RECEPTIVE-FIELD-CONSTRAINED STIMULUS OPTIMIZATION FOR HUMAN EARLY AND INTERMEDIATE VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

An ongoing challenge in sensory neuroscience is to characterize the feature dimensions encoded by cortical populations. Recent approaches probe feature selectivity in a data-driven way, by synthesizing a most-exciting-input (MEI) for a target neural population. While this approach has been successfully applied to human higher visual cortex using fMRI data, generating MEIs for early- and mid-level retinotopic visual areas requires additional modeling constraints due to small receptive field sizes. To address this challenge, we introduce two novel MEI generation frameworks, Receptive Field Diffusion for Visual Exploration (**RF-DiVE**) and Receptive Field Gradient Optimization (**RF-GO**). Both methods use a population receptive field (pRF)-constrained voxelwise encoding model; RF-DiVE combines this with a pretrained latent diffusion model, while RF-GO uses regularized gradient ascent. When applied to single voxels in retinotopically defined areas V1-hV4, using data from the Natural Scenes Dataset, we obtain MEIs that exhibit consistent structure within the pRF, suggesting selectivity for local features like contour, color, and texture. We systematically compare MEIs generated by RF-DiVE and RF-GO using two encoding backbones, performing *in-silico* validation of predicted responses to MEIs using independent encoding models. Across all methods and all visual areas, MEIs elicit higher model-predicted responses than the most activating natural images. We further find that the choice of generation framework and encoding backbone differentially affects MEI properties, including their visual appearance, structural interpretability, and cross-model generalizability. These results offer a new approach for performing data-driven characterization of spatial and feature selectivity across human visual cortex.

1 INTRODUCTION

Characterizing feature selectivity in the visual cortex has long been a central challenge in sensory neuroscience. Traditional approaches measure neural selectivity using small sets of hand-selected stimuli (e.g., oriented gratings, isolated images of faces or objects), which can bias results toward established hypotheses and fail to capture more novel aspects of selectivity. This issue may be especially pronounced when investigating intermediate stages of the visual system (e.g., V4), where the space of mid-level features encoded is high-dimensional, nonlinear, and unlikely to be fully described by interpretable dimensions. To overcome these challenges, recent approaches use stimulus optimization to probe neural feature selectivity in a more data-driven way. Stimulus optimization frameworks use computational encoding models to generate a most-exciting-input (MEI) for a target neural population measured with either invasive single-unit or fMRI recordings, using a forward encoding model combined with either gradient ascent or a generative model (Bashivan et al., 2019; Ponce et al., 2019; Walker et al., 2019; Cowley et al., 2026; Gu et al., 2022; Ratan Murty et al., 2021). A recent method (BrainDiVE) uses diffusion models (Rombach et al., 2022) to generate fMRI-based MEIs with more realistic semantic information compared to previous approaches (Henderson et al., 2026; Hwang et al., 2026; Luo et al., 2023). This diffusion-based approach has thus far been applied to regions of higher visual cortex, but has not yet been applied to early and intermediate visual areas.

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

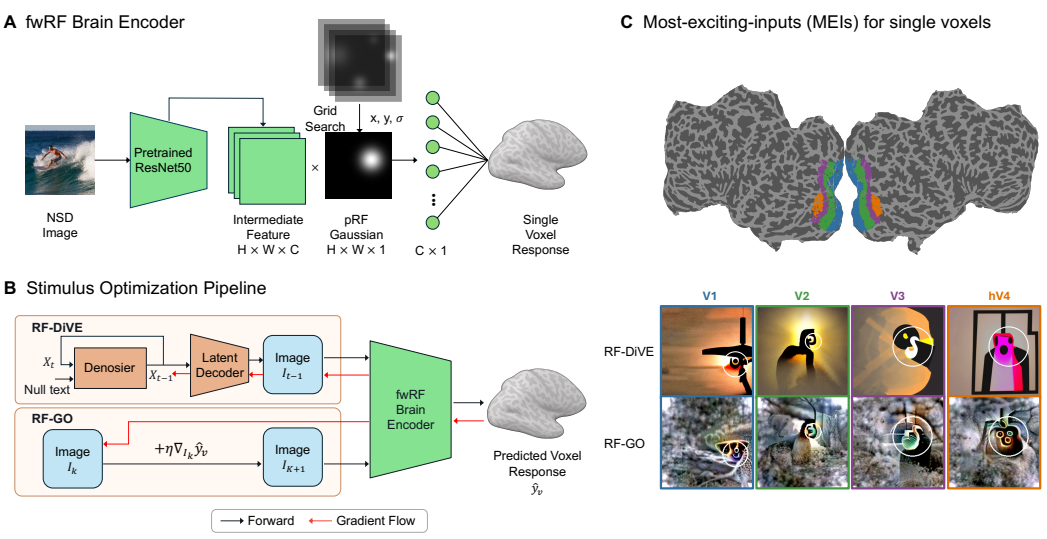

Figure 1: **Our stimulus optimization pipeline.** **(A)** Our encoding model predicts activity in early visual cortex using a deep CNN backbone, constrained by a Gaussian that models voxel population receptive field (pRF). **(B)** Visualization of our **RF-DiVE** and **RF-GO** frameworks. In RF-DiVE, the optimization process is constrained by a diffusion model, while RF-GO optimizes the image using MACO. **(C)** Examples of most-exciting-images (MEIs) for individual voxels in V1, V2, V3, and hV4. Each column corresponds to one visual area; white circle labels the voxel's pRF

Extending this approach to early and intermediate visual cortex requires new modeling strategies. In particular, early and intermediate visual responses are strongly shaped by retinotopic organization and small population receptive fields (pRFs; Dumoulin & Wandell (2008)), constraints which are not modeled in existing fMRI stimulus optimization methods. Further, neural selectivity varies widely across sub-populations in early and intermediate cortical regions, necessitating fine-grained analysis of response properties at the single voxel level, in contrast to previous approaches which have averaged across voxels in a region. Moreover, because early visual cortex regions are not strongly selective for semantic information, it is not yet clear whether the natural image prior imposed by a diffusion model is required in order to generate interpretable MEIs, or whether a gradient-based method can provide complementary insight.

Here, we introduce and compare two MEI-generation frameworks for single voxels in early and intermediate human visual cortex: receptive field diffusion for visual exploration (**RF-DiVE**), and receptive field gradient optimization (**RF-GO**). Both frameworks use a subject-specific, voxelwise feature-weighted receptive field (fwRF) encoding model (St-Yves & Naselaris, 2018) to guide stimulus optimization. By spatially pooling features of a deep convolutional neural network according to an estimated voxel-specific pRF, this model incorporates both feature and spatial selectivity to predict fMRI responses. RF-GO uses these predictions to directly optimize images through regularized gradient ascent, whereas RF-DiVE uses response gradients to guide a pretrained latent diffusion model (LDM) with a natural-image prior. Targeting individual voxels allows both approaches to probe variation in selectivity within cortical areas while accounting for differences in receptive-field location and size. Together, these frameworks extend model-guided stimulus optimization to the spatially specific and heterogeneous representations found across V1, V2, V3, and hV4. Our contributions are as follows:

- We introduce two new frameworks, RF-DiVE and RF-GO, enabling spatially-constrained MEI synthesis for individual voxels in early and intermediate retinotopic cortex.
- Applying our method to data from the Natural Scenes Dataset (NSD; Allen et al. (2022)), we generate single-voxel MEIs that exhibit coherent, spatially localized structure, appearing to depict mid-level properties such as orientation, curvature, color, and form.
- We provide *in-silico* validation of these MEIs using an independent test brain encoder and show MEIs consistently elicit higher predicted responses than natural-image controls.

- We provide a direct comparison of gradient-based (RF-GO) and diffusion-based (RF-DiVE) optimization methods for the same brain regions, which reveal differences in model generalization performance, visual interpretability, and low-level visual statistics, which also depend on the DNN backbone used.
- We use a human behavioral study to provide initial characterization of the visual properties captured by these MEIs, showing that RF-DiVE MEIs for hV4 voxels depict more three-dimensional form than those for V1 voxels, indicating our method can capture expected differences between early and intermediate visual cortex.

2 Related work

Feature selectivity in the visual hierarchy. In classic models, representational complexity is thought to increase along the primate visual cortex hierarchy, where V1 encodes basic features such as contrast, orientation, and retinotopic position (Hubel & Wiesel, 1962; Carandini et al., 2005), V4 encodes contour, shape, color, and texture (Pasupathy et al., 2020; Conway et al., 2007); and higher visual areas represent objects and semantic categories (Grill-Spector & Weiner, 2014; Desimone et al., 1984). However, the feature space represented in intermediate areas like V4 remains poorly understood, and recent work also shows that V1 can encode more complex properties such as curvature (Tang et al., 2018; Ding et al., 2026). Our work addresses these questions with a data-driven approach for characterizing feature selectivity in early and intermediate visual cortex.

Voxelwise encoding models for fMRI data. Forward encoding models (i.e., brain encoders) provide a powerful way to model neural activity, providing an image-computable mapping from stimulus to brain response by extracting features from deep neural network (DNN) models. Task-optimized DNNs have proven effective as feature extractors for brain encoders because of their hierarchical representations similar to the primate ventral visual stream (Naselaris et al., 2011; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Yamins & DiCarlo, 2016; Eickenberg et al., 2017; Schrimpf et al., 2018; Zhuang et al., 2021; Conwell et al., 2024). Standard encoding models linearly map DNN features to voxel responses, often collapsing spatial information by averaging. In contrast, feature-weighted receptive field models explicitly capture voxelwise spatial and feature selectivity through a pRF model (St-Yves & Naselaris, 2018). We adopt this framework with multiple pretrained visual DNN backbones.

Stimulus optimization frameworks. Stimulus optimization synthesizes MEIs by maximizing brain-encoder responses, providing a data-driven probe of neural selectivity. Prior work has used regularized gradient ascent (Bashivan et al., 2019; Walker et al., 2019; Willeke et al., 2026; Cowley et al., 2026; Khosla & Wehbe, 2026) and generative models such as GANs and diffusion models (Ponce et al., 2019; Gu et al., 2022; Ratan Murty et al., 2021; Luo et al., 2023; Henderson et al., 2026; Hwang et al., 2026). Strong natural-image priors imposed by generative models can improve interpretability, whereas weaker gradient-based priors may reveal selectivity outside the natural-image distribution. Related work has also perturbed existing images to modulate regional responses (Prince et al., 2026; García Cerdas et al., 2025). For gradient-based methods, recent closed-loop results further suggest strong backbone dependence, with adversarially-trained networks performing well for neural control (Prince et al., 2026). Here, we compare gradient- and diffusion-based MEIs in visual properties and *in silico* generalization using ADV-RN50 (Engstrom et al., 2019) and DINO-RN50 (Caron et al., 2021) brain encoders.

3 Method

In this section, we first describe our approach for constructing an image-computable pRF- constrained brain encoder, which can predict the response in early and intermediate visual fMRI voxels for image stimuli. Then, we introduce two frameworks that use the pRF-constrained brain encoder to generate MEIs for early and intermediate visual cortex voxels: diffusion-based RF-DiVE framework and gradient-based RF-GO framework.

pRF-constrained brain encoder. The brain encoder was implemented using a CNN-based fwRF model (St-Yves & Naselaris, 2018), as shown in Figure 1A. Specifically, the input image was first processed by a CNN to obtain intermediate feature maps with height, width, and channel dimensions

(h, w, c). The pRF for each voxel was modeled as a 2D isotropic Gaussian, with center (x, y) and size (σ) parameters. To extract pRF-specific features, we took a dot product between the pRF (scaled to size h x w) and the spatial dimensions of the CNN feature maps, achieving spatial pooling that reduced the features to a channel dimension (c). Features were concatenated across multiple layers in the channel dimension. The resulting feature representation was then linearly combined with channel weights and biases (estimated using ridge regression) to predict the response of each target voxel. The pRF parameters for each voxel were determined by grid search over a set of candidate pRFs in a log-polar grid (Henderson et al., 2023a;b); see Section 4.

Receptive Field Diffusion for Visual Exploration (RF-DiVE). RF-DiVE uses a pretrained LDM (Rombach et al., 2022) as in BrainDiVE (Luo et al., 2023). At each step of generation, the LDM latent representation is passed through the latent decoder to produce an image, which is then evaluated by a fwRF brain encoder. The predicted response of a target voxel was used as the optimization objective, and gradients were propagated through the brain encoder and latent decoder, perturbing the denoising procedure toward images that maximize the target voxel's predicted activity (Figure 1B).

Receptive Field Gradient Optimization (RF-GO). RF-GO uses encoder gradients to optimize image Fourier phase to maximize predicted voxel responses (Figure 1B). To avoid high-frequency artifacts common with gradient-based synthesis, MEIs were generated using Magnitude-Constrained Optimization (MACO, Fel et al. (2023)), which regularizes the image by fixing the Fourier magnitude spectrum to an ImageNet-derived natural-image template while optimizing only the Fourier phase.

4 Experiments

Dataset and partitions. All modeling was performed using the NSD dataset, a large-scale 7T fMRI dataset in which 8 participants each viewed up to 10,000 natural images (shown at 8.4° visual angle). We used 4 subjects (S1, S2, S5, and S7) who each viewed all 10k images each repeated 3 times. Out of the 10k images for each subject, 1000 images were shared across the 4 subjects and 9000 were unique. For each subject, we split the 10k images randomly in half resulting in 2 partitions (P1 and P2), where each set contains 4500 unique images and 500 shared images. Within each partition, the 4500 unique images were further split into training and nested held-out sets with ratio 9:1, and the 500 shared images were used as test images.

We used single-image β weights for each voxel computed using the GLMSingle pipeline (Prince et al., 2022), which we z-scored for each voxel within each session, then averaged over repeats of each image. To define ROIs, we used the NSD's subject-specific V1-hV4 definitions (combining dorsal and ventral subdivisions and two hemispheres), which were manually delineated from independent population receptive field (pRF) mapping data as described by Allen et al. (2022).

Brain encoders, backbones, and training. We constructed four brain encoder backbone types for each subject. The first two brain encoders ("generation models") were used for MEI generation, and were based on ADV-RN50 (Engstrom et al., 2019) and DINO-RN50 (Caron et al., 2021) backbones, each trained using the data from P1 only. A separate encoder ("ranking model") was constructed based on OpenCLIP-RN50 (Ilharco et al., 2021), also trained on P1, and was used to select the top MEIs across random seeds. Finally, an independent test encoder ("test model") was based on a different convolutional architecture, OpenCLIP ConvNeXt-Base (Ilharco et al., 2021), trained on P2, and was used to independently evaluate the model-predicted responses elicited by the MEIs. See Appendix A.1 for the overall R^2 of each model backbone.

Each encoding model was constructed using the fwRF framework (Section 3). Input images were transformed according to the preprocessing associated with the pretrained backbone before extracting activations of first ReLU and layer 1-4 for RN-50, stem and stage 1-4 for ConvNext-Base. For each voxel, feature maps were spatially pooled using a Gaussian pRF shared across layers, concatenated, standardized channel-wise, and mapped to predicted fMRI responses through a linear readout with an intercept. Readout weights were fitted using ridge regression. The optimal pRF position and size (x, y, σ), as well as regularization strength (λ) were selected based on minimum prediction error on the nested held-out set within the partition. This procedure for fitting the best pRF was performed separately for each of the generation and test models, so pRF could vary depending on backbone. The ranking encoders were fitted using the pRFs of their corresponding generation en-

coder. For each subject and each area, we selected the top 20 voxels based on the generation models' test-set R^2 within P1, which were then used for MEI generation. Generation models with different backbones (ADV/DINO) can select voxels differently, and their estimated pRFs for the same voxel can also vary (see Appendix A.5).



Figure 2: **Spatial and feature specificity of V1 MEIs. (A)** Top 5 MEIs for an example voxel using both frameworks. The MEIs show consistent patterns within the pRF (white circle, 3σ range of the Gaussian) and greater variation outside it. **(B)** Full-image and pRF-zoomed views of an example MEI. **(C–F)** pRF-zoomed top 5 natural images and MEIs for four example V1 voxels, shown for both ADV-RN50 and DINO-RN50 backbones. See Appendix A.11 for more examples.

Most-Exciting-Image (MEI) optimization. For RF-DiVE, we used Stable Diffusion v2.1 with a DPM-Solver multistep scheduler, 100 denoising steps, an empty text prompt, and classifier-free guidance disabled. At each step, gradients from a subject-specific encoder ("generation model") guided generation to maximize the predicted response of a single target voxel. Brain-guidance and self-attention-guidance scales were set to 300 and 0.75, respectively. For RF-GO, the Fourier phases were optimized for 100 steps using NAdam with a learning rate of 1.0. The objective was to maximize the predicted response of the target voxel. For both RF-DiVE and RF-GO, for each voxel, we generated 1,000 images with different random seeds.

To control for low-level brightness and contrast, during optimization, every image presented to the encoding model was projected to a full-image mean luma of $116/255$ and RMS luma contrast of $57.48/255$, corresponding to the mean luma and mean RMS luma contrast computed from natural images in the training sets. This full-image constraint was applied again before ranking and evaluation (Appendix A.2). Additional controls for luma and contrast within the pRF region itself are provided in the Appendix A.4. Unless otherwise specified, quantitative analyses were performed on the top 10 MEIs selected according to the predicted response based on the "ranking model" (see above section). Additional ablations using MEIs selected by the generation model are provided in the Appendix A.6 and A.7.



Figure 3: **Feature specificity of hV4 MEIs across generation encoders.** Top 5 natural images and MEIs are shown for two example hV4 voxels (**A-B** show the same voxel, **C-D** show the same voxel). Note the pRF is fit separately for each backbone and the rank number of voxels is assigned separately for each backbone. See Appendix A.11 for more examples, including from V2 and V3.

To provide a baseline comparison for the predicted response to MEIs, we used the same ranking model to identify the top-10 predicted natural images from a large image set. Natural images were either sampled from the NSD stimulus set (Allen et al., 2022), using P2 test set, or from 500 randomly sampled images from the LAION-fMRI stimulus set (Zerbe et al. (2026); Schuhmann et al. (2022)), with LAION-fMRI providing a complementary natural-image control.

5 Results

5.1 Generated MEIs show both spatial and feature specificity

Figure 2A shows five MEIs generated for an example V1 voxel using both RF-DiVE and RF-GO. The generated MEIs exhibit consistent feature patterns within the voxel's pRF (white circle), while remaining largely random outside it, demonstrating that the pRF constraint imposed by the encoding model led to spatially specific MEIs. Figure 2B illustrates a zoomed view of the pRF region within a full MEI; this zoomed view is used to display images in later figures.

Beyond spatial specificity, the MEIs also exhibit coherent structure that captures distinct visual features, including color, form, and texture. Comparing across methods, the MEIs resulting from RF-DiVE often appear more coherent and more colorful than the RF-GO MEIs (we discuss these differences in more detail in the next section), but we also observe shared structure in the MEIs for a given voxel that is common across both methods. Figure 2C shows MEIs that depict high-contrast yellow–black rectilinear and angular contours, Figure 2D shows multicolored bent lines and curves that create abstract patterns, Figure 2E shows clusters of dark dots on warm-colored backgrounds; and Figure 2F shows dense, pinkish or greenish speckled patterns. These patterns are evident in both the RF-DiVE and RF-GO MEIs, and similar features are also visible in the top natural images for each voxel. For example, the top images for the voxel in Figure 2C include a bright yellow railing, and the voxel in Figure 2F includes pink flowers with green stems. Notably, inspecting the top natural images alone does not always reveal a coherent set of shared features, which highlights how MEI generation may be needed to uncover the underlying features driving a neural population.

The MEIs generated for hV4 voxels capture additional intermediate-level visual structure. In Figure 3, we show MEIs generated for two example hV4 voxels using both DINO and ADV generation backbones, revealing feature content that is consistent across backbone choices. The voxel depicted in Figure 3A-B exhibits smooth, arch-like forms and saturated colors, while the voxel in Figure 3C-D exhibits orange–yellow radiating stripes and curved bands. Again, the top natural images show related features (top of a dollhouse for A-B, round fruits for C-D).

5.2 Generated MEIs produce higher predicted voxel response

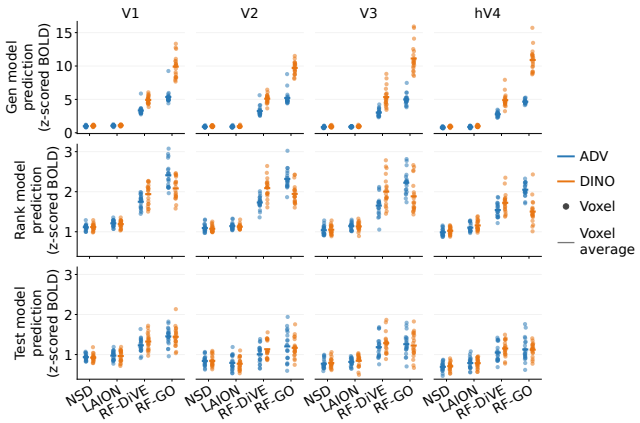

Figure 4: **Predicted responses to top 10 images for S1.** Top-10 ranking model-selected images for each image group were evaluated by the generation, ranking, and independent test encoders. Each point represents mean prediction across images for one voxel; horizontal bar = mean across voxels in S1. See Appendix A.6 for the other NSD subjects.

To rigorously assess whether the generated MEIs generalize beyond the models used for optimization, we perform *in-silico* evaluation using an independent test encoder with a different architecture, pretraining dataset, and NSD training partition. Figure 4 compares predicted voxel responses for the top-20 voxels (ranked by R^2) in each visual area in S1, elicited by RF-DiVE and RF-GO MEIs versus natural images from NSD and LAION-fMRI, using the generation, ranking, and independent test encoders. Figure 5 depicts the advantage for MEIs over natural images (predicted difference score), across all 4 NSD participants. Across these results, both RF-DiVE and RF-GO exhibit a predicted-response advantage over natural images from the stricter control dataset, LAION-fMRI. This difference is especially prominent when evaluated using the generation model, and is attenuated on the ranking and testing model. However, the predicted difference score remains positive for all participants, even in the most challenging case of generalization to the test model. Although there is variability across participants, this advantage is significantly greater than zero for the majority of subjects in all visual areas and for all generation methods (one-sample t-test; $p<0.01$; FDR corrected; see Appendix A.3, Table 1 for test statistics). These results indicate that the MEIs consistently elicit higher model-predicted responses than the evaluated natural image controls.

5.3 Effects of different encoder backbones and generation frameworks

A key strength of our study is the breadth of conditions under which MEIs were generated, spanning different generation methods, brain-encoder backbones, subjects, visual areas, and target voxels. We systematically analyze the effects of these factors on MEI properties and their predicted responses.

Figure 5 shows a comparison of the predicted difference score of MEIs over LAION-fMRI images, across different encoder backbones and generation frameworks. When evaluated using the generation encoder, RF-GO achieved larger difference scores than RF-DiVE with both backbones. However, this advantage was not consistently retained under the ranking or independent test encoder, suggesting that the learned naturalistic prior enforced by the RF-DiVE method may result in weaker activation of the generation model but improved generalization across backbones, compared to RF-GO.

We also observed effects of model backbone that depended on the generation framework. For RF-DiVE, DINO yielded larger scores than ADV under the generation and ranking encoders, and this tendency was retained in most independent-test comparisons, with DINO having consistently higher difference score than ADV for the majority of participants in all areas (see Appendix A.7). However, this pattern differed for RF-GO. For RF-GO, DINO also showed larger difference score than ADV on the generation encoder, but this pattern partially inverted on the ranking and testing model, with DINO having similar or worse difference score relative to ADV in all regions. This may indicate that the DINO backbone, when paired with the RF-GO method, is prone to model-specific biases that impair cross-model generalization, whereas the ADV model results in more robustly generalizable RF-GO MEIs; this is broadly consistent with other recent work (Prince et al., 2026). These patterns were similar across all regions tested.

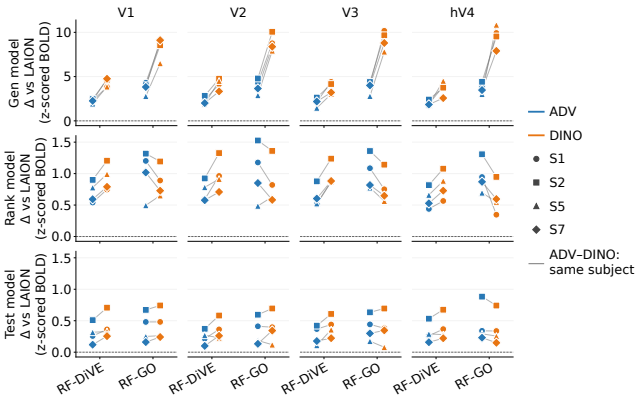

Figure 5: **Predicted difference score (advantage of MEIs over LAION-fMRI) across subjects.** For each voxel, we computed the predicted-response difference between MEIs and natural images. Points reflect the mean of these differences across voxels within each subject and ROI.

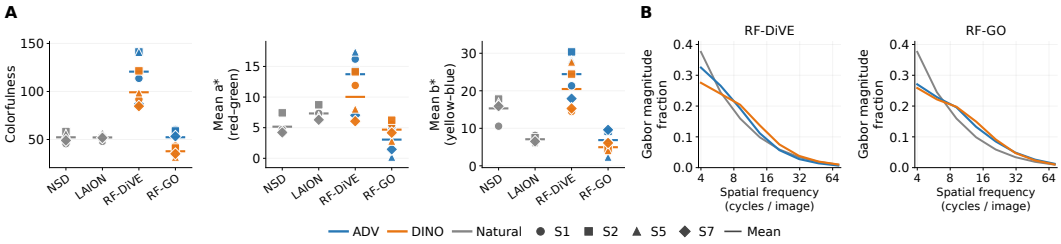

Figure 6: **Chromatic and spatial-frequency statistics for V1 MEIs. (A)** Colorfulness (Hasler & Suesstrunk, 2003) and CIELAB color statistics. **(B)** Normalized Gabor magnitude profiles across spatial frequencies, shown for RF-DiVE and RF-GO. Similar patterns were observed in V2, V3, and hV4.

Given these differences between methods and backbones in model generalization, we next examined differences in the basic visual properties of generated MEIs (see Appendices A.8 and A.9 for image quantification and additional comparisons). Comparing chromatic statistics across generation frameworks and natural-image controls (Figure 6A) shows that RF-DiVE MEIs exhibit higher colorfulness than RF-GO MEIs and natural images. RF-DiVE also shows higher subject-averaged a^* and b^* values, indicating a stronger tendency toward warm colors, particularly red and yellow. These differences are observed with both generation backbones. Figure 6B compares the spatial-frequency profiles of the generated and natural images using normalized Gabor response magnitudes. Relative to natural images, both RF-DiVE and RF-GO MEIs show a smaller relative contribution at the lowest measured spatial frequency and a larger contribution at intermediate frequencies. Within each generation framework, MEIs generated using DINO exhibit a further shift toward intermediate and higher spatial frequencies compared with ADV generated MEIs.

5.4 Feature differences across regions

The MEIs from each region depict varied mid-level features that could provide new insight into how feature coding evolves along the visual hierarchy. As a first step toward interpreting the MEI features, we conducted a human behavioral study comparing the perceived three-dimensional form of MEIs generated for V1 and hV4 voxels (see Appendix A.10). This design was motivated by past evidence that sensitivity to shape and form increases from V1 to hV4 (Connor et al., 2007; Pasupathy et al., 2020). In a forced-choice task, participants (n=32 in each of two experiments) more frequently judged hV4 RF-DiVE MEIs as exhibiting greater three-dimensional form than V1 MEIs, across both encoder backbones and viewing conditions. In contrast, no reliable hV4-over-V1 preference was observed for natural images from LAION. Results for RF-GO were less consistent across encoder backbones and viewing conditions, but showed a preference for hV4 in the ADV backbone in the full image condition. These findings suggest that RF-DiVE MEIs may more effectively capture interpretable differences between regions compared to natural images, again illustrating the benefit of MEI synthesis for capturing visual properties that are not readily detectable from natural images alone. Within the RF-GO method, the ADV backbone may also be more effective at capturing these interpretable properties than DINO. Further work is needed to confirm and expand these findings.

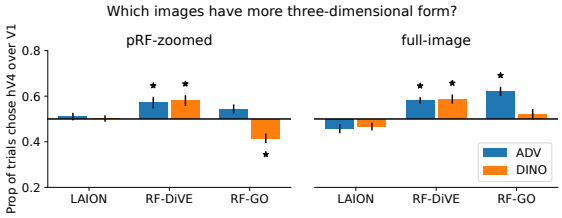

Figure 7: **Depiction of form in MEIs.** Proportion of trials in which participants selected the hV4 images as having more three-dimensional form than the V1 images, using pRF-zoomed images (left) or full images (right). Bar height and error bars indicate mean±SEM across 32 participants; asterisk indicates a significant difference from 0.50 (one-sample t-test; $p<0.01$).

6 Discussion

We introduced RF-DiVE and RF-GO, two frameworks for generating most-exciting-images for individual fMRI voxels in early and intermediate human visual cortex. Across V1-hV4, the resulting MEIs contained consistent structure within the estimated pRF and captured diverse features including contour, color, texture, and form. Moreover, MEIs generated by both methods produced higher predicted responses than the strongest natural-image controls across subjects, visual areas, and encoder backbones. These findings extend model-guided stimulus optimization from region-level analyses of higher visual cortex to spatially localized representations at the single-voxel level.

Our analyses also show that an MEI is shaped jointly by the target voxel, the brain-encoder backbone, and the optimization framework. RF-GO had a higher difference score under the generation encoder, but showed relatively weaker generalization to the independent test encoder; this effect was more pronounced for DINO than for ADV. The two frameworks also produced systematically different image statistics: RF-DiVE generated more colorful and warmer images, whereas DINO MEIs shifted toward higher spatial frequencies relative to ADV. At the same time, visual comparison of the MEIs across methods (Figures 2-3) does suggest structural elements such as curves and texture that are common across methods. These patterns suggest that combining multiple methods for MEI generation may be an effective strategy to separate the most robust, generalizable properties of neural tuning from those that are idiosyncratic to one method.

Several additional limitations suggest directions for future work. First, further validation of the generated MEIs will require closed-loop experiments with human participants. Second, our current quantitative analyses focus on relatively low-level visual properties, including Gabor features, color, edges, and curvature, and therefore provide limited insight into intermediate-level visual features such as texture, contour, and surface structure. Finally, our perceptual experiment was preliminary. Larger-scale behavioral studies, together with more systematic measurements of visual properties such as contour, texture, and shape, will be needed to determine which visual dimensions reliably differentiate cortical areas and generation methods.

AI USE STATEMENT

AI tools were used to assist with text polishing, code editing, computational job execution, and figure formatting. All code, scientific analyses, results, and interpretations were reviewed and verified by the authors.

ETHICS STATEMENT

In this work we used human participant data encompassing fMRI measurements from the Natural Scenes Dataset (Allen et al., 2022). The human participants procedures were approved by an appropriate Institutional Review Board as described in the original data paper. The data was shared in an anonymized format. We also collected new human behavioral data from online participants using Prolific. Participants provided informed consent, and the study procedures were approved by the Institutional Review Board at the authors' institution. Our research addresses basic science questions, and we do not anticipate any direct negative societal harm associated with these results.

REPRODUCIBILITY STATEMENT

All details needed to reproduce our data processing steps and modeling procedures are provided in the main text (see *Methods* and *Experiments*) and in the Appendix. We plan to publicly release the code in the future.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *The Journal of Neuroscience*, 2005. doi: 10.1523/JNEUROSCI.3726-05.2005.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Charles E. Connor, Scott L. Brincat, and Anitha Pasupathy. Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, 17:140–147, 2007. doi: 10.1016/j.conb.2007.03.002.

Bevil R. Conway, Sebastian Moeller, and Doris Y. Tsao. Specialized color modules in macaque extrastriate cortex. *Neuron*, 56(3):560–573, 11 2007. ISSN 0896-6273. doi: 10.1016/J.NEURON.2007.10.008. URL https://pubmed.ncbi.nlm.nih.gov/17988638/.

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

Benjamin R Cowley, Patricia L Stan, Jonathan W Pillow, and Matthew A Smith. Compact deep neural network models of the visual cortex. *Nature*, 652(8111):947–954, 2026.

R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4:2051–2062, 8 1984. doi: 10.1523/jneurosci.04-08-02051.1984.

Zhiwei Ding, Dat Tran, Kayla Ponder, Zhuokun Ding, Rachel Froebe, Lydia Ntanavara, Paul G. Fahey, Erick Cobos, Luca Baroni, Maria Diamantaki, Eric Y. Wang, Andersen Chang, Stelios Papadopoulos, Jiakun Fu, Taliah Muhammad, Christos Papadopoulos, Santiago A. Cadena, Alexandros Evangelou, Konstantin Willeke, Fabio Anselmi, Sophia Sanborn, Jan Antolik, Emmanouil Froudarakis, Saumil Patel, Edgar Y. Walker, Jacob Reimer, Fabian H. Sinz, Alexander S. Ecker, Katrin Franke, Xaq Pitkow, and Andreas S. Tolias. Functional bipartite invariance in mouse primary visual cortex receptive fields. *Nature Neuroscience*, 29(4):851–863, Apr 2026. ISSN 1546-1726. doi: 10.1038/s41593-026-02213-3. URL https://doi.org/10.1038/s41593-026-02213-3.

Serge O Dumoulin and Brian A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660, 2008.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, et al. Unlocking feature visualization for deep network with magnitude constrained optimization. *Advances in Neural Information Processing Systems*, 36:37813–37826, 2023.

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. In *International Conference on Learning Representations*, volume 2025, pp. 9040–9075, 2025.

Kalanit Grill-Spector and Kevin S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15:536–548, 6 2014. ISSN 1471-0048. doi: 10.1038/nrn3747.

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.

David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pp. 87–95. SPIE, 2003.

Margaret M Henderson, Michael J Tarr, and Leila Wehbe. Low-level tuning biases in higher visual cortex reflect the semantic informativeness of visual features. *Journal of Vision*, 23(4):8–8, 2023a.

Margaret M Henderson, Michael J Tarr, and Leila Wehbe. A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *The Journal of Neuroscience*, 43(22):4144–4161, 2023b.

Margaret M Henderson, Andrew F Luo, Sungjoon Park, Michael J Tarr, and Leila Wehbe. Diffusion-based stimulus optimization reveals functional organization across higher visual cortex. *bioRxiv*, pp. 2026–05, 2026.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160:106, 1 1962. ISSN 14697793. doi: 10.1113/JPHYSIOL.1962.SP006837.

Ethan Hwang, Hossein Adeli, Wenxuan Guo, Andrew Luo, and Nikolaus Kriegeskorte. In silico mapping of visual categorical selectivity across the whole brain. *Advances in Neural Information Processing Systems*, 38:164602–164646, 2026.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Meenakshi Khosla and Leila Wehbe. Higher visual areas act like domain-general filters with strong selectivity and functional specialization. *Nature Communications*, 17(1):7484, 2026.

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.

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

Anitha Pasupathy, Dina V. Popovkina, and Taekjun Kim. Visual functions of primate area V4. *Annual Review of Vision Science*, 6:363–385, September 2020. doi: 10.1146/annurev-vision-030320-041306.

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, 2019.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11, 2022. ISSN 2050084X. doi: 10.7554/ELIFE.77599.

Jacob S. Prince, Binxu Wang, Thomas Fel, Akshay V. Jagadeesh, Parisa A. Vaziri, George A. Alvarez, Margaret S. Livingstone, and Talia Konkle. Parametric neural control differentiates top neural network models of primate visual cortex. *bioRxiv*, 2026. doi: 10.64898/2026.08.16.745063. URL https://www.biorxiv.org/content/early/2026/08/20/2026.08.16.745063.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 2018.

Shiming Tang, Tai Sing Lee, Ming Li, Yimeng Zhang, Yue Xu, Fang Liu, Benjamin Teo, and Hongfei Jiang. Complex pattern selectivity in macaque primary visual cortex revealed by large-scale two-photon imaging. *Current Biology*, 28(1):38–48, 2018.

