Behavior Research Methods (2026) 58:336
https://doi.org/10.3758/s13428-026-03211-w

ORIGINAL MANUSCRIPT

# The SPOT Grid: Naturalistic images with predefined settings, patterns, objects, and textures

**Zall Hirschstein**[1] · **Mariam Aly**[2] · **Chris Baldassano**[1]

Received: 21 April 2026 / Accepted: 21 September 2026
© The Psychonomic Society, Inc. 2026

**Abstract**
In the natural world, higher-level visual features (e.g., scenes and objects) systematically covary with specific lower-level constituent features (e.g., textures and patterns). This natural covariation makes it difficult to isolate aspects of behavior or brain activity that are specific to a single level of the feature hierarchy. For example, a particular neural or behavioral response may reflect recognition of a "jungle" scene, or it may reflect sensitivity to the lower-level features – leaf textures, dense vertical lines, etc. – that constitute jungle scenes. To address this challenge, we developed a novel approach for partially decoupling higher- and lower-level features in naturalistic images. We employed a Stable Diffusion image generation framework to create a naturalistic image set in which setting–object pairings and texture–pattern pairings are factorially combined, enabling systematic manipulation of higher-level and lower-level image content. To validate our approach, we used state-of-the-art image-processing models to confirm that higher-level features are represented similarly regardless of their lower-level constituents and, conversely, that lower-level features are represented similarly regardless of the higher-level objects/settings they constitute. We then confirmed that human participants can readily and rapidly categorize these features in our generated images. We present the image generation framework and introduce the SPOT (Settings, Patterns, Objects, Textures) Grid, a publicly available stimulus set accompanied by image-specific metrics (upon request: forms.gle/X2HpzRYUGigFS1S27). This resource provides new opportunities for investigating hierarchical visual representations, feature binding, and abstraction across perception, behavior, and the brain.

**Keywords** Visual hierarchy · Naturalistic stimuli · Feature binding · Generative models · Visual perception · Stimulus generation

## Introduction

The natural visual world is hierarchically structured: high-level visual features are systematically constituted by specific lower-level features. This nested structure spans the spectrum of feature levels – from higher-order scene categories to spatial primitives like edges. Consider a jungle environment: jungle scenes reliably contain particular objects (foliage, vines) that exhibit textural regularities (leafy and bark surfaces, dappled light) which are, in turn, composed of reliable patterns of edges (organic striations, dense vertical lines). Likewise, kitchen environments reliably contain particular objects (appliances, tiles) with specific textures (reflective surfaces, ceramic materials) composed of dependable edge patterns (consistent corner radii, rectilinear form repetition). These examples illustrate that the identity of any given scene and its component objects is intrinsically linked to lower-level visual statistics in the natural world, posing a fundamental challenge for unraveling how visual systems process information across different levels of the feature hierarchy. We refer to this interwoven structure as *feature-level entanglement*. In this context, we define lower-level features as spatially local visual patterns including edge orientation, color, localized contrast, and textures (e.g., Portilla & Simoncelli, 2000), whereas higher-level features are the emergent, global semantic identities – including discrete objects and scene categories.

Mariam Aly and Chris Baldassano contributed equally to this work.

✉ Zall Hirschstein
zsh2109@columbia.edu

[1] Department of Psychology, Columbia University, 800 Uris Hall, 3022 Broadway, New York, NY 10027, USA

[2] Department of Psychology, University of California, Berkeley, CA 94720, USA

Published online: 23 October 2026

When attempting to disentangle how different levels of the visual hierarchy contribute to cognition, researchers face interpretive obstacles: responses attributed to high-level processing of scene or object identity may actually reflect sensitivity to co-varying low-level features, and vice versa. For instance, when a person or vision model categorizes a "jungle" scene, the specific visual features driving this behavior remain obscure: the response could reflect scene recognition, object detection, texture sensitivity, spatial frequency, etc. Likewise, differential neural activity driven by "jungle" versus "kitchen" images could originate from semantic, scene-level, object-level, textural, or edge-level differences, or complex combinations thereof. Indeed, neural representations of different scene categories differ as early as V1 (Park et al., 2011; Torralbo et al., 2013; Walther et al., 2009), indicating that the entanglement of higher- and lower-level visual features affects fundamental levels of visual processing. Entanglement also persists at later stages of visual processing: low-level image properties can be classified from activity in higher-order visual processing regions like inferotemporal cortex (Baldassi et al., 2013; Rice et al., 2014; see also Long et al., 2018).

Researchers increasingly use computational models, including convolutional deep neural networks (DNNs), to dissociate the contributions of specific feature levels to cognition. These investigations typically employ post-hoc analytical approaches: first collecting responses to natural images, then using computational models to extract feature representations or region maps varying in low-level, high-level, and/or semantic content. Behavior in, for example, memory encoding tasks (Bone et al., 2020; Morales-Torres et al., 2024) or attentional guidance paradigms (Hayes & Henderson, 2021, 2025; Pedziwiatr et al., 2021; Walter et al., 2024) can then be modeled in relation to these representations or regions. While these approaches will remain critical, experimental designs could benefit from moving beyond analytical decomposition to incorporate a priori stimulus control that preserves ecological richness (Nastase et al., 2020). Ideally, researchers could create stimulus sets in which multiple feature levels can each be systematically controlled in naturalistic stimuli, permitting a cleaner isolation of behavioral or neural effects related to specific levels of the visual feature hierarchy.

One approach for isolating feature levels is to "scramble" visual content. Techniques such as diffeomorphic warping or texture-synthesized scrambling (e.g., Freeman & Simoncelli, 2011; Stojanoski & Cusack, 2014) transform natural visual content until, for example, higher-level features like objects are unrecognizable, while maintaining basic color, texture, or edge statistics. Although effective for isolating low-level content, these methods yield stimuli that lack the ecological realism of the natural world. This creates a fundamental tradeoff: researchers must choose between well-controlled but scrambled images and realistic but fully entangled natural scenes.

A different approach manipulates images to control high-level and/or low-level features while maintaining naturalistic visual content. We took inspiration from such computer vision paradigms, including neural style transfer (Gatys et al., 2015). Style transfer leverages the representational hierarchy of convolutional DNNs to separate higher-level "content" (e.g., global layout and semantic identity) from lower-level "style" (e.g., local color and orientation statistics). While "style" is a broad term in computer vision, it is often operationally defined in terms of texture: style represents the spatially invariant summary statistics of visual features, which capture local patterns rather than larger-scale objects and geometry (Chen et al., 2025; Gatys et al., 2015). That is, image stylization involves manipulating the visual appearance and texture of an image while preserving its layout and content (Frenkel et al., 2025). The possibility of independent feature manipulation was furthered by generative adversarial networks (GANs), such as StyleGAN (Karras et al., 2019), which can, unsupervised, learn latent representations that separate higher-order structure (e.g., pose or object identity) from lower-level variation (e.g., skin texture, lighting). Recent evidence suggests that the internal representations of these generative models converge with neural activity in the higher-order visual cortex, implying that these models may extract high-level visual content from raw sensory input in a biologically plausible manner (Dado et al., 2023). While GANs have excelled at certain tasks, their limitations (e.g., compute demand, training specificity, mode collapse) and inferior image fidelity and diversity mean they have been surpassed in many respects by diffusion-based models (Dhariwal & Nichol, 2021). Critically, GANs often lack the granular, user-defined control required to factorially cross specific high-level and low-level visual features, limiting their use for reliably creating stimuli with particular desired features.

The goal of this work is to build on prior research by developing a framework that generates realistic images in which higher-level semantic content and lower-level visual statistics are partially decoupled through factorial manipulation of specific low- and high-level visual features. We accomplish this by systematically rendering different settings and objects (high-level features) in different textures with specified patterns (low-level features), allowing us to create groups of images optimized for either high-level or low-level similarity (Fig. 1). It is important to emphasize that *complete* decoupling of low-level and high-level visual features is not possible – a high-level visual representation (e.g., a coffee cup) is necessarily constituted by specific low-level features (e.g., the particular edges that make up its form). The goal of our approach is not to completely disentangle low- and high-level features, but instead to allow

A

| Setting | Object |
|---|---|
| underwater | teacup |
| jungle | computer |
| field | truck |
| desert | guitar |
| city | backpack |
| mountain | armchair |
| bedroom | plant |
| outerspace | church |

B

| Texture | Pattern |
|---|---|
| paper | diamonds |
| embroidery | vertical lines |
| oil paint | spiral |
| graffiti | dots |
| claymation | checkerboard |
| ink | big X |
| crayon | horizontal lines |
| stained glass | tilted lines |

**Fig. 1** Categories of features within the 8×8 image set. **A** Setting and object are always paired (e.g., if an image contains the setting "underwater" it also contains the object "teacup", row 1). **B** Texture and pattern are also always paired (e.g., if an image was made with the texture "paper" it was also made with the pattern "diamonds", row 1). (see Fig. S2 for the expanded set of features within the 10×10 grid)

precise control over factorial combinations of particular high- and low-level features.

Our approach addresses the lack of user-defined control in GAN frameworks by leveraging a new generation of diffusion-based models. These architectures not only generate highly realistic images but also offer modular extensibility that allows for remarkable command over semantic content and visual style. This capability has been augmented by recent innovations like Low-Rank Adaptation (LoRA; Hu et al., 2021), which enhances the quality and consistency of style rendering, and ControlNet (Zhang et al., 2023). ControlNet forces the generated image to follow a structural "blueprint" that the user specifies with an input image, allowing us to specify particular locations where edges should appear in the generated image. Even with these extensions, however, generated images can vary substantially in quality. This requires validation to ensure the images meet high-quality standards for use in studies of behavior and the brain.

To meet these goals, we developed the SPOT (Settings, Patterns, Objects, Textures) framework: a generative pipeline designed to produce naturalistic stimuli with systematically manipulated higher-level and lower-level visual features. We leveraged Stable Diffusion, integrated with LoRA and ControlNet, to synthesize an image set in which high-level features (settings and objects) are factorially crossed with low-level features (textures and patterns) – see Fig. 2 for visualization of this feature crossing. Initially, a large pool of 1594 images was produced within a 10×10 grid, crossing ten setting–object pairings with ten texture–pattern pairings. (For analyses of this full image set, see Supplementary Information). From this pool, we curated an "intermediate" 8×8 grid based on pilot data (see Supplementary Information), selecting the features and specific images associated with the best categorization performance in human observers. A final representative subset of 64 images (with one image randomly chosen per cell of the 8×8 grid) was then selected for additional validation (Fig. 2). This subset validation is meant to demonstrate the viability of selecting images from the intermediate 8×8 grid as necessary. To validate the SPOT grid and the framework's efficacy, we used four complementary computational models: two representing low-level visual features and two representing high-level visual features. These models confirmed that we successfully and systematically manipulated these feature levels across our image set. Finally, a human behavioral experiment with the 64-image set confirmed that participants could accurately and rapidly categorize each setting, object, texture, and pattern. The pattern of results across our modeling and human behavioral experiments was similar for the 64-image subset and the full set of 1594 images, supporting the flexible use of parts of the full image set in future experiments. These results establish the experimental viability of our SPOT generation framework and stimulus grid as a robust resource for future studies of visual behaviors and brain activity.

