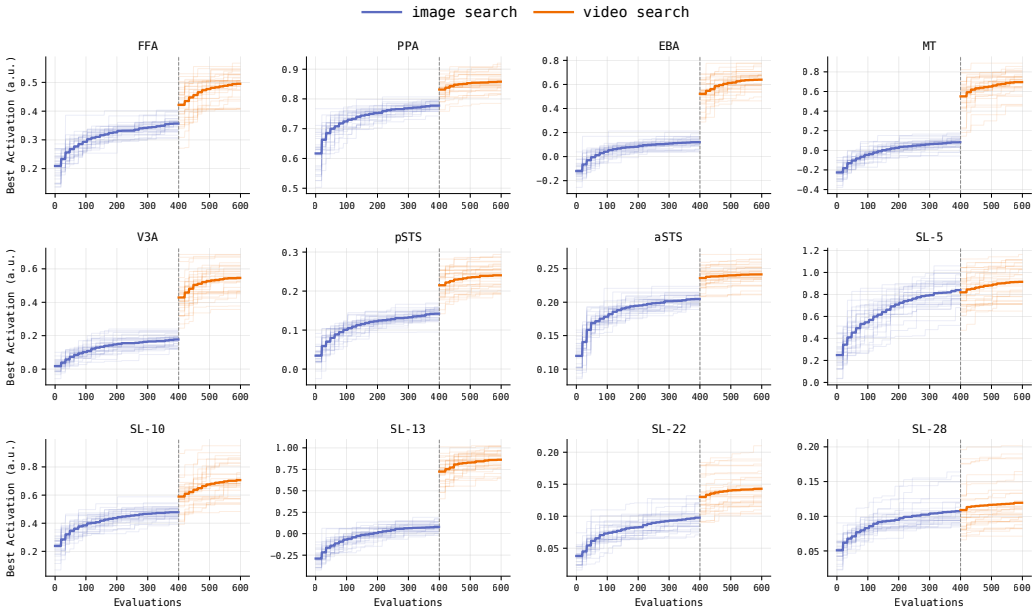

Figure S7: Two-stage search across iterations for additional ROIs.

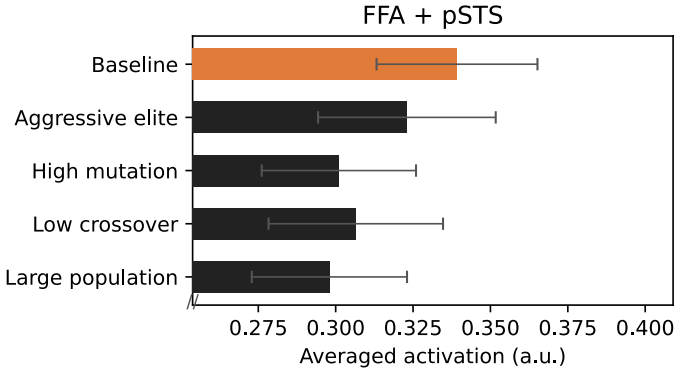

Figure S8: Hyperparameter ablation for the genetic search algorithm. Mean activation (a.u.) across FFA and pSTS, averaged over image and video search stages and 5 random seeds. Error bars show SEM. We compare five configurations: baseline (population size 20, mutation rate 0.2, crossover rate 0.5, elite fraction 0.3), high mutation (mutation rate 0.5), low crossover (crossover rate 0.2), large population (population size 50, elite fraction 0.2), and aggressive elite (mutation rate 0.1, crossover rate 0.6, elite fraction 0.5). Differences between configurations are not statistically significant across 5 seeds; we use the baseline throughout.

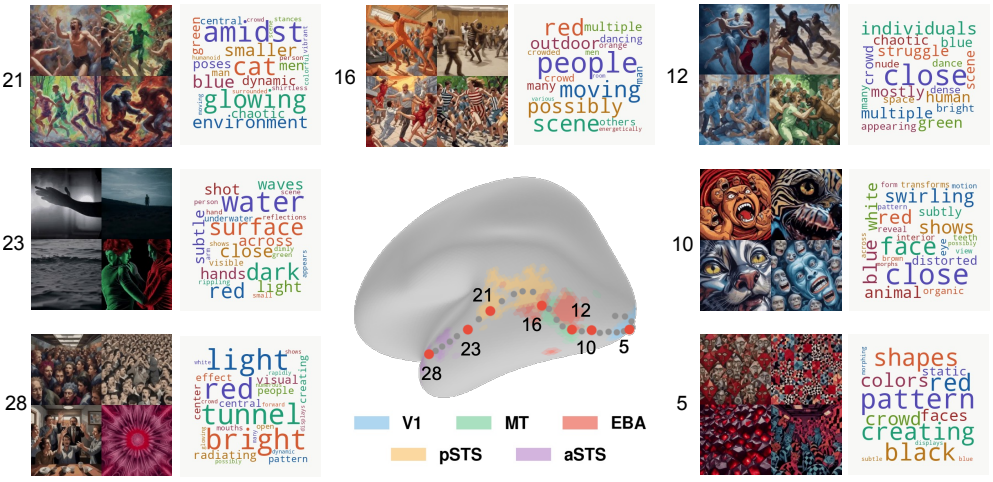

Figure S9: **Searchlight synthesis along the lateral visual stream in the left hemisphere.** Top synthesis stimuli (left) and word clouds from Gemini-generated descriptions (right) are shown for example searchlight clusters along the left-hemisphere lateral stream trajectory. Cluster indices are marked on the cortical surface. Feature selectivity follows a similar progression to the right hemisphere, transitioning from repeated colorful patterns in early visual cortex (clusters 5 and 10) to motion and body content in mid-level regions (clusters 12 and 16), and physical interaction content in mid-to-high-level regions (cluster 21). However, anterior STS clusters (clusters 23 and 28) show weaker social interaction selectivity compared to the right hemisphere, potentially reflecting hemispheric lateralization: the left STS is more strongly associated with language and auditory processing [22], which are absent from our model's training data, and these regions also exhibit lower noise ceilings in our dataset.