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, 2019.

Konstantin F Willeke, Kelli Restivo, Katrin Franke, Arne F Nix, Santiago A Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S Ecker, Fabian H Sinz, and Andreas S Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 supports topological organization. *eLife*, 15:RP109875, 2026. doi: 10.7554/eLife.109875.1.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

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

Josefine Zerbe, Johannes Roth, Maggie Mae Mell, Peer Herholz, Tomas Knapen, and Martin N. Hebart. Laion-fmri: A densely sampled 7t-fmri dataset providing broad coverage of natural image diversity, 2026. URL https://www.visionsciences.org/talk-sessions?id=642. Talk 25.11, Scene Perception Talk Session, Vision Sciences Society Annual Meeting, May 16, 2026.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A APPENDIX

Contents of this appendix

A.1 Encoding Model Performance 14
A.2 Luma mean and RMS-contrast constraints 19
A.3 Paired-voxel comparisons with natural images 20
A.4 Effects of Final Luma Constraints 25
A.5 Fitted pRF coverage 25
A.6 Predicted responses across subjects and selection conditions 32
A.7 Predicted-response advantages over LAION 40
A.8 Additional chromatic and orientation statistics 42
A.9 Additional structural image comparisons 47
A.10 Behavioral experiment to test differences between region MEIs 49
A.11 Additional MEIs 49

A.1 ENCODING MODEL PERFORMANCE

We report voxelwise encoding performance on 500 held-out test images for each subject and training partition. The reported R^2 values are not normalized by noise ceiling. Each figure contains a 4×4 grid: rows correspond to subjects S1, S2, S5, and S7, and columns to ADV-RN50, DINO-RN50, OpenCLIP-RN50, and OpenCLIP-ConvNeXt-Base. Points represent voxels within the retinotopically defined area with positive noise ceiling and finite evaluation values; these are not restricted to the voxels selected for MEI generation. Red dashed lines show ordinary least-squares fits, and black dashed lines indicate R^2 equal to noise ceiling. Panel annotations report the mean, median, and number of voxels. Finite observations outside the displayed axis limits remain included in the summary statistics and regression fits.

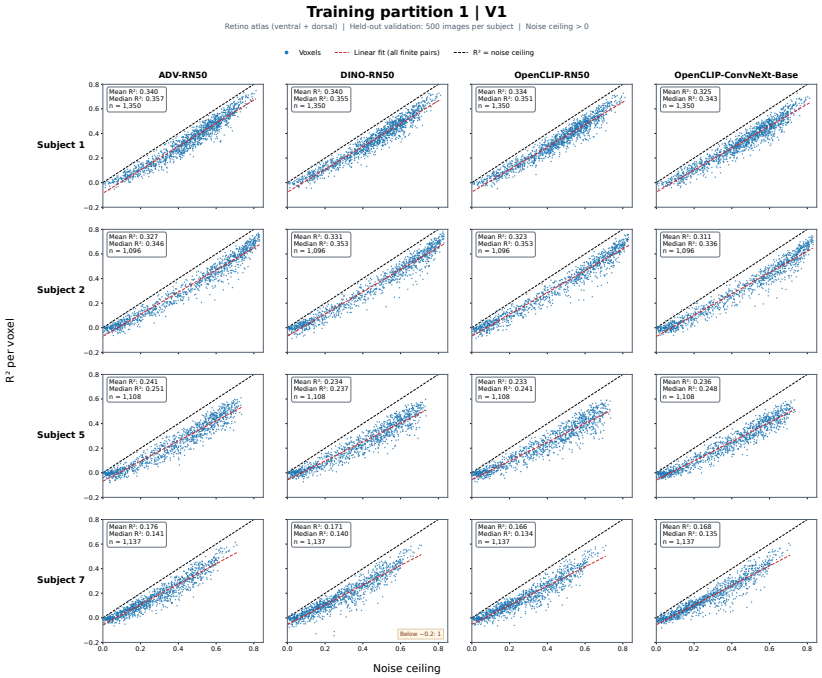

Figure 8: Held-out encoding performance in V1 for training partition 1. Voxelwise R^2 is plotted against noise ceiling.

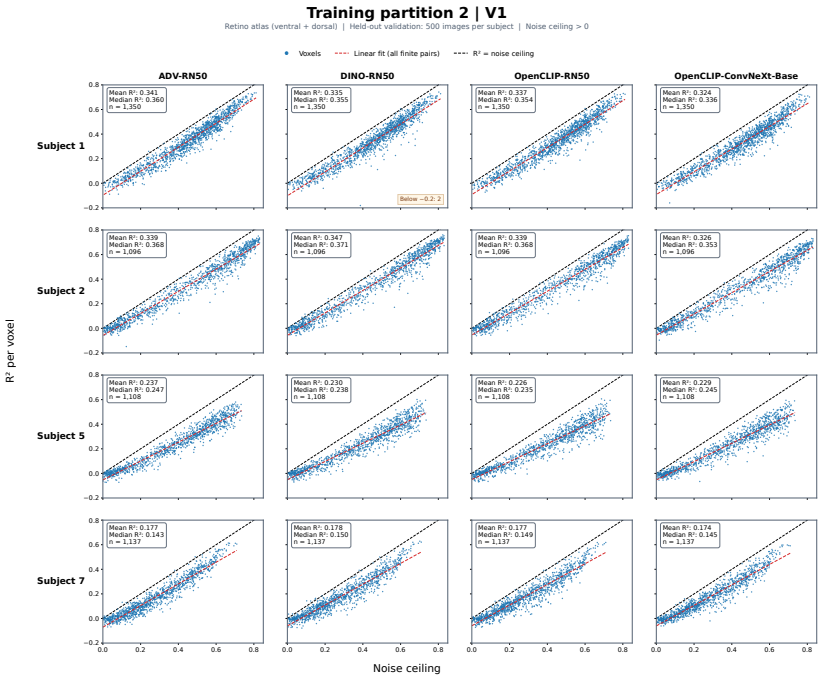

Figure 9: Held-out encoding performance in V1 for training partition 2. Voxelwise R^2 is plotted against noise ceiling.

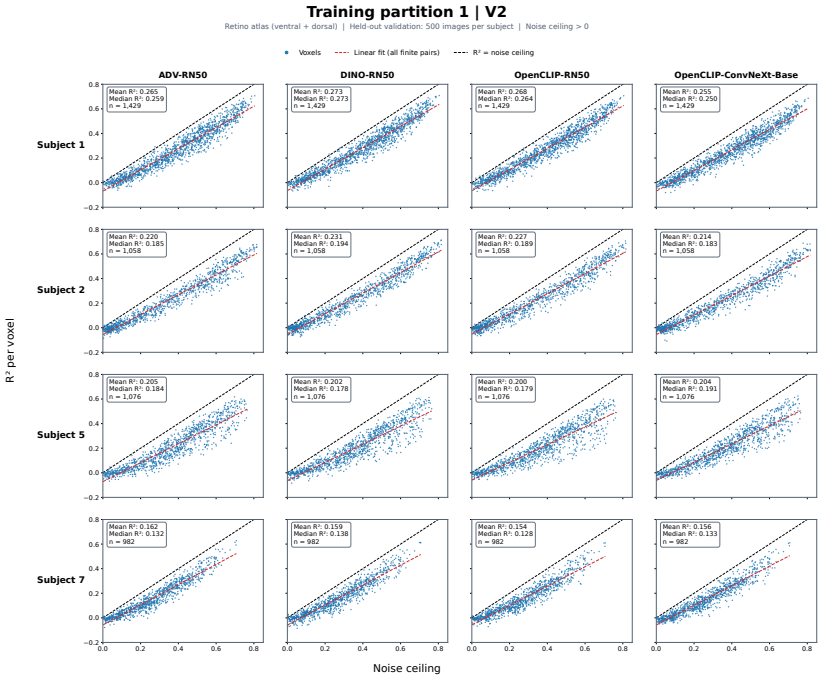

Figure 10: Held-out encoding performance in V2 for training partition 1. Voxelwise R^2 is plotted against noise ceiling.

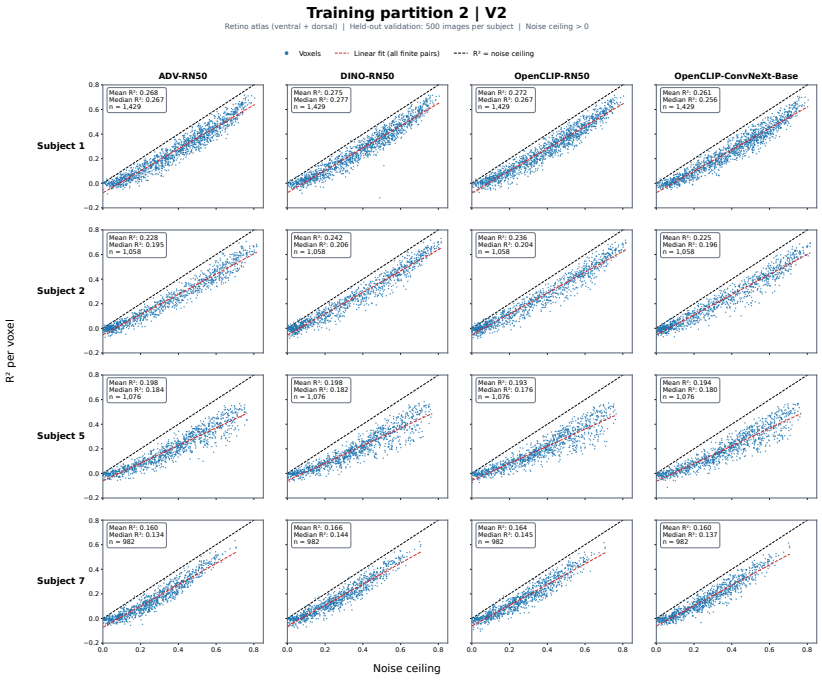

Figure 11: Held-out encoding performance in V2 for training partition 2. Voxelwise R^2 is plotted against noise ceiling.

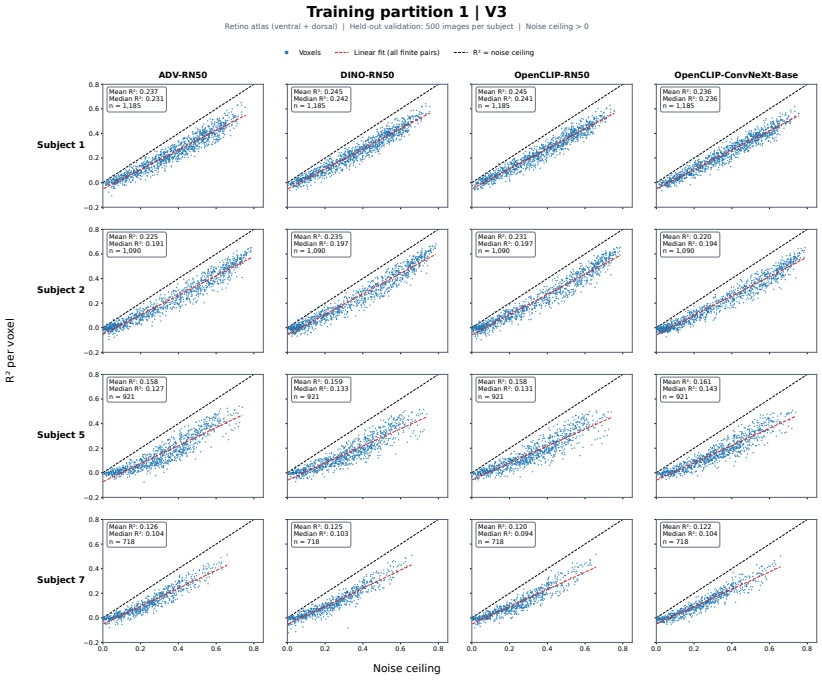

Figure 12: Held-out encoding performance in V3 for training partition 1. Voxelwise R^2 is plotted against noise ceiling.

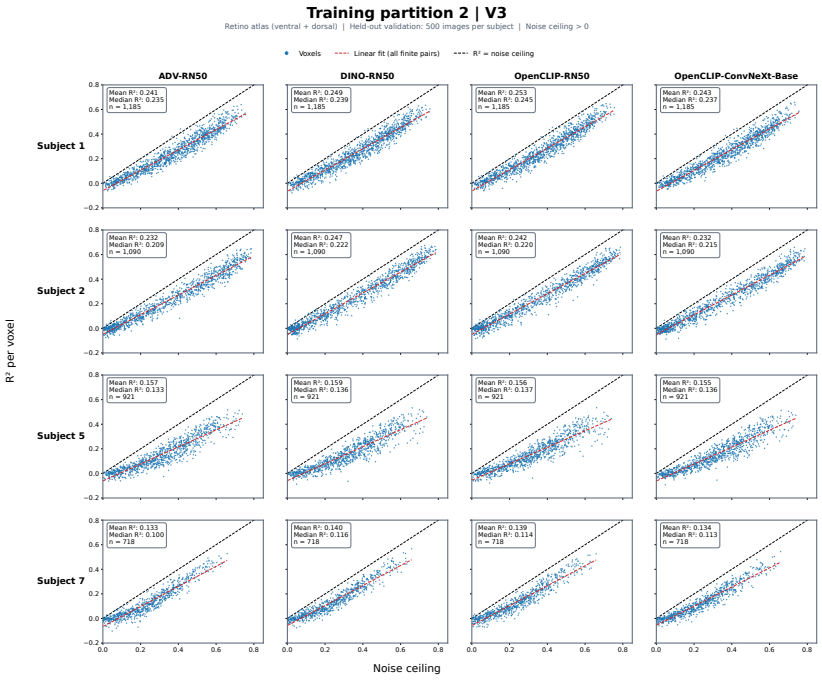

Figure 13: Held-out encoding performance in V3 for training partition 2. Voxelwise R^2 is plotted against noise ceiling.

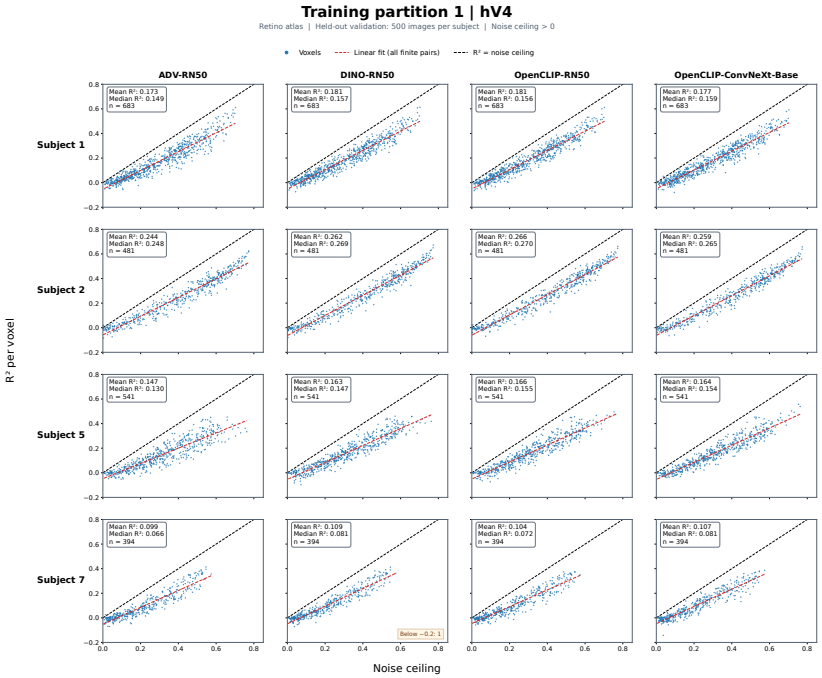

Figure 14: Held-out encoding performance in hV4 for training partition 1. Voxelwise R^2 is plotted against noise ceiling.

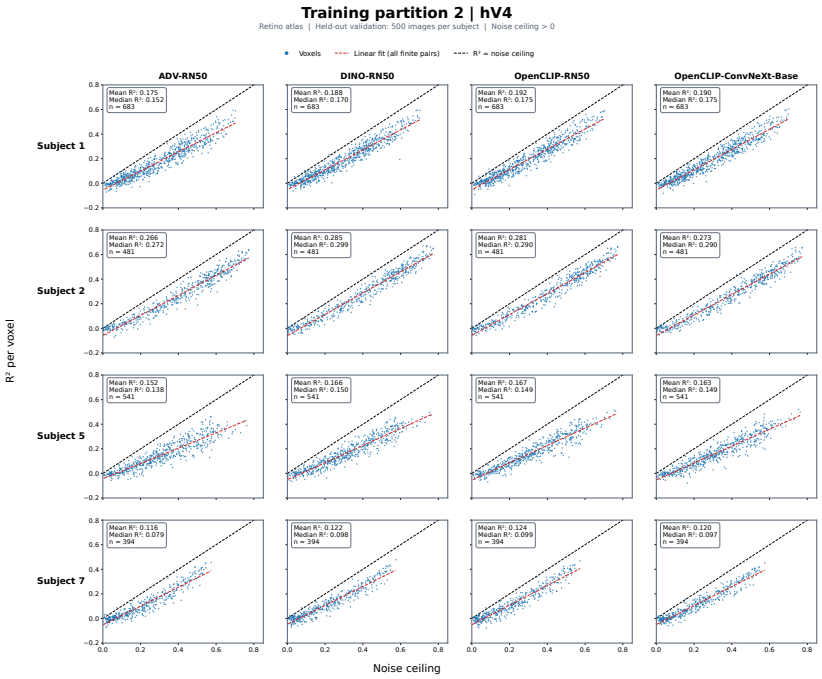

Figure 15: Held-out encoding performance in hV4 for training partition 2. Voxelwise R^2 is plotted against noise ceiling.

A.2 LUMA MEAN AND RMS-CONTRAST CONSTRAINTS

Luma definition and constraint statistics. For an RGB image $I \in [0,1]^{3\times H\times W}$, let Ω denote its $N = HW$ spatial locations. We computed luma directly from gamma-coded RGB values using Rec. 709 coefficients:

$$Y_p = \mathbf{a}^\mathsf{T} I_p, \qquad \mathbf{a} = (0.2126,\ 0.7152,\ 0.0722)^\mathsf{T}. \tag{1}$$

This quantity is distinct from linear-light luminance. We defined the chromatic residual as $D_p = I_p - Y_p \mathbf{1}_3$, where $\mathbf{1}_3 = (1,1,1)^\mathsf{T}$, so that $\mathbf{a}^\mathsf{T} D_p = 0$.

For a scalar field L and nonnegative spatial weights q satisfying $\sum_p q_p = 1$, we defined

$$\mu_q(L) = \sum_{p\in\Omega} q_p L_p, \qquad C_q(L) = \left[\sum_{p\in\Omega} q_p \left(L_p - \mu_q(L)\right)^2\right]^{1/2}. \tag{2}$$

Full-image statistics used uniform weights $g_p = 1/N$. For pRF-weighted statistics, the generator model's fitted pRF was mapped to image space and normalized to obtain weights w_p. These soft weights were used continuously, without conversion to a binary aperture. The target mean and RMS contrast were $\mu^\star = 116/255$ and $C^\star = 57.48/255$, respectively.

Projection during MACO optimization. At each MACO optimization step, the current raw image was projected before evaluation by the generator encoding model. Starting from $Z^{(0)} = Y$, the scalar-luma adjustment was

$$Z_p^{(k+1)} = \mathrm{clip}_{[0,1]}\left[\mu^\star + \frac{C^\star}{\max\left(C_g(Z^{(k)}), \epsilon\right)}\left(Z_p^{(k)} - \mu_g(Z^{(k)})\right)\right], \tag{3}$$

where $\epsilon = 10^{-8}$. Because scalar-luma clipping can perturb the statistics, the initial adjustment was followed by up to 50 compensation iterations. The loop terminated early when both scalar-luma statistics were within $3/255$ of their targets for all images in the batch.

The RGB image presented to the model was

$$P_g(I)_p = \mathrm{clip}_{[0,1]}\left[Z_p^{(\mathrm{final})}\mathbf{1}_3 + D_p\right]. \tag{4}$$

Thus, the original chromatic residual was restored with a fixed coefficient of one, rather than explicitly scaled with contrast. RGB clipping can nevertheless alter both chromatic residuals and luma statistics. At optimization step t, the predicted voxel response $\widehat{r}_v(P_g(I(\phi_t)))$ was evaluated, and gradients were propagated through the executed projection operations to the MACO Fourier-phase parameters ϕ_t.

After the 100th phase update, the image was rendered again without an additional projection, quantized to uint8, and saved losslessly as a reusable parent image. Both final branches below were derived independently from this same parent. The parent therefore omitted only the final projection; its synthesis still used the full-image projection described above.

Final constraint branches. Let Y and D now denote the luma and chromatic residual recomputed from the saved parent image. The *full-image* branch applied the iterative scalar-luma projection in equation 3, followed by RGB reconstruction and clipping as in equation 4.

The *full-image+pRF* branch instead fitted four parameters jointly. We defined the local adjustment envelope $E_p = (w_p/\max_{r\in\Omega} w_r)^{1/2}$ and used

$$Y_p^{\mathrm{local}} = Y_p + (e^\beta - 1)E_p\left(Y_p - \mu_w(Y)\right) + \delta E_p, \tag{5}$$

$$Z_p(\boldsymbol{\theta}) = \mathrm{clip}_{[0,1]}\left[\mu_g(Y^{\mathrm{local}}) + \eta + e^\gamma\left(Y_p^{\mathrm{local}} - \mu_g(Y^{\mathrm{local}})\right)\right], \tag{6}$$

$$I'_p(\boldsymbol{\theta}) = \mathrm{clip}_{[0,1]}\left[Z_p(\boldsymbol{\theta})\mathbf{1}_3 + D_p\right], \qquad \boldsymbol{\theta} = (\beta,\gamma,\delta,\eta). \tag{7}$$

Here, β and δ parameterized local contrast and mean adjustments, whereas γ and η parameterized global contrast and mean adjustments. All four parameters were fitted jointly because their effects on the target statistics are coupled. Each solver evaluation used the same parent Y and D, rather than recursively transforming the previous RGB output. Both spatial weightings used the same targets $\mu^\star$ and $C^\star$.

Numerical solution and acceptance. For the full-image+pRF branch, fitting statistics were evaluated on the clipped scalar-luma field $Z(\boldsymbol{\theta})$, before chromatic-residual restoration and RGB clipping. We minimized $\|\mathbf{r}(\boldsymbol{\theta})\|_2^2$, where

$$\mathbf{r}(\boldsymbol{\theta}) = \begin{bmatrix} (\mu_g(Z(\boldsymbol{\theta})) - \mu^\star)/C^\star \\ (\mu_w(Z(\boldsymbol{\theta})) - \mu^\star)/C^\star \\ \log(C_g(Z(\boldsymbol{\theta}))/C^\star) \\ \log(C_w(Z(\boldsymbol{\theta}))/C^\star) \end{bmatrix}. \tag{8}$$

The four-parameter fit used a GPU-batched Levenberg–Marquardt solver with an analytic Jacobian $J = \partial\mathbf{r}/\partial\boldsymbol{\theta} \in \mathbb{R}^{4\times4}$ and double-precision arithmetic. Parameter bounds were $\beta, \gamma \in [-8, 8]$ and $\delta, \eta \in [-1, 1]$, with at most 100 solver iterations per image. The full-image branch used the scalar-luma compensation procedure rather than this four-parameter solver.

Solver convergence and scalar-luma target checks were required before final RGB acceptance checks. Because chromatic-residual restoration and RGB clipping can alter luma statistics, satisfying the scalar-luma targets alone did not guarantee acceptance.

Final RGB outputs were quantized to uint8, and luma $\widehat{Y}$ was recomputed from these exact quantized RGB values normalized to $[0, 1]$. A candidate was accepted only if it passed the preceding numerical checks and satisfied

$$\max_{q\in\mathcal{Q}} \max\left\{\left|\mu_q(\widehat{Y}) - \mu^\star\right|, \left|C_q(\widehat{Y}) - C^\star\right|\right\} \le \frac{3}{255}, \tag{9}$$

where $\mathcal{Q} = \{g\}$ for the full-image branch and $\mathcal{Q} = \{g, w\}$ for the full-image+pRF branch. Ranking and subsequent evaluation used only accepted, quantized images, and retained images were saved as lossless uint8 RGB PNGs.

A.3 Paired-voxel comparisons with natural images

We compare the independent test encoder's predictions for MEIs and natural images separately for each subject, ROI, generation backbone, generation method, and final-constraint condition. Both image sets are selected by the ranking encoder. For each source, predictions form an $n \times 10$ array: we average the ten selected images within each voxel and pair the resulting means by global voxel ID. Defining their difference as d_v, we use the two-sided paired t statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad \bar{d} = \frac{1}{n}\sum_{v=1}^{n} d_v, \qquad s_d^2 = \frac{1}{n-1}\sum_{v=1}^{n}(d_v - \bar{d})^2,$$

with $n - 1$ degrees of freedom. Most conditions contain 20 paired voxels; incomplete diffusion results yield 18 or 19 in the remaining conditions. Natural images are unmodified. Benjamini–Hochberg (BH) correction is applied jointly to the 256 comparisons in Table 1. These are exploratory within-subject tests over selected voxels, not population-level tests over subjects. They assume independent voxel differences; spatial dependence is not accounted for.

Table 1: Paired-voxel tests of MEI responses against natural controls. Positive t favors MEIs; n is the paired-voxel count. Bold adjusted p values indicate significantly higher MEI responses after BH correction across all 256 comparisons. Results continue on the following pages.

				MEI − NSD		MEI − LAION	
ROI	Backbone	Method	n	t	p_{BH}	t	p_{BH}
S1 — Full image							
V1	ADV	RF-DiVE	20	8.67	$\mathbf{1.7\times10^{-7}}$	6.76	$\mathbf{3.8\times10^{-6}}$
		RF-GO	20	11.08	$\mathbf{7.9\times10^{-9}}$	11.47	$\mathbf{6.1\times10^{-9}}$
	DINO	RF-DiVE	20	11.33	$\mathbf{6.7\times10^{-9}}$	8.38	$\mathbf{2.5\times10^{-7}}$
		RF-GO	20	11.25	$\mathbf{7.2\times10^{-9}}$	10.81	$\mathbf{9.8\times10^{-9}}$
V2	ADV	RF-DiVE	20	4.40	$\mathbf{4.3\times10^{-4}}$	6.51	$\mathbf{5.9\times10^{-6}}$
		RF-GO	20	4.73	$\mathbf{2.1\times10^{-4}}$	7.20	$\mathbf{1.7\times10^{-6}}$
	DINO	RF-DiVE	20	9.67	$\mathbf{4.0\times10^{-8}}$	11.18	$\mathbf{7.2\times10^{-9}}$
		RF-GO	20	8.73	$\mathbf{1.5\times10^{-7}}$	13.13	$\mathbf{1.4\times10^{-9}}$
V3	ADV	RF-DiVE	20	8.82	$\mathbf{1.4\times10^{-7}}$	6.68	$\mathbf{4.4\times10^{-6}}$
		RF-GO	20	8.75	$\mathbf{1.5\times10^{-7}}$	8.54	$\mathbf{2.0\times10^{-7}}$
	DINO	RF-DiVE	20	10.06	$\mathbf{2.3\times10^{-8}}$	11.44	$\mathbf{6.1\times10^{-9}}$
		RF-GO	20	8.04	$\mathbf{4.3\times10^{-7}}$	6.75	$\mathbf{3.8\times10^{-6}}$
hV4	ADV	RF-DiVE	20	7.71	$\mathbf{7.6\times10^{-7}}$	7.68	$\mathbf{7.7\times10^{-7}}$
		RF-GO	20	6.62	$\mathbf{4.9\times10^{-6}}$	8.21	$\mathbf{3.2\times10^{-7}}$
	DINO	RF-DiVE	20	10.53	$\mathbf{1.3\times10^{-8}}$	12.26	$\mathbf{3.0\times10^{-9}}$
		RF-GO	20	11.20	$\mathbf{7.2\times10^{-9}}$	10.44	$\mathbf{1.4\times10^{-8}}$
S1 — Full image + pRF							
V1	ADV	RF-DiVE	20	1.59	0.136	0.86	0.415
		RF-GO	20	6.24	$\mathbf{9.7\times10^{-6}}$	6.07	$\mathbf{1.4\times10^{-5}}$
	DINO	RF-DiVE	20	4.63	$\mathbf{2.6\times10^{-4}}$	3.07	**0.007**
		RF-GO	20	7.08	$\mathbf{2.2\times10^{-6}}$	6.38	$\mathbf{7.5\times10^{-6}}$
V2	ADV	RF-DiVE	20	5.89	$\mathbf{1.9\times10^{-5}}$	6.73	$\mathbf{4.0\times10^{-6}}$
		RF-GO	20	7.42	$\mathbf{1.2\times10^{-6}}$	11.42	$\mathbf{6.1\times10^{-9}}$
	DINO	RF-DiVE	20	10.99	$\mathbf{8.5\times10^{-9}}$	10.60	$\mathbf{1.3\times10^{-8}}$
		RF-GO	20	11.57	$\mathbf{5.6\times10^{-9}}$	13.85	$\mathbf{8.8\times10^{-10}}$
V3	ADV	RF-DiVE	20	7.68	$\mathbf{7.7\times10^{-7}}$	6.09	$\mathbf{1.3\times10^{-5}}$
		RF-GO	20	7.57	$\mathbf{9.2\times10^{-7}}$	8.49	$\mathbf{2.1\times10^{-7}}$
	DINO	RF-DiVE	20	10.45	$\mathbf{1.4\times10^{-8}}$	10.91	$\mathbf{8.9\times10^{-9}}$
		RF-GO	20	7.79	$\mathbf{6.6\times10^{-7}}$	6.88	$\mathbf{3.1\times10^{-6}}$
hV4	ADV	RF-DiVE	20	7.55	$\mathbf{9.5\times10^{-7}}$	7.03	$\mathbf{2.3\times10^{-6}}$
		RF-GO	20	7.15	$\mathbf{1.9\times10^{-6}}$	9.28	$\mathbf{6.8\times10^{-8}}$
	DINO	RF-DiVE	20	10.22	$\mathbf{1.9\times10^{-8}}$	11.20	$\mathbf{7.2\times10^{-9}}$
		RF-GO	20	9.36	$\mathbf{6.1\times10^{-8}}$	8.35	$\mathbf{2.6\times10^{-7}}$

Table 1 (continued): S2.

				MEI − NSD		MEI − LAION	
ROI	Backbone	Method	n	t	p_{BH}	t	p_{BH}
S2 — Full image							
V1	ADV	RF-DiVE	20	8.54	$\mathbf{2.0\times10^{-7}}$	10.00	$\mathbf{2.5\times10^{-8}}$
		RF-GO	20	8.63	$\mathbf{1.8\times10^{-7}}$	10.68	$\mathbf{1.2\times10^{-8}}$
	DINO	RF-DiVE	20	13.48	$\mathbf{1.1\times10^{-9}}$	20.69	$\mathbf{4.4\times10^{-12}}$
		RF-GO	20	14.30	$\mathbf{8.1\times10^{-10}}$	19.06	$\mathbf{9.8\times10^{-12}}$
V2	ADV	RF-DiVE	20	5.97	$\mathbf{1.6\times10^{-5}}$	5.54	$\mathbf{3.8\times10^{-5}}$
		RF-GO	20	8.50	$\mathbf{2.1\times10^{-7}}$	7.67	$\mathbf{7.8\times10^{-7}}$
	DINO	RF-DiVE	20	9.00	$\mathbf{1.0\times10^{-7}}$	9.34	$\mathbf{6.2\times10^{-8}}$
		RF-GO	20	8.80	$\mathbf{1.4\times10^{-7}}$	9.54	$\mathbf{4.8\times10^{-8}}$
V3	ADV	RF-DiVE	20	7.81	$\mathbf{6.4\times10^{-7}}$	6.24	$\mathbf{9.7\times10^{-6}}$
		RF-GO	20	10.58	$\mathbf{1.3\times10^{-8}}$	11.76	$\mathbf{4.4\times10^{-9}}$
	DINO	RF-DiVE	19	8.98	$\mathbf{1.5\times10^{-7}}$	10.59	$\mathbf{1.9\times10^{-8}}$
		RF-GO	20	11.08	$\mathbf{7.9\times10^{-9}}$	13.38	$\mathbf{1.2\times10^{-9}}$
hV4	ADV	RF-DiVE	20	12.27	$\mathbf{3.0\times10^{-9}}$	12.23	$\mathbf{3.0\times10^{-9}}$
		RF-GO	20	10.90	$\mathbf{8.9\times10^{-9}}$	10.88	$\mathbf{8.9\times10^{-9}}$
	DINO	RF-DiVE	20	11.04	$\mathbf{8.1\times10^{-9}}$	13.78	$\mathbf{8.8\times10^{-10}}$
		RF-GO	20	7.07	$\mathbf{2.2\times10^{-6}}$	8.38	$\mathbf{2.5\times10^{-7}}$
S2 — Full image + pRF							
V1	ADV	RF-DiVE	20	0.08	0.939	1.77	0.100
		RF-GO	20	4.23	$\mathbf{6.2\times10^{-4}}$	6.43	$\mathbf{6.9\times10^{-6}}$
	DINO	RF-DiVE	20	3.08	**0.007**	5.69	$\mathbf{2.8\times10^{-5}}$
		RF-GO	20	4.99	$\mathbf{1.2\times10^{-4}}$	7.33	$\mathbf{1.4\times10^{-6}}$
V2	ADV	RF-DiVE	20	6.57	$\mathbf{5.4\times10^{-6}}$	5.84	$\mathbf{2.1\times10^{-5}}$
		RF-GO	20	8.51	$\mathbf{2.1\times10^{-7}}$	7.42	$\mathbf{1.2\times10^{-6}}$
	DINO	RF-DiVE	20	10.38	$\mathbf{1.5\times10^{-8}}$	9.40	$\mathbf{5.8\times10^{-8}}$
		RF-GO	20	9.05	$\mathbf{9.7\times10^{-8}}$	9.25	$\mathbf{7.0\times10^{-8}}$
V3	ADV	RF-DiVE	20	9.84	$\mathbf{3.2\times10^{-8}}$	7.82	$\mathbf{6.3\times10^{-7}}$
		RF-GO	20	11.83	$\mathbf{4.4\times10^{-9}}$	12.93	$\mathbf{1.6\times10^{-9}}$
	DINO	RF-DiVE	19	10.08	$\mathbf{3.6\times10^{-8}}$	12.26	$\mathbf{4.4\times10^{-9}}$
		RF-GO	20	12.97	$\mathbf{1.6\times10^{-9}}$	16.83	$\mathbf{6.1\times10^{-11}}$
hV4	ADV	RF-DiVE	20	12.18	$\mathbf{3.0\times10^{-9}}$	11.98	$\mathbf{3.8\times10^{-9}}$
		RF-GO	20	10.06	$\mathbf{2.3\times10^{-8}}$	10.19	$\mathbf{1.9\times10^{-8}}$
	DINO	RF-DiVE	20	12.47	$\mathbf{2.7\times10^{-9}}$	13.88	$\mathbf{8.8\times10^{-10}}$
		RF-GO	20	7.68	$\mathbf{7.7\times10^{-7}}$	9.22	$\mathbf{7.3\times10^{-8}}$

Table 1 (continued): S5.