## Methods

### Materials

We generated an image library by factorially crossing two "high-level" features (setting and object) with two "low-level" features (texture and pattern). For the purposes of

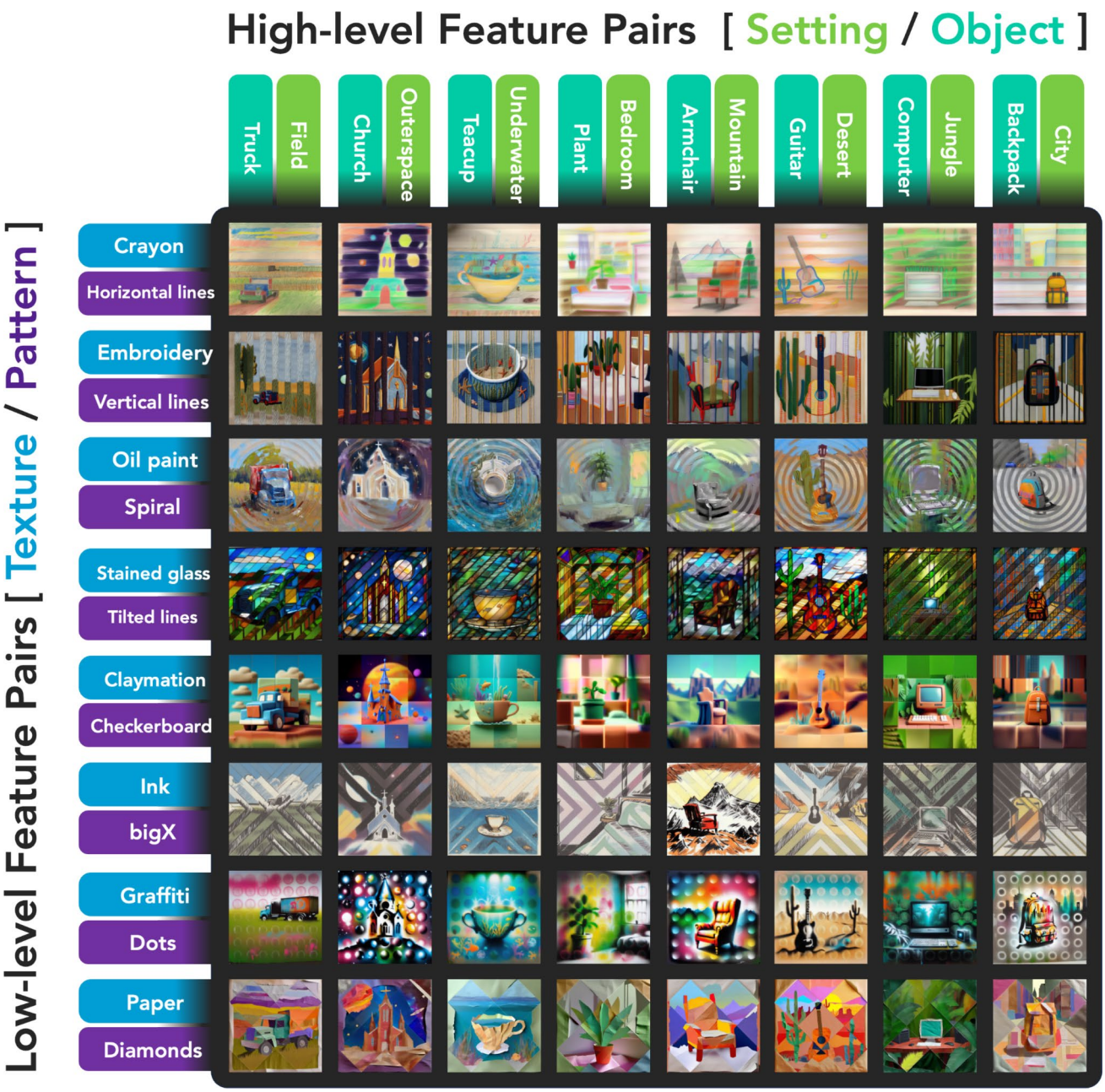



**Fig. 2** The 8 × 8 feature pair grid and the 64 images selected for additional validation (see Supplementary Information for full set of 1594 images). Feature pairs are on the axes and images are in the cells. The *x*-axis represents the high-level feature pairs: setting (*green*) and object (*teal*). For example, the first column of images all depict "truck" (object) in "field" (setting) rendered in various textures and patterns. The *y*-axis represents low-level feature pairs: texture (*blue*) and pattern (*purple*). For example, the first row of images all depict different objects and settings rendered as if drawn with crayon (texture) and containing large horizontal edges (pattern)

this framework, a "high-level feature pair" is a specific setting–object combination, and a "low-level feature pair" refers to a specific texture–pattern combination (for validation of these designations, see Supplement/Low-level Feature Separation). Lastly, "4-feature grouping" refers to the combination of a high-level feature pair and a low-level feature pair.

An initial 1594 images were generated within a 10 (high-level feature pairs) × 10 (low-level feature pairs) grid. This yielded 100 distinct 4-feature groupings that contain features consistent with their respective setting, object, texture, and pattern specification (Fig. S1). A series of pilot experiments were used to refine this to an optimized 8 × 8 grid, by dropping feature pairs associated

with worse behavioral labeling performance. All images used in the initial pilot studies, and their associated behavioral performance, are available upon request at forms.gle/X2HpzRYUGigFS1S27, and methods and results regarding the 10 × 10 grid are reported in Supplementary Information (Fig. S1–S7). The remainder of this paper focuses on images from the refined 8 × 8 grid.

## Image generation

Images were made with the generative latent diffusion model Stable Diffusion (Rombach et al., 2022) XL using the pretrained SDXL 1.0 model (Podell et al., 2023). Initial noisy latents were processed with SDXL's default refinement model for final denoising. An open-source user interface, Automatic1111 Stable Diffusion web UI (2022), was used as the generation platform (Fig. 3) to leverage its fine-grained

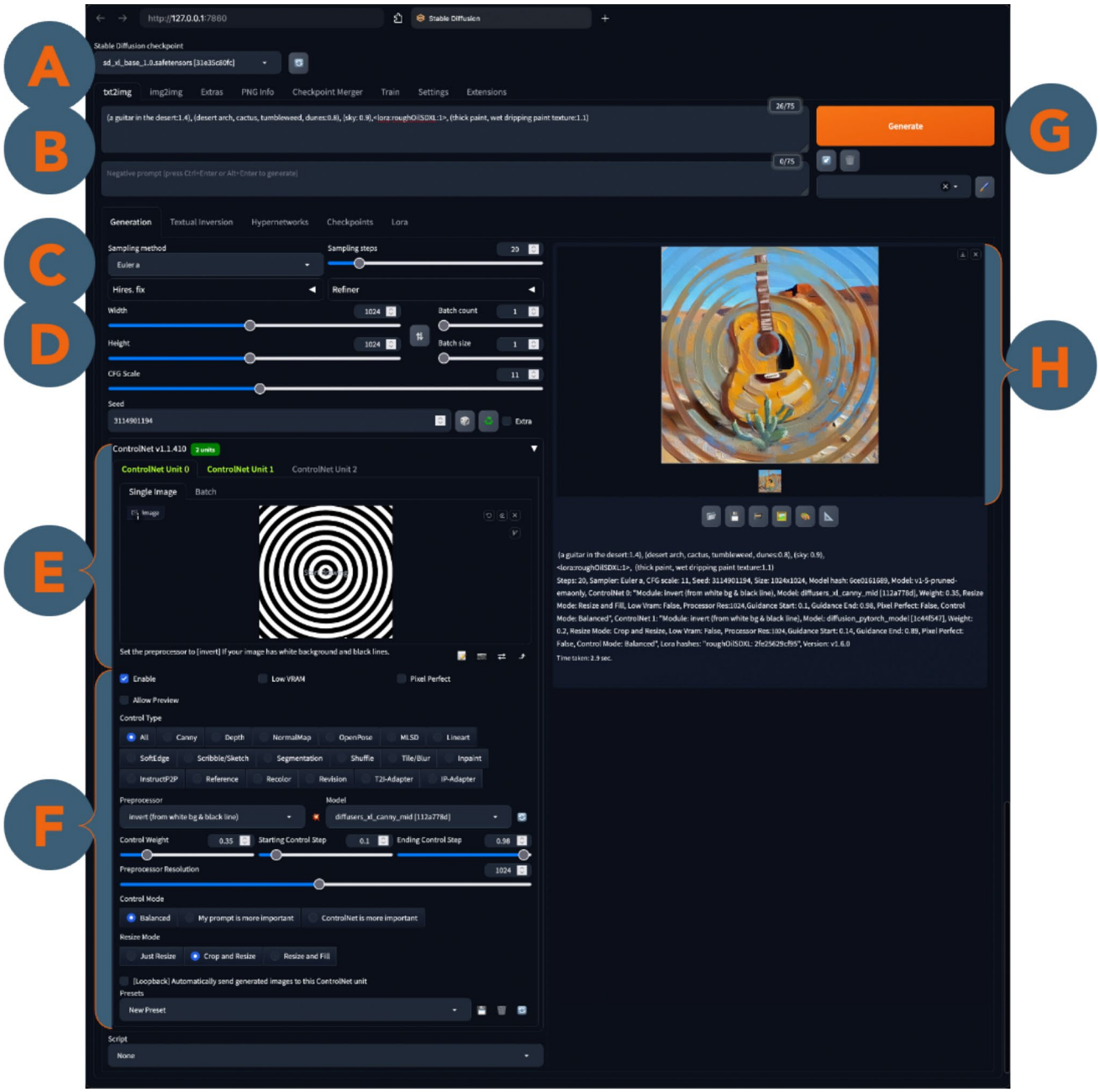



**Fig. 3** Screenshot of example generation within AUTOMATIC1111 Web UI. *Orange letters* are overlaid to denote panes within the interface: **A** Base model selection; **B** Prompt inputs; **C** Sampling parameters; **D** Basic model image parameters; **E** ControlNet input; **F** ControlNet parameters; **G** Generate button; **H** Output gallery from the last generation event

parameter control and integration with its extension library. All images were manually curated at the point of generation and generated with a resolution of 1024 × 1024 pixels. To maximize the subjective recognizability of the four target features, generation parameters were iteratively tuned. While specific settings were optimized at the image level, we primarily employed either the Euler Ancestral or DPM + + 2 M Karras sampling method, typically over 20 sampling steps. Other hyperparameters, such as classifier-free guidance scale (CFG) – which modulates how closely the generation process follows the text prompt – were adjusted per 4-feature grouping and image. Precise metadata for every stimulus, including seeds and specific sampler settings, is embedded in the images, permitting reproduction and modification of any image. A given image is exactly reproducible if the same seed and parameters are used in the same version of the web user interface. If the seed changes while other parameters remain constant, image generation varies; see Fig. S4 for an example of image-generation variance within 4-feature groupings.

## Features

Each image was generated to contain four distinct feature types: setting, object, texture, and pattern (Fig. 1). These four feature types were selected to span the representational hierarchy of the ventral visual stream, from local features in early visual cortex to semantic identity in the temporal lobe (Grill-Spector & Malach, 2004).

Individual features within a given feature type were chosen to minimize both functional and perceptual overlap with other features, as these form the basis for categorization and brain representations (Greene et al., 2016; Martin et al., 2018). For example, "fork" and "chopsticks" have similar functions, and thus only one of those objects would be considered for use. Likewise, despite having very different functions, "spoon" and "shovel" are perceptually similar, and thus only one of those objects would have been considered.

We fixed the pairing of object and setting and the pairing of texture and pattern to reduce the number of possible feature combinations (Fig. 1), while maintaining our goal of manipulating higher-level and lower-level features separately. For example, images with the setting "underwater" always contain the object "teacup." Texture and pattern were likewise paired: if an image was made with the texture "paper," it was also made with the pattern "diamonds." Pairing settings/objects and textures/patterns allowed us to achieve our goal of crossing high-level and low-level feature groupings without the complexity of manipulating each of the four feature types separately. If a given research question requires each feature type to be manipulated separately, however, this approach is feasible in the SPOT framework (see Supplementary Information/Low-level feature separation; Fig. S8).

## High-level feature manipulation

Manipulation of the setting and object depicted in each image was achieved with the Stable Diffusion prompt input. For example, to generate a picture with the setting "field" and object "truck," the prompt might be "a truck in a field." More complex prompt addenda were sometimes used to refine the final setting and object; for instance, "truck in a field" might be appended with "(blue sky and clouds:0.7), (farm:0.2)" to ensure that images contained consistent feature representation across images with the same setting and object, depending on the other variables active in the generation process (see Fig. 4).

Numbers at the end of a parenthetical prompt element reflect parameter weight. These weights control the influence of prompt components on the image. For example, "(farm:0.2)" indicates that farm-related elements should be present, but only subtly. A weight of 1.0 represents the default prompt-component influence – i.e., without parentheses and a weight, 1.0 is the implicit parameter weight. Occasionally, negative prompts were also used. Entered as a separate input, negative prompts guide the model away from specified, unwanted visual elements. For example, if components active in the generation process have biased the model towards producing colorless images, entering ("black and white: 0.6") into the *negative* prompt input might ensure that the model produces a colored image.