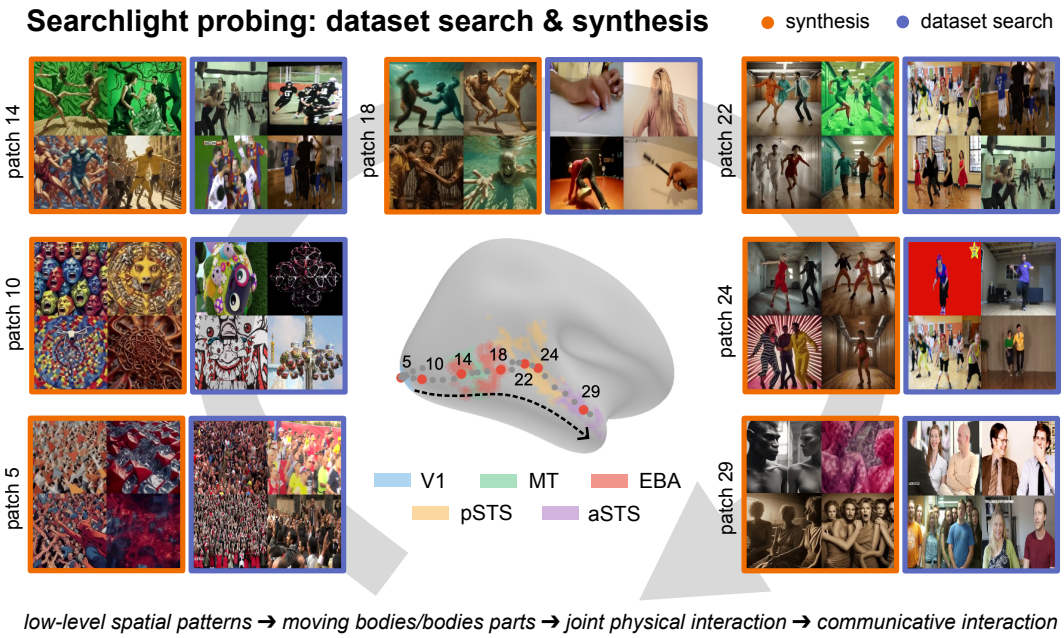

Figure S10: **Searchlight: synthesis vs. dataset-search stimuli.** Here, we present the same searchlight and synthesis results as in Fig. 5, with an additional comparison to the top stimuli identified through dataset search in Moments in Time. Across lateral-stream ROIs, the stimuli reveal a progression in the level and complexity of social features.

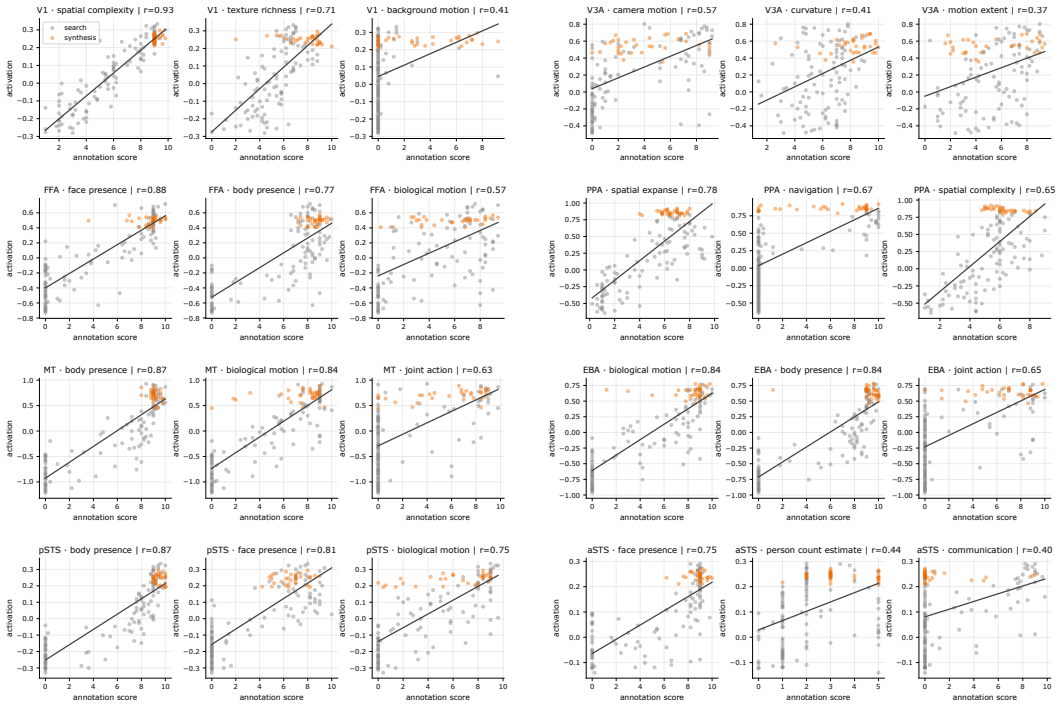

Figure S11: Correlation between model-predicted activation and annotated visual properties across all ROIs. For each ROI, the three visual properties with the highest Pearson correlation are shown. Gray points represent naturalistic videos sampled uniformly across the full range of predicted activation; orange points represent synthesis stimuli.