				MEI − NSD		MEI − LAION	
ROI	Backbone	Method	n	t	p_{BH}	t	p_{BH}
S5 — Full image							
V1	ADV	RF-DiVE	20	5.76	$\mathbf{2.5\times10^{-5}}$	8.27	$\mathbf{2.9\times10^{-7}}$
		RF-GO	20	3.67	**0.002**	5.50	$\mathbf{4.1\times10^{-5}}$
	DINO	RF-DiVE	20	10.52	$\mathbf{1.3\times10^{-8}}$	10.90	$\mathbf{8.9\times10^{-9}}$
		RF-GO	20	5.87	$\mathbf{2.0\times10^{-5}}$	7.66	$\mathbf{7.8\times10^{-7}}$
V2	ADV	RF-DiVE	18	3.08	**0.008**	3.48	**0.004**
		RF-GO	20	1.85	0.087	2.67	**0.017**
	DINO	RF-DiVE	19	2.79	**0.014**	3.83	**0.002**
		RF-GO	20	0.94	0.375	2.10	0.055
V3	ADV	RF-DiVE	19	2.14	0.052	1.92	0.077
		RF-GO	20	3.87	**0.001**	3.06	**0.008**
	DINO	RF-DiVE	19	6.64	$\mathbf{5.9\times10^{-6}}$	5.49	$\mathbf{5.1\times10^{-5}}$
		RF-GO	20	0.69	0.510	0.97	0.363
hV4	ADV	RF-DiVE	20	4.09	$\mathbf{8.4\times10^{-4}}$	5.59	$\mathbf{3.5\times10^{-5}}$
		RF-GO	20	2.57	**0.022**	3.26	**0.005**
	DINO	RF-DiVE	20	3.58	**0.003**	6.23	$\mathbf{9.8\times10^{-6}}$
		RF-GO	20	3.49	**0.003**	4.01	$\mathbf{9.9\times10^{-4}}$
S5 — Full image + pRF							
V1	ADV	RF-DiVE	20	4.19	$\mathbf{6.8\times10^{-4}}$	8.19	$\mathbf{3.4\times10^{-7}}$
		RF-GO	20	0.78	0.457	2.23	**0.043**
	DINO	RF-DiVE	20	7.54	$\mathbf{9.5\times10^{-7}}$	10.50	$\mathbf{1.4\times10^{-8}}$
		RF-GO	20	1.97	0.070	3.55	**0.003**
V2	ADV	RF-DiVE	18	2.71	**0.017**	3.14	**0.007**
		RF-GO	20	1.50	0.160	2.31	**0.037**
	DINO	RF-DiVE	19	2.40	**0.031**	3.69	**0.002**
		RF-GO	20	0.00	1.000	1.29	0.224
V3	ADV	RF-DiVE	19	2.12	0.054	1.86	0.085
		RF-GO	20	4.54	$\mathbf{3.2\times10^{-4}}$	3.57	**0.003**
	DINO	RF-DiVE	19	6.64	$\mathbf{5.9\times10^{-6}}$	6.26	$\mathbf{1.2\times10^{-5}}$
		RF-GO	20	0.52	0.616	0.84	0.423
hV4	ADV	RF-DiVE	20	3.68	**0.002**	5.15	$\mathbf{8.5\times10^{-5}}$
		RF-GO	20	2.56	**0.022**	3.26	**0.005**
	DINO	RF-DiVE	20	3.38	**0.004**	5.38	$\mathbf{5.3\times10^{-5}}$
		RF-GO	20	3.20	**0.006**	3.73	**0.002**

Table 1 (continued): S7.

				MEI − NSD		MEI − LAION	
ROI	Backbone	Method	n	t	p_{BH}	t	p_{BH}
S7 — Full image							
V1	ADV	RF-DiVE	20	3.67	$\mathbf{0.002}$	3.96	$\mathbf{0.001}$
		RF-GO	20	4.55	$\mathbf{3.1\times10^{-4}}$	6.35	$\mathbf{8.0\times10^{-6}}$
	DINO	RF-DiVE	20	9.57	$\mathbf{4.6\times10^{-8}}$	9.46	$\mathbf{5.4\times10^{-8}}$
		RF-GO	20	6.43	$\mathbf{6.8\times10^{-6}}$	6.74	$\mathbf{3.9\times10^{-6}}$
V2	ADV	RF-DiVE	19	0.92	0.382	3.40	$\mathbf{0.004}$
		RF-GO	20	2.37	$\mathbf{0.032}$	4.03	$\mathbf{9.6\times10^{-4}}$
	DINO	RF-DiVE	19	7.67	$\mathbf{1.1\times10^{-6}}$	9.18	$\mathbf{1.2\times10^{-7}}$
		RF-GO	20	8.17	$\mathbf{3.4\times10^{-7}}$	7.37	$\mathbf{1.3\times10^{-6}}$
V3	ADV	RF-DiVE	19	5.41	$\mathbf{5.9\times10^{-5}}$	5.41	$\mathbf{5.9\times10^{-5}}$
		RF-GO	20	5.86	$\mathbf{2.0\times10^{-5}}$	4.56	$\mathbf{3.1\times10^{-4}}$
	DINO	RF-DiVE	19	6.73	$\mathbf{5.2\times10^{-6}}$	8.16	$\mathbf{5.1\times10^{-7}}$
		RF-GO	20	8.45	$\mathbf{2.3\times10^{-7}}$	6.91	$\mathbf{3.0\times10^{-6}}$
hV4	ADV	RF-DiVE	20	5.97	$\mathbf{1.6\times10^{-5}}$	5.50	$\mathbf{4.2\times10^{-5}}$
		RF-GO	20	4.87	$\mathbf{1.6\times10^{-4}}$	5.29	$\mathbf{6.4\times10^{-5}}$
	DINO	RF-DiVE	20	6.85	$\mathbf{3.2\times10^{-6}}$	7.68	$\mathbf{7.7\times10^{-7}}$
		RF-GO	20	4.15	$\mathbf{7.4\times10^{-4}}$	4.54	$\mathbf{3.2\times10^{-4}}$
S7 — Full image + pRF							
V1	ADV	RF-DiVE	20	0.68	0.510	1.12	0.293
		RF-GO	20	1.21	0.255	1.94	0.074
	DINO	RF-DiVE	20	6.06	$\mathbf{1.4\times10^{-5}}$	6.48	$\mathbf{6.2\times10^{-6}}$
		RF-GO	20	3.59	$\mathbf{0.002}$	4.08	$\mathbf{8.5\times10^{-4}}$
V2	ADV	RF-DiVE	19	−0.18	0.862	2.01	0.066
		RF-GO	20	0.93	0.379	2.61	$\mathbf{0.020}$
	DINO	RF-DiVE	19	5.84	$\mathbf{2.5\times10^{-5}}$	6.94	$\mathbf{3.6\times10^{-6}}$
		RF-GO	20	6.02	$\mathbf{1.5\times10^{-5}}$	5.73	$\mathbf{2.6\times10^{-5}}$
V3	ADV	RF-DiVE	19	5.10	$\mathbf{1.1\times10^{-4}}$	5.33	$\mathbf{6.9\times10^{-5}}$
		RF-GO	20	5.27	$\mathbf{6.6\times10^{-5}}$	4.16	$\mathbf{7.2\times10^{-4}}$
	DINO	RF-DiVE	19	7.06	$\mathbf{3.0\times10^{-6}}$	9.90	$\mathbf{4.6\times10^{-8}}$
		RF-GO	20	8.44	$\mathbf{2.3\times10^{-7}}$	6.25	$\mathbf{9.7\times10^{-6}}$
hV4	ADV	RF-DiVE	20	5.78	$\mathbf{2.4\times10^{-5}}$	4.75	$\mathbf{2.0\times10^{-4}}$
		RF-GO	20	4.62	$\mathbf{2.7\times10^{-4}}$	5.08	$\mathbf{9.8\times10^{-5}}$
	DINO	RF-DiVE	20	7.43	$\mathbf{1.2\times10^{-6}}$	7.82	$\mathbf{6.3\times10^{-7}}$
		RF-GO	20	4.03	$\mathbf{9.5\times10^{-4}}$	4.21	$\mathbf{6.5\times10^{-4}}$

A.4 Effects of Final Luma Constraints

We compared images before final adjustment with two final-constraint conditions: Full, which controls full-image mean luma and RMS contrast, and Full+pRF, which additionally controls these statistics within the generator's soft pRF. The targets were 116 for mean luma and 57.48 for RMS contrast, with a tolerance of ± 3 in 0–255 units. For generated images, Parent denotes the saved endpoint before final adjustment, not synthesis without constraints.

Selected image identities were held fixed across conditions, retaining the intersection of images that passed both final projections. Natural images underwent the corresponding adjustments. Responses were evaluated using the independent test encoder, without re-ranking after adjustment. The predicted-response advantage of MEIs over natural-image controls generally persisted under both constraint conditions.

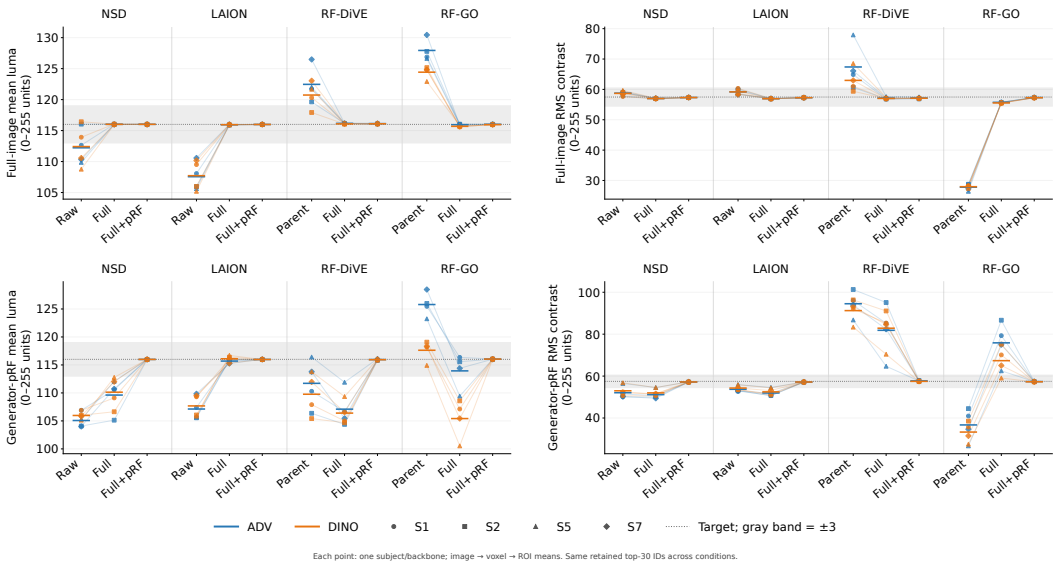

Figure 16: Verification of full-image and pRF-weighted luma constraints. Each point summarizes one subject and generation backbone; horizontal bars indicate means across subjects. Statistics were averaged over images, voxels, and ROIs in sequence. Dotted lines and shaded bands indicate targets and tolerances. Measurements use saved-image resolution; subsequent encoder-input resizing can change RMS contrast.

For the following response comparisons, each point represents one voxel's mean prediction over the retained images from its top-10 selection. Lines connect the same voxel across final-constraint conditions, and horizontal bars indicate means across voxels. Blue and orange denote ADV and DINO generation backbones, respectively. The retained image count can be below ten when a projection fails.

A.5 Fitted pRF coverage

Each subject is shown with ADV above DINO. Within each backbone block, columns show V1, V2, V3, and hV4; the upper row shows generator pRFs, also used by the refitted ranker, and the lower row shows the test encoder's own pRFs for the same 20 voxel IDs. Dots mark centers and circles have radius 2σ, where σ is the fitted Gaussian width. Colors identify voxels within each ROI. Dashed squares mark the $8.4°$-wide input field; the displayed field is $12.6°$ wide (1.5 times the input width). These are fitted feature-pooling pRFs, not full CNN effective receptive fields.

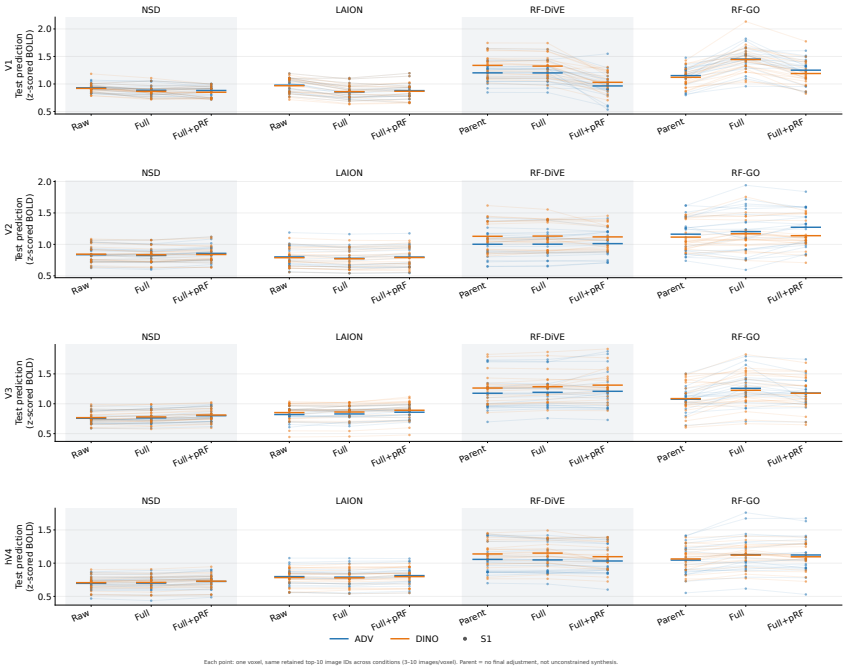

Figure 17: Independent test-encoder predictions across final-constraint conditions for S1. Rows show V1, V2, V3, and hV4; image sources comprise NSD, LAION-fMRI, RF-DiVE, and RF-GO.

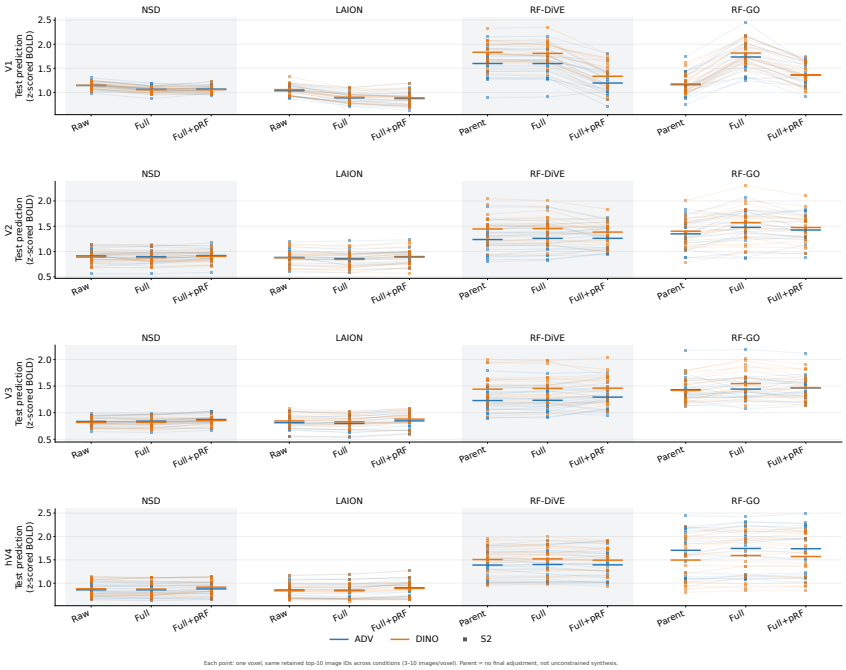

Figure 18: Independent test-encoder predictions across final-constraint conditions for S2, using the same layout and conventions as Figure 17.

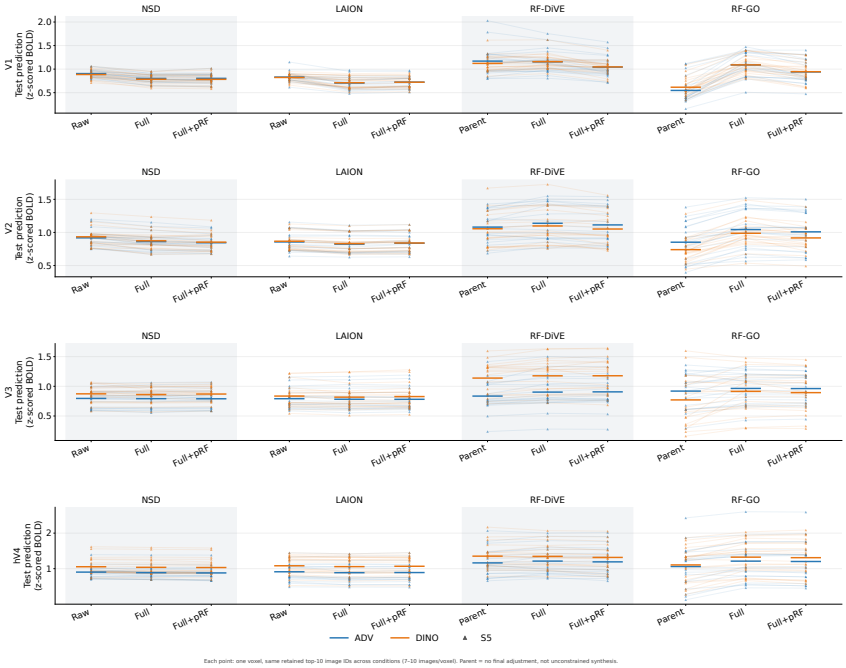

Figure 19: Independent test-encoder predictions across final-constraint conditions for S5, using the same layout and conventions as Figure 17.

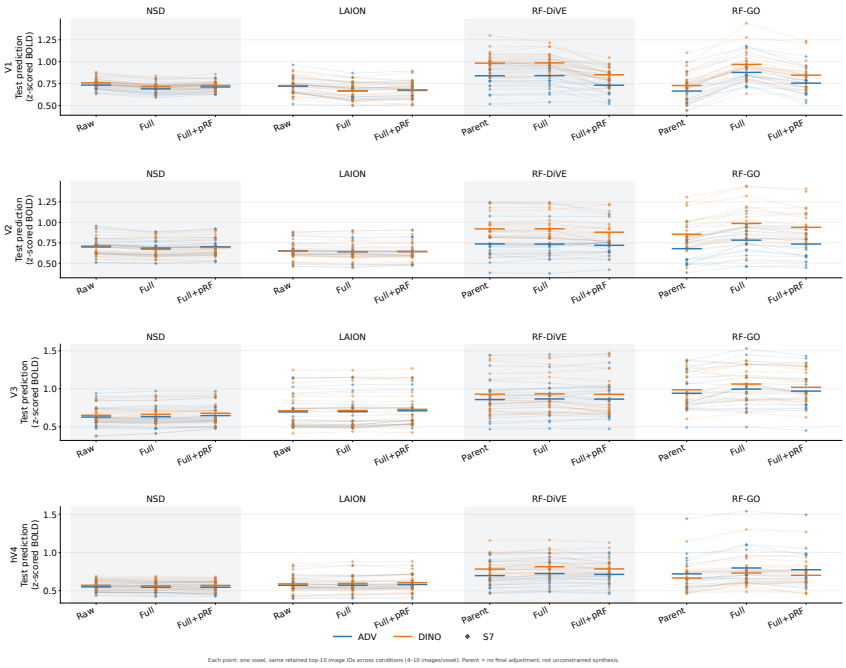

Figure 20: Independent test-encoder predictions across final-constraint conditions for S7, using the same layout and conventions as Figure 17.

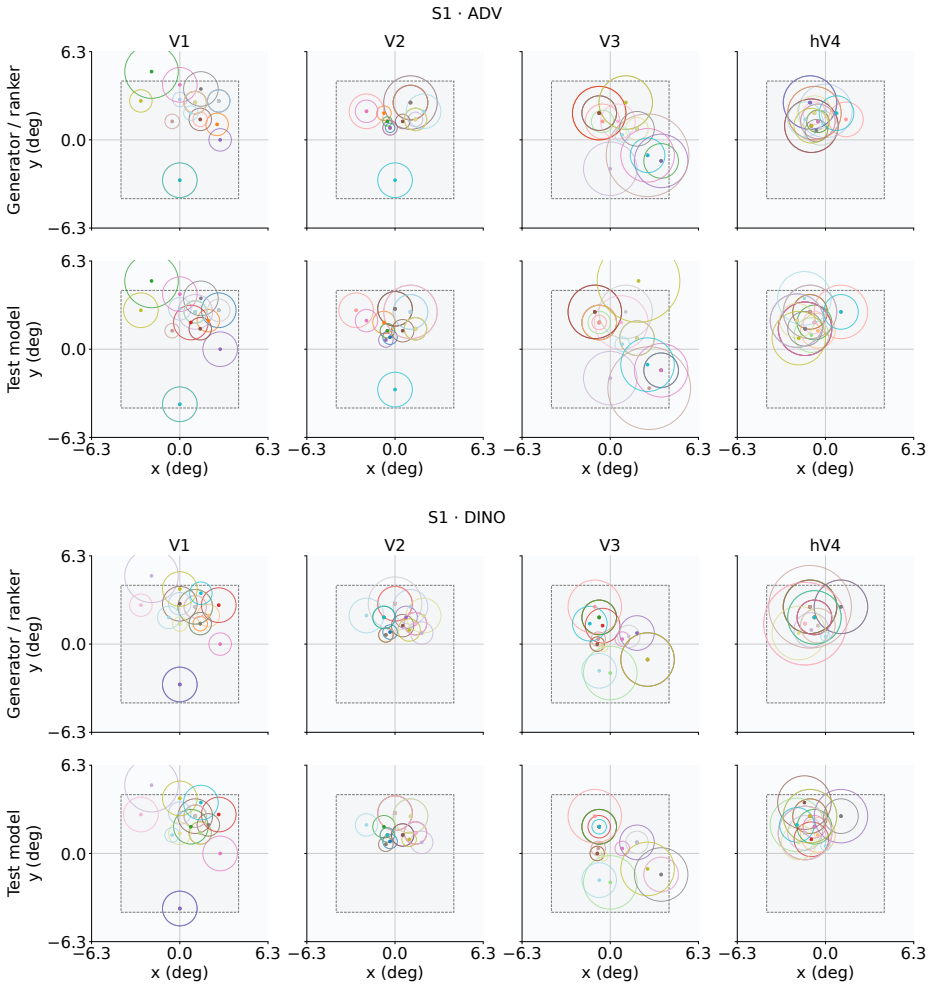

Figure 21: S1 pRF coverage: ADV above DINO.

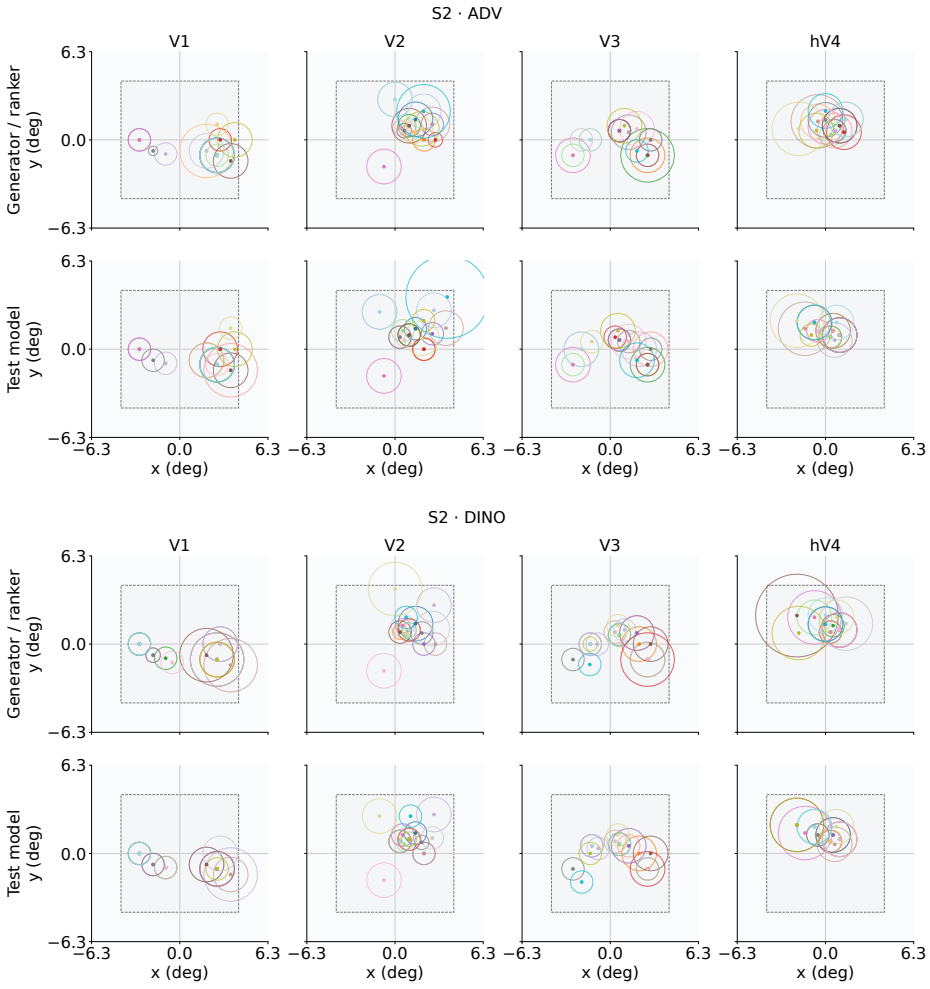

Figure 22: S2 pRF coverage: ADV above DINO.

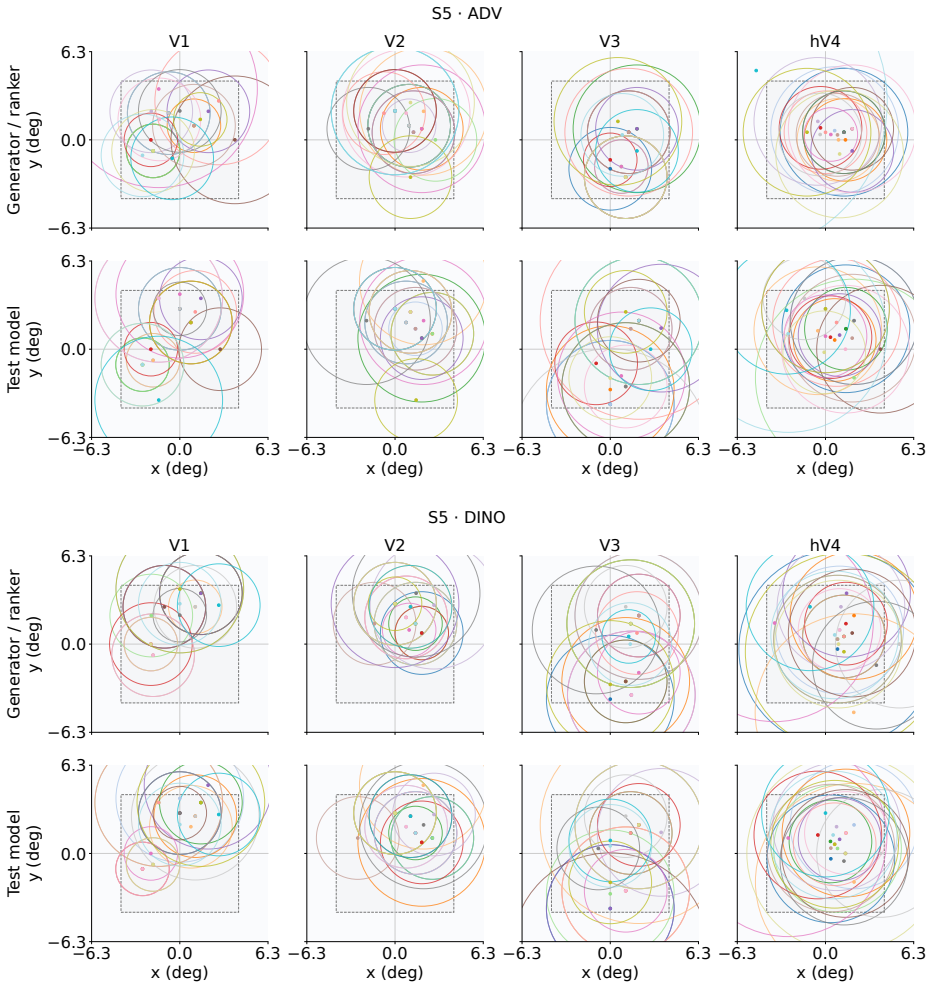

Figure 23: S5 pRF coverage: ADV above DINO.

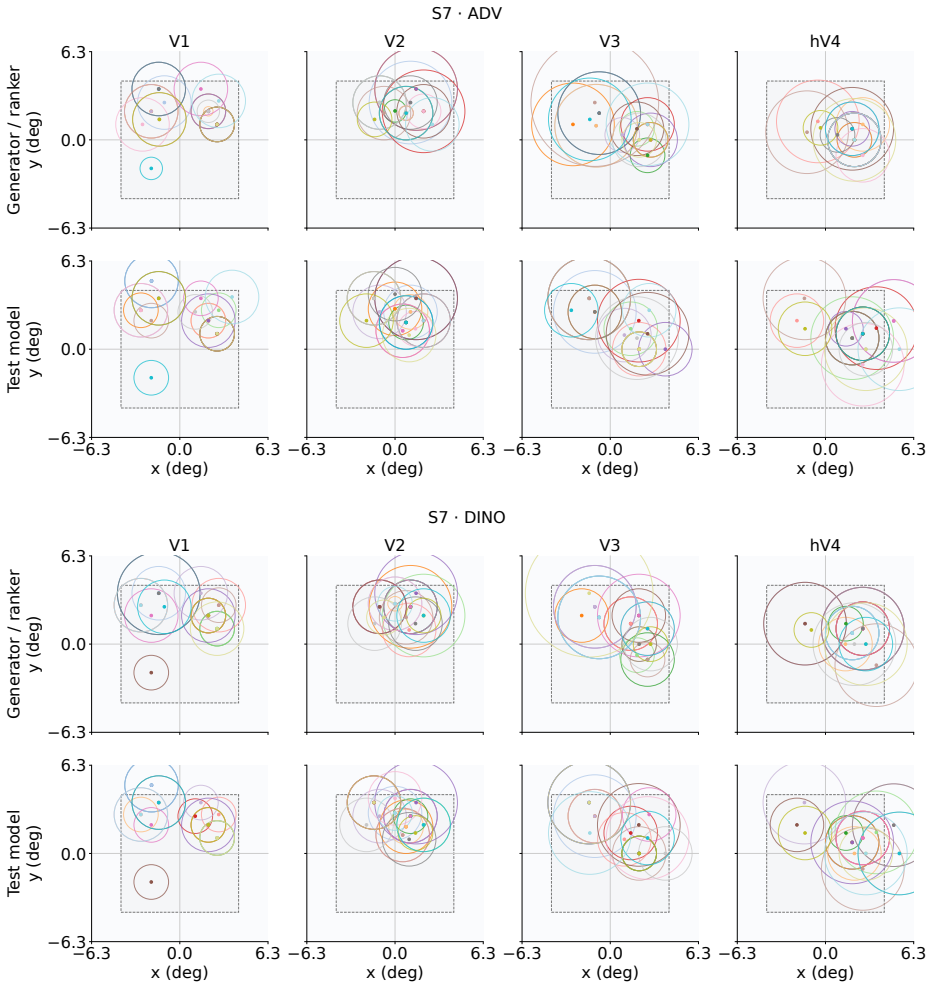

Figure 24: S7 pRF coverage: ADV above DINO.

A.6 PREDICTED RESPONSES ACROSS SUBJECTS AND SELECTION CONDITIONS

Rows show generation, ranking, and independent test encoder predictions; columns show V1, V2, V3, and hV4. Each point is one voxel's mean prediction across ten selected images, and short horizontal bars indicate means across available voxels. Blue denotes ADV and orange denotes DINO, including their corresponding natural-image readouts. Images from every source are selected by either the ranking or generation encoder, as specified in each caption; the same selected images are then evaluated by all three encoders. Full-image constraints control global mean luma and RMS contrast; full-image + pRF constraints additionally control the corresponding pRF-weighted statistics. Natural controls remain unmodified in both cases. The S1 ranker-selected, full-image result is shown in the main text and is not repeated here.

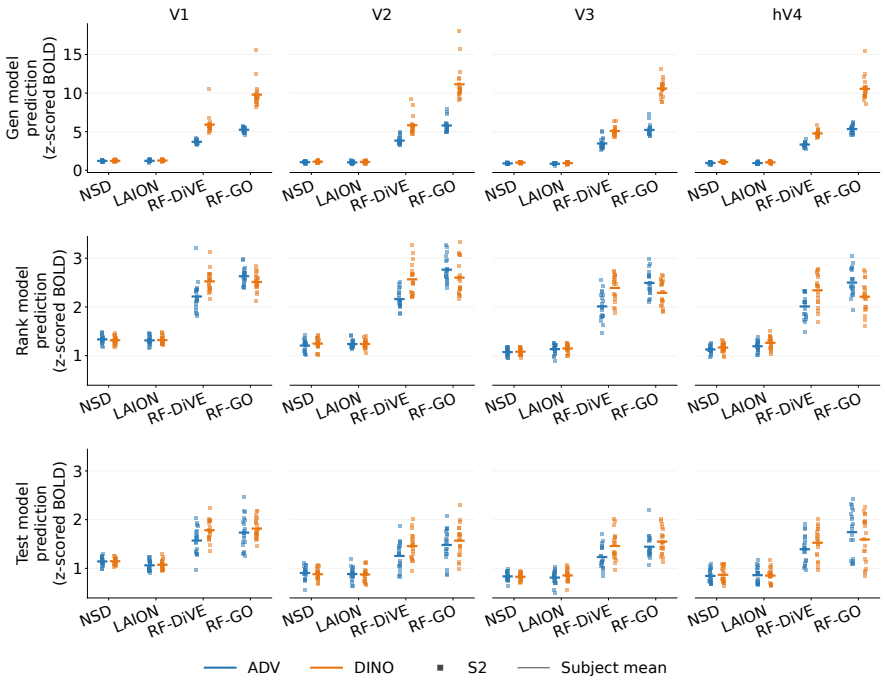

Figure 25: S2 responses: ranker-selected top-10 images; full-image constraints.

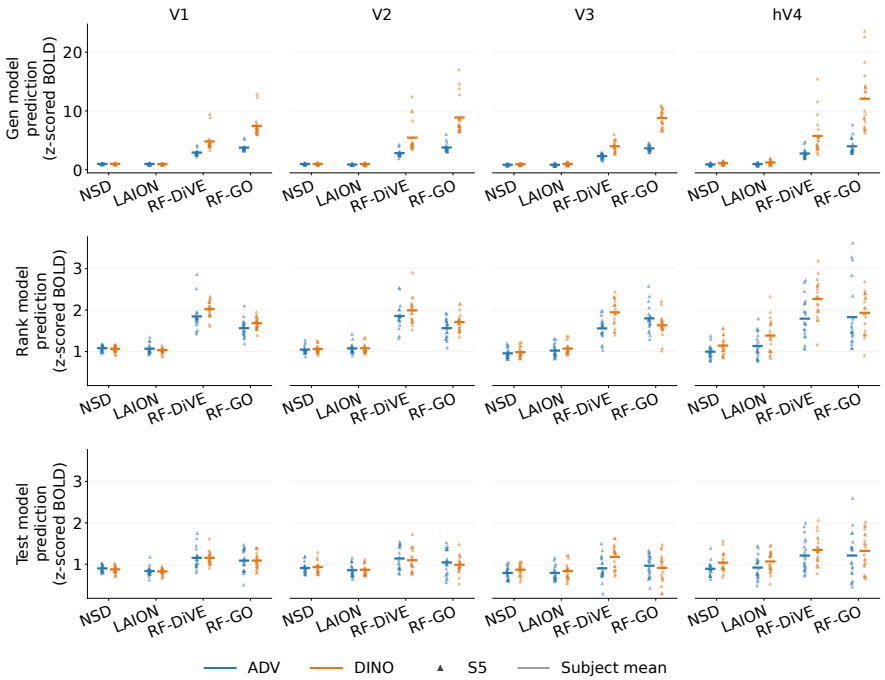

Figure 26: S5 responses: ranker-selected top-10 images; full-image constraints.