## Low-level feature manipulation

Modulating image textures and patterns posed a greater technical challenge, as standard text-to-image AI pipelines are not natively designed for low-level feature control. To modify texture, we employed Low-Rank Adaptation models (LoRAs; Hu et al., 2021). LoRAs fine-tune the SDXL base model's cross-attention layers to depict specific image styles with minimal additional parameters. Because style and content are strongly connected in image models, it can be challenging to manipulate image style while preserving semantic content (Frenkel et al., 2025). To mitigate this, we manually selected publicly available, pre-trained SDXL LoRAs from Civitai and Hugging Face to create images with maximally differentiable textures. Typically, we increased the LoRA weight parameter until just before the other features depicted in the images (e.g., its object) began to become scrambled or absent, to maximize the amount of texture present in the final images. Some LoRAs require activation text in the prompt to trigger their use; in most cases, we added text to the prompt to trigger the LoRA and/or refine the appearance of the given textures in the image (Fig. 4).

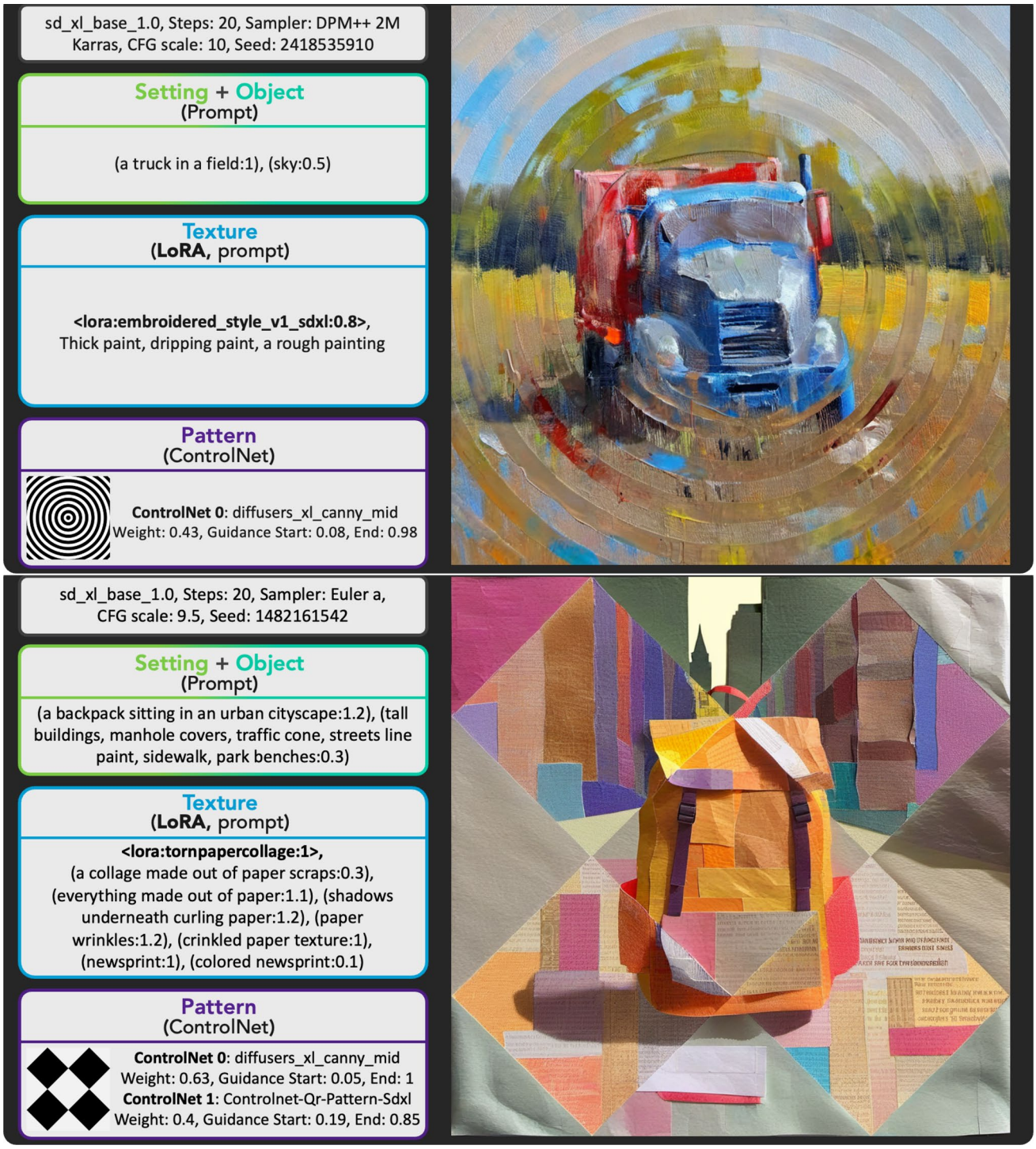



**Fig. 4** Two exemplar images (*right*) and relevant parameters used in their respective generation processes (*left*). Some images were created with fairly simple feature parameters (*top*), while others had more specifications (*bottom*)

To modulate patterns, we employed ControlNet (Zhang et al., 2023) to add low-level edges at specific, consistent locations within images. ControlNet is a neural network architecture that can add spatially localized input conditions to a pretrained text-to-image diffusion model by controlling each level of the U-net within the base model. We employed Canny edge detection (Canny, 1986; Zhang, 2023) as a conditioning method within the ControlNet implementation. Occasionally, an additional ControlNet model called OpticalPattern (Nacholmo, 2024) was used for additional, more

luminance-based pattern control (Fig. 4, bottom). Pattern templates (i.e., the user-imported detectmap images; Fig. 5B, right) were generated manually in Adobe Illustrator and exported at a resolution of 1024×1024 pixels before using the "invert" preprocessor method within AUTOMATIC1111 to prepare an inverted outline-style reference input. ControlNet weights, control mode, and guidance parameters varied across images depending on other active variables, but control mode was typically "balanced," and guidance always spanned the bulk of the generation process (e.g., 0.05–0.98). Like LoRA weight parameters, we increased or extended the ControlNet weight parameters until just before other features of the images began to become scrambled or absent, to maximize the appearance of the pattern edges in the resultant images.

We selected these specific low-level manipulations – texture and pattern – to address the fundamental challenge of manipulating low-level image structure while minimizing impact on high-level/semantic content. While textures provide a means to vary the local statistical properties of a scene, the pattern manipulation (via ControlNet) allows for the enforcement of specific edge orientations and spatial distributions. By imposing a consistent "edge backbone," we create a stimulus set in which images in different high-level categories can share a specific structural blueprint of edge features (e.g., a spiral pattern applied to both a "jungle" and a "bedroom" scene). This approach enables a level of control over low-level features in naturalistic images that is often lost when using unmanipulated image sets.

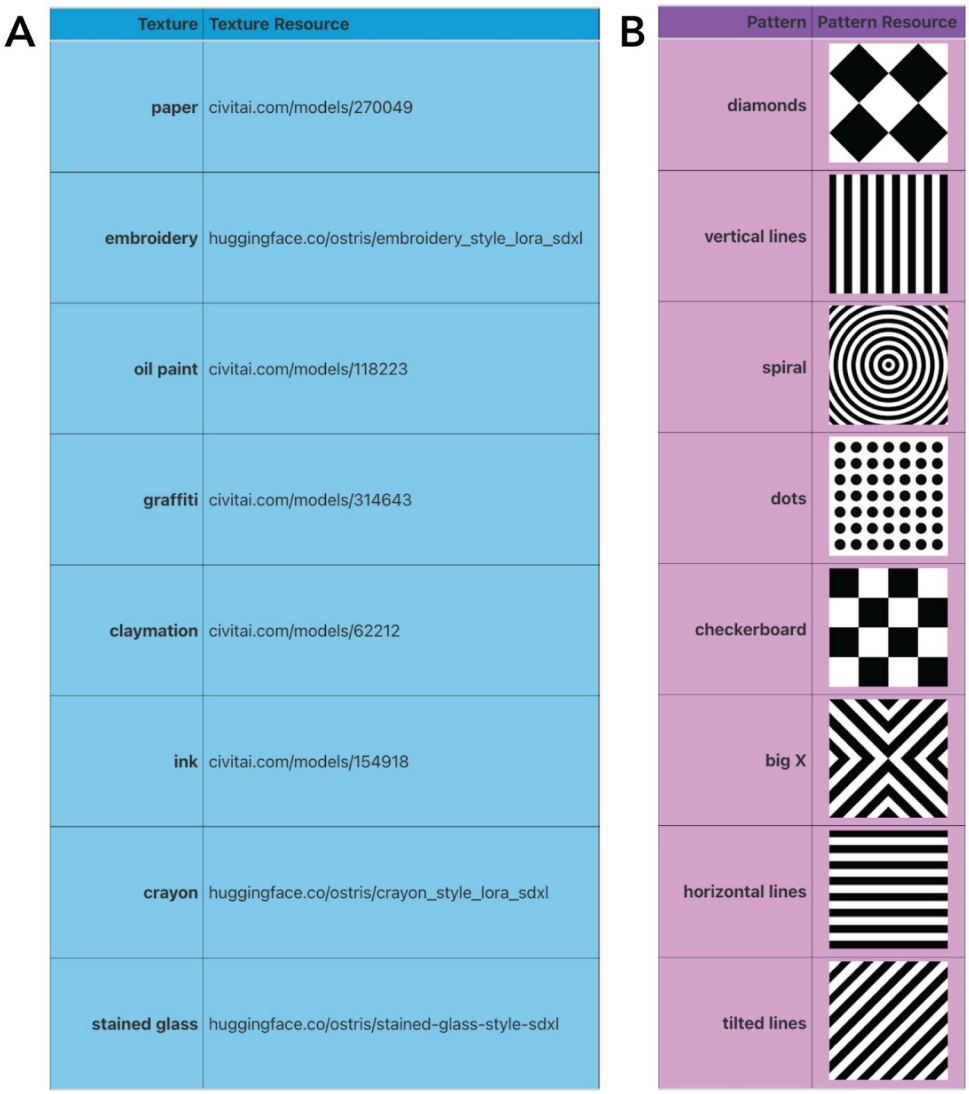



**Fig. 5** Low-level feature pairs. **A** Textures and **B** patterns from the 8×8 grid and the resources used to implement them (see Fig. S3 for the expanded set of low-level feature pairs set within the 10×10 grid)

## Modeling

Before presenting the images to human participants, we used computational models to assess whether we were successful in systematically manipulating high-level and low-level visual features. These models served complementary roles: (1) indexing the degree to which the low-level (texture and pattern) content of the images remained similar when high-level features (object and setting) varied, and (2) indexing the degree to which the high-level content of the images remained similar when low-level features varied (see Supplementary Information for a third modeling approach that evaluates holistic perceptual similarity between images). The methods for using each model share some common steps, which are outlined below, followed by specific details for each model.

We used the EarlyVis model from the img2fmri package (Bennett & Baldassano, 2023) and a Gabor wavelet pyramid (Kay et al., 2008) to capture low-level edge-, texture-, and orientation-specific features characteristic of early stages of the visual cortex. By employing both a data-driven neural network and a formal analytical model, respectively, we ensured that validation of our low-level feature manipulation is not idiosyncratic to a particular model architecture. We used the Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021) and Florence-2 (Xiao et al., 2023) to capture high-level, semantic category-level information. By using both a contrastive dual-encoder and a generative encoder-decoder (respectively), we ensured our high-level feature validation was not idiosyncratic to a single model's training objective. While the internal architectures varied, all models followed a standardized analytical pipeline. For each model, images were input into the respective model to generate feature representations from the images. Then, for each of the models, we compared cosine similarity values between image representations that (a) shared the same low-level features but varied in high-level features (i.e., same texture/pattern, different object/setting); (b) shared the same high-level features but varied in low-level features (i.e., same object/setting, different texture/pattern); or (c) shared neither of our manipulated features (baseline similarity). This allowed us to examine how manipulating features across the image grid influenced both low- and high-level visual representations.