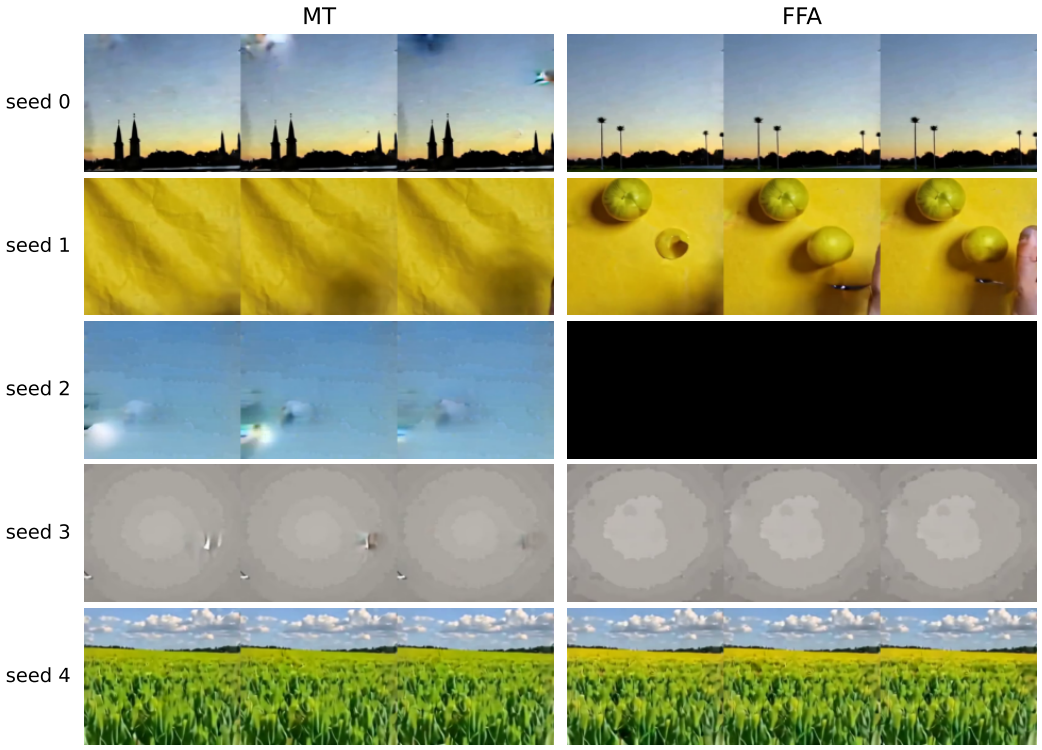

Figure S12: Additional examples of synthetic videos from BrainDiVE, using the same models specified in Suppl. B.

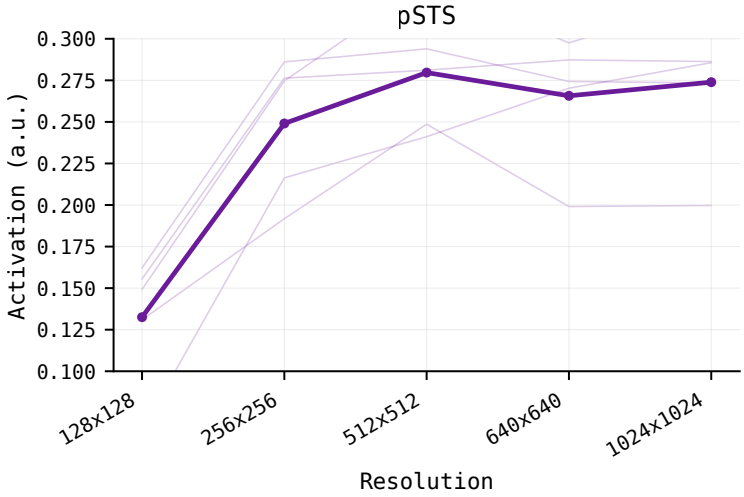

Figure S13: Best predicted activations obtained by running NEvo at different video generation resolutions, averaged over 5 seeds. Our default setting, described in Suppl. C.1, uses a resolution of 256×256.