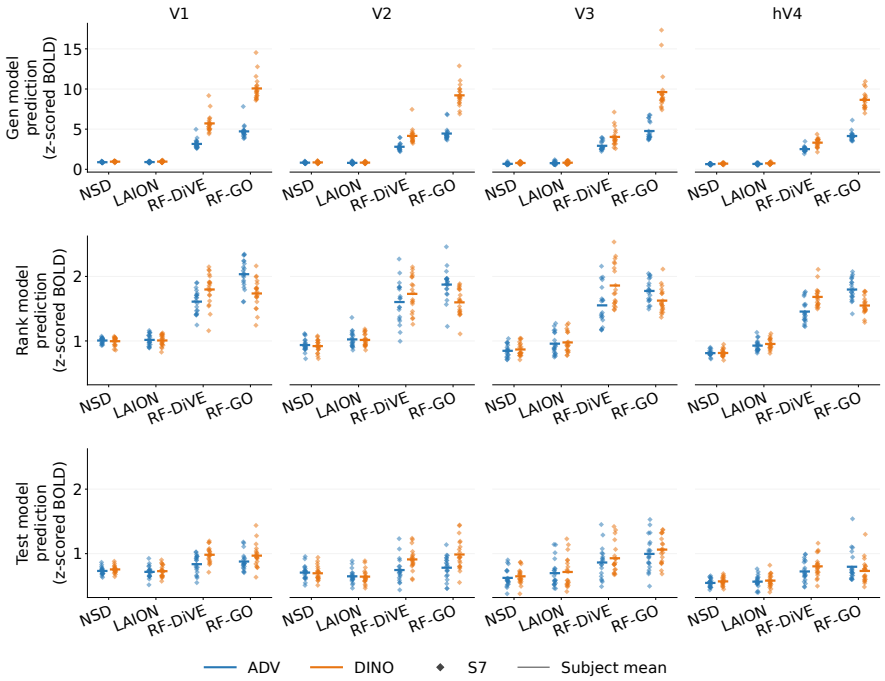

Figure 27: S7 responses: ranker-selected top-10 images; full-image constraints.

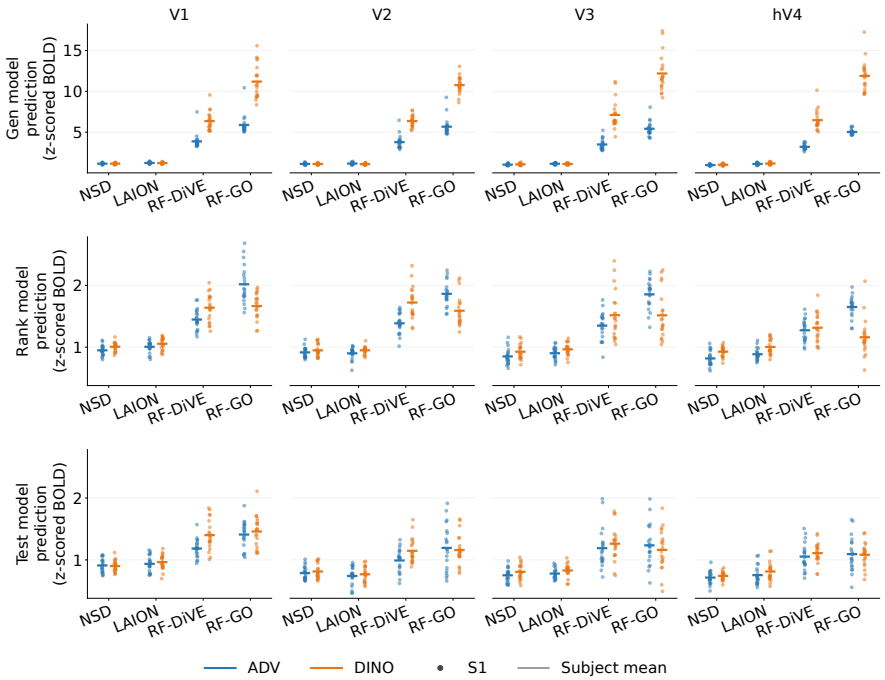

Figure 28: S1 responses: generator-selected top-10 images; full-image constraints.

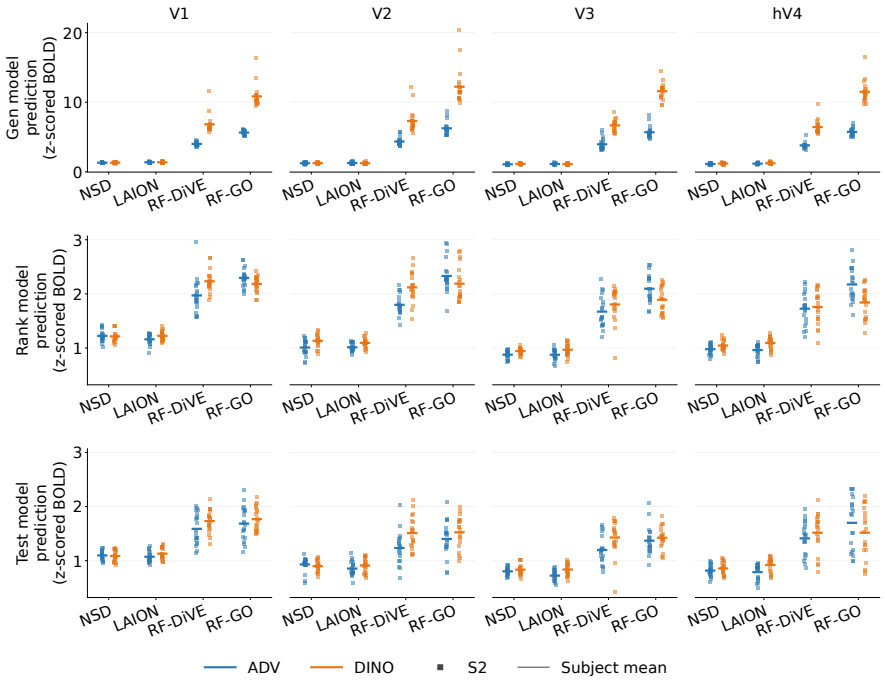

Figure 29: S2 responses: generator-selected top-10 images; full-image constraints.

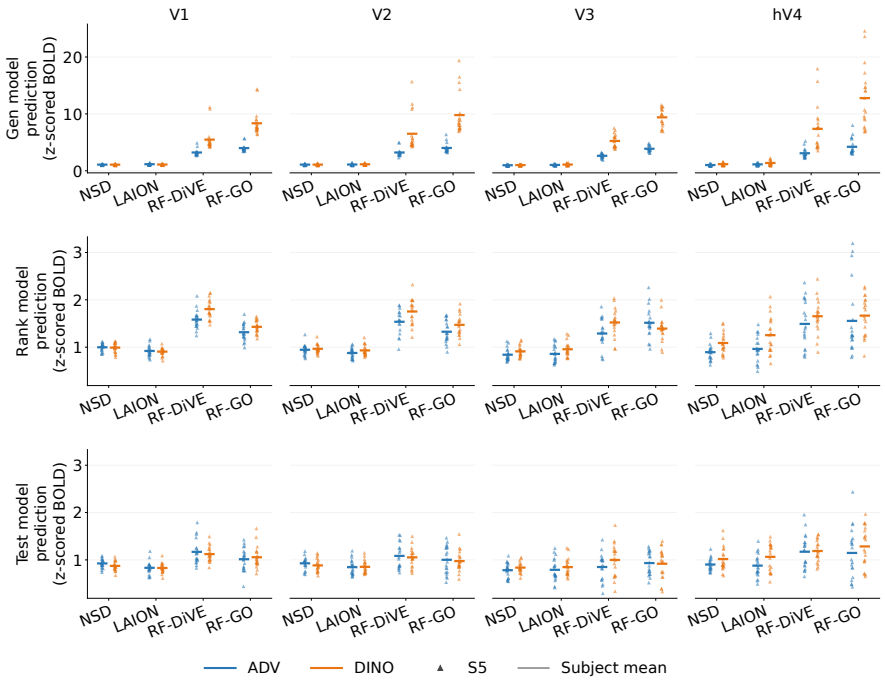

Figure 30: S5 responses: generator-selected top-10 images; full-image constraints.

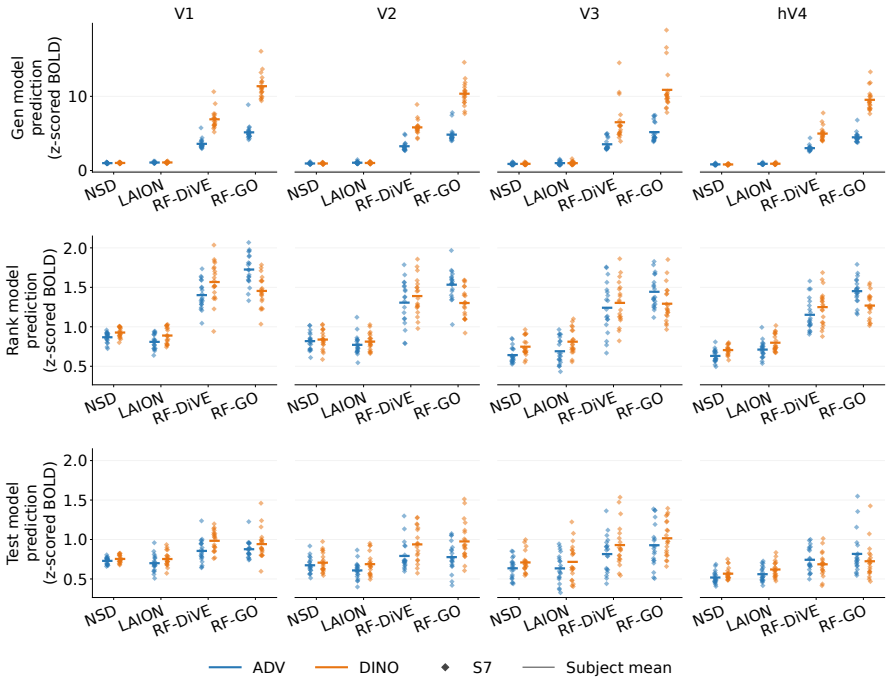

Figure 31: S7 responses: generator-selected top-10 images; full-image constraints.

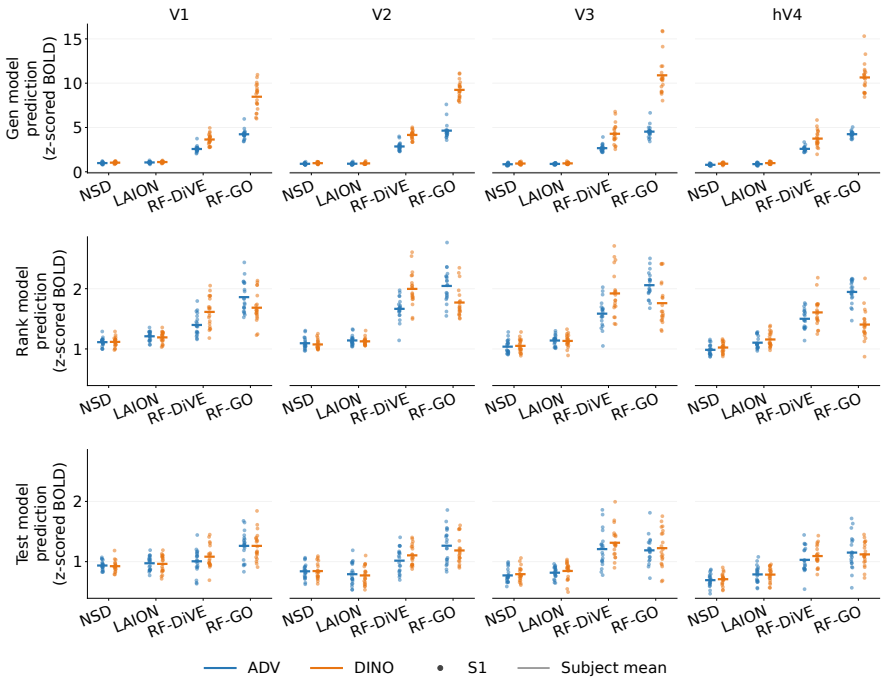

Figure 32: S1 responses: ranker-selected top-10 images; full-image + pRF constraints.

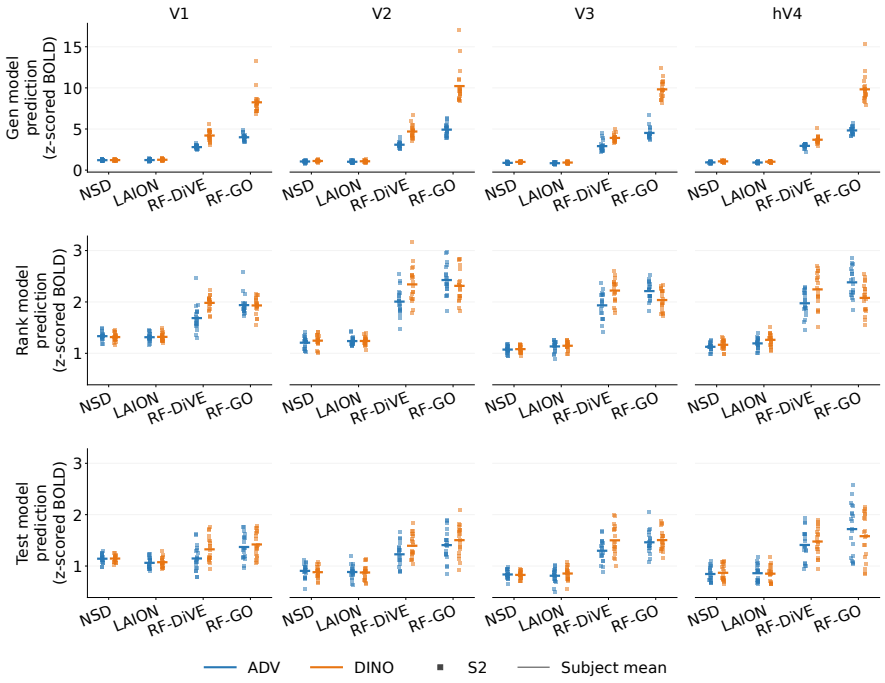

Figure 33: S2 responses: ranker-selected top-10 images; full-image + pRF constraints.

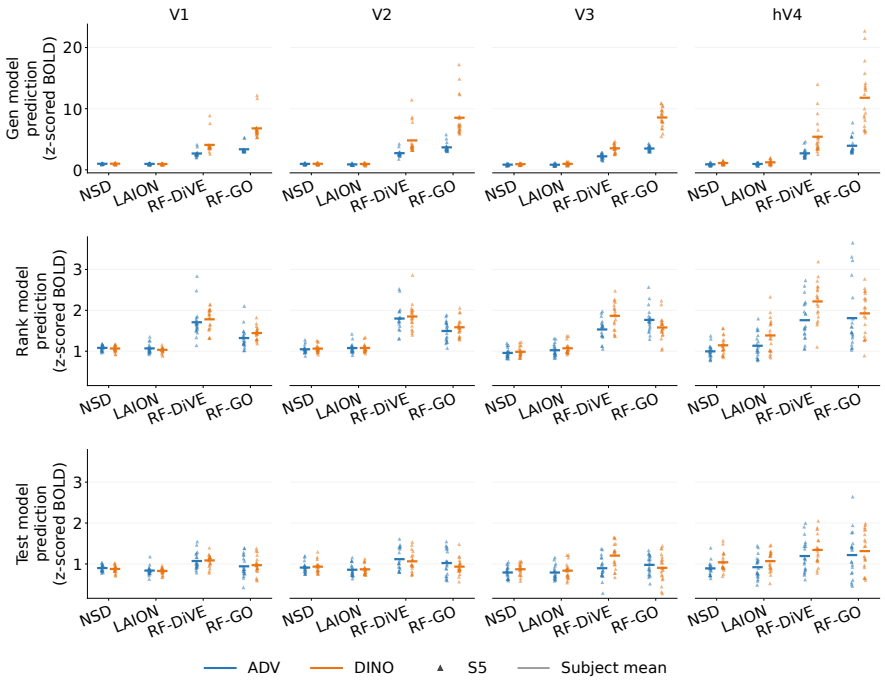

Figure 34: S5 responses: ranker-selected top-10 images; full-image + pRF constraints.

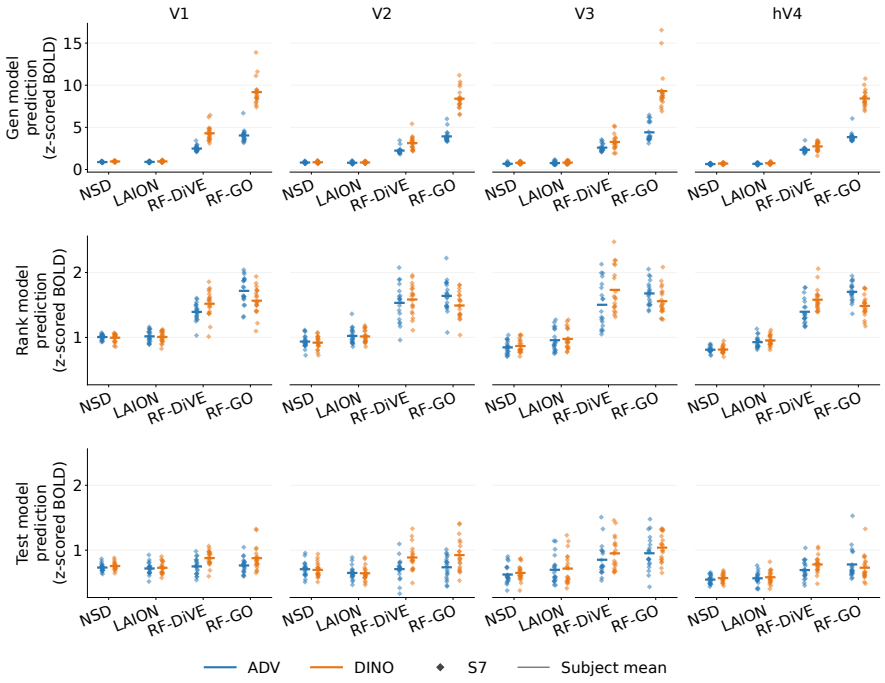

Figure 35: S7 responses: ranker-selected top-10 images; full-image + pRF constraints.

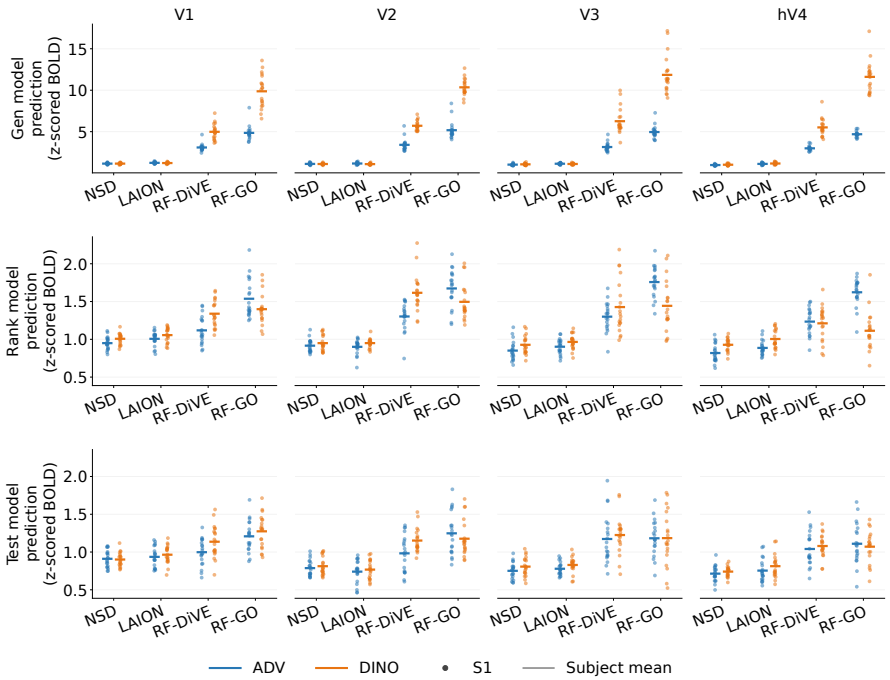

Figure 36: S1 responses: generator-selected top-10 images; full-image + pRF constraints.

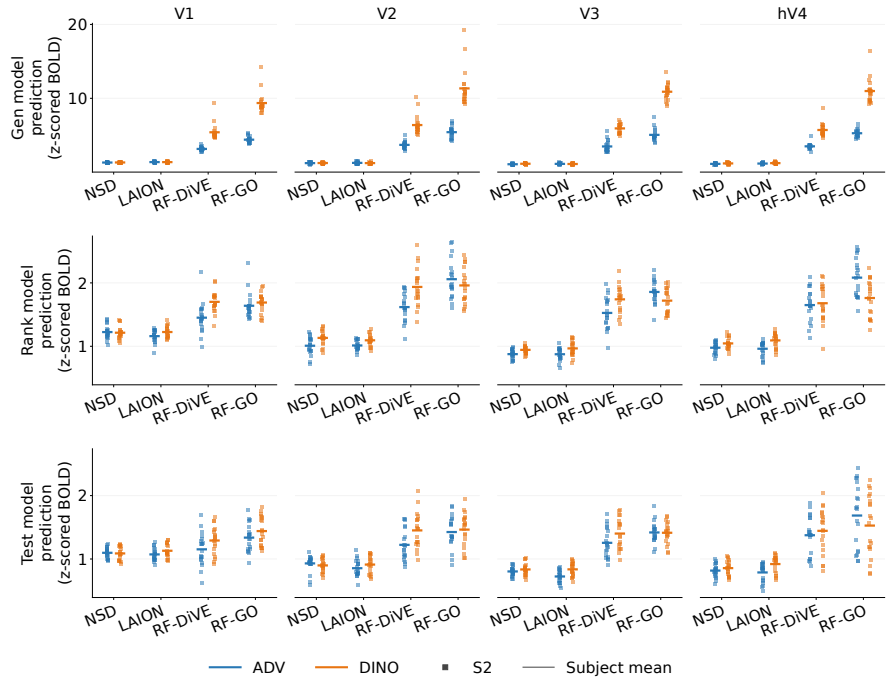

Figure 37: S2 responses: generator-selected top-10 images; full-image + pRF constraints.

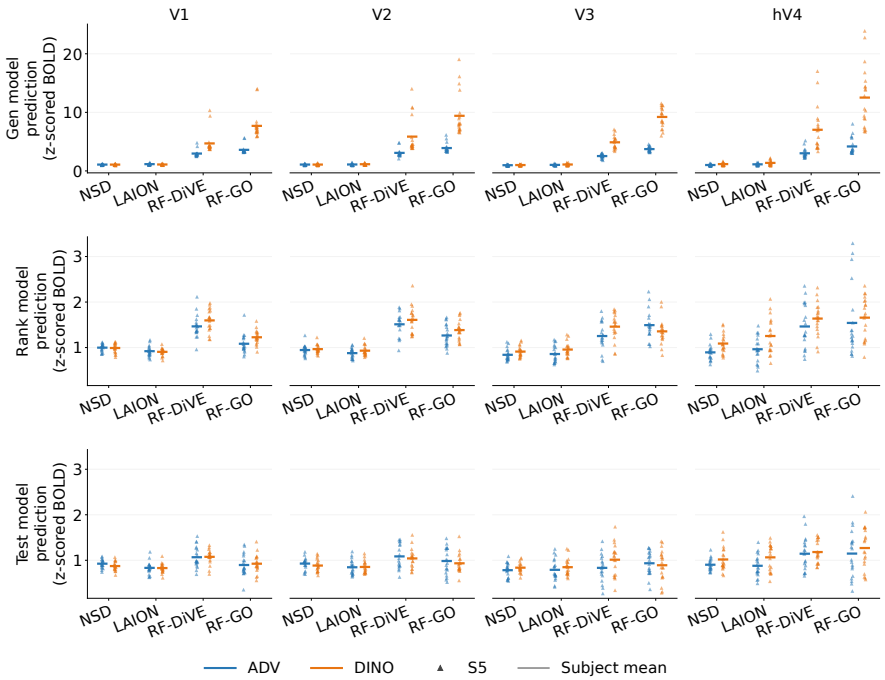

Figure 38: S5 responses: generator-selected top-10 images; full-image + pRF constraints.

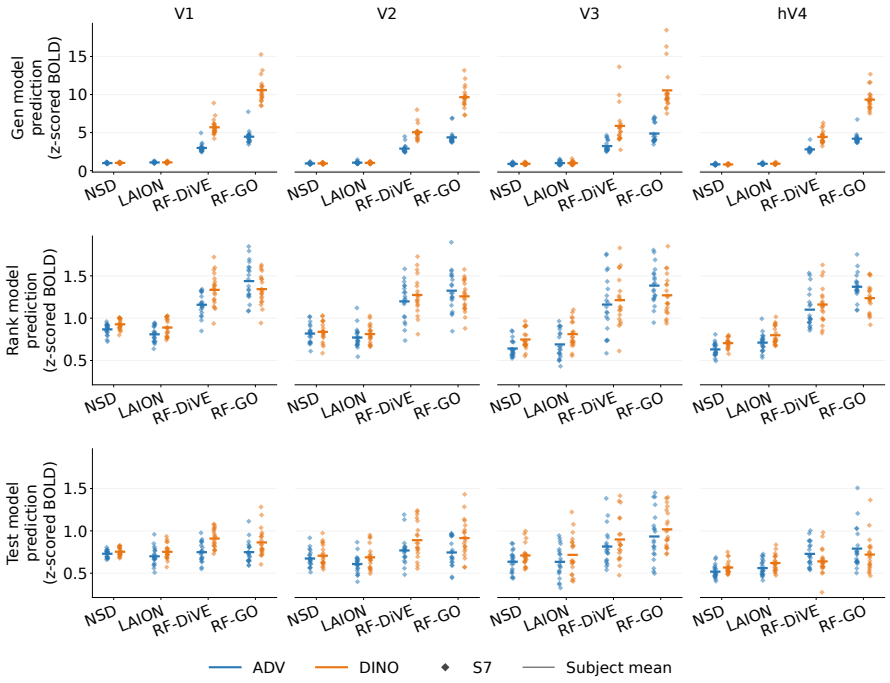

Figure 39: S7 responses: generator-selected top-10 images; full-image + pRF constraints.

A.7 PREDICTED-RESPONSE ADVANTAGES OVER LAION

For each voxel, we subtract the mean prediction for the selected LAION images from that for the selected MEIs, using ten images per source and the same evaluation encoder. Points show equal-weight means of these voxel-wise differences within each subject and ROI. Positive values favor MEIs. Color denotes backbone; circles, squares, triangles, and diamonds denote S1, S2, S5, and S7, respectively. Lines connect backbone summaries within a subject and method; the backbone-specific voxel selections need not be identical. The ranker-selected, full-image result appears in the main text; its three complementary conditions follow.

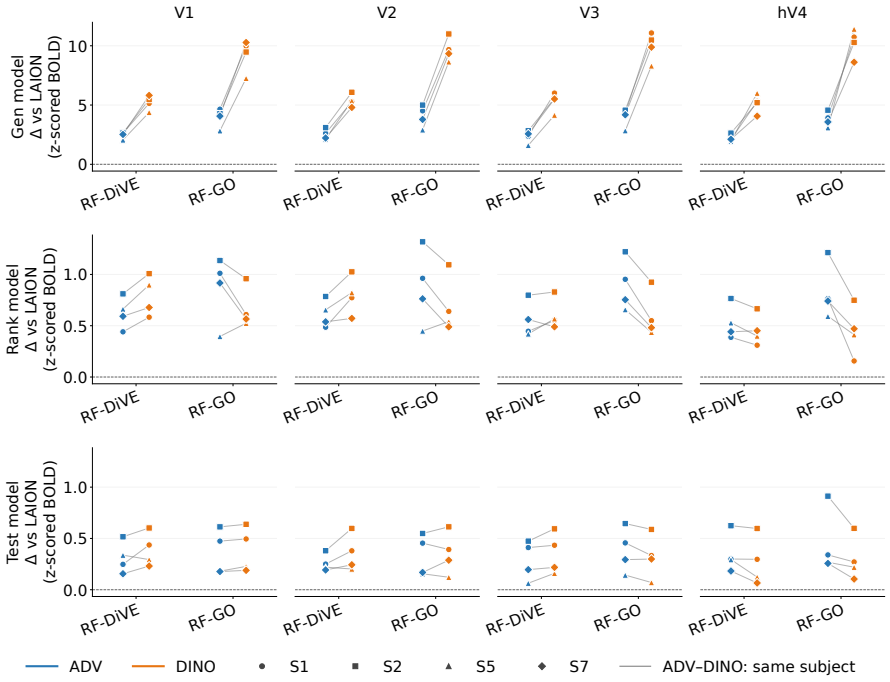

Figure 40: MEI–LAION response differences: generator-selected top-10 images; full-image constraints.

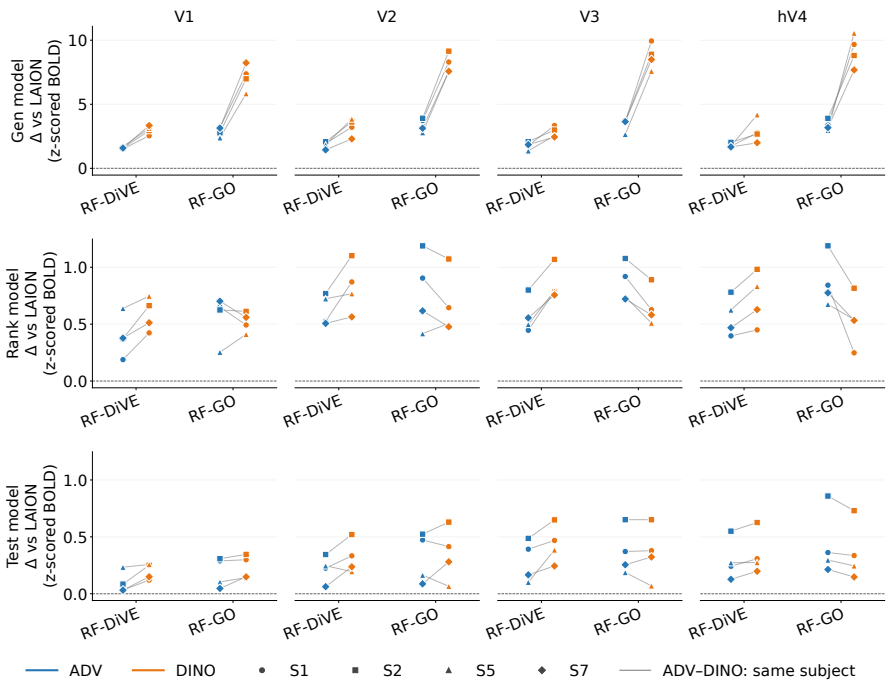

Figure 41: MEI–LAION response differences: ranker-selected top-10 images; full-image + pRF constraints.

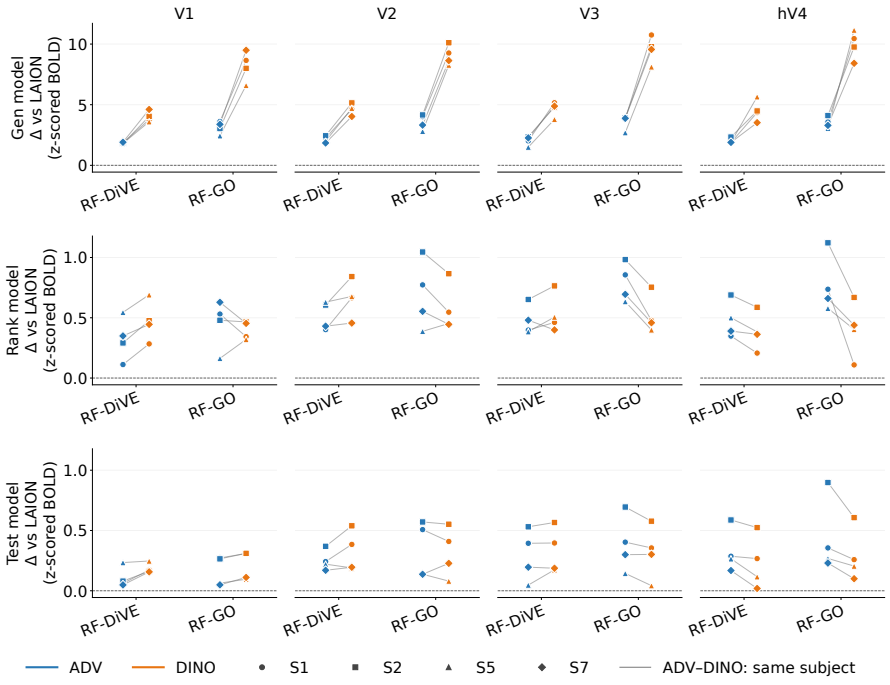

Figure 42: MEI–LAION response differences: generator-selected top-10 images; full-image + pRF constraints.

A.8 ADDITIONAL CHROMATIC AND ORIENTATION STATISTICS

We analyzed the top-10 ranker-selected images per voxel, using the full-image final-constraint condition for generated images and unmodified natural images. CIELAB L^*, a^*, b^* (D65) and HSV saturation were computed per pixel and averaged with normalized generator-pRF weights w_p. Saturation was $S_p = (\max_c I_{pc} - \min_c I_{pc})/\max_c I_{pc}$, with $S_p = 0$ for black pixels; thus the image statistic was $\sum_p w_p S_p$. For orientation, unit-L^2-normalized Gabor filters were applied to the full image before weighted pooling. For pooled magnitudes $A_i(f,\theta)$, each image profile was $p_i(\theta) = \sum_f A_i(f,\theta)/\sum_{f,\theta} A_i(f,\theta)$. Statistics were averaged over images within voxel, then available matched voxels within subject (18–20 per subject, backbone, and ROI). These are descriptive stimulus statistics, not direct neuronal tuning estimates.

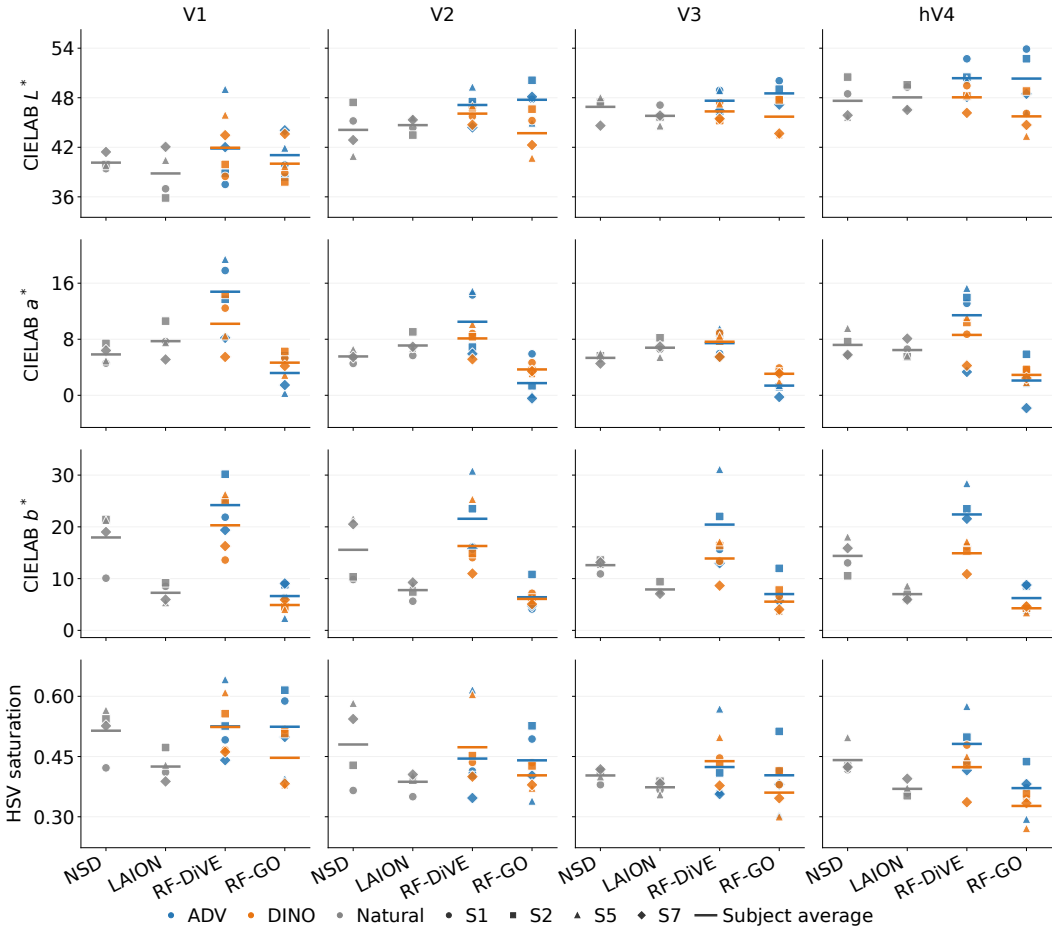

Figure 43: CIELAB components and HSV saturation across subjects and visual areas. Points denote subject means; short bars denote means across subjects. Natural references pool the two backbone-conditioned summaries.