### Modeling of low-level visual content

To model low-level visual content, we employed two complementary approaches: a deep neural network-based prediction of early visual cortical responses and a structural decomposition of images using Gabor wavelets.

We first used the EarlyVis model from the img2fmri package (Bennett & Baldassano, 2023). img2fmri is a Python package of artificial deep neural network models that predict group-level fMRI BOLD responses to images. The model's architecture is based on the pretrained ResNet-18, and it was trained using data from the open-source BOLD5000 project (Chang et al., 2019). The EarlyVis model's output reflects predictions of neural responses located near the calcarine sulcus (from early visual regions, e.g., V1 and V2; see Chang et al., 2019 for definitional details).

Procedures outlined in the img2fmri documentation were followed to generate voxel-wise activation patterns for each image. We then mean-centered each voxel's activity across images. We assessed the similarity of predicted early visual cortex responses for image pairs by computing the cosine similarity between their activation vectors. Pairwise correlations between all image representations were subsequently sorted by whether the images shared low-level features only, high-level features only, or neither.

To complement the EarlyVis model, we performed a Gabor wavelet pyramid analysis (Carlin, 2023). This allowed us to quantify local orientation and spatial-frequency energy in the images, providing a structural, multi-scale metric of low-level information (e.g., Kay et al., 2008; Naselaris et al., 2015). We converted images to grayscale and mean-centered luminance within each image. We used a Gabor bank consisting of six scales ranging from 2 to 64 cycles per field of view (logarithmically spaced) at eight orientations (0°, 22.5°, 45°, …, 157.5°) and two phases (0 and $\pi/2$) to compute quadrature pairs for local energy extraction; for each spatial scale, the stride was adjusted (N/f) to ensure appropriate sampling density across the pyramid. The resulting energy vector was calculated by squaring and summing the quadrature phase responses to all spatially localized energy maps across all scales and orientations. Because naturalistic images inherently share similar statistical structures (1/f power spectrum, global luminance statistics), raw similarity scores between these vectors are often near ceiling; to isolate variance driven by our feature manipulations, we subtracted the grand mean energy vector (across the entire image set) from each individual image vector. As before, pairwise correlations between all image representations were then sorted by whether the images shared low-level features only, high-level features only, or neither.

### Modeling of high-level visual content

Modeling high-level visual similarity is a challenge as it involves capturing more abstract distinctions that may not be reflected in pixel-wise visual representations. To ensure our approach was not biased by a specific model's training objective, we leveraged two different architectures: CLIP (Contrastive Language-Image Pretraining; Radford et al., 2021), a contrastive dual-encoder; and Florence-2 (Florence-2-large-no-flash-attn; Xiao et al., 2023), a generative encoder-decoder.

We first employed CLIP, a deep neural network trained on large-scale image–text pairs to learn joint visual and semantic representations. CLIP is pre-trained to align visual and linguistic information in a shared vector space via a contrastive learning task that requires the model to distinguish correct image-caption pairs from a large set of incorrect examples. At test time, the model can therefore identify the textual description that most closely aligns with a given image's embedding. We extracted textual descriptions of the images from the CLIP model (Radford et al., 2021; ViT-B/32 variant, implemented via the open_clip Python package) before vectorizing them with SBERT (Sentence BERT; Reimers & Gurevych, 2019; specifically the all-MiniLM-L6-v2 model. As with the EarlyVis model, this CLIP + SBERT procedure (henceforth summarized as "CLIP") yielded a single vector for each image. To isolate the variance unique to our images, we applied a cocktail-blank demeaning procedure by computing a global mean vector across all images – representing the "baseline" semantic content of the stimulus set – and subtracted it from each individual image vector. To evaluate the similarity of visual-semantic representations across images, we computed pairwise cosine similarity between these vectors. We then sorted similarity values by whether the images shared only low-level features, only high-level features, or neither.

To verify CLIP results, we employed Florence-2 (Xiao et al., 2023). Unlike the contrastive learning of CLIP, Florence-2 uses a generative, autoregressive approach. It does not "search" for a match in a pre-existing embedding space; instead, it uses a transformer-based decoder to generate a sequence of tokens from scratch based on the visual input – a reconstructive process in which the model builds a description. For each image in our stimulus set, we used the <CAPTION> task prompt within Florence-2 to generate a concise description. Captions were then tokenized and converted into high-dimensional vectors using the same pre-trained SBERT model (Reimers & Gurevych, 2019). As with the CLIP analysis, we applied cocktail-blank demeaning by subtracting the global mean vector from each individual image vector, and computed pairwise cosine similarities between each resultant image vector. These similarity values were subsequently sorted by whether the images shared low-level features only, high-level features only, or neither.

## Model statistics

To test whether the models were significantly sensitive to our feature manipulations, we performed permutation-based significance tests on pairwise similarity scores between images. For each model, image pairs' vectors were categorized into three conditions: sharing high-level features only, sharing low-level features only, or sharing neither (baseline). (See Supplementary Information for an evaluation of the whole image set, for which similarity metrics from images that share all four features are additionally possible.) We then performed permutation tests ($N = 10{,}000$ iterations) to compare the mean similarities across these three conditions. For each comparison (e.g., Shared High-Level vs. Neither Shared), we calculated an observed difference in means. To generate a null distribution, we pooled the similarity scores from the two conditions being compared and randomly reassigned them to two surrogate groups of the original sizes. For each iteration, we calculated the difference in means between these randomly partitioned groups. A two-tailed $p$ value was calculated from the proportion of shuffled trials for which the absolute difference in means met or exceeded our observed result, assessing whether images' representational similarity was significantly driven by our targeted feature manipulations or driven by baseline noise or model bias. To account for multiple comparisons across conditions, all $p$ values for modeling statistics were adjusted using Bonferroni correction.

## Experiment

Initial behavioral pilots of the images from the $10 \times 10$ grid (1594 images) were run to verify that participants ($N = 88$) were generally able to identify features. These pilot results are reported in Supplementary Information. On the basis of pilot data, four feature pairs were removed due to labeling performance. The result was a smaller $8 \times 8$ feature pair grid (Fig. 2). Additionally, data from this initial pilot indicated that some images were associated with poorer feature identification than others. We removed images for which (1) any participant mislabeled one of the four features and (2) mean response times were more than 2 SD above the grand mean. This resulted in a set of 484 images within the $8 \times 8$ grid.

64 images from this subselected set (one image for each 4-feature grouping within the $8 \times 8$ grid) were selected for robust validation and used in the current experiment. In this experiment, each participant judged each feature (setting, object, texture, pattern) once for each of these 64 images from the $8 \times 8$ grid. Some of the remaining 420 images were used for practice and test phases within the main experiment.

## Participants

Participants were recruited from Prolific (Prolific.co) until 30 participants completed the full experiment. Participants were required to be based in the United States, have English as their primary language, and be 18–60 years old (11 female, 19 male; mean age = 37.8; 70% White, 16.7% Black, 6.7% Asian, 6.7% other). Participants who completed the full experiment were compensated a fixed amount ($12.50) at a rate of $15/h for an estimated participation time of 50 min (median completion time after data collection was 48.5 min).

An additional eight participants did not pass a test within three attempts (see Methods/Experiment/Feature Practice and Testing) and so did not continue to the main phase of the experiment (their data are not reported here); they were compensated at the same rate as other participants but based on a reduced estimated completion time.

The sample size of $N = 30$ was determined a priori using pilot data (see Supplementary Information) via the Python Pingouin statistical package. Specifically, we determined the generalized $\eta^2$ effect sizes for the main effect of feature type on accuracy and response time from all pooled pilot data using a repeated-measures ANOVA. Using these empirical effect sizes, required sample sizes were calculated (via pg.power_anova) with an alpha of 0.05 and 80% statistical power. The accuracy data (pilot $\eta^2 = 0.34$) yielded a minimum sample size of $N = 7$; for the response time data (pilot $\eta^2 = 0.18$), a minimum sample size of $N = 14$ was required. To ensure sufficient power, we selected a final sample size of $N = 30$.

## Procedure

Upon recruitment, participants were forwarded via URL to the survey platform Qualtrics for consent processes, before being forwarded via URL to the experiment platform Pavlovia (pavlovia.org). Pavlovia is a website for hosting experiments developed using PsychoJS (the JavaScript version of PsychoPy; Peirce et al., 2019), which was used to implement this experiment.

## Feature training

Participants were first told to maximize their screen brightness and sit a normal distance (~25 inches) from their screen. They were then informed that their task would be to label four features within images: the "setting," "primary object," "style" (texture), and "pattern." They were told the images were created with AI, and that each image was explicitly generated to contain these four feature types. Participants were first given a text list of the possible labels (features) for settings, followed by the analogous lists for objects, then textures, and finally patterns.

We selected settings and objects from common categories that are likely intuitively familiar to our participants. The textures we selected and patterns we created are, however, not necessarily intuitively labelable. To ensure a high degree of mapping consistency between visual features and linguistic labels, the specific labels for our textures (e.g., "claymation" vs. "oil paint") and patterns (e.g., "checkerboard" vs. "diamonds") were taught to our participants. This ensured that the subsequent classification data reflected the participants' perceptual discrimination of the features rather than their familiarity with the provided labels. In total, this additional label-training procedure for textures and patterns took about four and half minutes. To that end, we pre-generated a set of "training" images using the same procedures, textures, and patterns as for the main stimulus set. Each of these training images (Fig. 6A, bottom) was produced with the object "cat" and the setting "rooftop," as these are not features present in the $8 \times 8$ feature grid.

After viewing the text texture list, eight texture-training images were presented simultaneously with their labels. Participants viewed these training images for a minimum of 60 s before they could advance. Participants then viewed each individual texture training image, one by one, for a minimum of 10 s. These images were shown in a video depicting a magnified view of a portion of the image to better highlight its texture, accompanied by a text description of the texture (Fig. 6A, bottom left).

The same framework was then used to introduce the pattern labels. First, eight pattern-training images were presented simultaneously (with their pattern labels) for at least 60 s for study. After the training images for patterns were viewed simultaneously, participants viewed each individual pattern training image one by one for a minimum of 10 s. The images were shown alongside the actual pattern template (Fig. 5B, right) in a video. The videos demonstrated the alignment between the pattern and the generated example image by having red lines appear over the major edges of the black-and-white pattern image before being horizontally translated across the screen onto the example image (Fig. 6A, bottom right).

Before proceeding to the next phase of the experiment, participants were explicitly told that settings and objects were paired, and that textures and patterns were paired, but that no other feature types were reliably paired. This was done to minimize behavioral variability that may occur if some participants realize the feature pairings and others do not. Lastly, before the testing phase, participants were told that they would select the label to a given image's feature by pressing a number on their keyboard.