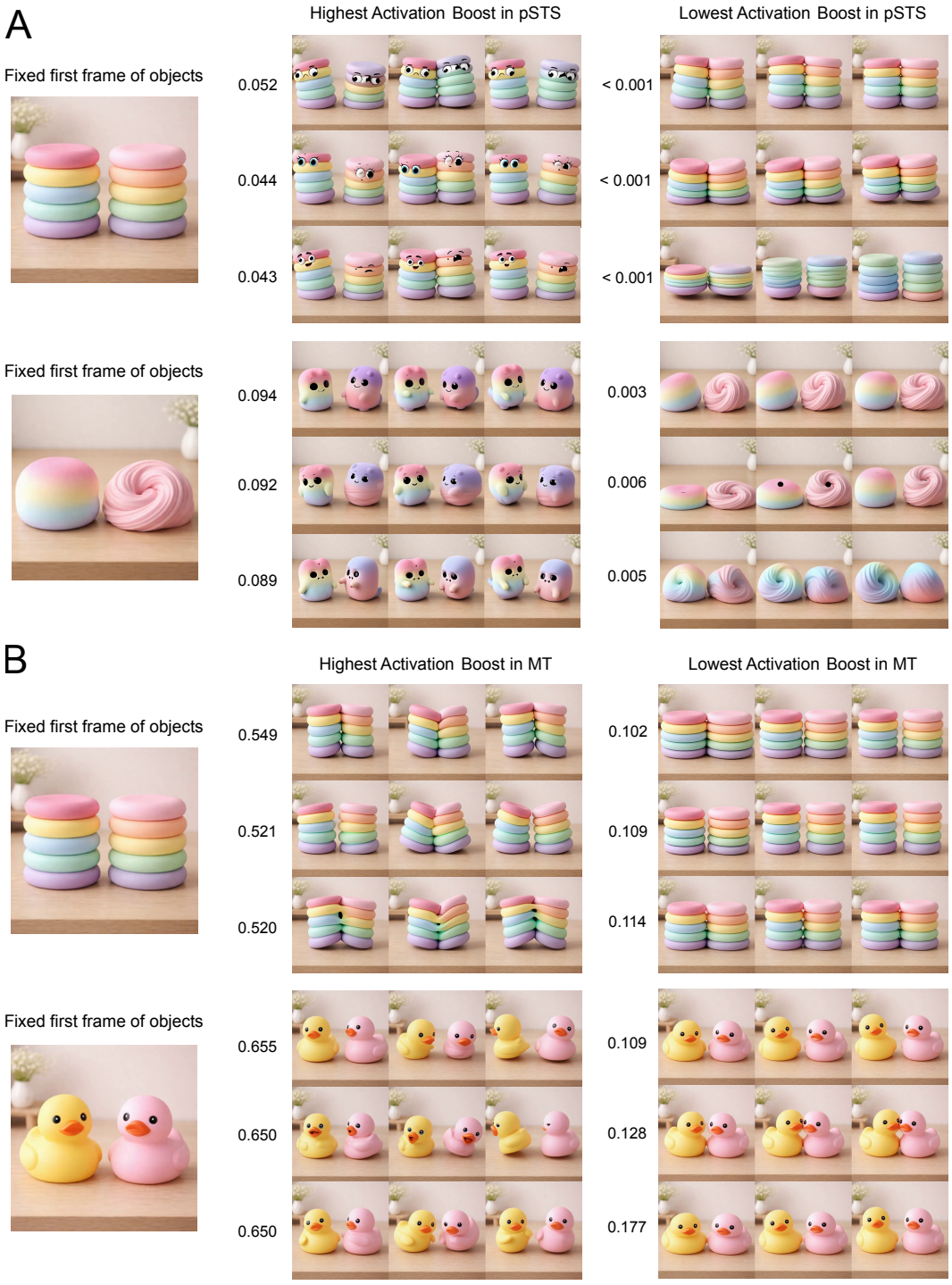

Figure S14: **Additional controlled-synthesis examples for pSTS and MT across multiple non-naturalistic anchors.** Extended results for the controlled synthesis experiment in Fig. 6, showing different objects: plasticine stacks, deformable blobs, and rubber ducks. For each anchor, we run for multiple seeds and show the three videos with the highest predicted activation boosts (middle columns) and the three with the lowest boosts (right columns). Values indicate predicted activation improvements (a.u.) over the static first-frame baseline. **A. pSTS optimization.** High-boost videos consistently introduce face-like features and coordinated, interactive movements between objects, regardless of the underlying anchor, whereas low-boost videos retain the static appearance with minimal dynamics or social structure. **B. MT optimization.** High-boost videos exhibit motion-rich dynamics such as translation, oscillation, and deformation of the objects, while low-boost videos show little coherent motion.

Table S6: Summary of the image prompt search space (614 options, 30 categories). Each category contributes one discrete gene to the prompt. Example options are representative samples from each category.

Group	Category	# Options	Example options
Rendering	RenderType	25	photorealistic, cinematic still, 35mm film photo, infrared photography
	Lighting	25	dramatic chiaroscuro, golden hour warm, neon light, cold moonlight
	Style	24	surrealist dreamscape, hyperrealism, gothic dark, cyberpunk noir
	Color	33	vivid saturated, dominant red, teal and orange, deep black shadows
Low-level	Pattern	25	high contrast checkerboard, radial starburst, fractal branching
	SpatialFrequency	17	high frequency texture, horizontal sine grating, random phase pattern
	EdgeDensity	10	dense sharp edges, sparse isolated edges, soft blurry edges
	Surface	25	wet glistening skin, cold brushed metal, translucent membrane
Face & Body	FaceContent	28	direct eye contact, baby face, angry face, multiple faces
	FaceViewpoint	11	direct eye contact, looking to side, unfocused distant gaze
	FaceOcclusion	13	full face visible, face behind mask, face in shadow
	BodyContent	17	full body front, hands close-up, headless body, body in fog
	BodyAction	25	standing still, reaching toward viewer, dancing, stumbling forward
	BodyPartFocus	10	full body, hands and arms, legs and gait, torso posture, back and spine
Social	AgentType	23	adult person, robot, moving triangle, silhouette person
	SocialCue	20	eye contact, pointing gesture, beckoning gesture, looming dominant
	SocialDynamics	21	one chasing other, working together, ignoring each other
	IntentionSignal	12	reaching for goal, helping another person, hiding something
	AgentNumber	5	one person, two people, three people, crowd of people
Scene	SceneCategory	29	city street, hospital room, abandoned building, snowy landscape
	SceneGeometry	18	deep perspective, narrow corridor, bird eye view, mirrored room
	SceneOccupancy	7	completely empty, densely packed, single object only
	ScenePhysics	20	frozen mid-air, swirling fluid vortex, gravity defying float
Object	ObjectForm	21	everyday object, floating object, tangled objects, melting object
	Texture	22	rough coarse surface, reptile scales, cracked ice, honeycomb cells
	HandObjectInteraction	29	hand throwing object, hand painting surface, hand kneading dough
Other	Subject	29	embracing couple, figure in doorway, crowd from above
	Framing	20	extreme close-up, fisheye distortion, symmetrical framing
	Mood	25	uncanny dread, primal fear, tender warmth, violent rage
	MotionSourceType	25	spinning wheel, rotating gear, flowing water, falling leaves
Total		**614**	30 categories

Table S7: Summary of the video motion prompt search space (163 options, 11 categories). Used in the I2V stage where an anchor image provides visual content and the search optimizes motion and interaction dynamics.