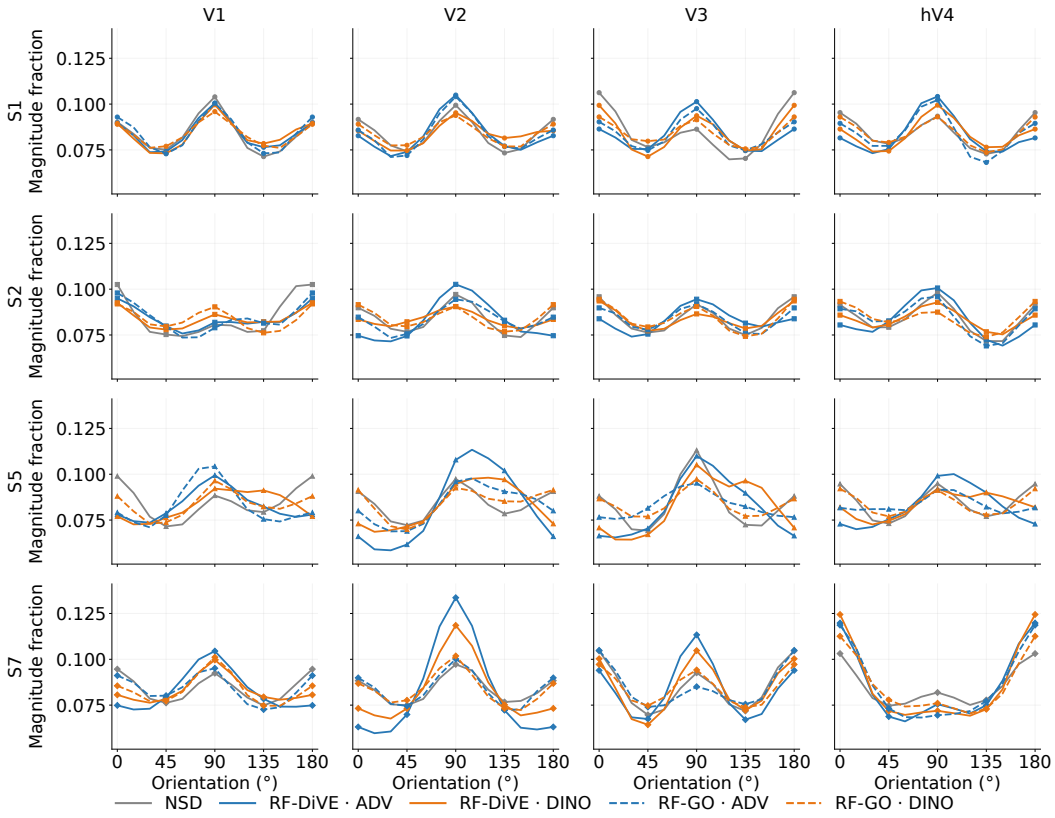

Figure 44: Normalized Gabor orientation profiles for each subject and visual area, using generator-pRF pooling. Solid and dashed colored lines denote RF-DiVE and RF-GO; gray denotes the pooled NSD reference.

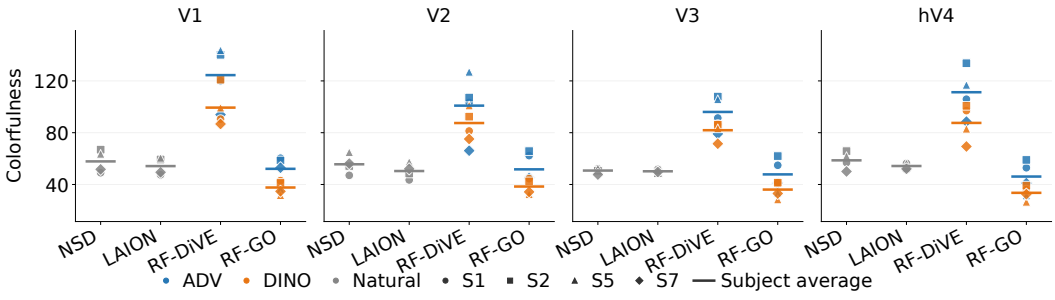

Figure 45: Colorfulness across subjects and visual areas, using generator-pRF weights. Points denote subject means; bars denote means across subjects.

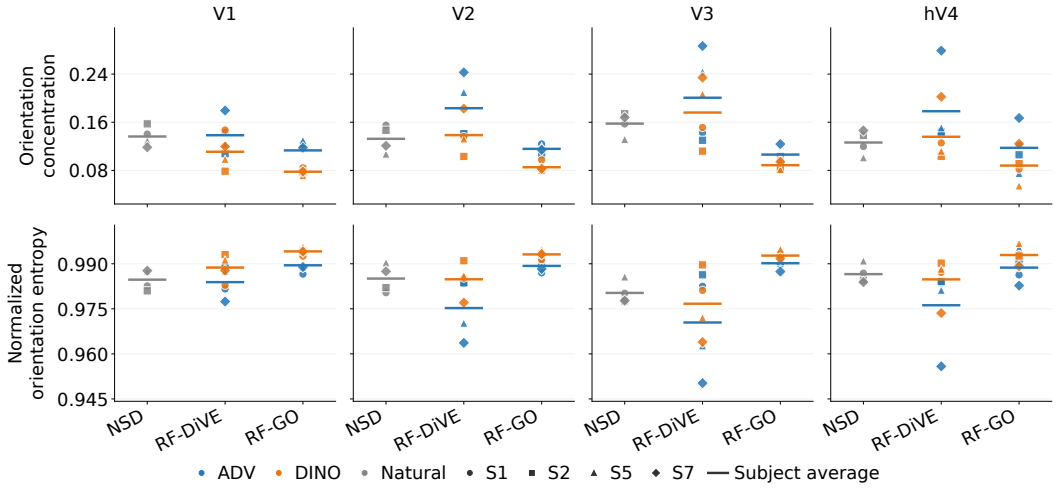

Figure 46: Per-image orientation concentration and normalized entropy, using generator-pRF weights. Points denote subject means; bars denote means across subjects.

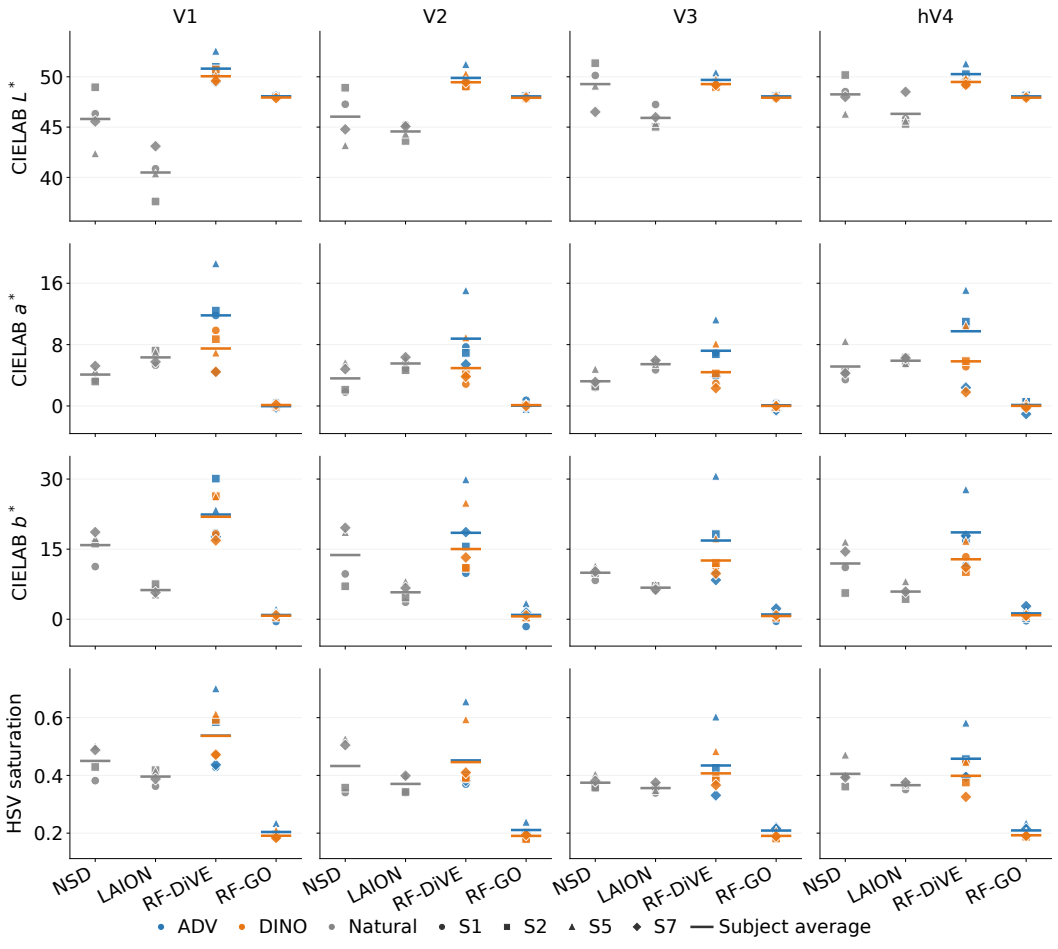

Figure 47: CIELAB components and HSV saturation with uniform full-field analysis weights. Points denote subject means; bars denote means across subjects.

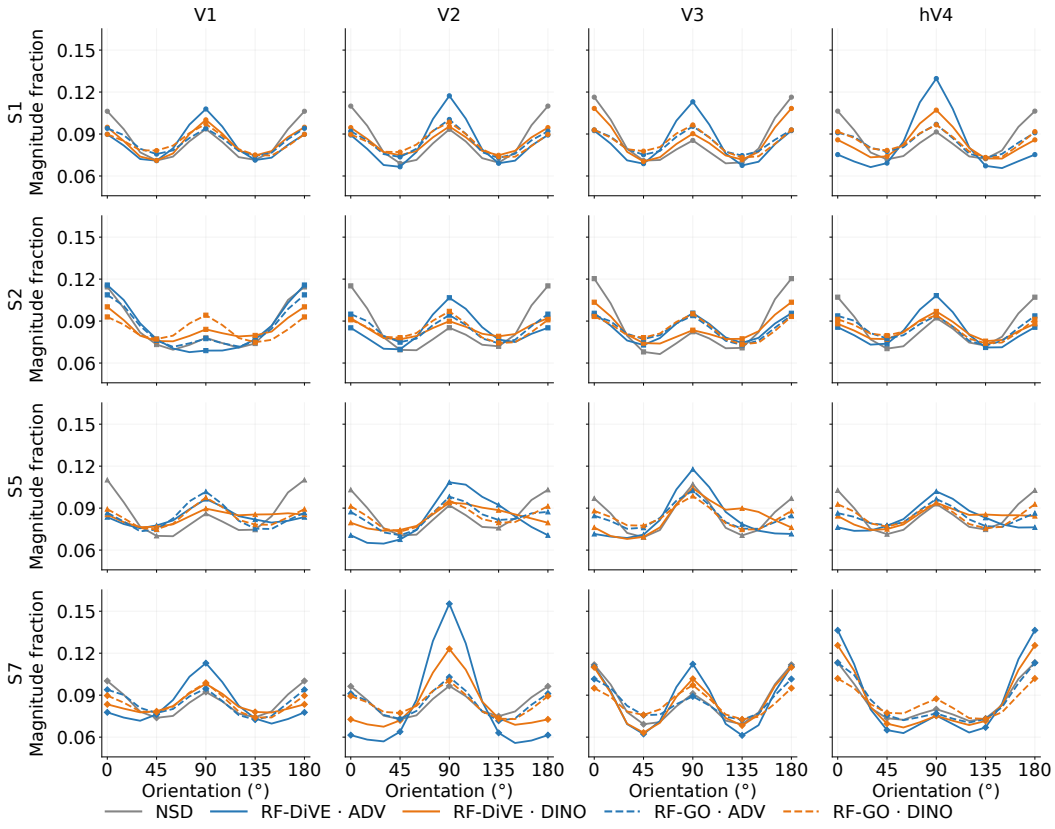

Figure 48: Normalized full-field Gabor orientation profiles for each subject and visual area. Solid and dashed colored lines denote RF-DiVE and RF-GO; gray denotes NSD.

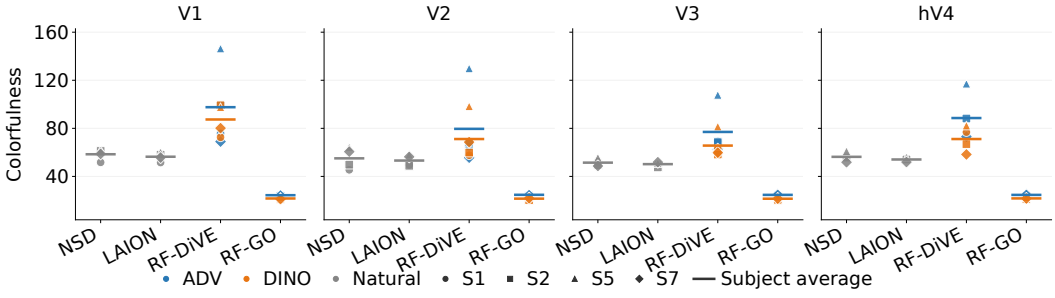

Figure 49: Colorfulness across subjects and visual areas, using uniform full-field weights. Points denote subject means; bars denote means across subjects.

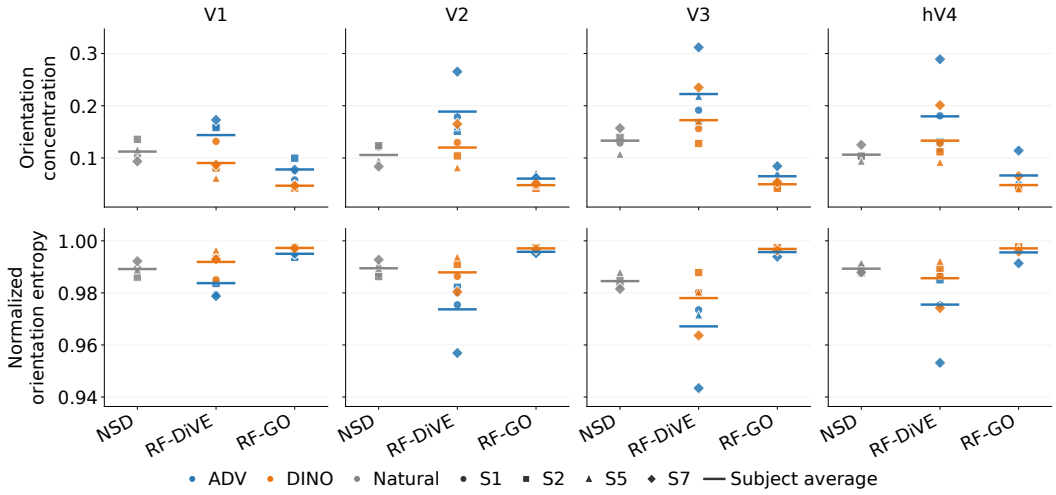

Figure 50: Per-image orientation concentration and normalized entropy with full-field weights. Points denote subject means; bars denote means across subjects.

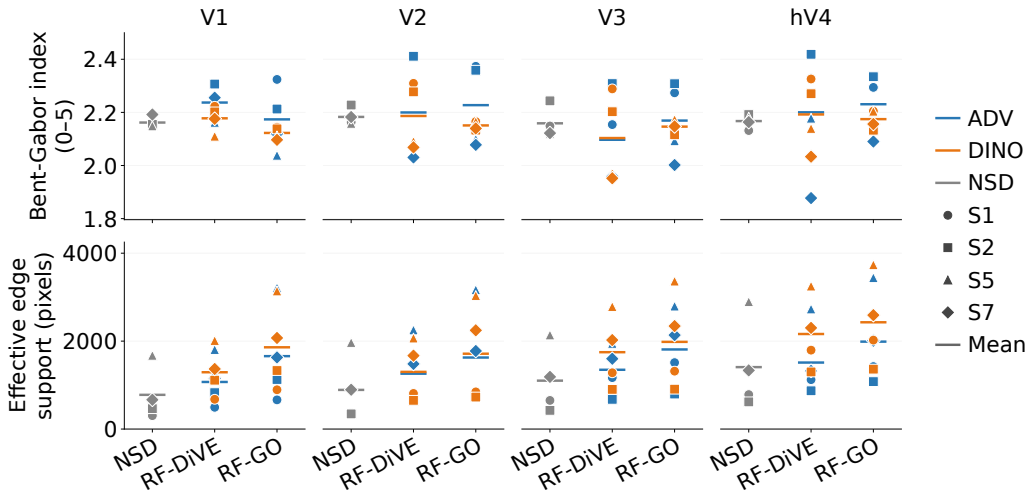

Figure 51: Bent-Gabor curvature index and effective edge support across visual areas. Points denote subject means of voxel-wise top-10 image statistics; bars denote four-subject means. Statistics use generator soft-pRF weights.

A.9 Additional structural image comparisons

We summarize ranker-selected top-10 images from the full-image final-constraint condition, with unmodified NSD images as natural controls. These descriptive comparisons supplement the main chromatic and spatial-frequency analyses.

Structural image statistics. For each voxel, the ranker-selected images form an array of shape $[10, 224, 224, 3]$ with RGB in $[0, 1]$. Filters are applied to each full image before pooling with normalized generator soft-pRF weights w_p ($\sum_p w_p = 1$). Canny edge fraction is $\sum_p w_p E_p$, where E_p is the binary edge map (Rec.709 luma; smoothing $\sigma = 1$ pixel; thresholds $0.1/0.2$). GLCM joint entropy is $-\sum_{a,b} P_{ab} \log_2 P_{ab}$, averaged over offsets $(0,1), (1,0), (1,1), (1,-1)$, using 32 luma levels and symmetric co-occurrence probabilities weighted by $\sqrt{w_p w_{p+d}}$. The bent-Gabor index is $\sum_p w_p M_p q_p / \sum_p w_p M_p$, where $q_p \in \{0, \ldots, 5\}$ is the winning bend-bin index among 144 filters and M_p selects pixels above the whole-image 90th-percentile Roberts edge magnitude (BT.601 grayscale). Effective edge support is $(\sum_p w_p M_p)^2 / \sum_p (w_p M_p)^2$. Metrics are averaged over selected images within voxel and then over voxels; the four-subject summaries weight subjects equally. These are descriptive image statistics, not measurements of neural tuning.

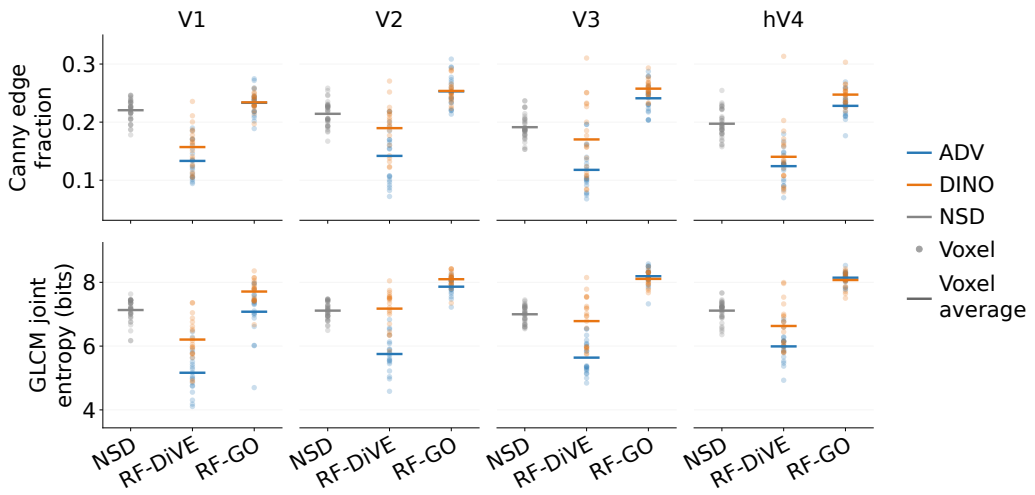

Figure 52: Canny edge fraction and GLCM joint entropy in S1. Pale points denote voxel-wise top-10 image means; bars denote means across 20 voxels per backbone and ROI. Statistics use generator soft-pRF weights.

A.10 BEHAVIORAL EXPERIMENT TO TEST DIFFERENCES BETWEEN REGION MEIS

To evaluate which perceptual attributes differ between MEIs targeted to V1 and hV4, we collected human behavioral data using an online experiment conducted on Pavlovia, with participants recruited through Prolific (London, UK). We recruited 35 distinct participants for each of two task versions, and following exclusions due to incomplete data and quality control, we retained 32 participants in each task version. All participants provided informed consent, in accordance with a human subject protocol approved by the Institutional Review Board at the authors' institution.

In the experiment, participants performed trials of a two-alternative forced choice task comparing images that were targeted to V1 or hV4. On each trial, 10 images from a V1 voxel and 10 images from a hV4 voxel were shown, each presented in a 2x5 grid on the left and right sides of the screen. Below the images, a text question was presented, which asked: "Which images have more three-dimensional form?". Participants responded with keyboard buttons to indicate the left or right image group. The side (left/right) on which V1 and hV4 were shown was counterbalanced across trials. Participants were given a maximum of 10s to respond. On each trial, the images compared were always generated using the same method (RF-DiVE, RF-GO, or top LAION-fMRI images ranked by the "ranking model" encoder), and generated using the same DNN backbone, and were generated from voxels in the same NSD participant. Thus, the comparison isolated the effect of V1 and hV4 within an NSD participant and method. The pairing of which V1 voxel was compared to which hV4 voxel on each trial was randomly assigned using a different order within each participant. The two task versions (see above) referred to whether the participants viewed the full image or a cropped version that isolated the region within 3σ of the voxel's Gaussian pRF.

A.11 ADDITIONAL MEIS

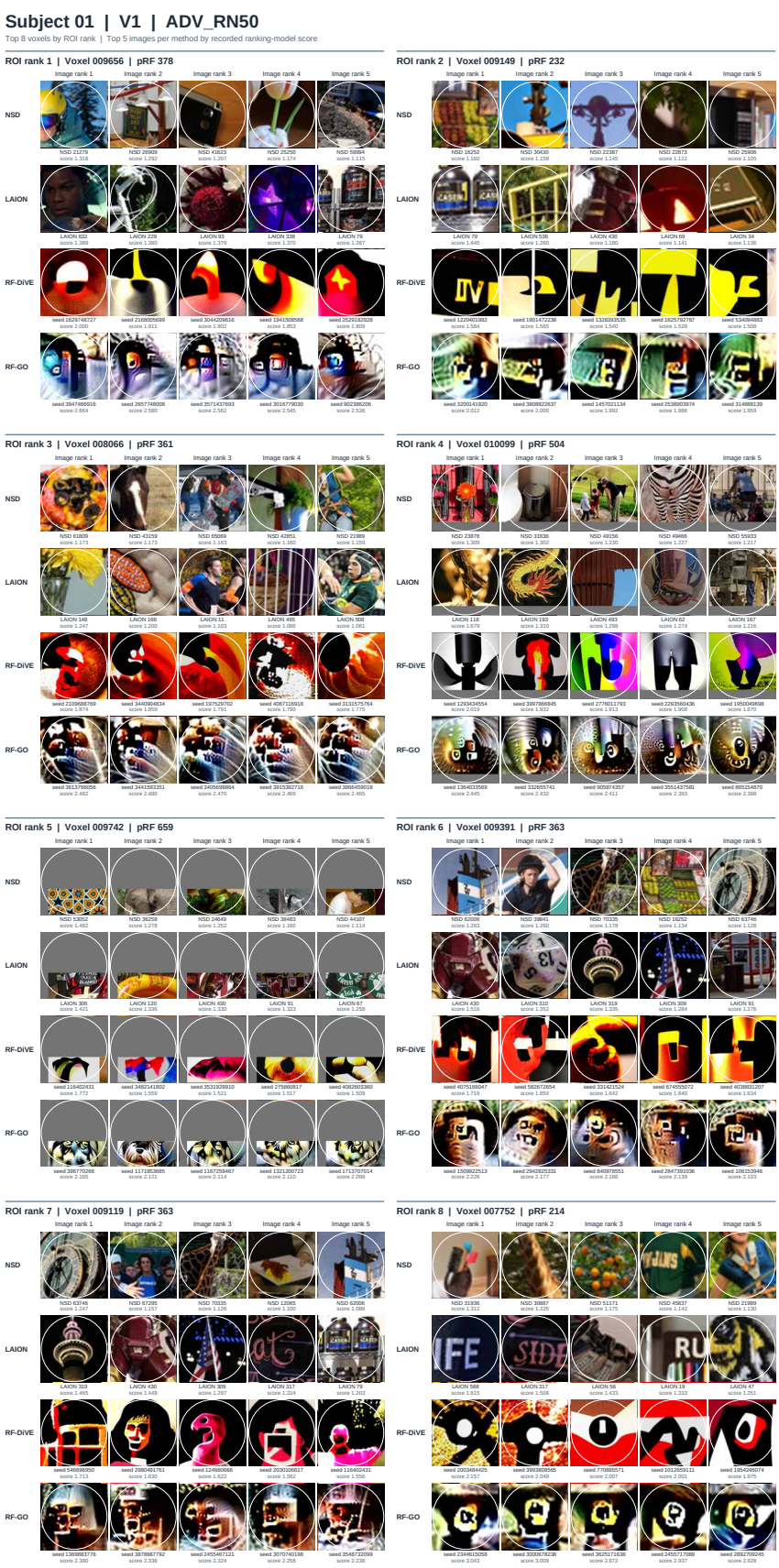

Figure 53: MEIs of top-8 voxels for Subject 01, visual area V1, using the ADV-RN50 brain-encoder backbone.

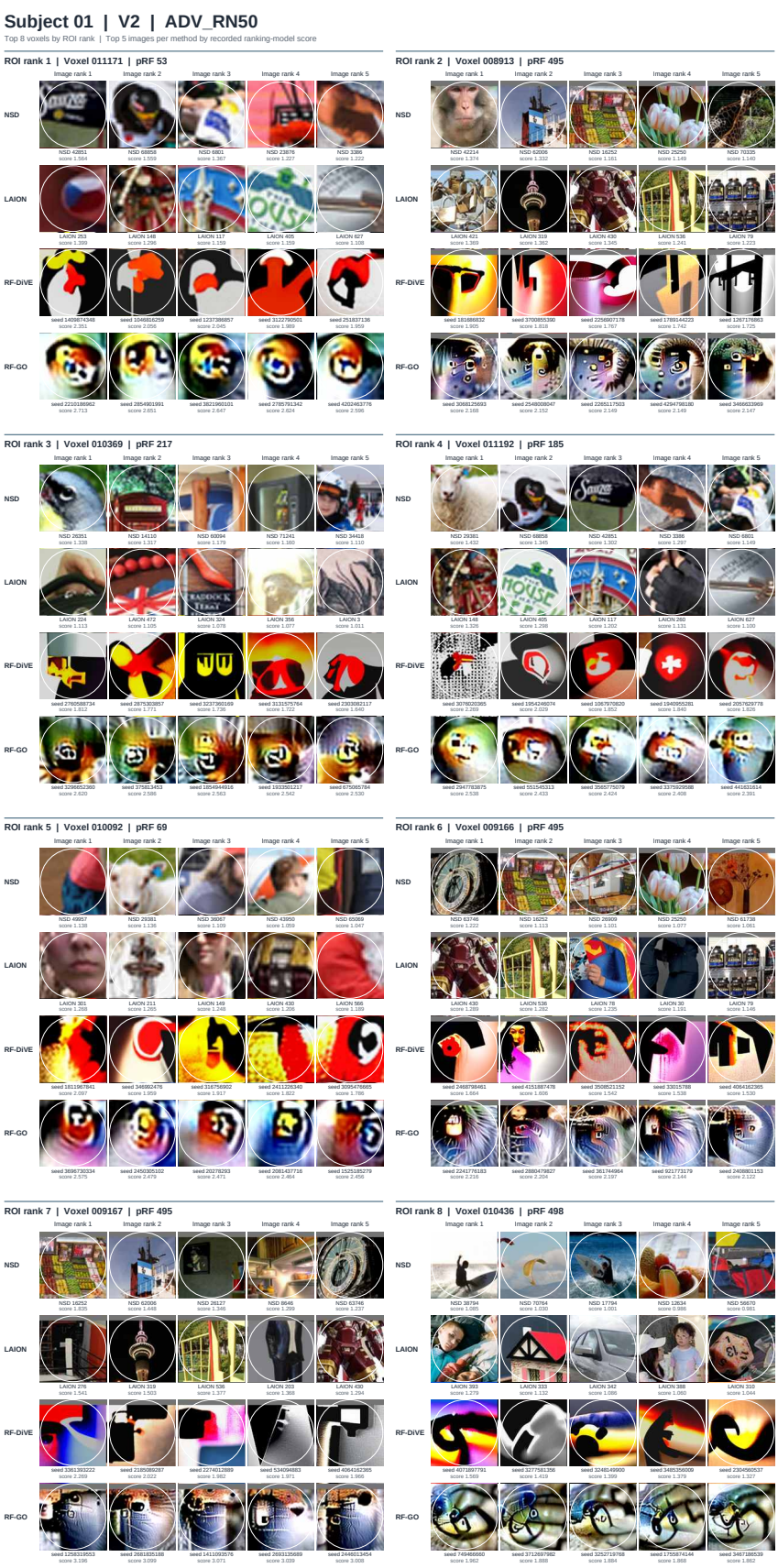

Figure 54: MEIs of top-8 voxels for Subject 01, visual area V2, using the ADV-RN50 brain-encoder backbone.

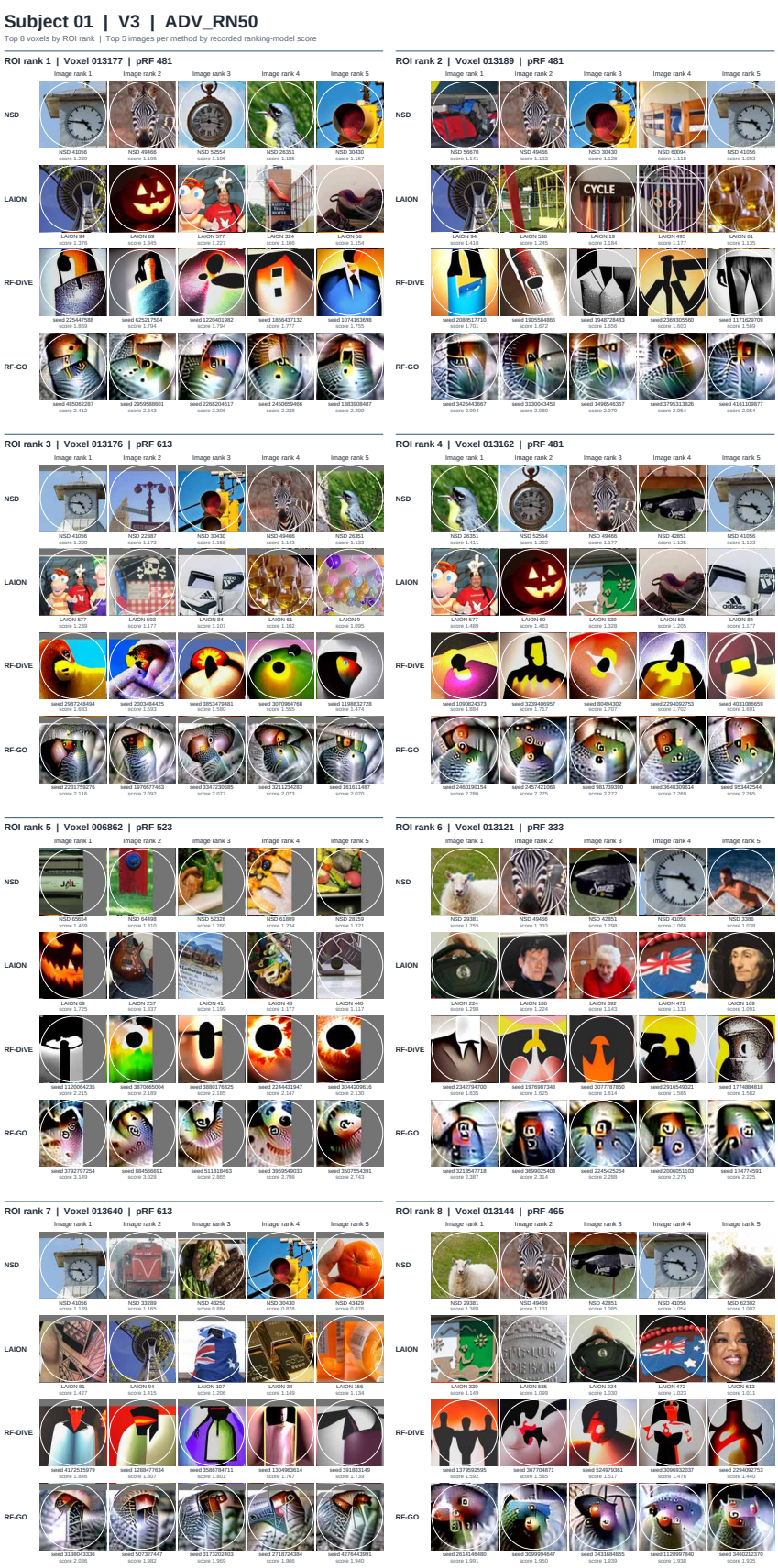

Figure 55: MEIs of top-8 voxels for Subject 01, visual area V3, using the ADV-RN50 brain-encoder backbone.