## Feature practice and testing

Participants then entered a brief practice phase (Fig. 6B, top) to familiarize them with the general procedures. An image from the set of 420 images reserved for practice and testing was presented centrally with a feature report prompt (e.g., "Report: Setting"). All eight label options for the given feature type were presented below the prompt, alongside the keyboard number (1–8) assigned to each label for that participant. Label order was randomly determined for each participant and for each feature type when their participation commenced. In these eight practice trials, participants were allowed up to 20 s to respond. Participants also received

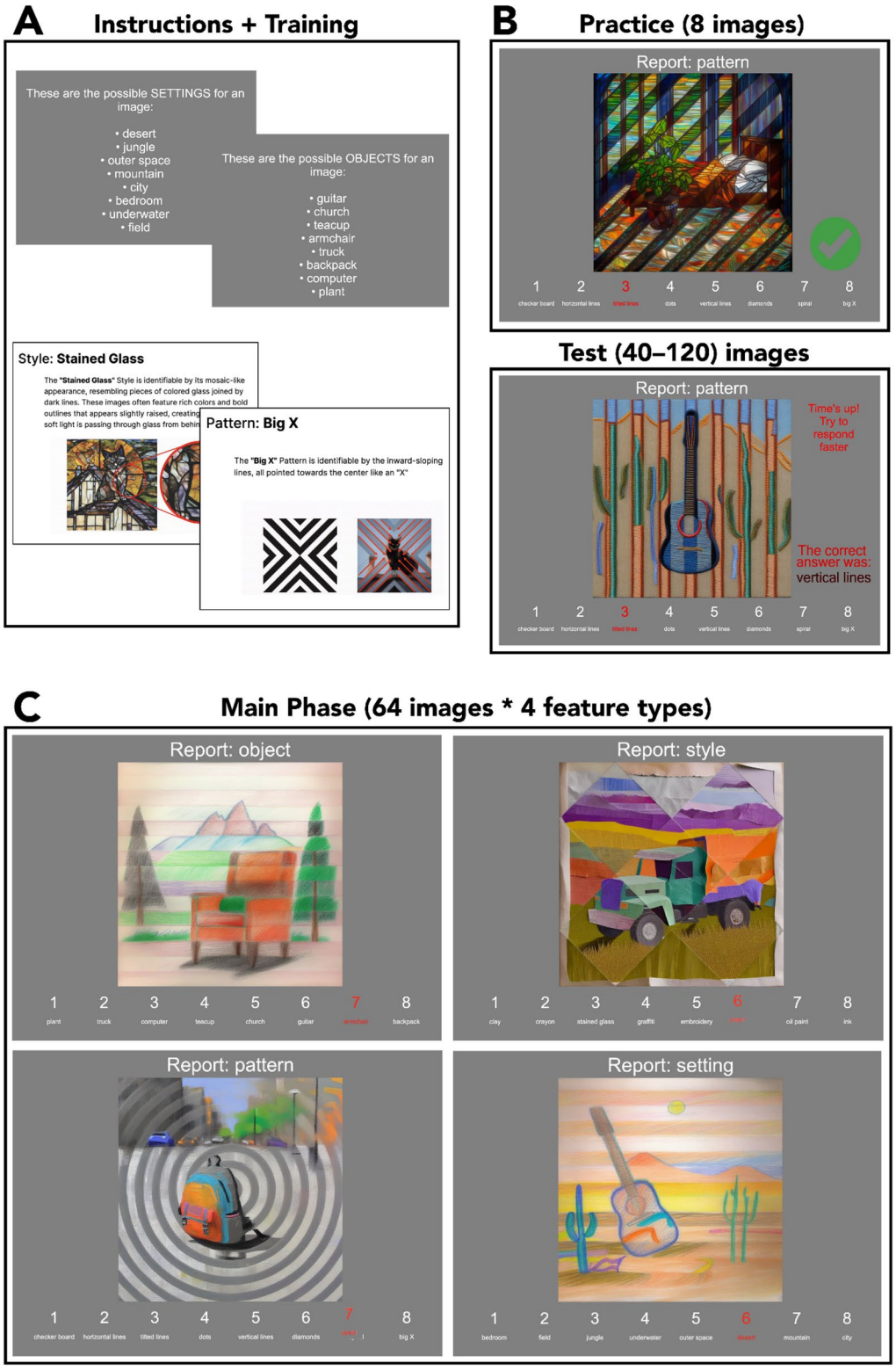



**Fig. 6** Behavioral experiment procedure. **A** Instructions and training phase. Participants were introduced to the task before receiving lists of feature labels for each feature type. To ensure equivalent labeling performance across feature types, participants were given additional low-level label training by being shown texture-training ("style") example images (*bottom left*) and pattern-training example images (*bottom right*). **B** Participants completed a very brief practice section (*top*) before being tested on their labeling performance (*bottom*), for which participants completed 1–3 test phases of 40 images each. **C** Main phase of the experiment

feedback about their label selection after pressing a button corresponding to a label (i.e., "Correct!" if correct, or, e.g., "The correct answer was City"). Participants repeated this procedure for the eight practice images (each feature pair present in one practice image). Each feature type was prompted twice, consecutively and in pairs, but otherwise randomly determined (e.g., setting, setting, texture, texture, pattern, pattern, object, object). This brief practice phase was only implemented to familiarize participants with responding; regardless of their responses, participants proceeded to a testing phase.

The general procedure for the testing phase (Fig. 6B, bottom) was similar to the practice phase. Participants again had 20 s to respond to a given image, but at the 10-s mark, red text appeared saying "Time's up! Try to respond faster." Participants were still allowed to respond until the image had been on-screen for 20 s, but responses were only considered correct for the purposes of the test if they were provided within 10 s of image onset. Participants repeated this testing procedure for 40 images (drawn from the 420-image set reserved for practice and testing) per test phase. Each feature type was tested in ten-image miniblocks. The order of feature miniblocks was randomly determined per test. The first 15 participants in our dataset were required to have correctly responded to $\geq 38$ of test trials to pass; for the second half of data collection, this threshold was dropped to $\geq 34$ to reduce the testing phase failure rate.

After each test, a 60-s break was provided alongside an encouraging message and performance feedback on the prior test. Participants were only permitted three attempts to pass the test. After passing the test, participants entered the main phase of the experiment. Thus, some participants only viewed 40 testing images (plus the eight practice images) before proceeding to the main phase, while others viewed 80 or 120. Of the participants for whom data are reported: 43.3% completed the testing phase on their first try; 46.7% completed the testing phase on their second attempt. If a participant did not pass by the end of the third test phase, they were informed that their participation would be terminated, and were forwarded via URL back to Prolific for partial compensation at the promised rate (calculated based on median time to fail all three tests).

### Image feature classification

Upon passing the test, participants began the main phase of the experiment (Fig. 6C). These trials were almost exactly like the testing trials: again, an image from the image set was presented centrally above a feature prompt (e.g., "*Report: Setting*") with the eight feature labels displayed below alongside their corresponding key number assignment. This main phase comprised four blocks in which each of the images from the 64-image set was displayed once. Each block was assigned a feature type prompt, such that all 64 images' setting, object, texture, and pattern were tested in separate blocks. Block order was randomized across participants. Unlike the test phase, images were presented for only 10 s per judgment, and no feedback was presented as to the participants' accuracy on a given judgment.

### Behavioral modeling

We analyzed participants' performance using linear mixed-effects models (LMMs) for response times and generalized linear mixed-effects models (GLMMs) for accuracy. This analysis was conducted using R, version 4.0.5 (R Core Team, 2022), with the lmer and glmer functions from the lme4 package (Bates et al., 2015). Response times (RT) for correct trials were log-transformed to satisfy normality assumptions and modeled using an LMM. Accuracy was modeled using a binomial GLMM with a logit link function.

Both models included feature type (Setting, Object, Texture, Pattern) as a fixed factor and a random intercept for participants to account for individual differences in baseline performance. We selected these random effects by starting with a maximal model (with all random effects specified) and then testing for over-parameterization with a principal component analysis of the random-effects variance–covariance estimates (using the rePCA function of the lme4 package) to determine the random effects that the data were capable of supporting (Bates et al., 2015). Excluded random-effects terms were removed based on least explained variance until the model converged. The Akaike information criterion indicated that removing terms identified by the PCA improved goodness of fit. Predictor estimate significance (based on conditional F-tests with Kenward–Roger approximation for the degrees of freedom) was computed using the *tab_model* function from the sjPlot package (Lüdecke, 2018; Wickham, 2018).

We employed effect coding for the feature type predictors, so the model intercept represents the grand mean (the average log RT for latency; the average log-odds for accuracy) across conditions. Individual coefficients represent the deviation of each specific feature type from that grand mean. For the RT model, predicted values were exponentiated to report effects as back-transformed differences (BTD) in seconds. Results are reported as beta estimates, standard errors, and *p* values, with odds ratios provided for the accuracy model to facilitate interpretation.

## Results

### Image modeling

We first tested how our visual-feature manipulations influenced image representations in the 64-image set selected

for additional behavioral testing (see Supplementary Information for modeling results across all 1594 images). To do so, we examined similarity metrics between image representations from four computational models. For low-level image representation analysis, we employed the img2fmri EarlyVis model and a Gabor wavelet pyramid analysis; for high-level image representation analysis, we utilized CLIP and Florence-2 models. For each model, we generated similarity matrices based on their vector outputs and compared image pairs based on whether they shared high- or low-level feature pairs (see Fig. 7).

The img2fmri EarlyVis image representations (Fig. 7, top left) showed differential sensitivity to our feature type manipulations. Representational similarity was significantly higher for image pairs sharing only texture/pattern features compared to baseline (those sharing no features; $\Delta = 0.310$; permutation test, $p < 0.0001$). Image pairs that shared only setting/object features also had significantly higher correlations than those with no shared features ($\Delta = 0.159$; $p < 0.0001$). Critically, however, a direct comparison revealed that correlations were much higher for images sharing only texture/pattern than for those sharing only setting/object ($\Delta = 0.151$; $p < 0.0001$). This suggests that although EarlyVis is sensitive to our high-level feature manipulation (perhaps because settings/objects are necessarily reliably composed of particular low-level features), it is disproportionately sensitive to our low-level feature manipulations, consistent with an architecture modeling early visual cortex representations.

The Gabor Wavelet Pyramid Analysis (Fig. 7, bottom left) yielded a similar pattern of results. Image pairs sharing only texture/pattern features showed significantly higher similarity compared to the baseline of no features shared ($\Delta = 0.164$; $p < 0.0001$). While image pairs sharing only

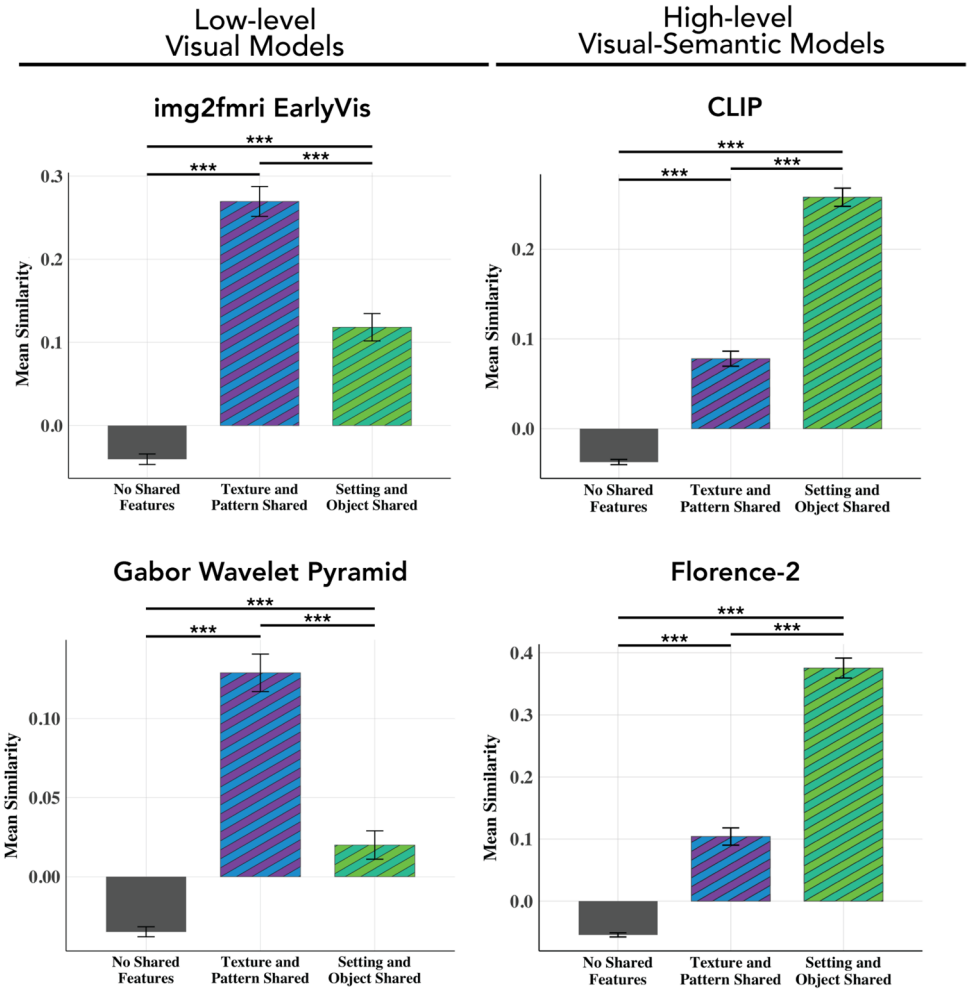



**Fig. 7** Image modeling results on the subselected 64-image set from the 8×8 grid. *Left*: Low-level visual models. *Right*: High-level visual-semantic models. For each model, average image pair representational similarity is plotted for (1) images that share neither setting/object nor texture/pattern ("no shared features"), (2) images that share only texture/pattern features, and (3) images that share only setting/object features. *Error bars* represent SEM; ***$p < .0001$

setting/object features also exhibited higher similarity than the baseline ($\Delta = 0.055$; $p < 0.0001$), the Gabor wavelet representations were significantly more sensitive to our texture/pattern feature manipulation than to our setting/object feature manipulation ($\Delta = 0.109$; $p < 0.0001$). These results validate that our low-level feature manipulations successfully captured the local orientation and spatial frequency energy targeted by structural models of low-level visual content.