Category	**# Options**	**Example options**
MotionSourceType	25	person walking, animal moving, bouncing ball, exploding particles
MotionStrength	9	barely moving, moderate steady motion, very fast intense, building up speed
MotionScale	7	single body part, whole scene moving, background only moves
PrimaryAction	25	walking, dancing, oscillating, merging together, collapsing inward
Interaction	20	chasing other, working together, passing object, blocking the path
SocialContingency	8	no contingency, back and forth, moving in sync, call and response
EventStructure	12	chase sequence, cause and effect, buildup and release, parallel events
CameraMotion	11	static camera, tracking shot, zoom in, crane shot rising
TemporalRhythm	10	smooth continuous, sudden bursts, long pause then move
AgentNaturalness	12	realistic person, geometric shape, point light dots, shadow only
HandObjectInteraction	24	hand throwing and releasing, hand spinning object, hand kneading repeatedly
Total	**163**	11 categories

Table S8: Full image prompt search space.

Category	Options
RenderType	photorealistic, cinematic still, national geographic photo, extreme macro photo, drone top-down view, 35mm film photo, digital concept art, dark fantasy art, unreal engine render, double exposure photo, fashion editorial photo, infrared photography, underwater photography, medical illustration, security camera footage, vintage film photo, polaroid snapshot, long exposure photo, tilt shift photo, scanner photography, pinhole camera photo, thermal imaging photo, x-ray photograph, electron microscope image, night vision footage
Subject	portrait face, direct eye contact, smiling face, fearful wide eyes, angry bared teeth, embracing couple, mother and newborn, hostile crowd, two figures fighting, dancer mid-motion, sprinting athlete, hands reaching out, falling figure, body silhouette, predator close-up, snake striking, long dark corridor, spiral staircase, crashing wave, volcanic eruption, lone figure walking, crowd from above, child playing, elderly person sitting, figure in doorway, person in window, hands clasped together, figure underwater, person in fog
Lighting	dramatic chiaroscuro, golden hour warm, harsh midday sun, rim light edge, backlit silhouette, volumetric god rays, neon light, strobe freeze frame, softbox studio, single spotlight, cold moonlight, eerie underlighting, natural window light, dappled forest light, emergency red light, candlelight warm flicker, fluorescent office light, overcast flat light, firelight orange glow, dawn pink light, deep shadow low key, high key bright white, colored gel light, light through blinds, underwater caustic light
Mood	overwhelming awe, uncanny dread, eerie unease, euphoric joy, primal fear, serene calm, epic grandeur, deep melancholy, ominous threat, suspenseful tension, disgust revulsion, urgent danger, hypnotic trance, body horror, sensory overload, quiet loneliness, tender warmth, playful energy, cold detachment, obsessive fixation, nostalgic wistfulness, claustrophobic panic, dreamlike confusion, violent rage, peaceful serenity
Color	vivid saturated, muted desaturated, monochromatic, sepia warm, teal and orange, complementary contrast, deep black shadows, neon electric, earthy brown tones, soft pastel, rich jewel tones, warm orange tones, cool purple tones, high contrast split, faded washed out, overexposed bright, underexposed dark, duotone two colors, acid green neon, deep navy blue, black and white, red and green, blue and yellow, dominant red, dominant blue, dominant green, dominant yellow, dominant orange, dominant purple, pink and cyan, gold and black, red and black, white and silver
Style	surrealist dreamscape, high realism, cyberpunk noir, vaporwave aesthetic, Bauhaus geometric, baroque theatrical, psychedelic, hyperrealism, pop art, street art graffiti, impressionist painterly, minimalist clean, gothic dark, futurist dynamic, brutalist concrete, art nouveau organic, expressionist distorted, flat design, glitch art, ukiyo-e woodblock, retro sci-fi, industrial gritty, ethereal soft, noir black white
Framing	centered subject, extreme close-up, wide establishing shot, dutch angle tilt, vanishing point, bird's eye view, worm's eye view, first-person view, over the shoulder, reflection in water, fisheye distortion, rule of thirds, subject at edge, two-shot framing, silhouette framing, foreground blur, through frame shot, symmetrical framing, deep depth of field, shallow depth of field

Table S8: Full image prompt search space.