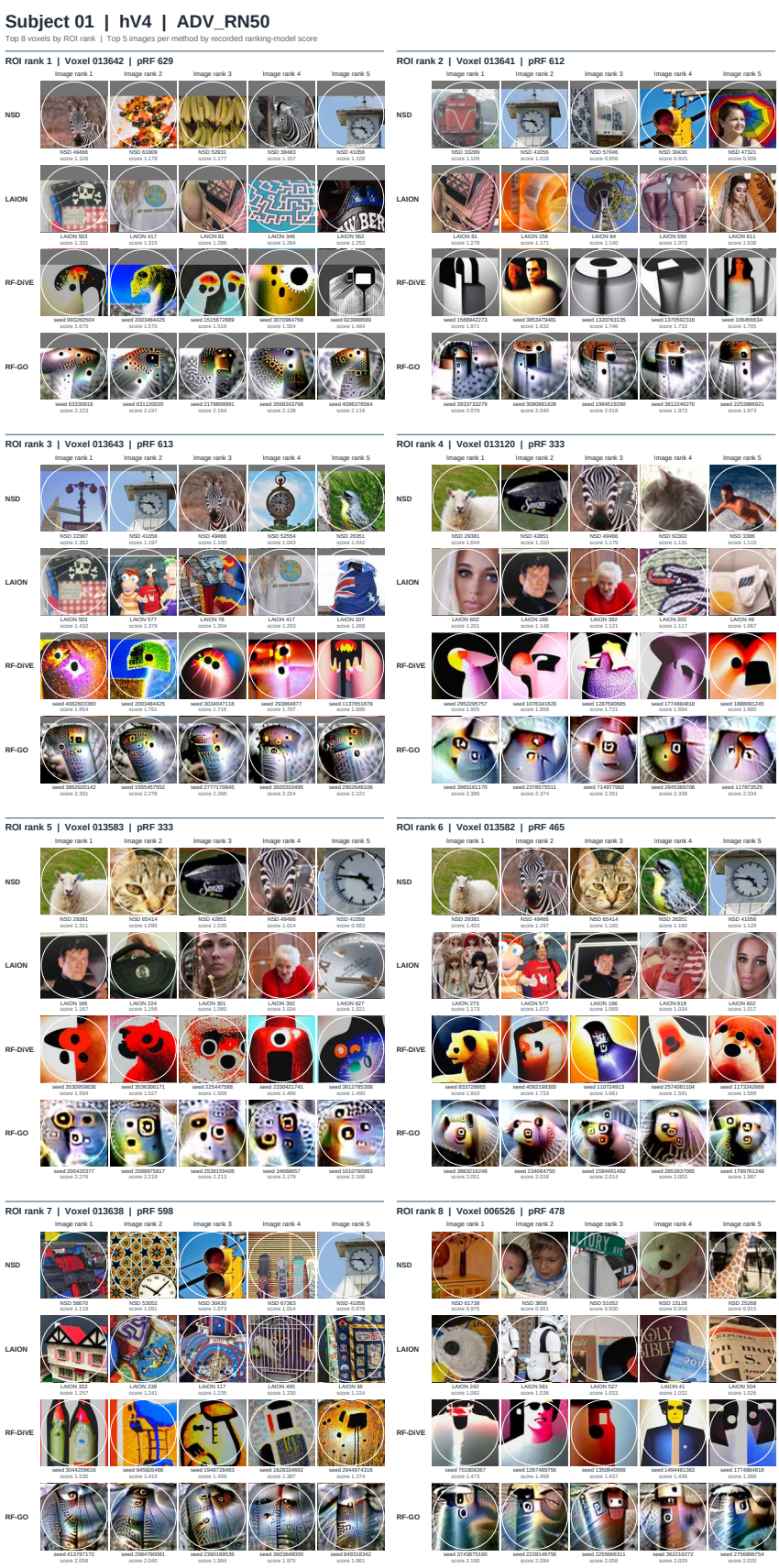

Figure 56: MEIs of top-8 voxels for Subject 01, visual area hV4, using the ADV-RN50 brain-encoder backbone.



Figure 57: MEIs of top-8 voxels for Subject 01, visual area V1, using the DINO-RN50 brain-encoder backbone.

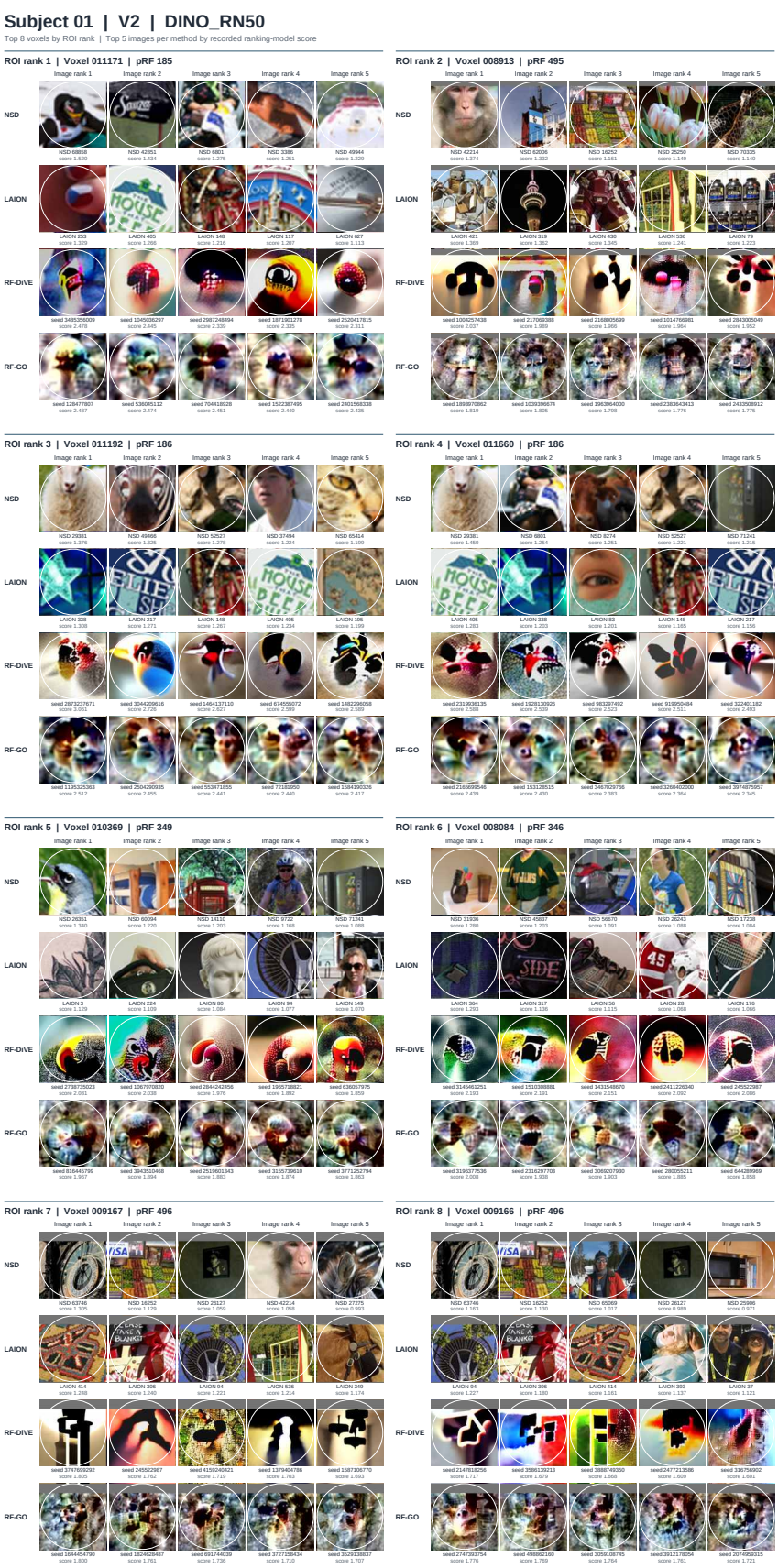

Figure 58: MEIs of top-8 voxels for Subject 01, visual area V2, using the DINO-RN50 brain-encoder backbone.

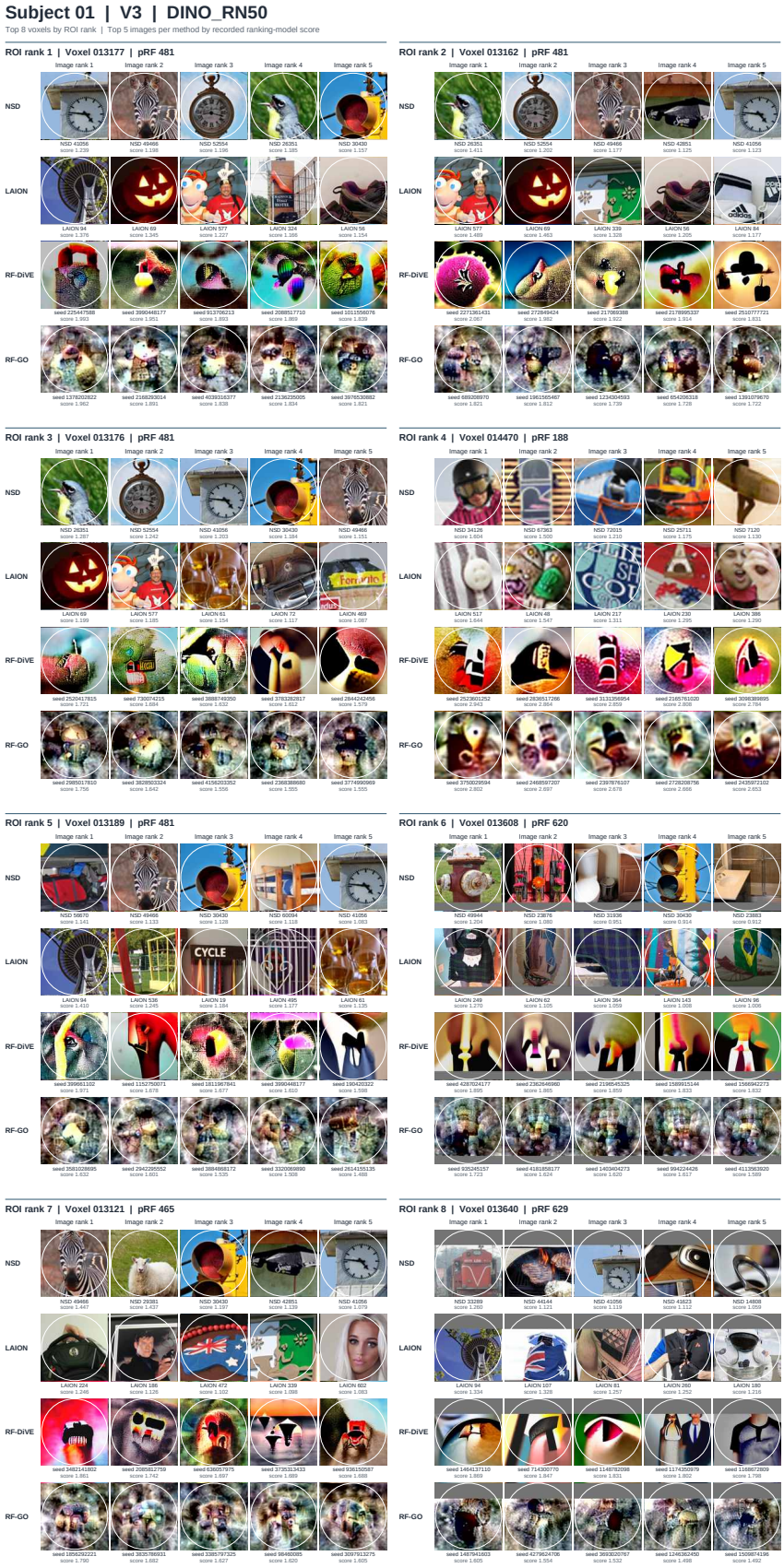

Figure 59: MEIs of top-8 voxels for Subject 01, visual area V3, using the DINO-RN50 brain-encoder backbone.

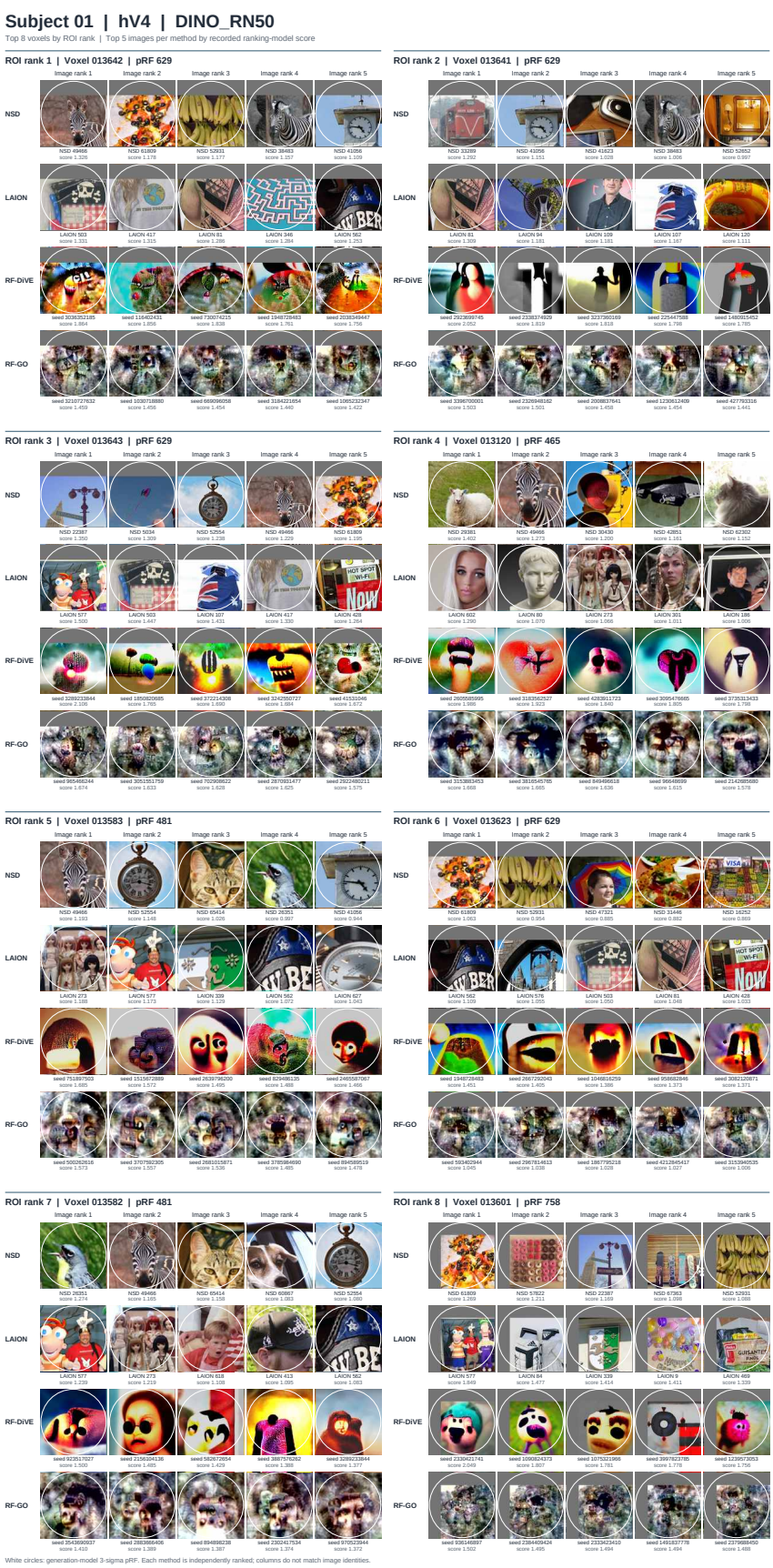

Figure 60: MEIs of top-8 voxels for Subject 01, visual area hV4, using the DINO-RN50 brain-encoder backbone.

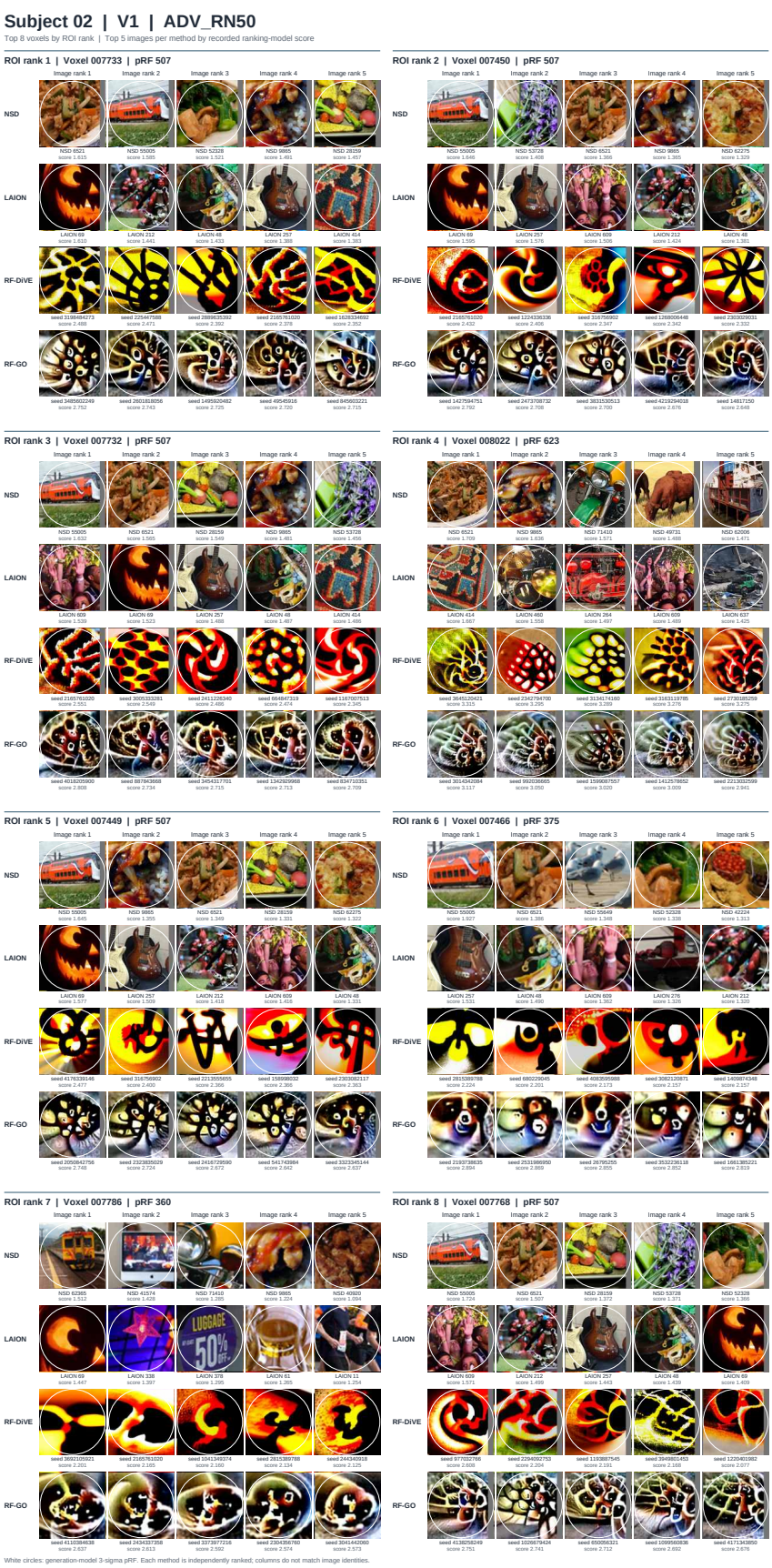

Figure 61: MEIs of top-8 voxels for Subject 02, visual area V1, using the ADV-RN50 brain-encoder backbone.

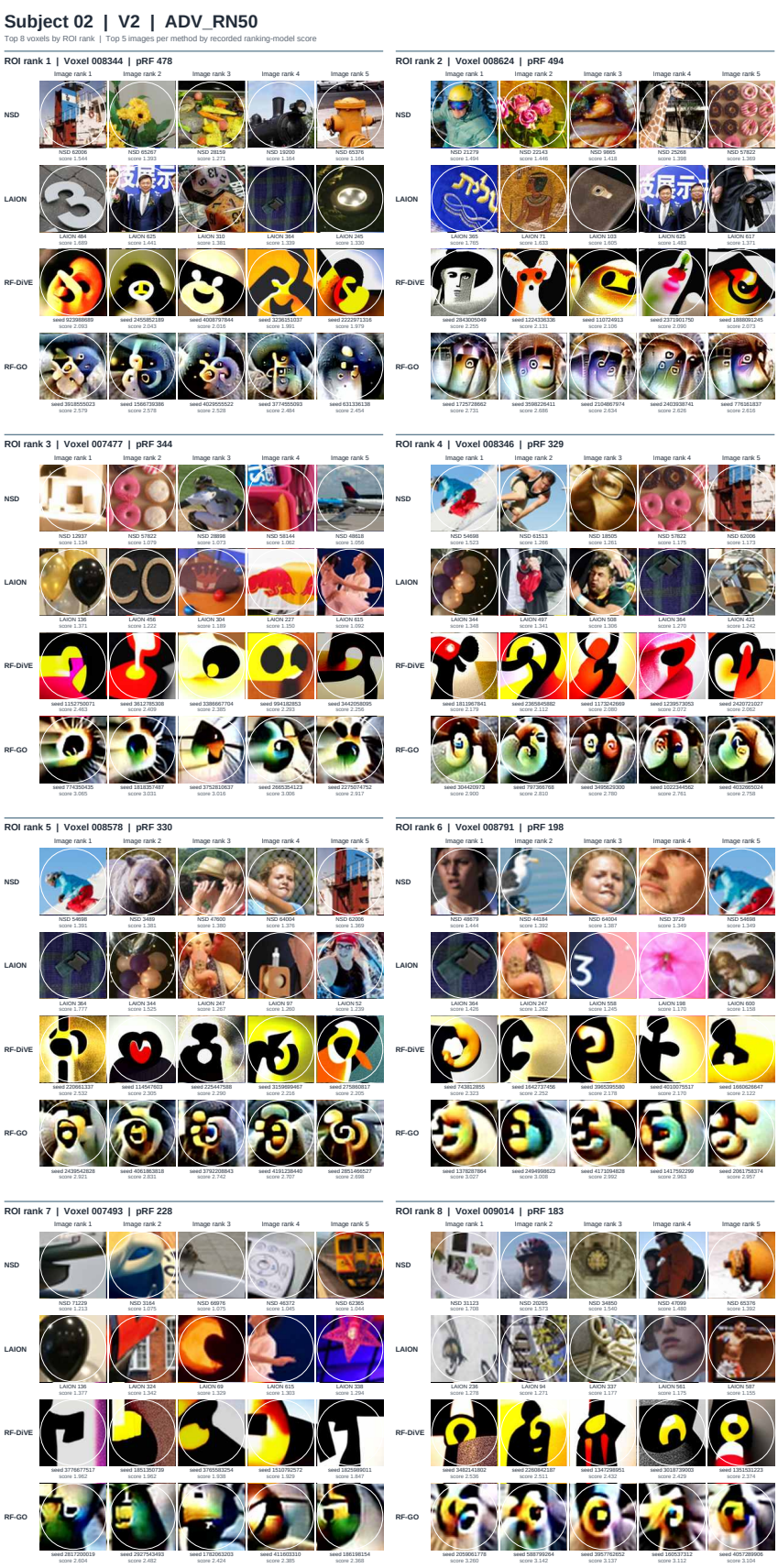

Figure 62: MEIs of top-8 voxels for Subject 02, visual area V2, using the ADV-RN50 brain-encoder backbone.

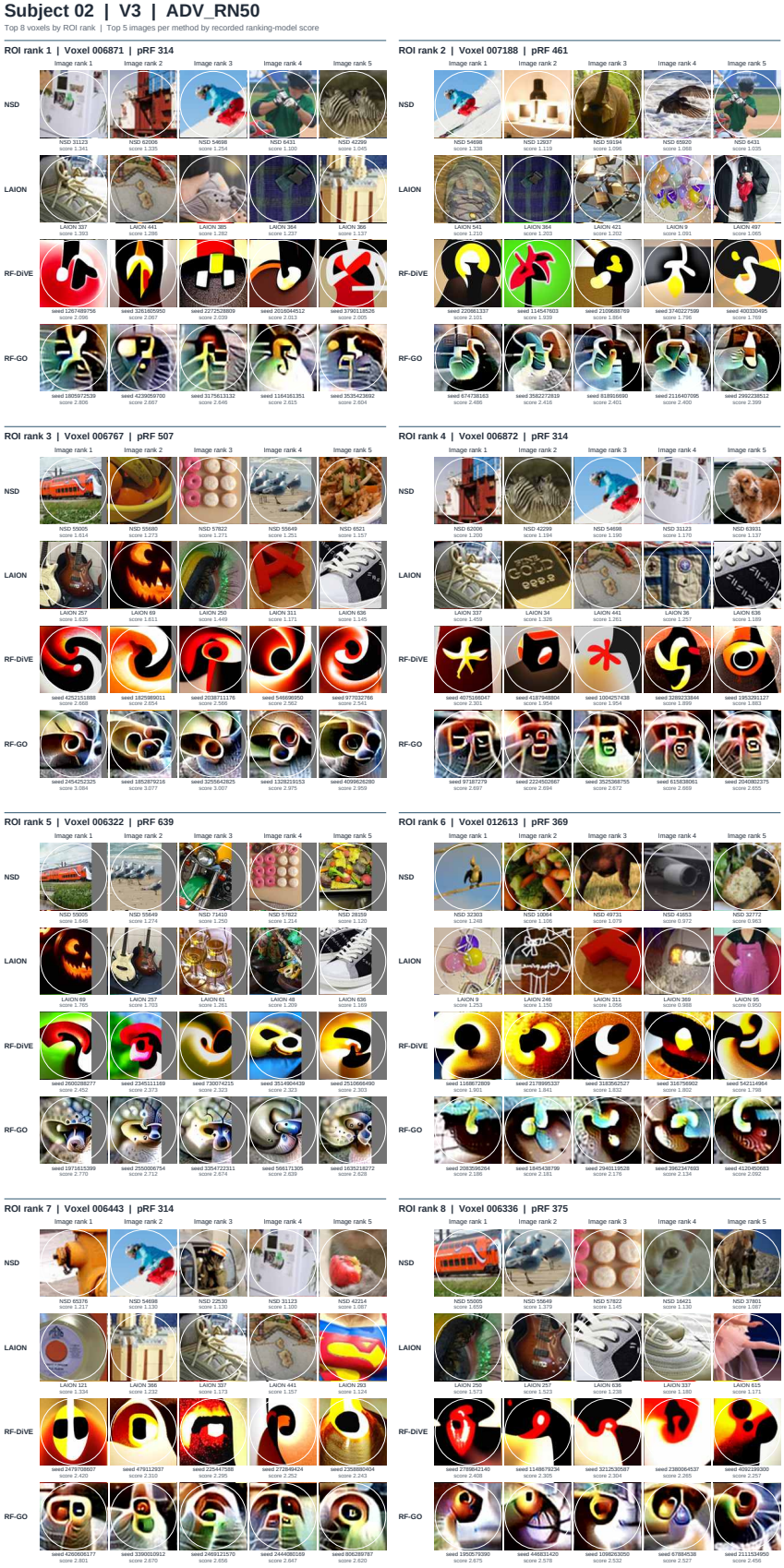

Figure 63: MEIs of top-8 voxels for Subject 02, visual area V3, using the ADV-RN50 brain-encoder backbone.

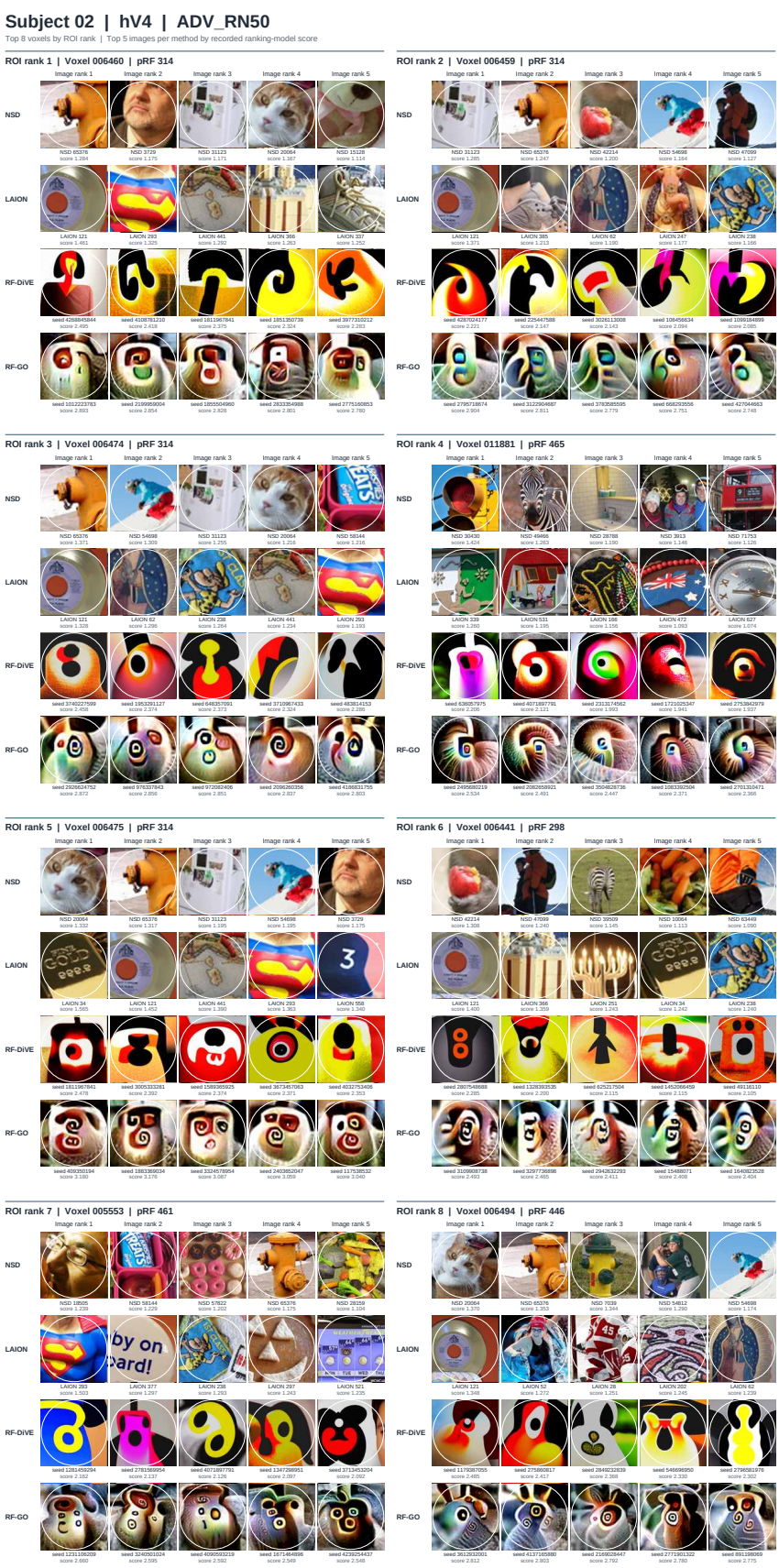

Figure 64: MEIs of top-8 voxels for Subject 02, visual area hV4, using the ADV-RN50 brain-encoder backbone.

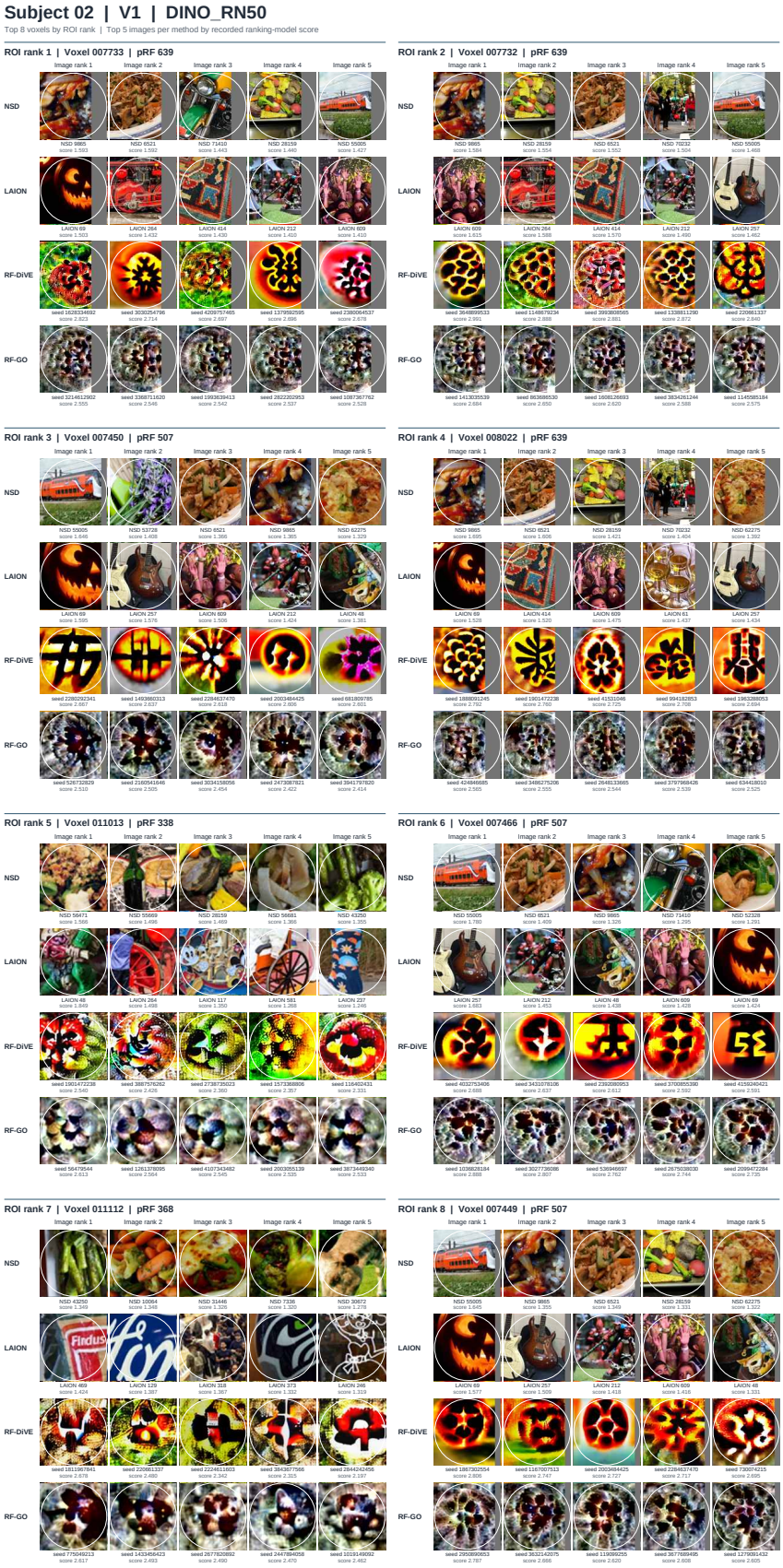

Figure 65: MEIs of top-8 voxels for Subject 02, visual area V1, using the DINO-RN50 brain-encoder backbone.

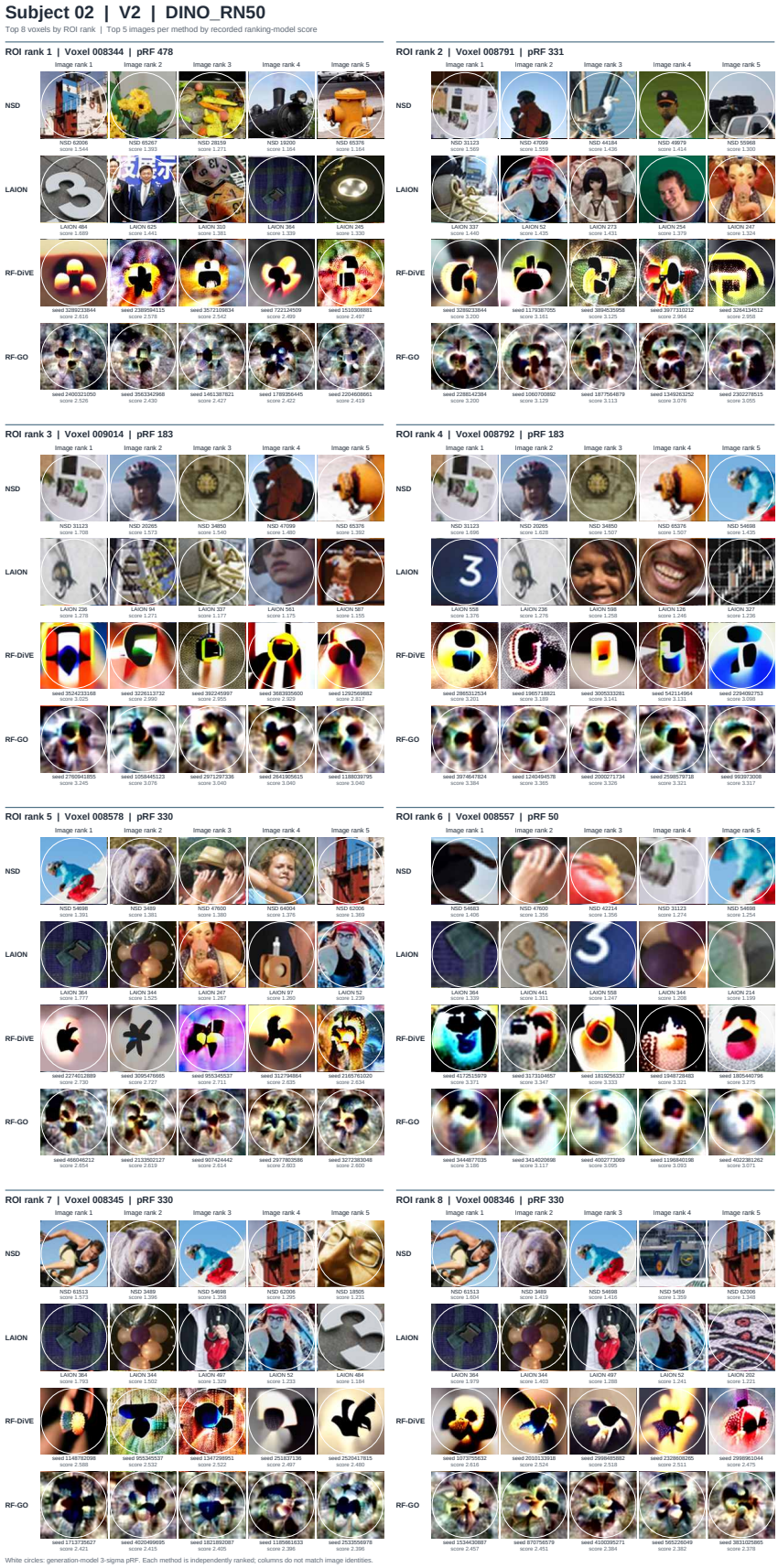

Figure 66: MEIs of top-8 voxels for Subject 02, visual area V2, using the DINO-RN50 brain-encoder backbone.