In contrast, CLIP representations (Fig. 7, top right) exhibited the reverse pattern. Image pairs sharing only setting/object features had substantially higher cosine similarity than baseline ($\Delta = 0.295$; $p < 0.0001$). Image pairs sharing texture/pattern features also showed elevated similarity over baseline ($\Delta = 0.115$; $p < 0.0001$). Critically, however, image pairs that only shared setting/object features were more similar than those that only shared texture/pattern ($\Delta = 0.180$; $p < 0.001$). This suggests that although CLIP captures some elements of our low-level feature manipulation, it is primarily sensitive to our high-level feature manipulation, which aligns with CLIP's objective to encode high-level visual-semantic associations over lower-level ones.

The Florence-2 representations (Fig. 7, bottom right) yielded a similar preference for our high-level feature manipulation. While sensitivity to our texture/pattern feature manipulation was present relative to baseline ($\Delta = 0.158$; $p < 0.0001$), image pairs sharing only setting/object features showed a much larger increase in similarity compared to baseline ($\Delta = 0.430$; $p < 0.0001$), as well as to those sharing only texture/pattern features ($\Delta = 0.271$; $p < 0.0001$). This confirms that Florence-2's generative, caption-based architecture is highly tuned to how we implemented our high-level feature manipulation.

Together, these results validate that our feature pair manipulations in the SPOT Grid modulated distinct levels of visual representation. The EarlyVis and Gabor wavelet models were most sensitive to manipulations of texture and pattern, whereas CLIP and Florence-2 were primarily driven by manipulations of setting and object. This double dissociation confirms that our image generation process successfully manipulated low- and high-level visual content with some independence. These results are strikingly consistent with our supplementary analysis of the 1594-image set from the initial $10 \times 10$ grid (Fig. S6), indicating that this double-dissociation is a robust characteristic of the entire SPOT Grid.

Image representations from a third model, DreamSim (Fu et al., 2023), trained on human similarity judgments and designed to represent images "holistically," exhibited comparable sensitivity across both feature manipulations – suggesting that our feature manipulations may have altered low- and high-level visual content in a balanced manner; see Supplementary Information; Fig. S7.

## Human behavior

In the human behavior validation experiment of the 64-image subset, participants labeled image features with a high degree of accuracy (raw grand mean of participant-level accuracies = 96.8%, SD = 0.18), and responded quickly (raw grand mean of participant-level response times [RTs] = 2.28 s, SD = 1.17).

Participants' accuracy varied by feature type (Fig. 8, left). The grand mean log-odds of a correct response ($b = 3.55$, SE = 0.08, $z = 47.55$, $p < .001$) corresponded to a model-estimated accuracy of 97.2%. Relative to

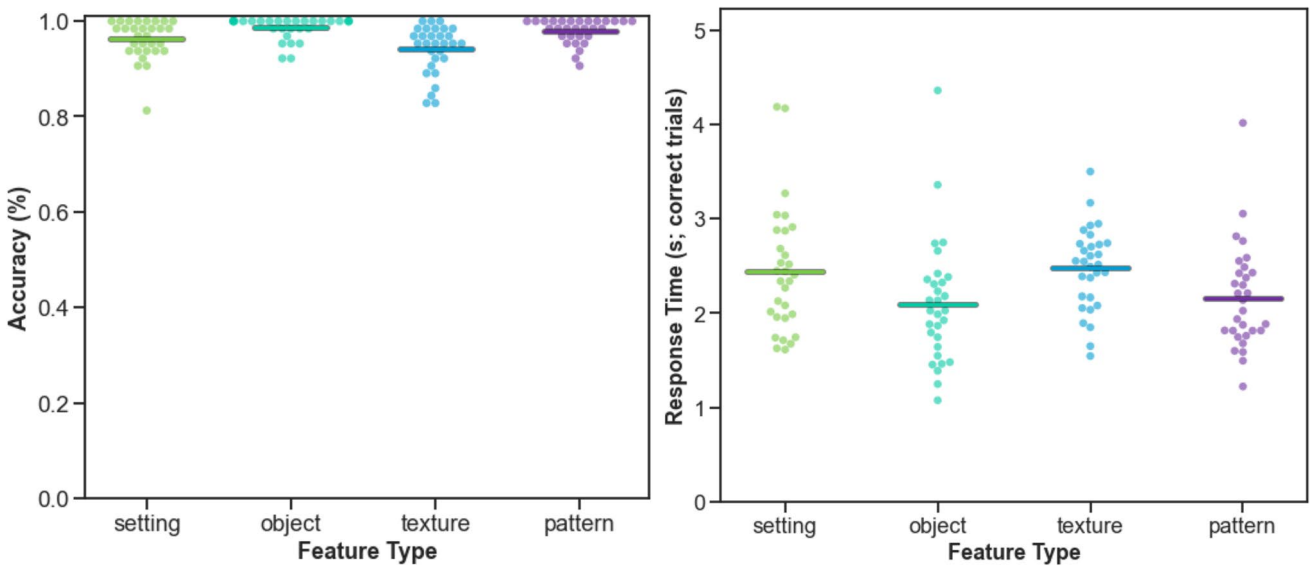



**Fig. 8** Feature judgment accuracy (*left*) and response time (for correct trials; *right*) plotted by feature type

this average: accuracy was higher for object judgments ($b = 0.74$, SE = 0.16, $z = 4.68$, $p < .001$, odds ratio = 2.10; model-estimated accuracy = 98.6%) and pattern judgments ($b = 0.33$, SE = 0.14, $z = 2.40$, $p = .016$, odds ratio = 1.39; model-estimated accuracy = 98.0%); accuracy was lower for setting judgments ($b = -0.30$, SE = 0.11, $z = -2.68$, $p = .007$, odds ratio = 0.74; model-estimated accuracy = 96.3%) and texture judgments ($b = -0.77$, SE = 0.10, $z = -7.55$, $p < .001$, odds ratio = 0.46; model-estimated accuracy = 94.2%).

Participants' response times also varied by feature type (Fig. 8, right). A linear mixed-effects model on log-transformed RTs revealed a geometric grand mean RT of approximately 2.05 s. Relative to this grand mean: RTs were faster for object ($b = -0.082$, SE = 0.008, $z = -10.23$, $p < .001$; back-transformed difference [BTD] = 0.16 s faster) and pattern judgments ($b = -0.054$, SE = 0.008, $z = -6.69$, $p < .001$; BTD = 0.11 s faster); RTs were slower for setting ($b = 0.062$, SE = 0.008, $z = 7.64$, $p < .001$; BTD = 0.13 s slower) and style judgments ($b = 0.074$, SE = 0.008, $z = 9.09$, $p < .001$; BTD = 0.16 s slower).

To better understand the nature of participants' errors at the feature level, we constructed a confusion matrix illustrating the percentage of trials in which each true feature was labeled as each possible response (Fig. 9). Overall, participants exhibited few confusions across all features, with diagonal entries on the matrix (correct identifications) mostly ranging from 95 to 99%. Confusions were generally sporadic and did not reveal systematic mislabeling between features. Some deviation was observed in texture judgments, where accuracy ranged between 88 and 97%. No feature pair was confused at rates exceeding 3%, except for one: images depicting the texture "graffiti" were mislabeled as "oil paint" in 8% of trials. This elevated confusion rate was primarily driven by three participants (out of 30) who consistently confused these textures. When these three participants were excluded in a post-hoc analysis, the confusion rate between "graffiti" and "oil paint" dropped to 3%, aligning

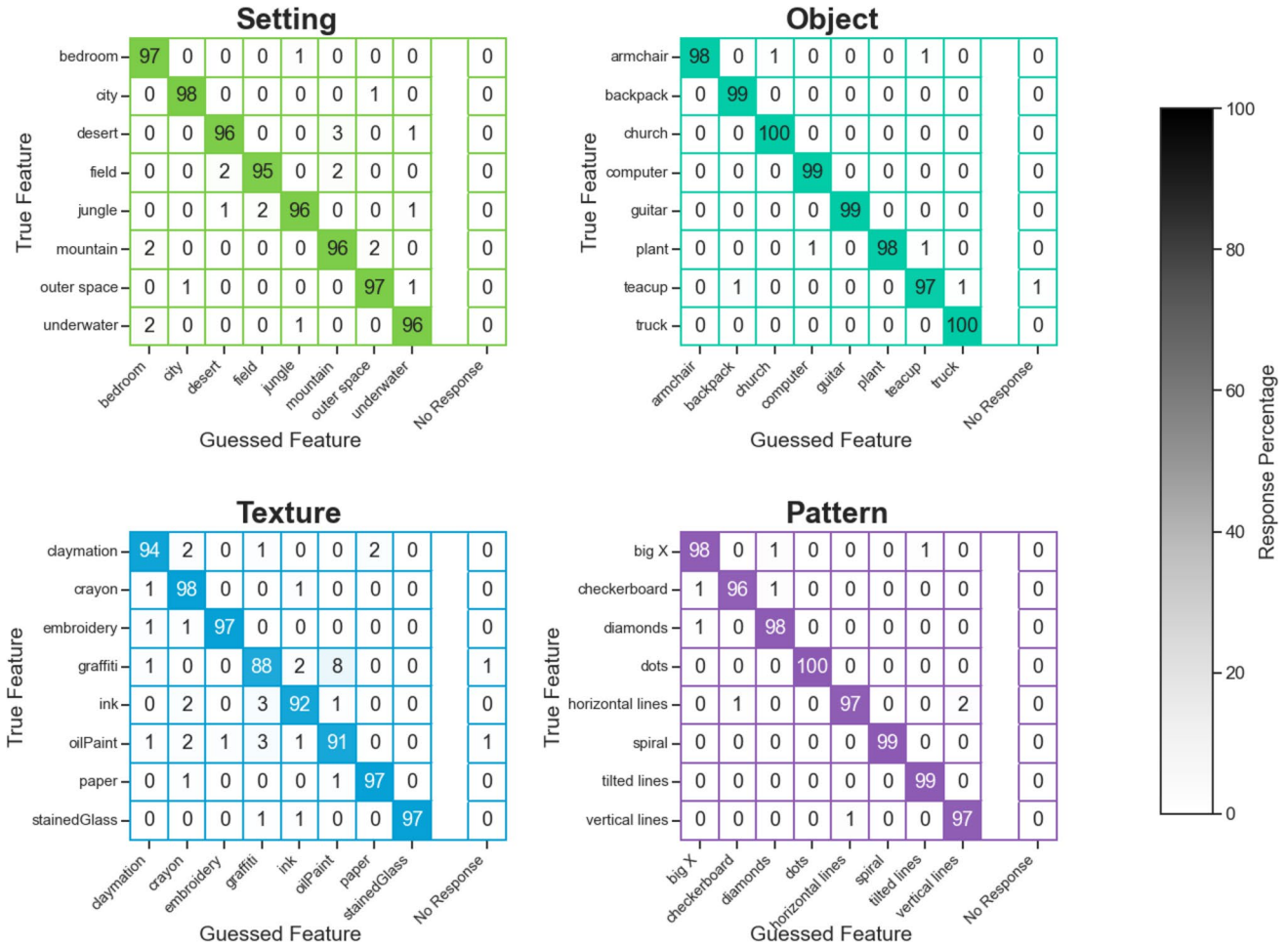



**Fig. 9** Confusion matrices for each feature type. Matrices depict the percentage of trials for which participants guessed a feature (*x*-axis) given the true feature the image depicted (*y*-axis). Percentages are rounded to the nearest integer

with confusion rates seen for other features. This suggests that while a few participants can confuse these categories, this confusion is not pervasive across participants. Future users of this image set might consider additional training focusing on these textures to reduce such errors, if ceiling performance is desired. (See Fig. S5 for human behavior results from the pilot, which includes the expanded features from the $10 \times 10$ grid.)