Category	Options
Pattern	tessellated hexagons, recursive mandala, circuit board traces, concentric ripple rings, fractal branching, sine wave interference, geometric grid, high contrast checkerboard, vertical bar grating, horizontal bar grating, diagonal stripes, random dot noise, op art illusion, spiral vortex pattern, radial starburst, honeycomb cells, Voronoi cell pattern, tangled line field, moiré interference, Truchet tile pattern, reaction diffusion spots, dendritic branching, web or net pattern, brick wall pattern, fish scale pattern
Surface	wet glistening skin, cold brushed metal, rough weathered stone, soft velvet fabric, cracked dry earth, sheer silk fabric, translucent membrane, heavy film grain, liquid chrome mercury, flowing molten lava, wispy smoke, subsurface skin scatter, polished white bone, rusted corroded metal, frosted ice surface, rough tree bark, smooth river pebble, worn leather texture, bubble foam surface, sandy desert texture, moss covered stone, shattered broken glass, peeling painted wall, woven basket texture, oily iridescent surface
ScenePhysics	completely frozen still, slow gentle drift, explosive sudden burst, slow motion, frozen mid-air, swirling fluid vortex, impact collision, falling descent, chaotic turbulence, rhythmic pulsing, time lapse, gravity defying float, rapid spinning, elastic bouncing, shockwave expanding, cascading collapse, spiral inward pull, zigzag erratic path, smooth gliding, violent shaking
FaceContent	close-up portrait, side profile face, three-quarter face, upside-down face, eyes closed face, direct eye contact, looking away, baby face, elderly face, dog face, cat face, monkey face, cartoon face, mannequin face, neutral expression, angry face, fearful face, smiling face, surprised face, crying face, multiple faces, no face, face in mirror, face in crowd, blurred face, painted face, face with glasses, face half shadow
FaceViewpoint	direct eye contact, slightly looking away, looking to side, looking down, looking up, eyes closed, wide open eyes, squinting eyes, gaze at object, eyes darting, unfocused distant gaze
FaceOcclusion	full face visible, half face occluded, face behind mask, face in shadow, face behind glass, digital avatar face, face out of focus, face at distance, face behind bars, face under water, face behind smoke, face with sunglasses, face behind hair
BodyContent	full body front, full body side, full body from behind, torso and arms, legs and feet, hands close-up, feet close-up, body outline only, headless body, body in action, resting body, body in shadow, body underwater, body in fog, body from above, body curled up, body stretched out
BodyAction	standing still, reaching toward viewer, reaching for object, walking, running, jumping, falling, crouching, punching or kicking, hugging, pointing, waving, throwing, pushing, pulling, dancing, crawling, sitting, lying down, climbing, spinning around, stumbling forward, recoiling backward, kneeling down, stretching out
BodyPartFocus	full body, hands and arms, legs and gait, torso posture, shoulders and neck, feet and ankles, fingers close-up, back and spine, hip and waist, single hand
AgentType	adult person, baby or toddler, chimpanzee or monkey, dog, cat, bird, moving triangle, moving circle, moving square, stick figure, robot, hand only, masked person, silhouette person, spider, insect, horse, deer, fish, shark, snake, bear, wolf

Table S8: Full image prompt search space.

Category	Options
SocialCue	eye contact, joint attention, pointing gesture, reaching toward person, turning toward person, turning away, mirroring posture, open arms welcome, crossed arms closed, protective gesture, beckoning gesture, pushing away, bowing submissive, looming dominant, shrugging gesture, nodding agreement, shaking head no, covering face, hands on hips, leaning in close
SocialDynamics	working together, competing against each other, one person dominant, threatening aggressive, playing together, parent and child, one chasing other, moving in sync, ignoring each other, hugging affectionate, fighting, teaching showing, imitating copying, greeting handshake, walking away, alone single person, comforting each other, negotiating bargaining, celebrating together, mourning together, strangers passing
IntentionSignal	reaching for goal, ambiguous unclear action, trying and failing, accidental slip, helping another person, blocking another person, asking or requesting, offering or giving, hiding something, searching for something, warning someone, inviting someone
AgentNumber	one person, two people, three people, small group, crowd of people
SceneCategory	living room, outdoor landscape, city street, forest, beach, kitchen, bedroom, office, hallway corridor, rooftop, cave, mountain, desert, underwater, parking lot, subway station, playground, hospital room, church or temple, stadium or arena, train or bus, elevator interior, stairwell, empty warehouse, abandoned building, snowy landscape, swamp or marsh, library interior, market or bazaar
SceneGeometry	deep perspective, flat surface view, low ceiling room, wide open panorama, bird eye view, looking up view, cluttered room, empty sparse room, narrow corridor, doorway framing, window framing, symmetrical room, staircase, curved arching space, mirrored room, split level space, open field horizon, tight enclosed corner
SceneOccupancy	completely empty, single object only, sparsely populated, moderately occupied, densely packed, one person only, crowded no space
ObjectForm	everyday object, abstract geometric shape, organic natural object, tool or instrument, household item, food item, vehicle, furniture, clothing item, plant or flower, transparent glass object, reflective shiny object, broken damaged object, single isolated object, multiple objects, oversized giant object, tiny miniature object, floating object, melting object, tangled objects, stacked objects
Texture	fine detailed texture, rough coarse surface, smooth glossy surface, human skin, flat solid color, fur, fabric, wood grain, stone surface, brushed metal, water ripples, sand, tree bark, reptile scales, feathers, velvet, bubble wrap, honeycomb cells, tangled hair, crumpled paper, cracked ice, woven fiber
SpatialFrequency	low frequency blobs, mid frequency edges, high frequency texture, white noise field, pink noise field, horizontal sine grating, vertical sine grating, diagonal sine grating, checkerboard pattern, random phase pattern, coarse grid pattern, fine dot pattern, blurred soft field, sharp crisp edges, mixed frequency noise, radial frequency pattern, concentric ring grating
EdgeDensity	sparse isolated edges, moderate edge density, dense sharp edges, edge-free flat field, soft blurry edges, high contrast edges, broken fragmented edges, curved smooth edges, jagged rough edges, layered overlapping edges

Table S8: Full image prompt search space.

Category	Options
MotionSourceType	spinning wheel, rotating gear, flowing water, falling leaves, rising smoke, bouncing ball, swinging pendulum, flapping bird wings, waving flag, drifting clouds, rolling vehicle, exploding particles, hand movement, fish swimming, rotating fan blades, falling snow, rolling waves, flickering flame, blowing grass, spinning top, dripping water, crawling insect, slithering snake, galloping horse, tumbling rocks
HandObjectInteraction	hand gripping tool, hand pouring liquid, hand writing or drawing, hand typing on keyboard, hand throwing object, hand catching object, hand cutting or slicing, hand molding clay, hand playing instrument, hand turning knob, hand pulling rope, hand pushing button, hand stacking blocks, hand tearing paper, hand folding fabric, hand stirring bowl, hand picking up object, hand placing object down, hand opening container, hand closing container, hand threading needle, hand striking match, hand squeezing object, hand painting surface, hand swiping screen, hand flipping page, hand kneading dough, hand snapping fingers, hand holding nothing

Table S9: Full video motion search space.