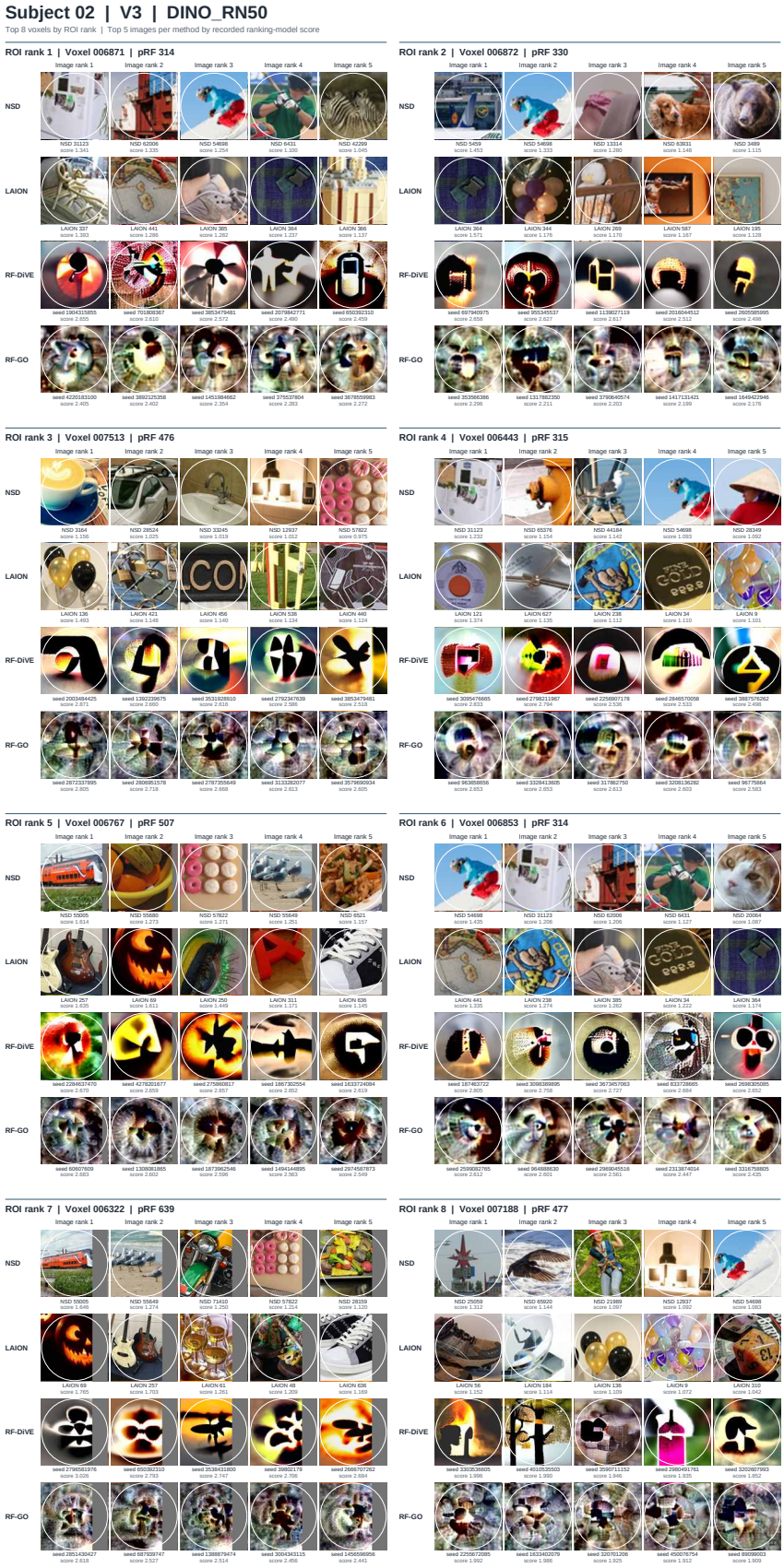

Figure 67: MEIs of top-8 voxels for Subject 02, visual area V3, using the DINO-RN50 brain-encoder backbone.

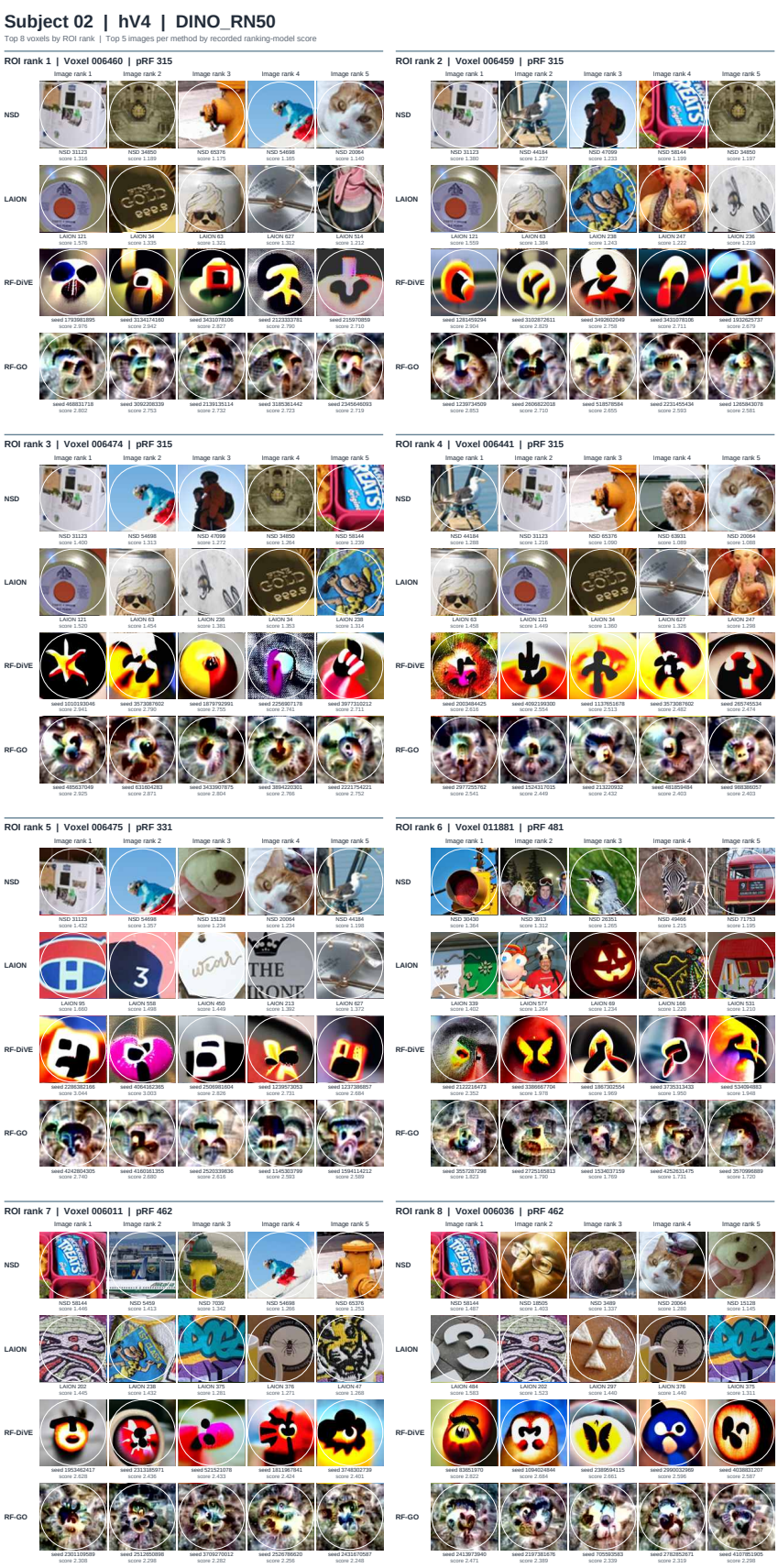

Figure 68: MEIs of top-8 voxels for Subject 02, visual area hV4, using the DINO-RN50 brain-encoder backbone.

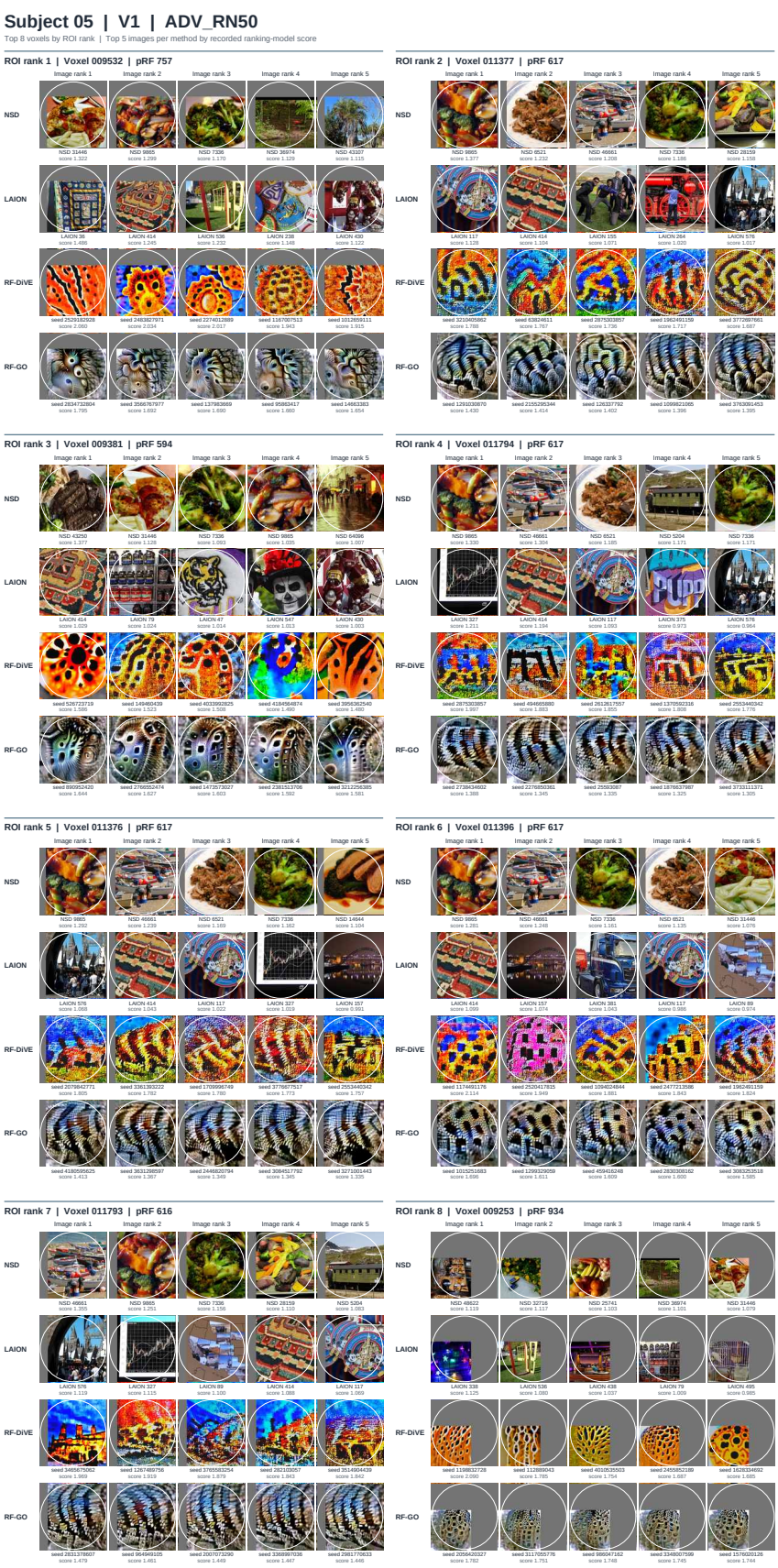

Figure 69: MEIs of top-8 voxels for Subject 05, visual area V1, using the ADV-RN50 brain-encoder backbone.

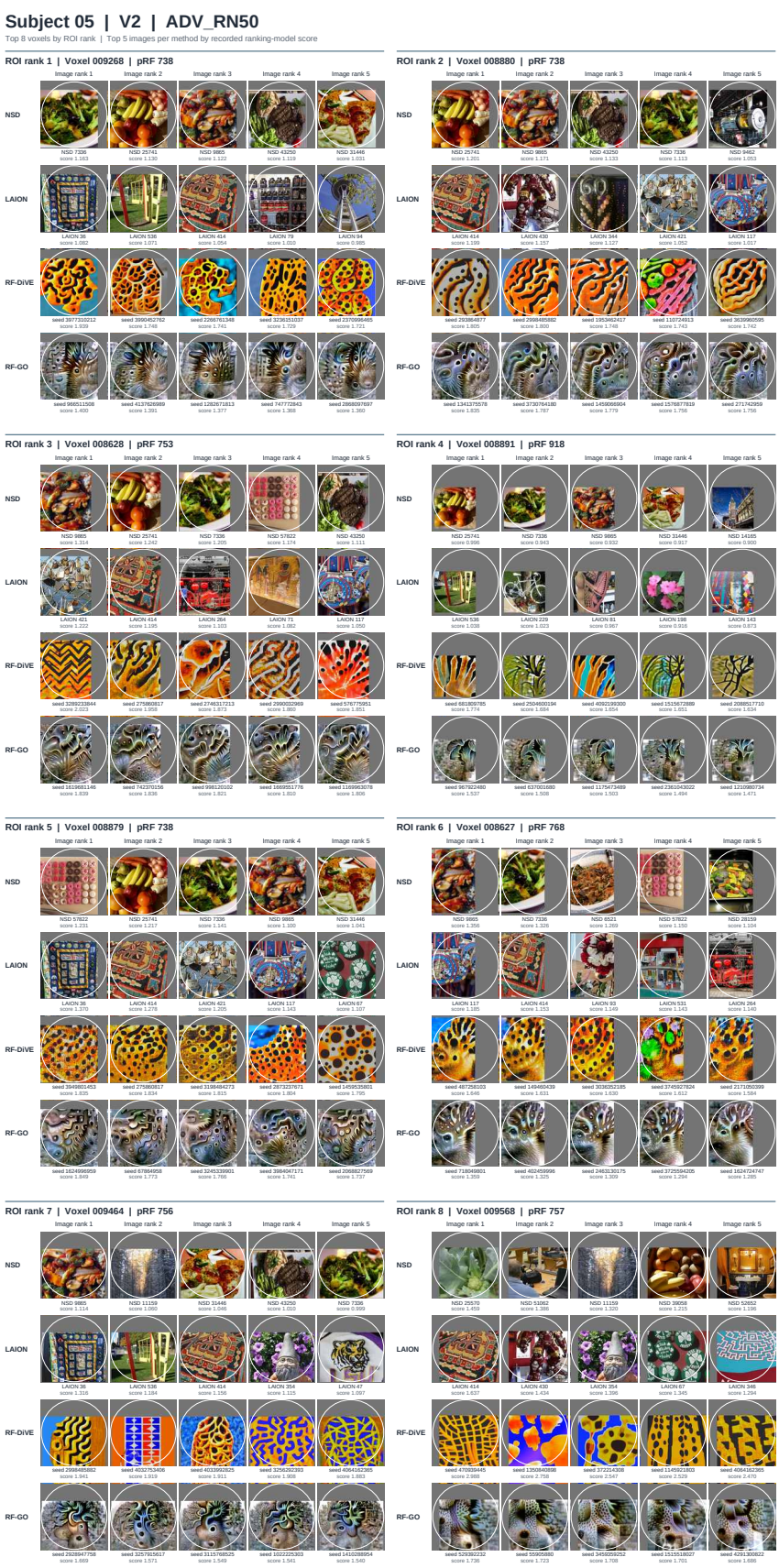

Figure 70: MEIs of top-8 voxels for Subject 05, visual area V2, using the ADV-RN50 brain-encoder backbone.

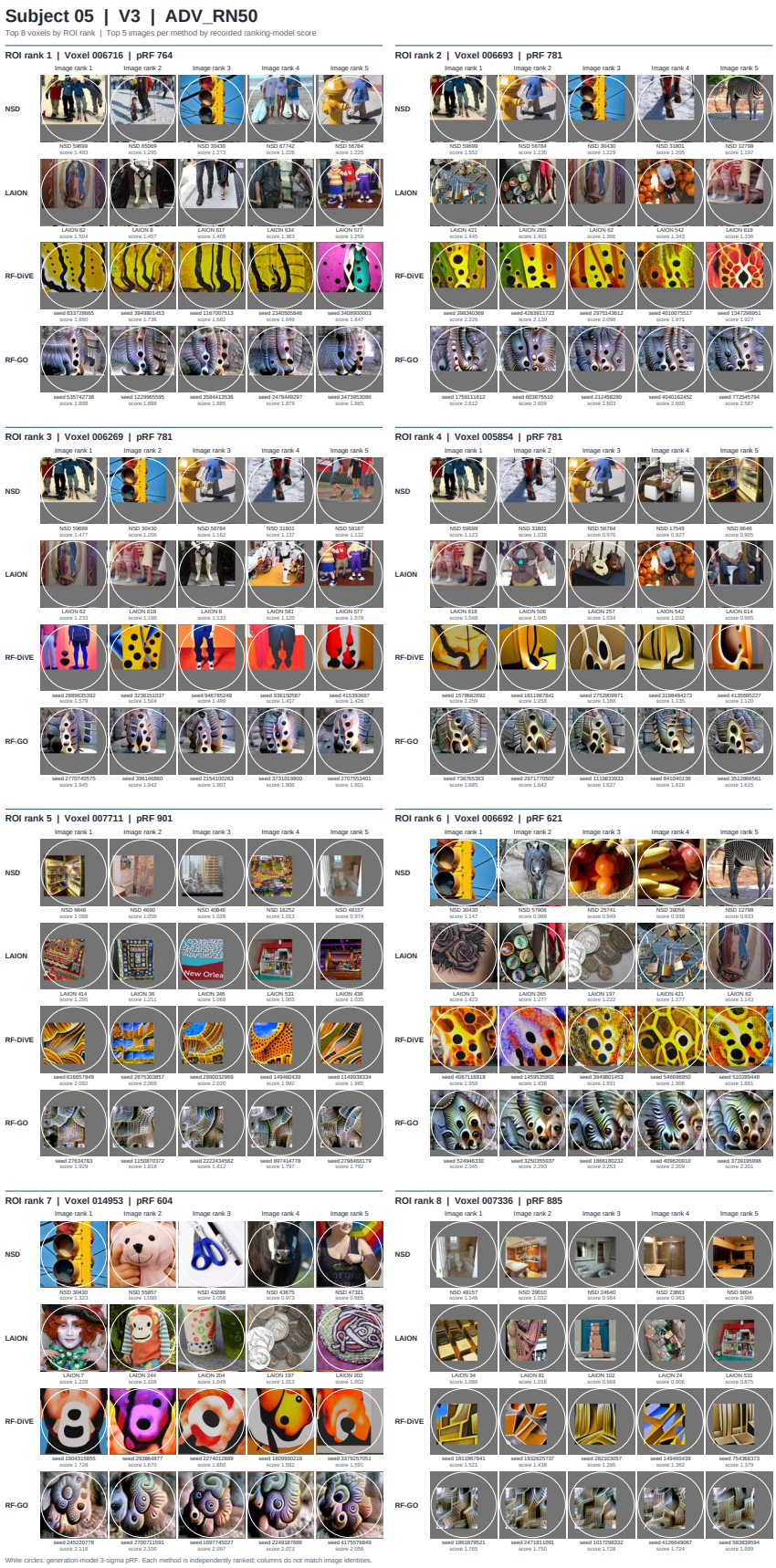

Figure 71: MEIs of top-8 voxels for Subject 05, visual area V3, using the ADV-RN50 brain-encoder backbone.

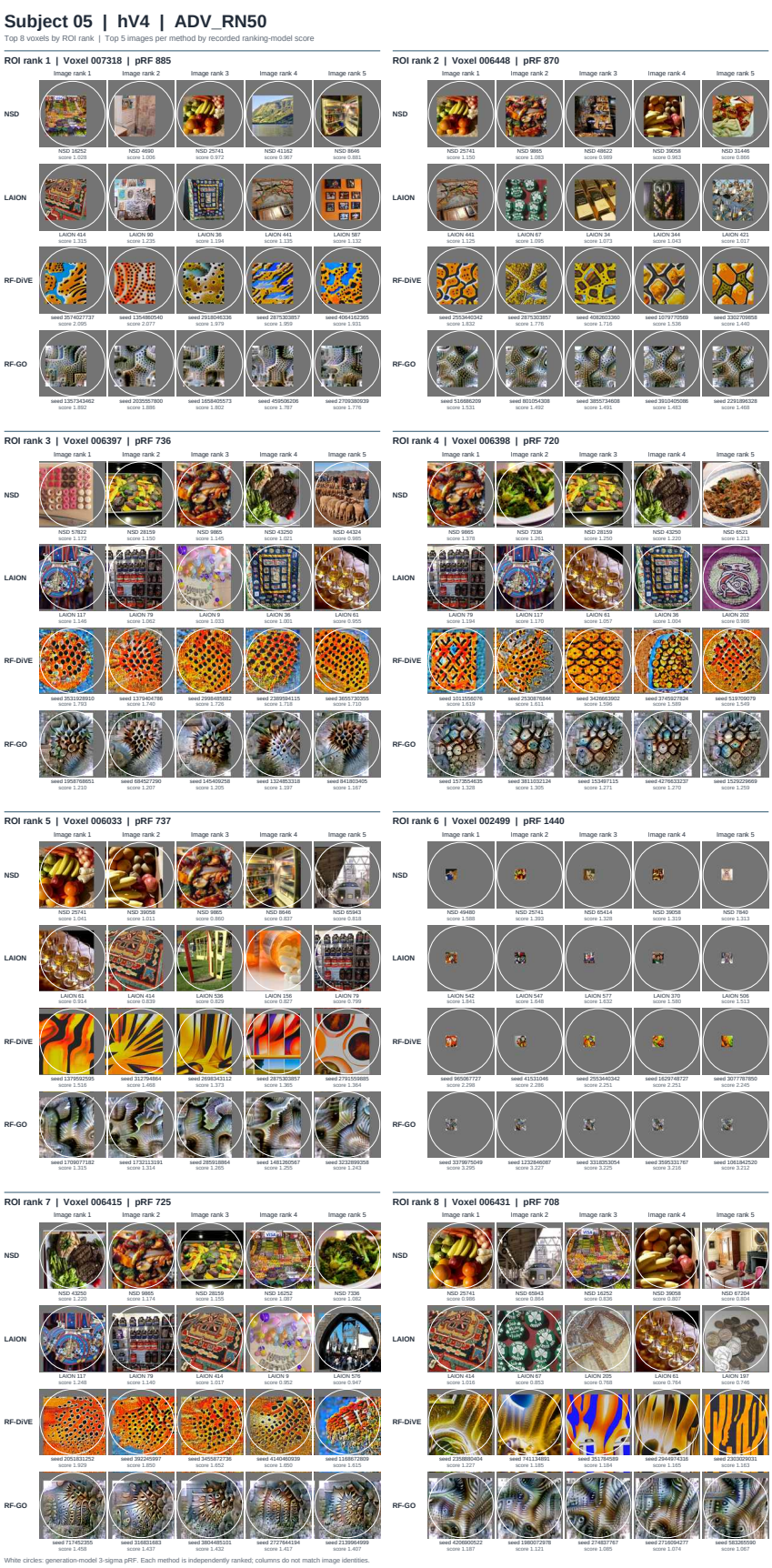

Figure 72: MEIs of top-8 voxels for Subject 05, visual area hV4, using the ADV-RN50 brain-encoder backbone.

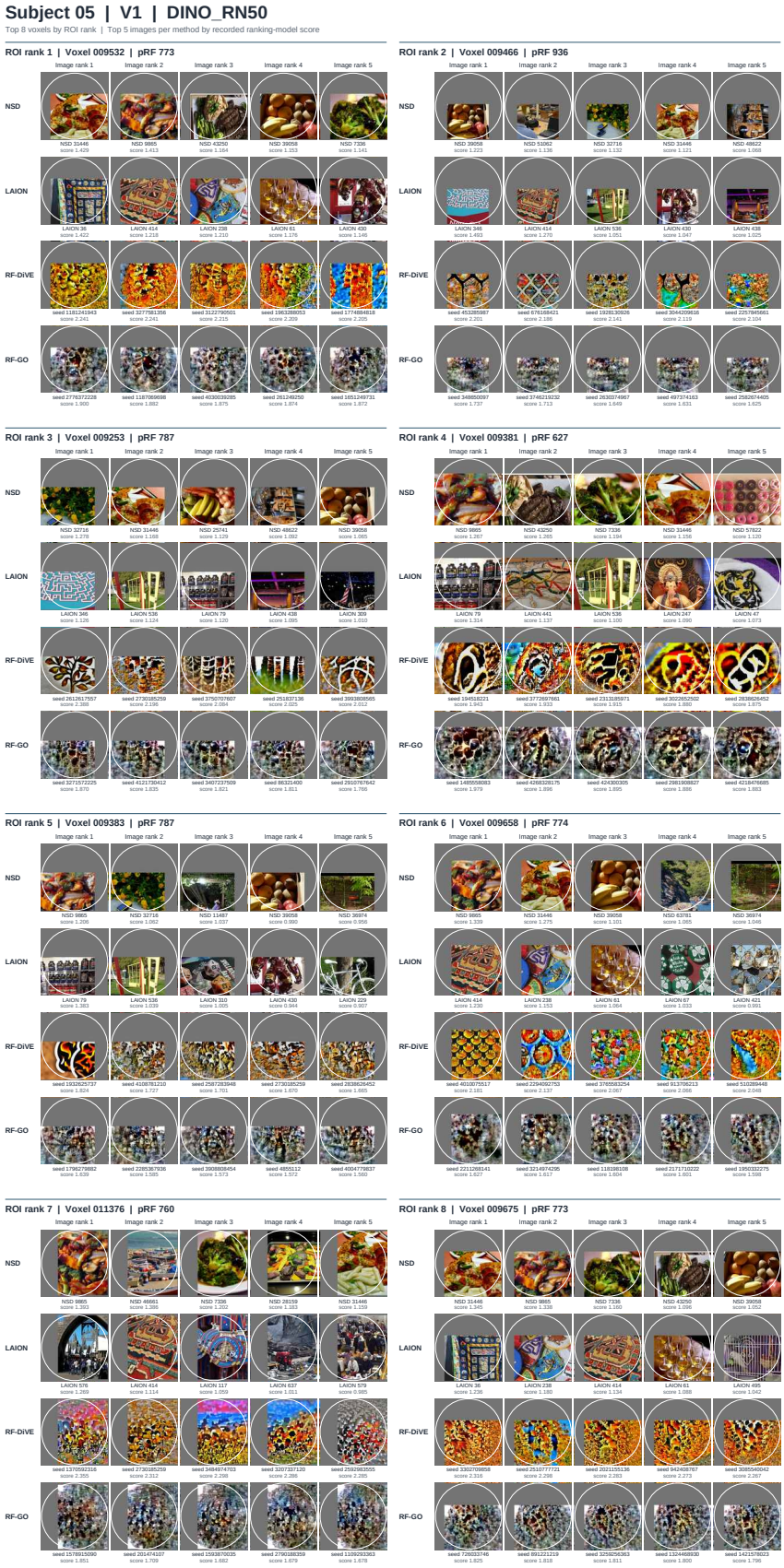

Figure 73: MEIs of top-8 voxels for Subject 05, visual area V1, using the DINO-RN50 brain-encoder backbone.

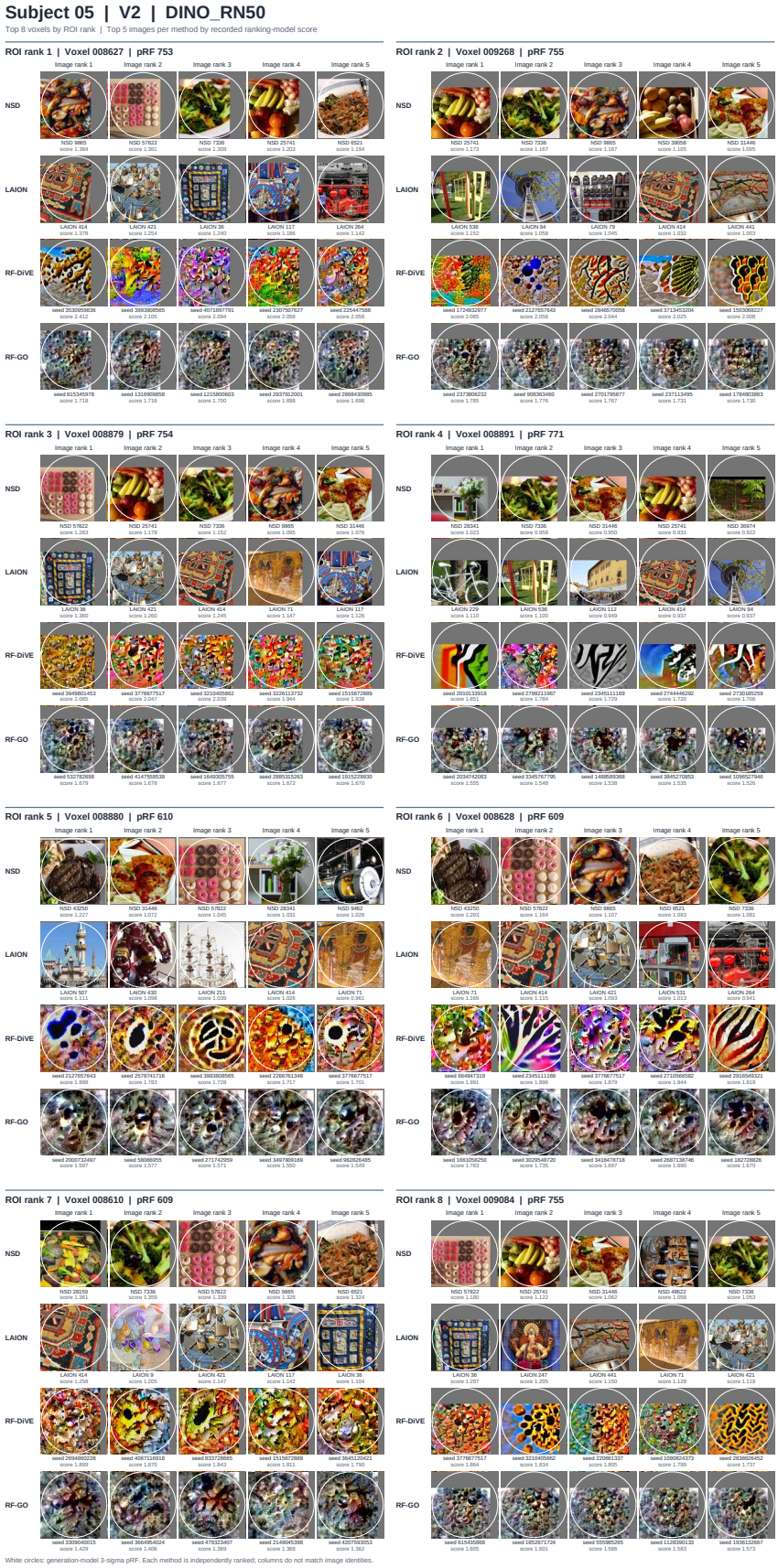

Figure 74: MEIs of top-8 voxels for Subject 05, visual area V2, using the DINO-RN50 brain-encoder backbone.

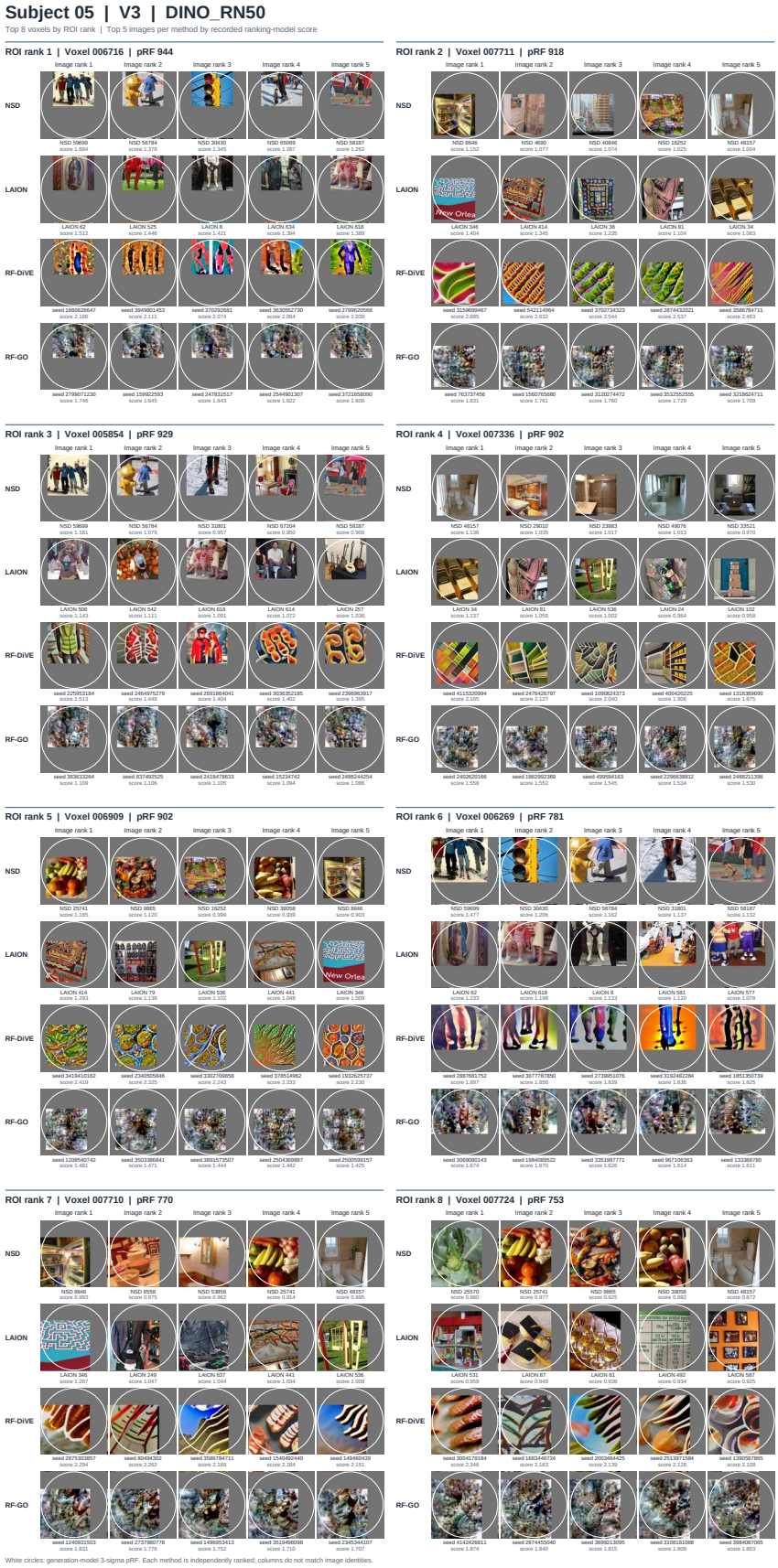

Figure 75: MEIs of top-8 voxels for Subject 05, visual area V3, using the DINO-RN50 brain-encoder backbone.