## Discussion

A central challenge in the study of natural vision is that features at different levels of the visual hierarchy are "entangled" in natural scenes: higher-level visual features (like settings and objects) are systematically constituted by specific lower-level features (like textures and edge patterns); a "jungle" is not just a semantic label, but a predictable collection of leafy textures and particular edge distributions. As such, independent manipulation of these different types of visual features remains unattainable in naturalistic scene images. Historically, researchers have faced a methodological compromise: use scrambled, "unnatural" stimuli to isolate low-level features, or use rich, natural images at the cost of experimental control. The SPOT (Settings, Patterns, Objects, Textures) Grid aims to mitigate this tradeoff. By leveraging the modular architecture of Stable Diffusion – specifically the structural guidance of ControlNet and the stylistic precision of LoRA – we factorially crossed high-level setting–object feature pairs with low-level texture–pattern feature pairs to partially decouple these features in naturalistic scene images.

Our validation confirmed that our generation process could separately target low-level or high-level features and further showed that the manipulated features were perceptually identifiable. Computational models of low-level visual processing (img2fmri EarlyVis, Gabor wavelet pyramid) and of high-level/semantic image properties (CLIP, Florence-2) revealed preferential tuning to our low- and high-level features, respectively, confirming that our manipulations targeted the intended levels of the hierarchy: setting/object features were represented similarly regardless of their lower-level texture/pattern features, and texture/pattern features were represented similarly regardless of the higher-level settings/objects they constitute. Furthermore, human behavioral data demonstrated rapid and accurate identification of all four feature types in each scene image with minimal confusion, confirming the perceptual intelligibility and differentiability of features within the images.

Traditional approaches typically face a choice between using artificial stimulus modulations or natural images. Artificial stimulus modulations (e.g., phase scrambling, texture synthesis) can isolate features, but they disrupt ecological validity and potentially mask real-world dynamics (Balas, 2012; Felsen & Dan, 2005; Hasson et al., 2010; Mante et al., 2008; Olshausen & Field, 2005; Simoncelli & Olshausen, 2001). In contrast, natural images preserve ecological validity but exhibit uncontrolled co-variation between feature types, requiring post hoc computational modeling to probe specific feature-level contributions. While both methodological approaches remain key to understanding how feature levels contribute to cognition, our framework provides a new "bridge" method that maintains naturalistic "anchors" – identifiable real-world objects and settings, coherent 3D spatial layouts, object–background segmentation, etc. – while allowing factorial control over features at different levels of the visual hierarchy. This lets researchers build hypothesis-relevant feature manipulations into their experimental designs while retaining much of the richness of natural scenes.

The SPOT framework offers potential value in cognitive domains where separating feature-level information is desired, such as memory, attentional guidance, concept learning, and aesthetic preference (Bone et al., 2020; Hayes & Henderson, 2021, 2025; Huebner & Gegenfurtner, 2012; Iigaya et al., 2021; Morales-Torres et al., 2024; Pedziwiatr et al., 2021; Walter et al., 2024; Wutz et al., 2018). It may also benefit research on the mechanics of perceptual systems (e.g., Reddy et al., 2021; Richter et al., 2024; Stecher et al., 2025).

Our work shows the growing utility of generative models for precise experimental control. We invert the typical logic of using neural networks to model perception and instead use the model's latent structure to synthesize stimuli with highly targeted properties along the feature hierarchy. Much of the prior work has focused on generative adversarial networks (GANs) (Cooper et al., 2023; Dado et al., 2023; Goetschalckx et al., 2021; Karras et al., 2019; Son et al., 2022). For example, Son et al. (2022) used GANs to systematically control image properties, creating a "scene wheel" tool that transitions visual properties between realistic scenes along a circular continuum. Advanced GANs can disentangle scene features (Yang et al., 2021), but much of their use has focused on generating face stimuli across disentangled dimensions like pose, facial expression, age, etc. (Kammoun et al., 2022). GANs, however, face limitations in terms of computational resources, image resolution, mode collapse, limited image variety, and image artifacts (Gui et al., 2020; Pettini et al., 2024); for these and other reasons, GANs have been surpassed in many respects by the kind of diffusion-based models used herein (Dhariwal & Nichol, 2021).

While diffusion-based models have resulted in some extremely large generated image sets – e.g., DiffusionDB, with 14 million images (Wang et al., 2023) – they have thus far received less attention in cognitive domains. Relevant diffusion-based pipelines have been used to generate visual

anagrams by holding local image parts constant and varying their global arrangement (Boger & Firestone, Forthcoming; Doshi et al., 2025), which functionally keeps textures constant across images depicting different objects. This image set, however, lacks naturalistic scene images, may have a limited range of textures it can represent, and cannot maintain spatially consistent low-level edge information across image transitions. Another recent set used the SDXL base model and varied low-level properties within a given semantic category (Pettini et al., 2024), but did not allow comparisons across categories with matched low-level features. We used the same diffusion base and added LoRA and ControlNet extensions, enabling systematic manipulation of both style and edge properties.

High-level visual concepts are commonly operationalized as object and scene categories, given their ecological relevance and the difficulty of predicting them from raw pixels (DiCarlo et al., 2012), but the choice of which low-level features to target for our stimuli was less clear. In classical visual processing hierarchies, early-to-intermediate representations encompass many properties, ranging from contrast and local spatial frequency to summary statistics, surface curvature, and color distributions (Cant & Goodale, 2007; Freeman & Simoncelli, 2011; Groen et al., 2017; Long et al., 2018). Within the SPOT Grid framework, our operationalization of "low-level" features was deliberately focused on two major, complementary axes: spatial layouts of edges ("patterns") and surface properties ("textures"). This selection was driven primarily by the theoretical relevance of these features' correspondence to representations in early visual cortex: our "pattern" manipulation enforced localized edge contours corresponding to early V1-like orientation features (Canny, 1986; Hubel & Wiesel, 1962; Walther et al., 2011), whereas "textures" modulated surface statistics and color distributions known to be represented in more anterior visual regions like V2 and V4 (Freeman & Simoncelli, 2011; Movshon & Simoncelli, 2014; Portilla & Simoncelli, 2000; Winawer & Witthoft, 2015). Additionally, these two features can be directly controlled through extensions of contemporary diffusion models (ControlNet residual spatial conditioning and LoRA cross-attention fine-tuning). Nevertheless, this choice constitutes a deliberate design boundary for the SPOT Grid images, and our image set does not independently isolate nor systematically manipulate every low-level visual property (e.g., unmanipulated global spatial frequency bands or isolated chromaticity channels). Researchers seeking to isolate specific alternative feature dimensions may need to apply post-hoc image filtering or extend the SPOT framework using different extension models that manipulate those alternative features.

The SPOT grid manipulates two pairs of four features to limit factorial complexity and provide proof of concept. Our framework does, however, have the potential to enable manipulation of more than just two feature levels at once (see Supplementary Information/Low-level feature separation for separate manipulations of texture and pattern). For example, this framework may be able to expand the current two-dimensional grid into a four-dimensional grid that independently crosses setting, object, texture, and pattern – or even additional features. Probing the visual hierarchy this way could lead to better models of cognition. However, more diverse feature sets and neuroimaging validation are needed before we can make strong claims about how well these methods can disentangle visual composition. As generative tools improve, however, they may allow researchers to create flexible, naturalistic stimuli that make these studies increasingly precise.

Despite its strengths, the SPOT framework has limitations. Our approach reduces, but does not fully disentangle, the inherent coupling between high-level and low-level visual features. While our framework offers greater compositional control than conventional image sets, it does not – and foreseeably cannot – eliminate the inherent coupling between high- and low-level visual features. Scene categories are, to some degree, inseparable from their composite low-level attributes. Jungle scenes, for example, are consistently defined by foliage, leafy textures, organic striations, green hues, tangled edge patterns, etc. Likewise, kitchen scenes are defined by particular objects, surface types, and structured layouts. At least some of these low-level properties must be present to evoke the perception of their associated high-level category (Walther et al., 2011). This interdependence reflects how the visual world is structured. As such, even when we carefully manipulate lower-level properties of scene images, deep interdependence between feature levels remains.

Crucially, while we factorially manipulate textures and edge patterns, other features – especially mid-level features – may remain entangled with high-level identity. These mid-level features include color prototypicality, global form (coarse spatial layout), scale and proportion, volumetric contour (3D mass), affordances, characteristic silhouette (2D outline), etc. (Bonner & Epstein, 2017; Leopold & Afraz, 2021; Long et al., 2016; Ostwald et al., 2008; Zeki & Marini, 1998). An image of a "backpack," for example, necessarily maintains certain characteristics to remain identifiable, even when its low-level properties (specific edges and textures) are altered. An image of a "backpack," for example, is presumably inextricable from many of its affordances: its capacity for containment as a hollow volume, its portability, graspability, and wearability. These affordances persist in our "backpack" stimuli regardless of the specific low-level properties that we manipulate, i.e., the texture in which the image is rendered or how the image's edge densities and orientations vary. For an object to be perceivable as a backpack, such affordances associated with mid-level visual

representations must, at least in part, be present – illustrating how high-level identity is constrained by mid-level structural features.

Further, residual coupling between low-level and high-level features may be more pronounced for some categories than others. Future work might achieve greater decoupling by selecting categories with more flexible visual attributes, such as replacing specific objects with more abstract functional categories. For example, one might replace "backpack" with "object for carrying" to increase low-level variability within the high-level category. This specification might also reduce the entanglement of mid-level affordances like "wearability" – e.g., a bucket is an "object for carrying" but cannot be worn – though it would not necessarily reduce the entanglement of other affordances like graspability (e.g., a bucket can still be grasped like a backpack). Our approach is flexible enough to accommodate such an approach, and this ability to control and cross features systematically represents a methodological advance that can support further attempts to manipulate information at different levels of the visual hierarchy.

We used generative models off-the-shelf to demonstrate the viability of this framework. While this approach affords ease of use, future work could benefit from advancing more integrated techniques. For instance, researchers could leverage perception models in a closed-loop manner to systematically tune generation parameters toward maximal between-feature distances and minimal within-feature noise or overlap. Neuroimaging studies using these stimuli would also help clarify which brain regions are selectively responsive to setting, object, texture, and edge-level information – and how such information is integrated or segregated across the cortical hierarchy.

We encourage researchers to consider several factors when using the SPOT Grid or framework. While objects and settings are relatively intuitive and familiar categories to label, participants are less likely to label textures and patterns spontaneously. Our participants underwent category training to ensure reliable labeling of textures and patterns, and we provide these procedures publicly (upon request: forms.gle/X2HpzRYUGigFS1S27). Researchers using these images should consider the demands of their paradigm – for example, whether ceiling-level labeling is desired – and adapt any training or stimulus selection accordingly. Second, although feature accuracy and representational differences were robust on average, variability exists at the level of individual images. Researchers may wish to pretest, filter, or supplement stimuli from our image sets depending on their specific goals and experimental designs. Lastly, we found that for some combinations of components in the generation process, intelligible scene images could not be reliably produced. This suggests that the pipeline's quality and generalizability depend on the representational flexibility of the underlying elements. For example, if a LoRA was trained on a very small or unique set of images (e.g., paintings from 16th-century Italy), it plausibly impedes the diffusion models' ability to reliably depict a given object outside of that training set (e.g., airplanes).

## Conclusion

We introduced a novel stimulus generation framework that enables researchers to partially disentangle high- and low-level visual features in naturalistic images. This SPOT Grid framework provides a validated, adaptable, and relatively off-the-shelf tool for probing how various types of compositional visual information are encoded across computational models, human behavior, and the brain. This framework opens new opportunities for precise, controlled, and naturalistic experimentation in visual cognition.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-026-03211-w.