Category	Options
MotionSourceType	person walking, person running, animal moving, spinning wheel, rotating gear, flowing water, falling leaves, rising smoke, bouncing ball, swinging pendulum, flapping bird wings, waving flag, drifting clouds, rolling vehicle, exploding particles, falling snow, rolling ocean waves, flickering flame, blowing grass field, spinning top, dripping water, crawling insect, galloping horse, tumbling rocks, rotating windmill
MotionStrength	barely moving, very subtle motion, gentle slow motion, moderate steady motion, strong fast motion, very fast intense, intermittent bursts, building up speed, slowing to stop
MotionScale	single body part, one person moving, two entities moving, whole scene moving, camera moving only, background only moves, foreground only moves
PrimaryAction	walking, running, jumping, dancing, fighting, reaching, grasping, rotating, bouncing, sliding, flowing, scattering, falling, rising, spinning, crawling, swaying, exploding outward, collapsing inward, stretching, shrinking, splitting apart, merging together, flickering, oscillating
Interaction	no interaction, mutual attention, looking at object, copying each other, working together, competing against, caring for other, threatening other, chasing other, fleeing from other, touching each other, passing object, approaching together, moving apart, circling each other, colliding, pulling apart, pushing together, following behind, blocking the path
SocialContingency	no contingency, weak contingency, clear response, back and forth, taking turns, moving in sync, mirroring movement, call and response
EventStructure	continuous process, single event, repeating cycle, approach then contact, chase sequence, cause and effect, escalating sequence, two things at once, transformation sequence, buildup and release, start stop start, parallel events
CameraMotion	static camera, slow pan left, slow pan right, tracking shot, zoom in, zoom out, orbiting around, handheld shaky, slow tilt up, slow tilt down, crane shot rising

Table S9: Full video motion search space.

Category	Options
TemporalRhythm	smooth continuous, sudden bursts, rhythmic periodic, chaotic irregular, slow then fast, fast then slow, pulsing in beat, long pause then move, constant acceleration, constant deceleration
AgentNaturalness	realistic person, cartoon character, geometric shape, robot or android, animal agent, stick figure, shadow only, silhouette only, point light dots, puppet or marionette, masked figure, faceless figure
HandObjectInteraction	hand gripping and lifting, hand pouring liquid, hand throwing and releasing, hand catching object, hand cutting or slicing, hand molding or shaping, hand playing instrument, hand turning or twisting, hand pulling toward body, hand pushing away, hand stacking or building, hand tearing apart, hand stirring or mixing, hand picking up, hand placing down carefully, hand opening or closing, hand squeezing and releasing, hand painting or brushing, hand swiping across, hand kneading repeatedly, hand snapping fingers, hand rolling object, hand spinning object, no hand interaction

Table S10: Visual properties annotated by Gemini-2.5-Flash, organized by functional category. All properties use a 0–10 integer scale unless otherwise noted.

Category	Property	Definition
Low-level	spatial_complexity	Degree of visual clutter or density of fine-grained patterns in the spatial layout (0 = nearly empty; 10 = densely packed).
	color_richness	Diversity and saturation of colors present (0 = monochromatic/desaturated; 10 = wide variety of vivid colors).
	texture_richness	Richness and variation of surface textures (0 = smooth/plain; 10 = highly detailed and varied).
	curvature	Prevalence of curved versus angular shapes and contours (0 = all straight/angular; 10 = predominantly rounded/curved).
Scene	spatial_expanse	Spatial scale of the depicted environment (0 = extremely confined/close-up; 10 = wide open landscape).
	navigation	Degree to which the video conveys first-person movement through space (0 = none; 10 = strong traversal cue).
Motion	motion_extent	Overall magnitude and spatial scale of visual motion in the frame (0 = completely static; 10 = large-scale continuous motion).
	biological_motion	Presence of fluid, coordinated whole-body movement (0 = absent/stiff; 10 = dynamic coordinated motion).
	camera_motion	Degree of camera movement during the clip (0 = static; 10 = rapid large-scale camera motion).
	background_motion	Motion of the scene background independent of foreground subjects (0 = static; 10 = continuous large-scale background motion).
Action	object_directedness	Extent to which a person performs sustained actions on a physical object (0 = none; 10 = clear object-directed action).
	goal_directed_action	Extent to which an action is directed toward a specific functional goal with a clear outcome (0 = none; 10 = clear goal-directed action).
Human / Body	face_presence	Visibility and prominence of human faces in the frame (0 = no face; 10 = face(s) occupying large portion of frame).
	body_presence	Visibility and prominence of human bodies or body parts (0 = none; 10 = body/bodies occupying large portion of frame).
	person_count_estimate	Number of distinct individuals present (integer: 0 = none, 1–4 = exact count, 5 = five or more).
Social	agent_distance	Physical proximity between individuals (0 = very close; 10 = spatially distant; 0 if <2 agents).
	facingness	Extent to which individuals face each other (0 = back-to-back; 10 = directly face-to-face; 0 if <2 agents).
	physical_contact	Degree of physical touch between individuals (0 = no contact; 10 = full-body contact; 0 if <2 agents).
	joint_action	Extent of engagement in a joint action together (0 = entirely independent; 10 = fully joint action; 0 if <2 agents).
	synchronized_movement	Degree of temporal synchrony between individuals' movements (0 = asynchronous; 10 = precisely synchronized; 0 if <2 agents).
	communication	Extent of active communication between individuals via speech, gesture, or eye contact (0 = none; 10 = salient; 0 if <2 agents).
Other	text_presence	Visual prominence of text or written characters in the frame (0 = none; 10 = text is the dominant visual element).