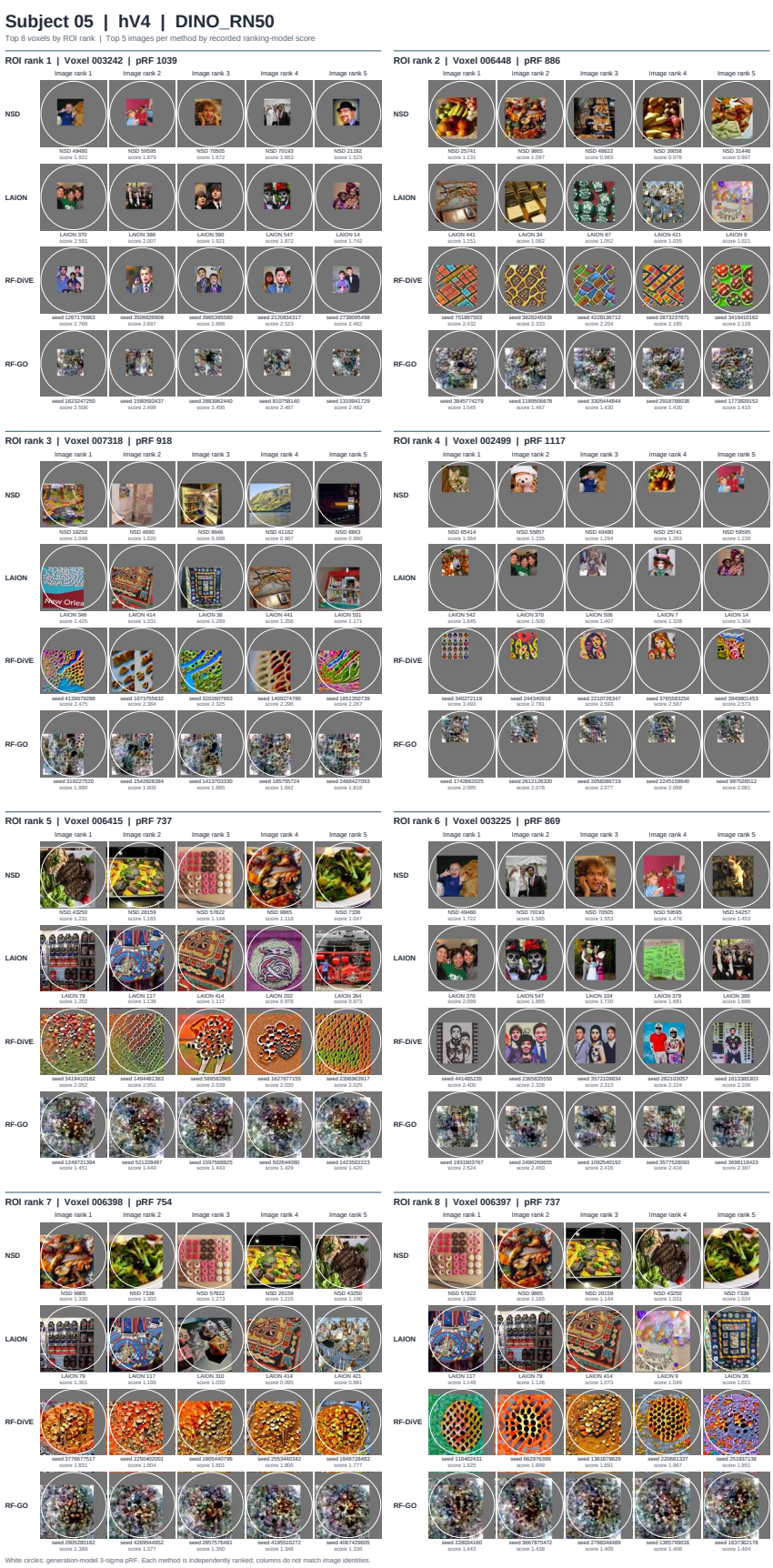

Figure 76: MEIs of top-8 voxels for Subject 05, visual area hV4, using the DINO-RN50 brain-encoder backbone.

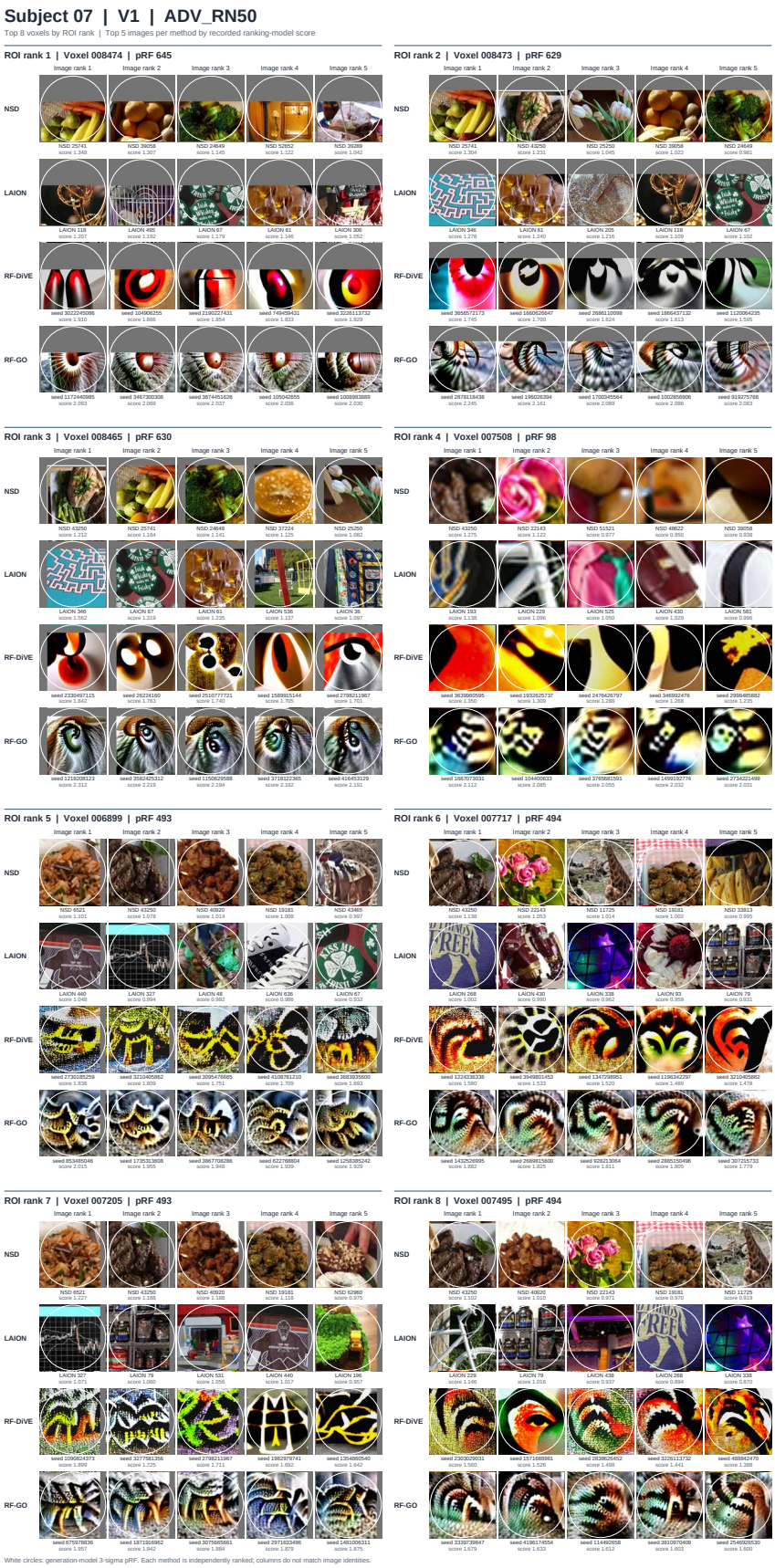

Figure 77: MEIs of top-8 voxels for Subject 07, visual area V1, using the ADV-RN50 brain-encoder backbone.

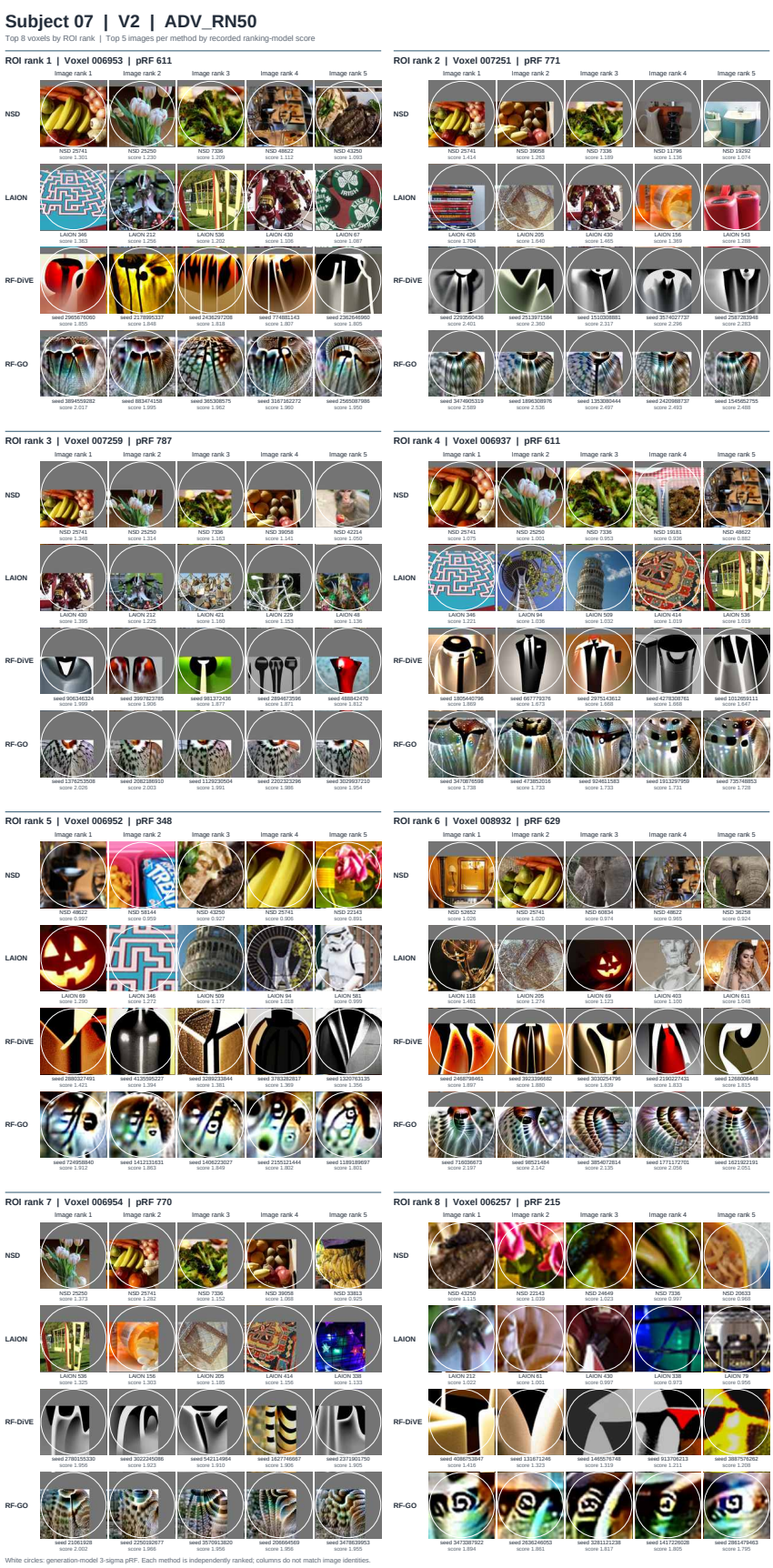

Figure 78: MEIs of top-8 voxels for Subject 07, visual area V2, using the ADV-RN50 brain-encoder backbone.

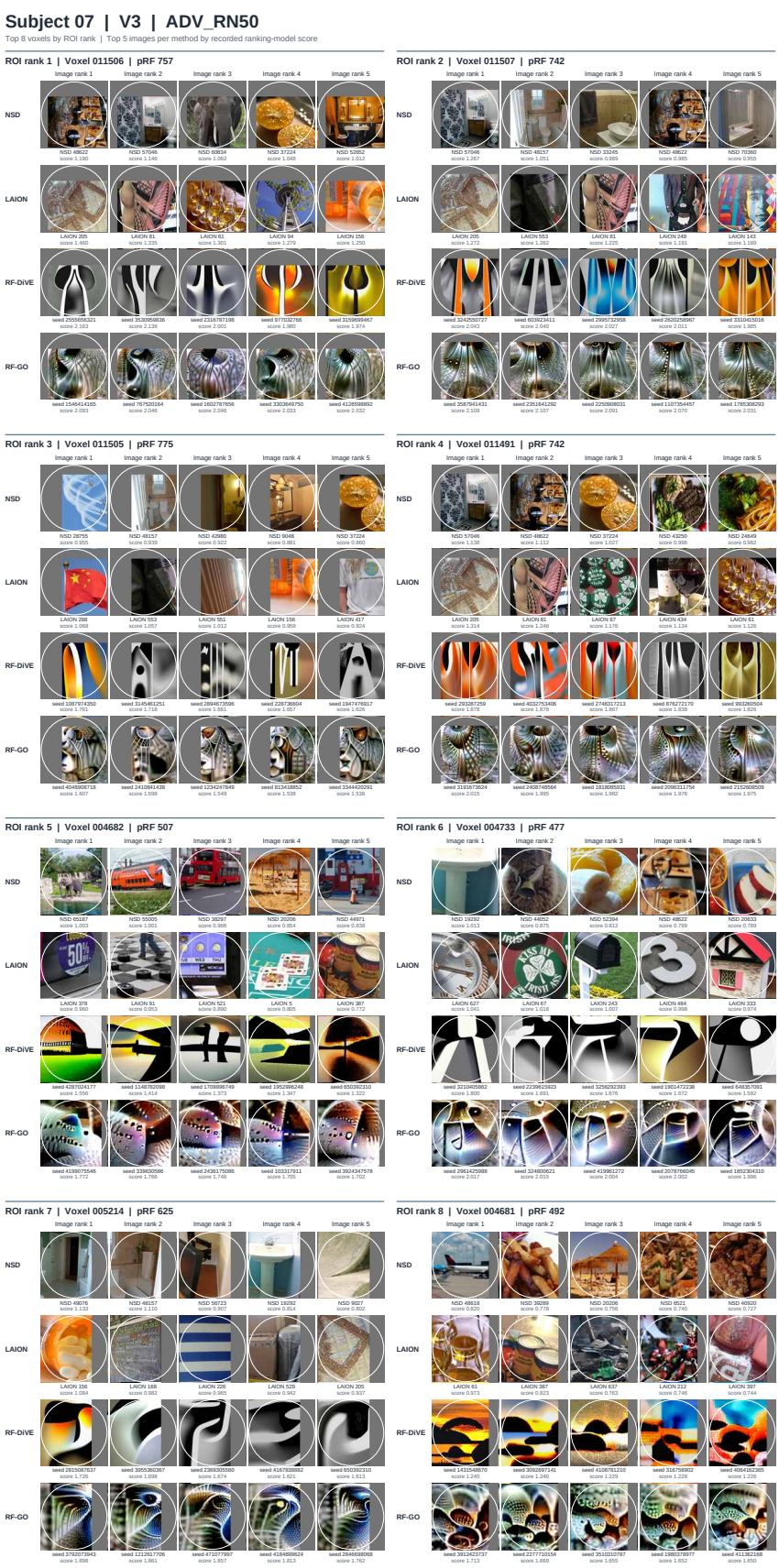

Figure 79: MEIs of top-8 voxels for Subject 07, visual area V3, using the ADV-RN50 brain-encoder backbone.

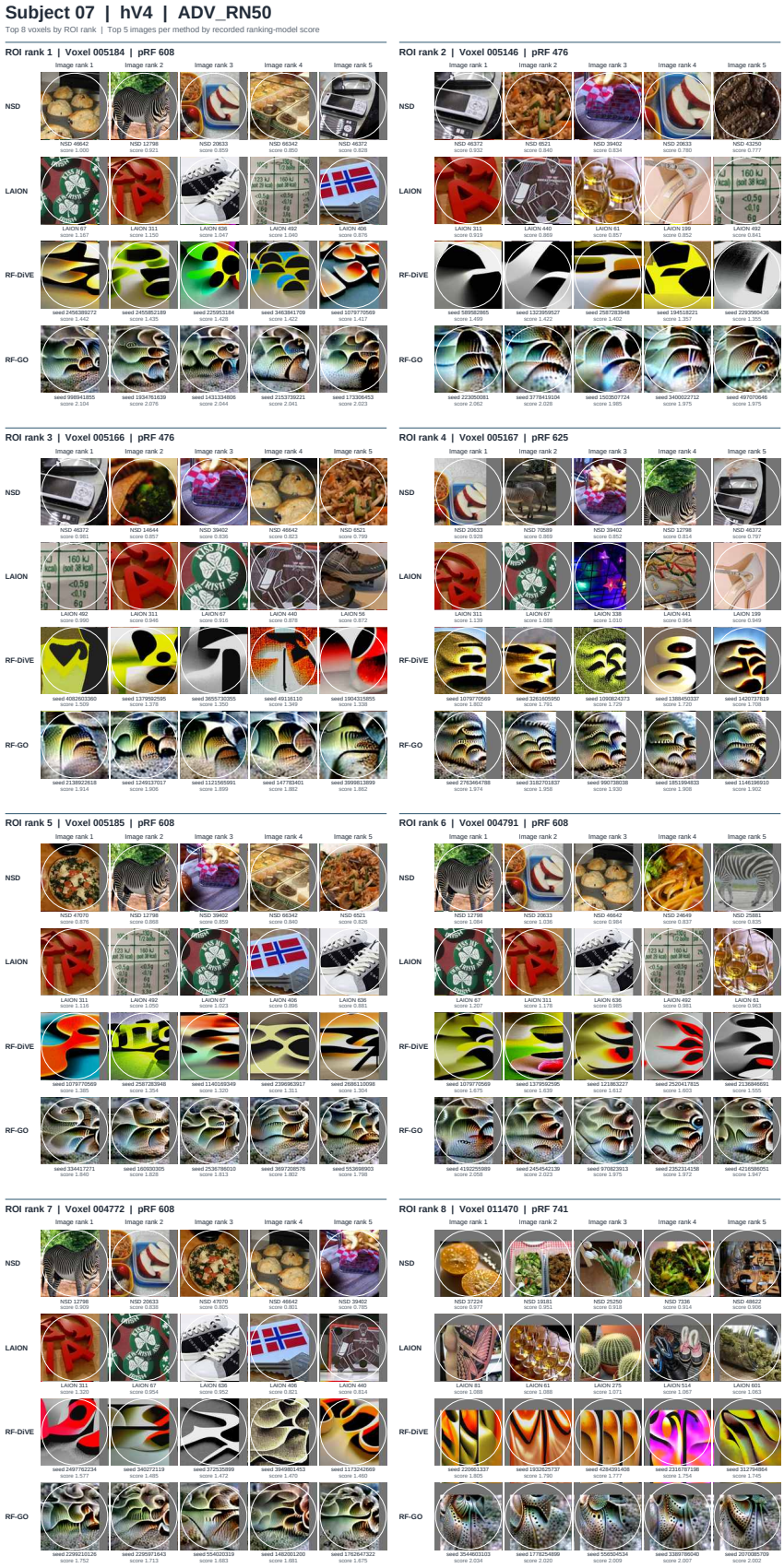

Figure 80: MEIs of top-8 voxels for Subject 07, visual area hV4, using the ADV-RN50 brain-encoder backbone.

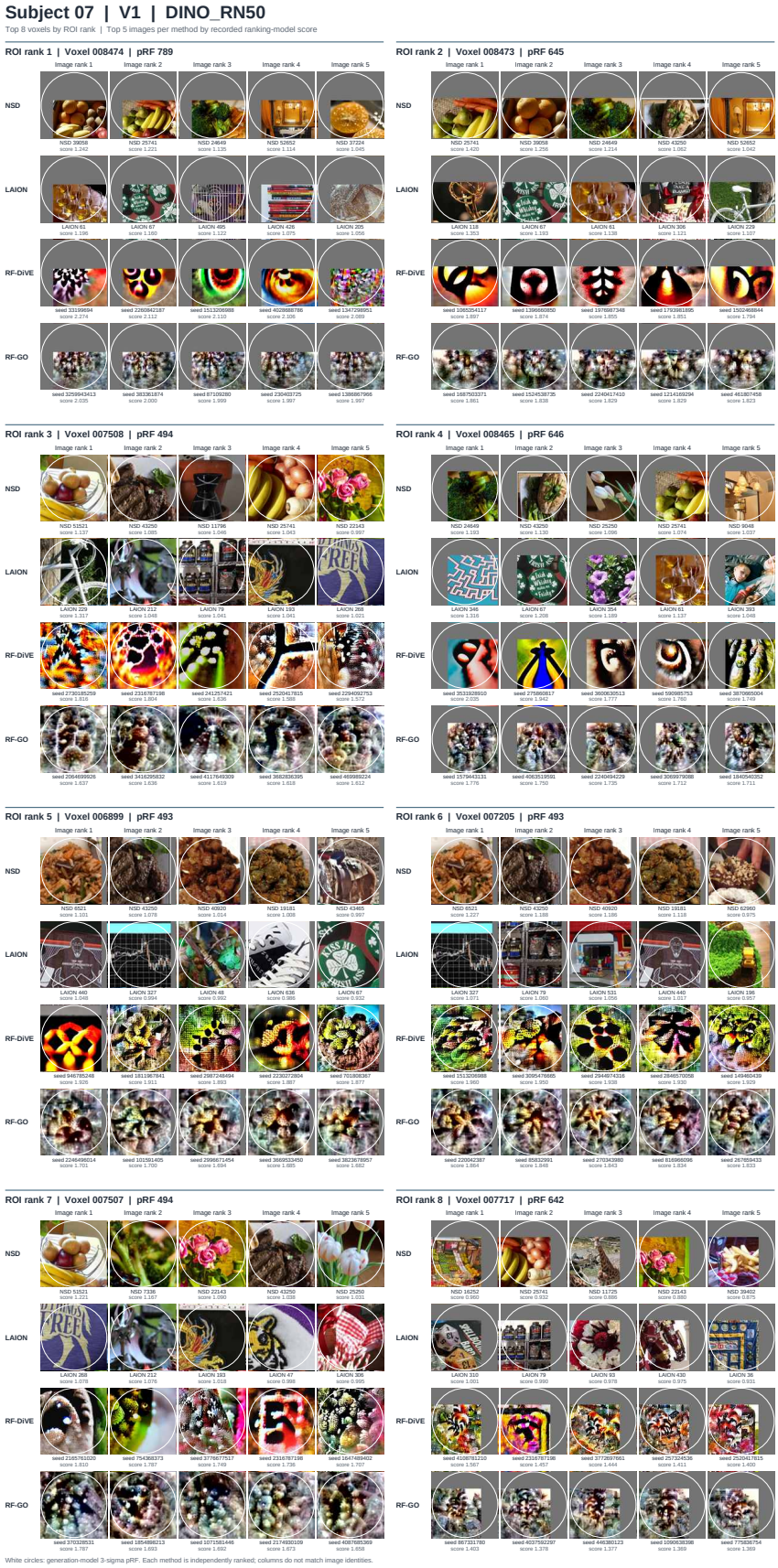

Figure 81: MEIs of top-8 voxels for Subject 07, visual area V1, using the DINO-RN50 brain-encoder backbone.

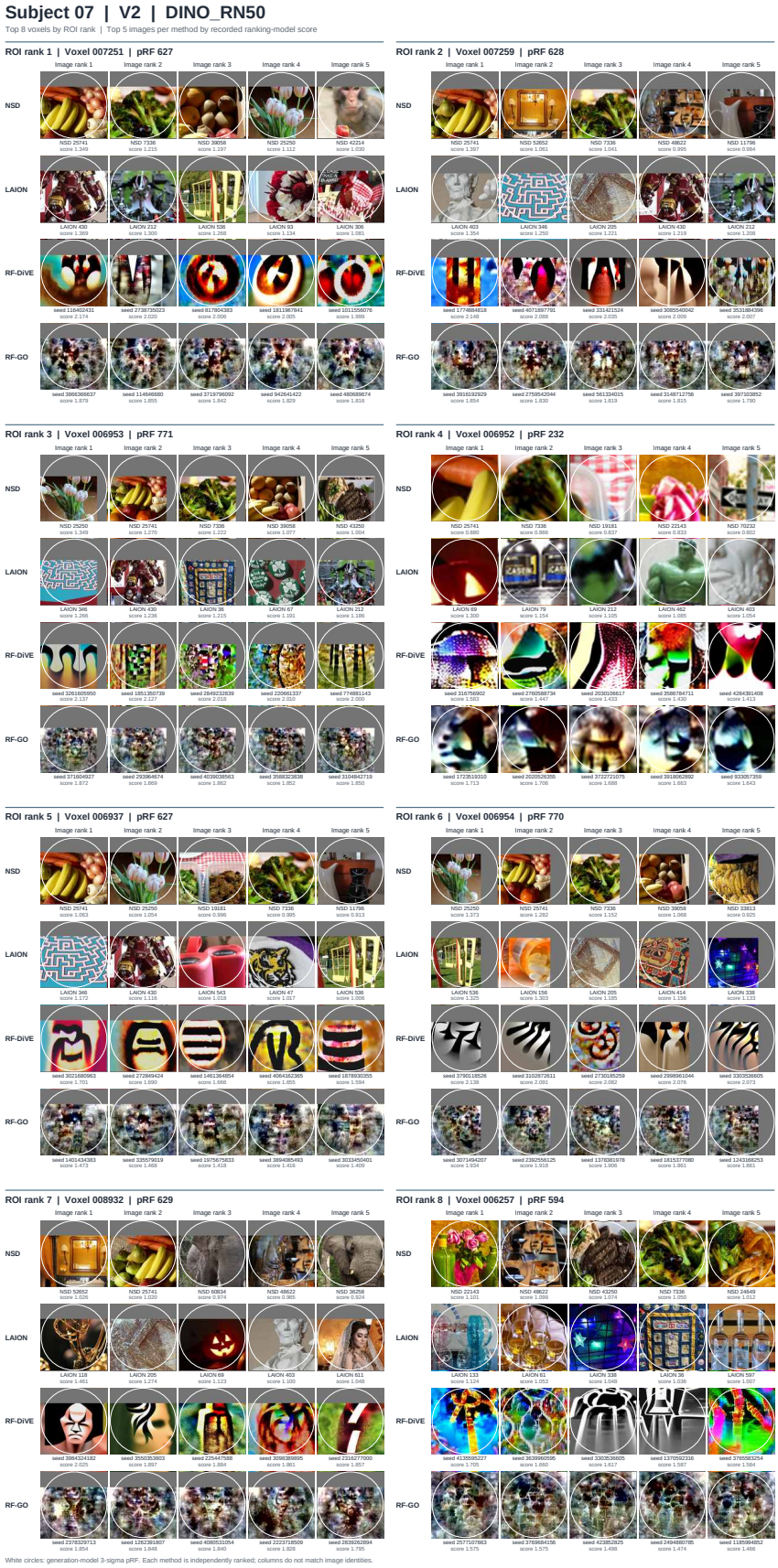

Figure 82: MEIs of top-8 voxels for Subject 07, visual area V2, using the DINO-RN50 brain-encoder backbone.

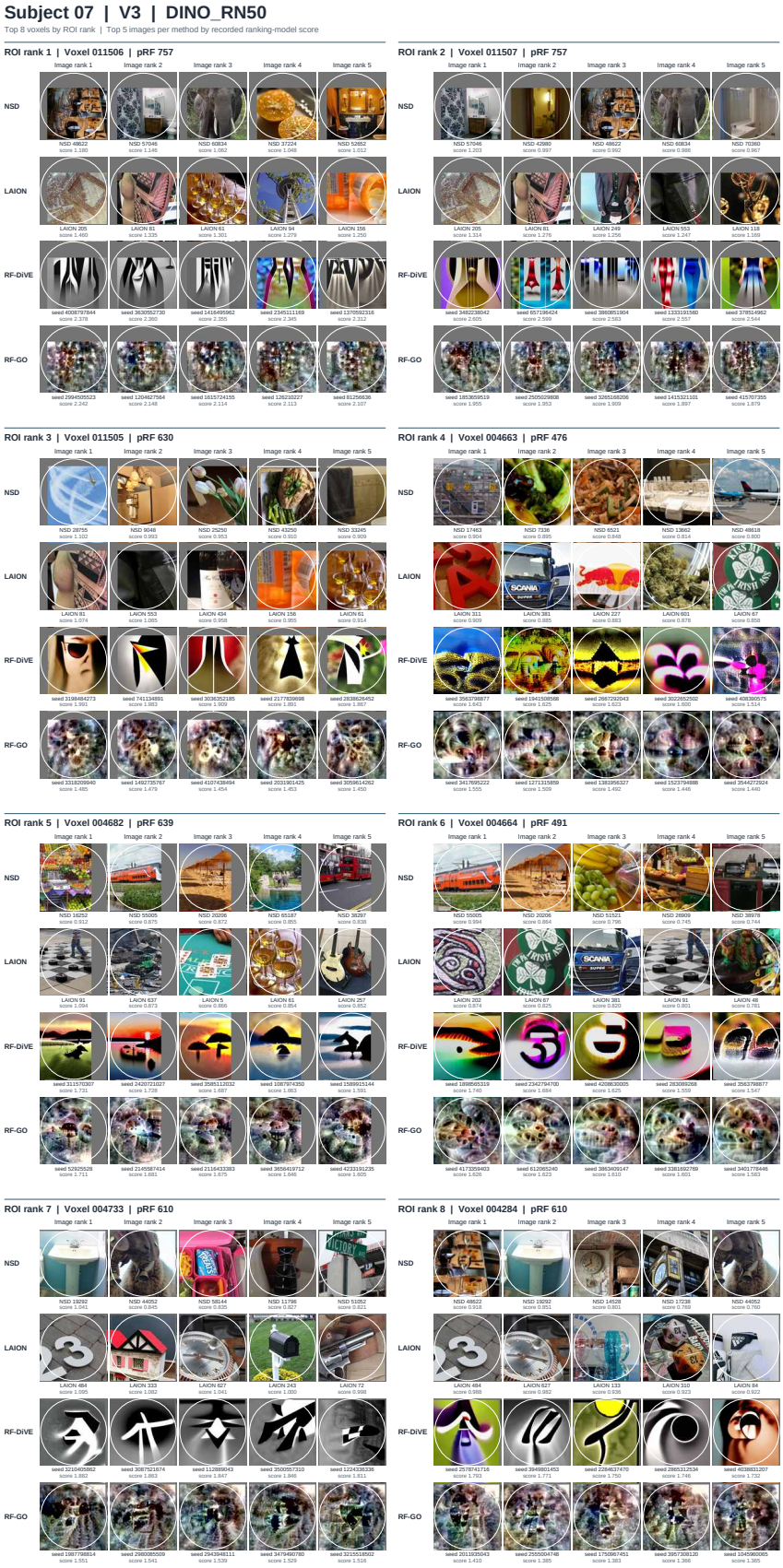

Figure 83: MEIs of top-8 voxels for Subject 07, visual area V3, using the DINO-RN50 brain-encoder backbone.

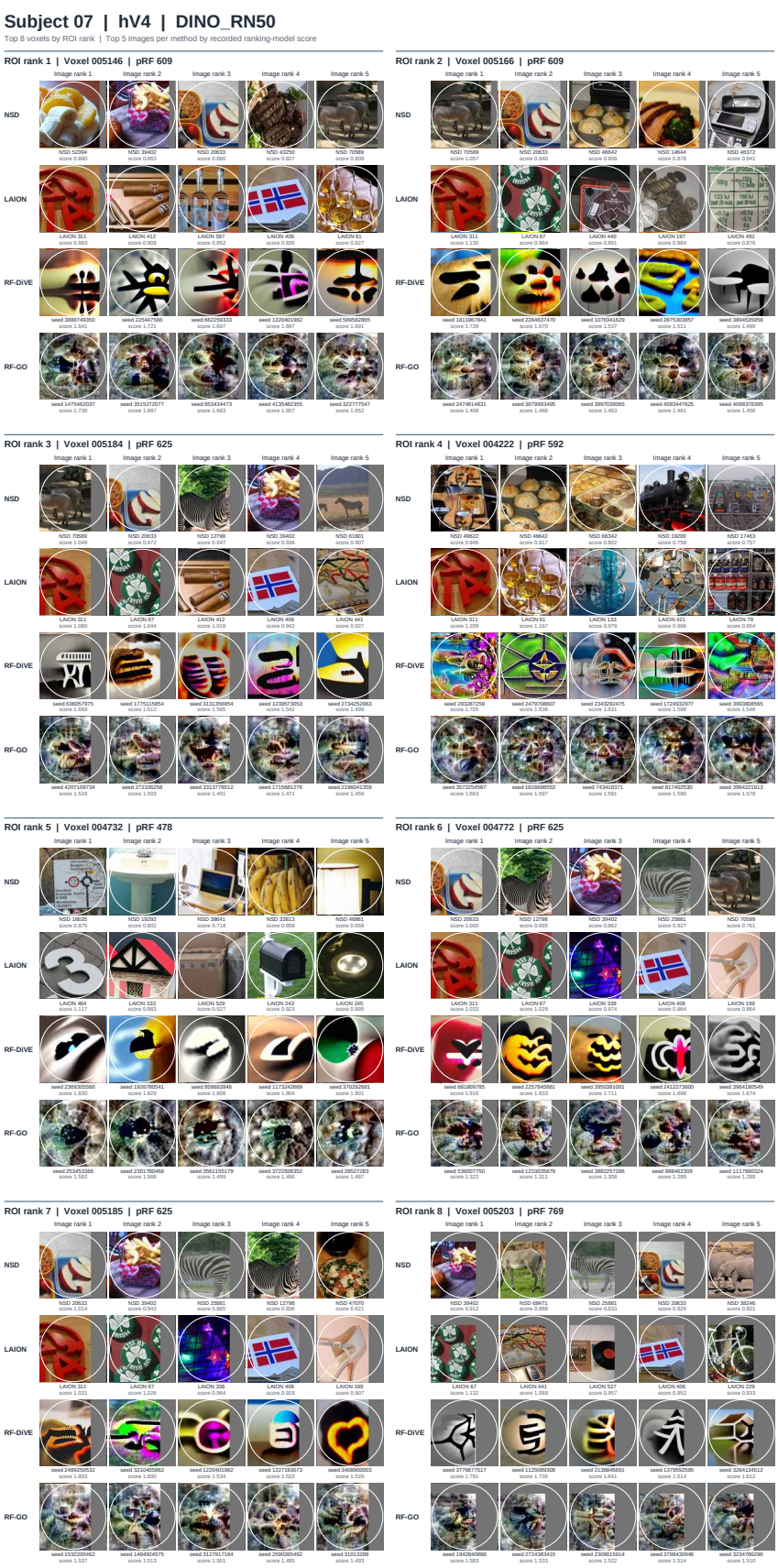

Figure 84: MEIs of top-8 voxels for Subject 07, visual area hV4, using the DINO-RN50 brain-encoder backbone.