**Acknowledgements** This work was funded by a National Institutes of Health Research Project Grant (R01EY034436 to M.A. and C.B). We thank the Alyssano Group for valuable feedback on this project.

**Authors' contributions** **Zall Hirschstein**: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – original draft, Writing – review and editing; **Mariam Aly**: Conceptualization, Methodology, Funding Acquisition, Resources, Project Administration, Supervision, Writing – original draft, Writing – review and editing; **Chris Baldassano**: Conceptualization, Methodology, Funding Acquisition, Resources, Project Administration, Supervision, Writing – original draft, Writing – review and editing.

**Funding** National Institutes of Health Research Project Grant R01EY034436 to M.A. and C.B.

**Data Availability** Materials, data, and experiment code are available upon request at forms.gle/X2HpzRYUGigFS1S27.

## Declarations

**Conflicts of interest/Competing interests** None.

**Ethics approval** This study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Institutional Review Board (IRB) in the Human Research Protection Office of Columbia University.

**Consent to participate** Informed consent was obtained from all participants included in this study.

**Consent for publication** Not applicable.

**Open practices statement** None of the experiments were preregistered.

## References


AUTOMATIC1111. (2022). *Stable Diffusion Web UI* [Python]. https://github.com/AUTOMATIC1111/stable-diffusion-webui (Original work published 2022)

Balas, B. (2012). Contrast negation and texture synthesis differentially disrupt natural texture appearance. *Frontiers in Psychology, 3*, 515. https://doi.org/10.3389/fpsyg.2012.00515

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., & Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Computational Biology, 9*(8), Article e1003167. https://doi.org/10.1371/journal.pcbi.1003167

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bennett, M., & Baldassano, C. (2023). Img2fmri: A Python package for predicting group-level fMRI responses to visual stimuli using deep neural networks. *Aperture Neuro*. https://doi.org/10.52294/001c.87545

Boger, T., & Firestone, C. (Forthcoming). *Visual anagrams reveal high-level effects with 'identical' stimuli*.

Bone, M. B., Ahmad, F., & Buchsbaum, B. R. (2020). Feature-specific neural reactivation during episodic memory. *Nature Communications, 11*, Article 1945. https://doi.org/10.1038/s41467-020-15763-2

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences of the United States of America, 114*(18), 4793–4798. https://doi.org/10.1073/pnas.1618228114

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*(6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Cant, J. S., & Goodale, M. A. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex, 17*(3), 713–731. https://doi.org/10.1093/cercor/bhk022

Carlin, J. (2023). *Jooh/gwp* [Python]. https://github.com/jooh/gwp (Original work published 2018)

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data, 6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Chen, Y., Shao, G., Shum, K. C., Hua, B.-S., & Yeung, S.-K. (2025). Advances in 3D neural stylization: A survey. *International Journal of Computer Vision, 133*(8), 5026–5061. https://doi.org/10.1007/s11263-025-02403-9

Cooper, P. S., Colton, E., Bode, S., & Chong, T.T.-J. (2023). Standardised images of novel objects created with generative adversarial networks. *Scientific Data, 10*(1), Article 575. https://doi.org/10.1038/s41597-023-02483-7

Dado, T., Papale, P., Lozano, A., Le, L., Wang, F., van Gerven, M., ... Güçlü, U. (2023). *Brain2GAN: Feature-disentangled neural coding of visual perception in the primate brain* (p. 2023.04.26.537962). bioRxiv. https://doi.org/10.1101/2023.04.26.537962

Dhariwal, P., & Nichol, A. (2021). *Diffusion Models Beat GANs on Image Synthesis* (arXiv:2105.05233). arXiv. https://doi.org/10.48550/arXiv.2105.05233

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Doshi, F. R., Fel, T., Konkle, T., & Alvarez, G. (2025). *Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models* (arXiv:2507.00493). arXiv. https://doi.org/10.48550/arXiv.2507.00493

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience, 8*(12), 1643–1646. https://doi.org/10.1038/nn1608

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201. https://doi.org/10.1038/nn.2889

Frenkel, Y., Vinker, Y., Shamir, A., & Cohen-Or, D. (2025). Implicit style-content separation using B-LoRA. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision – ECCV 2024* (Vol. Vol. 15068, pp. 181–198). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72684-2_11

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). *DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data* (arXiv:2306.09344). arXiv. https://doi.org/10.48550/arXiv.2306.09344

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). *A Neural Algorithm of Artistic Style* (arXiv:1508.06576). arXiv. https://doi.org/10.48550/arXiv.1508.06576

Goetschalckx, L., Andonian, A., & Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences, 25*(9), 788–801. https://doi.org/10.1016/j.tics.2021.06.006

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology. General, 145*(1), 82–94. https://doi.org/10.1037/xge0000129

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience, 27*(Volume 27, 2004), 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society b: Biological Sciences, 372*(1714), Article 20160102. https://doi.org/10.1098/rstb.2016.0102

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2020). *A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications* (arXiv:2001.06937). arXiv. https://doi.org/10.48550/arXiv.2001.06937

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences, 14*(1), 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science, 32*(8), 1262–1270. https://doi.org/10.1177/0956797621994768

Hayes, T. R., & Henderson, J. M. (2025). DeepMeaning: Estimating and interpreting scene meaning for attention using a vision-language transformer. *Open Mind, 9*, 1020–1036. https://doi.org/10.1162/opmi.a.6

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huebner, G. M., & Gegenfurtner, K. R. (2012). Conceptual and visual features contribute to visual memory for natural images. *PLoS ONE, 7*(6), Article e37575. https://doi.org/10.1371/journal.pone.0037575

Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2021). Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nature Human Behaviour, 5*(6), 743–755. https://doi.org/10.1038/s41562-021-01124-6

Kammoun, A., Slama, R., Tabia, H., Ouni, T., & Abid, M. (2022). Generative adversarial networks for face generation: A survey. *ACM Computing Surveys, 55*(5), 94:1-94:37. https://doi.org/10.1145/3527850

Karras, T., Laine, S., & Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks* (arXiv:1812.04948). arXiv. https://doi.org/10.48550/arXiv.1812.04948

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352–355. https://doi.org/10.1038/nature06713

Leopold, D. A., & Afraz, A. (2021). Neurophysiology: The three-dimensional building blocks of object vision. *Current Biology, 31*(1), R9–R11. https://doi.org/10.1016/j.cub.2020.10.064

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology. General, 145*(1), 95–109. https://doi.org/10.1037/xge0000130

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America, 115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Lüdecke, D. (2018). sjPlot: Data visualization for statistics in social science. *R Package Version*, *2*(1).

Mante, V., Bonin, V., & Carandini, M. (2008). Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron, 58*(4), 625–638. https://doi.org/10.1016/j.neuron.2008.03.011

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife, 7*, Article e31873. https://doi.org/10.7554/eLife.31873

Morales-Torres, R., Wing, E. A., Deng, L., Davis, S. W., & Cabeza, R. (2024). Visual recognition memory of scenes is driven by categorical, not sensory, visual representations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 44*(21), Article e1479232024. https://doi.org/10.1523/JNEUROSCI.1479-23.2024

Movshon, J. A., & Simoncelli, E. P. (2014). Representation of naturalistic image structure in the primate visual cortex. *Cold Spring Harbor Symposia on Quantitative Biology, 79*, 115–122. https://doi.org/10.1101/sqb.2014.79.024844

Nacholmo. (2024). *Controlnet-qr-pattern-sdxl* [Computer software]. https://huggingface.co/Nacholmo/controlnet-qr-pattern-sdxl

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage, 105*, 215–228. https://doi.org/10.1016/j.neuroimage.2014.10.018

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage, 222*, Article 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding v1? *Neural Computation, 17*(8), 1665–1699. https://doi.org/10.1162/0899766054026639

Ostwald, D., Lam, J. M., Li, S., & Kourtzi, Z. (2008). Neural coding of global form in the human visual cortex. *Journal of Neurophysiology, 99*(5), 2456–2469. https://doi.org/10.1152/jn.01307.2007

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience, 31*(4), 1333–1340. https://doi.org/10.1523/JNEUROSCI.3885-10.2011

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition, 206*, Article 104465. https://doi.org/10.1016/j.cognition.2020.104465

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pettini, L., Bogler, C., Doeller, C., & Haynes, J.-D. (2024). *Synthesis and Perceptual Scaling of High Resolution Natural Images Using Stable Diffusion* (arXiv:2410.13034). arXiv. https://doi.org/10.48550/arXiv.2410.13034

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... Rombach, R. (2023). *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis* (arXiv:2307.01952). arXiv. https://doi.org/10.48550/arXiv.2307.01952

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70. https://doi.org/10.1023/A:1026553619983

R Core Team. (2022). R: A language and environment for statistical computing□. □*R Foundation for Statistical Computing, Vienna, Austria.* □

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Reddy, L., Cichy, R. M., & VanRullen, R. (2021). Representational content of oscillatory brain activity during object recognition: Contrasting cortical and deep neural network hierarchies. *eNeuro*. https://doi.org/10.1523/ENEURO.0362-20.2021

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience, 34*(26), 8837–8844. https://doi.org/10.1523/JNEUROSCI.5265-13.2014

Richter, D., Kietzmann, T. C., & de Lange, F. P. (2024). High-level visual prediction errors in early visual cortex. *PLoS Biology, 22*(11), Article e3002829. https://doi.org/10.1371/journal.pbio.3002829

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. https://doi.org/10.48550/arXiv.2112.10752

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Son, G., Walther, D. B., & Mack, M. L. (2022). Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *Behavior Research Methods, 54*(1), 444–456. https://doi.org/10.3758/s13428-021-01630-5

Stecher, R., Cichy, R. M., & Kaiser, D. (2025). Decoding the rhythmic representation and communication of visual contents. *Trends in Neurosciences, 48*(3), 178–188. https://doi.org/10.1016/j.tins.2024.12.005

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision, 14*(12), 6. https://doi.org/10.1167/14.12.6

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PLoS ONE, 8*(3), Article e58594. https://doi.org/10.1371/journal.pone.0058594

Walter, K., Manley, C. E., Bex, P. J., & Merabet, L. B. (2024). Visual search patterns during exploration of naturalistic scenes are driven by saliency cues in individuals with cerebral visual impairment. *Scientific Reports, 14*(1), Article 3074. https://doi.org/10.1038/s41598-024-53642-8

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*(34), 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences of the United States of America, 108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., & Chau, D. H. (2023). *DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models* (arXiv:2210.14896). arXiv. https://doi.org/10.48550/arXiv.2210.14896

Wickham, H. (2018). *Ggplot2: Elegant graphics for data analysis*. Springer.

Winawer, J., & Witthoft, N. (2015). Human V4 and ventral occipital retinotopic maps. *Visual Neuroscience, 32*, Article E020. https://doi.org/10.1017/S0952523815000176

Wutz, A., Loonis, R., Roy, J. E., Donoghue, J. A., & Miller, E. K. (2018). Different levels of category abstraction by different dynamics in different prefrontal areas. *Neuron, 97*(3), 716-726. e8. https://doi.org/10.1016/j.neuron.2018.01.009

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., ... Yuan, L. (2023). *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*. arXiv. https://doi.org/10.48550/arXiv.2311.06242

Yang, C., Shen, Y., & Zhou, B. (2021). Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision, 129*(5), 1451–1466. https://doi.org/10.1007/s11263-020-01429-5

Zeki, S., & Marini, L. (1998). Three cortical stages of colour processing in the human brain. *Brain: A Journal of Neurology, 121*(Pt 9), 1669–1685. https://doi.org/10.1093/brain/121.9.1669

Zhang, L. [@lllyasviel]. (2023). *SDXL control collection* (Version main) [Hugging Face]. https://huggingface.co/lllyasviel/sd_control_collection/blob/main/diffusers_xl_canny_mid.safetensors

Zhang, L., Rao, A., & Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models* (arXiv:2302.05543). arXiv. https://doi.org/10.48550/arXiv.2302.05543

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.