Table S11: Test-retest reliability of Gemini-2.5-Flash annotations across five repeated runs. Reliability was evaluated using the ICC(C,1) single-measure consistency formulation [54], with interpretation thresholds of < 0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, and > 0.90 excellent [47].

Category	Property	ICC(C,1)	95% CI	Interpretation
Low-level	spatial_complexity	0.935	[0.929, 0.941]	Excellent
	color_richness	0.938	[0.932, 0.943]	Excellent
	texture_richness	0.876	[0.865, 0.887]	Good
	curvature	0.881	[0.870, 0.891]	Good
Scene	spatial_expanse	0.957	[0.953, 0.961]	Excellent
	navigation	0.910	[0.902, 0.918]	Excellent
Motion	motion_extent	0.895	[0.886, 0.904]	Good
	biological_motion	0.943	[0.938, 0.948]	Excellent
	camera_motion	0.887	[0.877, 0.896]	Good
	background_motion	0.767	[0.748, 0.784]	Good
Action	object_directedness	0.945	[0.939, 0.949]	Excellent
	goal_directed_action	0.914	[0.906, 0.921]	Excellent
Human / Body	face_presence	0.975	[0.973, 0.977]	Excellent
	body_presence	0.976	[0.974, 0.978]	Excellent
	person_count_estimate	0.978	[0.975, 0.980]	Excellent
Social	agent_distance	0.892	[0.882, 0.901]	Good
	facingness	0.939	[0.933, 0.944]	Excellent
	physical_contact	0.938	[0.932, 0.943]	Excellent
	joint_action	0.940	[0.934, 0.945]	Excellent
	synchronized_movement	0.913	[0.905, 0.921]	Excellent
	communication	0.927	[0.920, 0.933]	Excellent
Other	text_presence	0.931	[0.924, 0.937]	Excellent

Table S12: Pearson correlation between annotated visual property scores and model-predicted activation for each ROI. Values are computed over $n = 100$ naturalistic videos sampled uniformly across the predicted activation range, combined with $n = 32$ NEvo-synthesized videos per ROI.

Property	V1	V3A	FFA	PPA	MT	EBA	pSTS	aSTS
face_presence	−0.047	−0.331	0.881	−0.699	0.413	0.599	0.815	0.755
body_presence	−0.295	−0.226	0.770	−0.703	0.873	0.837	0.869	0.312
biological_motion	−0.329	−0.148	0.570	−0.490	0.837	0.843	0.747	−0.599
spatial_complexity	0.930	0.317	−0.434	0.651	−0.129	−0.431	−0.545	−0.225
spatial_expanse	0.039	−0.434	−0.674	0.780	−0.217	−0.411	−0.286	−0.261
joint_action	0.219	−0.088	0.342	−0.244	0.634	0.646	0.732	0.070
facingness	0.104	−0.132	0.438	−0.231	0.625	0.640	0.688	0.336
person_count_estimate	0.209	−0.190	0.420	−0.049	0.553	0.579	0.658	0.438
curvature	0.309	0.405	0.343	−0.730	0.174	0.253	0.168	−0.142
texture_richness	0.713	0.260	−0.357	0.162	−0.108	−0.160	−0.512	−0.181
synchronized_movement	0.279	−0.056	0.341	−0.121	0.498	0.538	0.590	−0.196
physical_contact	0.117	−0.052	0.291	−0.304	0.584	0.522	0.486	−0.008
motion_extent	−0.114	0.372	0.034	0.210	0.511	0.570	0.296	−0.800
communication	0.111	−0.152	0.446	−0.233	0.121	0.269	0.405	0.401
camera_motion	−0.219	0.573	0.065	0.599	0.346	0.183	−0.254	−0.469
background_motion	0.406	0.272	−0.225	0.541	−0.038	−0.228	−0.372	−0.250
object_directedness	−0.401	−0.098	−0.034	−0.418	0.105	−0.092	−0.105	−0.605
goal_directed_action	−0.350	−0.138	−0.097	−0.378	0.381	0.203	−0.015	−0.544
agent_distance	0.095	−0.292	0.037	0.212	0.159	0.291	0.386	0.382
navigation	0.041	0.189	−0.263	0.669	0.091	−0.146	−0.269	−0.225
color_richness	−0.013	0.127	0.166	−0.289	0.243	0.099	0.294	−0.192
spatial_expanse	0.039	−0.434	−0.674	0.780	−0.217	−0.411	−0.286	−0.261
text_presence	−0.101	−0.243	−0.154	−0.090	−0.269	−0.481	−0.345	0.025

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims present *NEvo* as a model-guided framework for dynamic stimulus synthesis, highlight its value for studying dynamic visual processing, and frame its outputs as testable predictions for future work.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The discussion section clearly identifies the limitations and potential impacts of *NEvo*.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper clearly describes the proposed algorithm and the models used. A dedicated supplementary section addresses reproducibility, and the paper is accompanied by code and additional video examples.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: The code and additional data are provided. The main text and supplement describe the proposed algorithm and implementation details.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details are provided in the supplement.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Definitions are provided for all error bars, and statistical significance is reported for the relevant comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details are provided in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The discussion section clearly describes both potential positive and negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code, data, models used in the paper are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper provides code and video examples, with accompanying documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [Yes]

    Justification: The details are given in the supplementary sections.

    Guidelines:

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

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The online behavioral experiment (Suppl. F) was conducted on Prolific under a protocol approved by the Johns Hopkins University Homewood Institutional Review Board (HIRB00009730). Participants only viewed short video clips and provided ratings, and no risks beyond those of everyday computer use were anticipated. All participants gave informed consent electronically before participating.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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