GENERATING METAMERS OF HUMAN SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human vision combines low-resolution "gist" information from the visual periphery with sparse but high-resolution information from fixated locations to construct a coherent understanding of a visual scene. In this paper, we introduce ***MetamerGen***, a tool for generating scenes that are aligned with latent human scene representations. ***MetamerGen*** is a latent diffusion model that combines peripherally obtained scene gist information with information obtained from scene-viewing fixations to generate image metamers for what humans understand after viewing a scene. Generating images from both high and low resolution (i.e. "foveated") inputs constitutes a novel image-to-image synthesis problem, which we tackle by introducing a dual-stream representation of the foveated scenes consisting of DINOv2 tokens that fuse detailed features from fixated areas with peripherally degraded features capturing scene context. To evaluate the perceptual alignment of ***MetamerGen*** generated images to latent human scene representations, we conducted a same-different behavioral experiment where participants were asked for a "same" or "different" response between the generated and the original image. With that, we identify scene generations that are indeed *metamers* for the latent scene representations formed by the viewers. ***MetamerGen*** is a powerful tool for understanding scene understanding. Our proof-of-concept analyses uncovered specific features at multiple levels of visual processing that contributed to human judgments. While it can generate metamers even conditioned on random fixations, we find that high-level semantic alignment most strongly predicts metamerism when the generated scenes are conditioned on viewers' own fixated regions.

1 INTRODUCTION

Understanding the latent representation of a scene formed by humans after viewing remains a fundamental unanswered challenge in cognitive science (Epstein & Baker, 2019; Bonner & Epstein, 2021; Malcolm et al., 2016; Võ, 2021). What is clear is that humans represent coherent scenes by a mixture of "gist" information encoded from peripheral vision (Potter, 1975; Greene & Oliva, 2009) with high-resolution but sparse information that humans extract during their scene viewing fixations (Larson & Loschky, 2009; Larson et al., 2014; Eberhardt et al., 2016). Related recent work on scene perception has focused on the concept of object and scene metamers—generated stimuli that, although physically different from originals, cannot be discriminated as different by humans when viewed under constrained experimental conditions. (Freeman & Simoncelli, 2011; Balas et al., 2009; Rosenholtz et al., 2012). Understanding scene metamerism is important because metamers tell us the level of misalignment between an actual and generated image that is tolerated by humans and judged to be the same. Generated scenes that fail to become metamers also reveal the details that are important to a scene's representation and that, if changed, result in the generation being detected. However, although several paradigms have been used to identify metamers (e.g., same-different tasks, A/B/X tasks, oddity judgment tasks) (Rosenholtz, 2020), this work on scene perception used simple generative models to synthesize textures and shapes that were shown in behavioral experiments to be metameric with what humans perceive in their visual periphery when the eye position is fixed. These paradigms, however, were not designed to study how post-gist changes in fixation affect scene metamerism or what objects a person believes to exist in their blurred peripheral view of a scene, which are the problems we engage.

Inspired by these previous studies showing that generated textures and shapes can become metamers for human scene *perception*, we introduce ***MetamerGen***, a state-of-the-art generative model that extends the metamer generation approach to human scene *understanding*. Rather than seeking to generate simple patterns that share low-level statistics with peripheral vision, ***MetamerGen*** better captures a post-gist level of representation reflecting multiple free-viewing fixations. We see this topic as closer to scene understanding because we are seeking to generate a hypothesis for what a person believes to be in their peripheral vision, and henceforth we will use the term *scene metamer* to refer to two scenes that have an equivalent understanding. Our approach combines a gist-level scene representation extracted from peripherally blurred pixels with higher-resolution and fixation-specific "foveal" representations corresponding to scene-viewing fixations. Scene gist and the objects fixated during viewing are therefore used to generate in the non-fixated blurred pixels a scene context that is aligned with what a human understands to be in their peripheral vision.

We not only show that many of the scenes generated by ***MetamerGen*** are metamers for human scene understanding, we also model the dynamic evolution of this understanding by leveraging the capability of a latent diffusion model (Rombach et al., 2022) to generate photorealistic images from diverse conditioning signals (Sohl-Dickstein et al., 2015; Zhang et al., 2023; Ramesh et al., 2022). Because ***MetamerGen*** is a latent diffusion model (Stable Diffusion; Rombach et al., 2022), we can use each viewing fixation as a conditioning signal to obtain an incremental fixation-by-fixation understanding of a scene (Figure 1).

To adapt the Stable Diffusion model to our task of generating a scene in blurred peripheral pixels, we introduce a dual-stream representation of foveated scenes (i.e., ones with a high-resolution center and blurred periphery) using a self-supervised image encoder (DINOv2) (Caron et al., 2021; Oquab et al., 2024; Darcet et al., 2024). We utilize an adapter-based framework (Mou et al., 2023), where we condition a pre-trained text-to-image diffusion model on fixation-grounded features extracted by DINOv2 feature representations obtained at each of the fixation locations. We complement the fixation representations with peripheral information, adding a second source of conditioning that uses DINOv2 tokens extracted from a blurred-out version of the same image, capturing the context.

Our conditioning mechanism allows us to generate plausible scene hypotheses from a variable information input, where more foveal glimpses of a scene will lead to a richer DINOv2 representation that enables ***MetamerGen*** to generate increasingly plausible and contextually appropriate content at the non-fixated scene locations, analogous to how human scene understanding becomes more elaborate with more viewing fixations. We see ***MetamerGen*** as a tool for generating fixation-specific scene understanding hypotheses that cognitive scientists can test in behavioral studies.

We integrated ***MetamerGen*** into a same-different behavioral paradigm and conducted experiments to identify the generated scenes that are metamers for human scene understanding. In our paradigm, participants viewed a scene for a variable number of fixations (i.e., gaze contingent), followed by a 5-second delay (during which ***MetamerGen*** generated a scene from the viewing behavior) and then briefly viewed a second scene (200 msec). Their task was to judge whether this second scene was the same or different from the first. We define a scene metamer as a generation that a participant judges to be the same as the real scene that was first viewed. Our post-hoc analysis showed that while all features throughout the visual hierarchy contributed to the understanding of a scene, high-level semantic features emerged as the strongest predictors of scene understanding metamers.

2 Preliminaries

2.1 Image generation using latent generative models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) comprise two opposing processes—a diffusion process that gradually corrupts data and a denoising process that restores information. The diffusion process relies on Gaussian noise of increasing intensity at every step, while the denoising process uses a learned denoiser model to reverse the degradation. By iterating this process, starting from random Gaussian noise, diffusion models generate new samples.

Latent diffusion models (LDMs) (Rombach et al., 2022) reduce the overall cost by applying the diffusion processes in the latent space of a variational autoencoder (VAE) (Kingma & Welling, 2013). Stable Diffusion (Rombach et al., 2022) uses a pre-trained VAE that spatially compresses images

$8\times$ with its encoder and decompresses latent diffusion samples with the corresponding decoder. The denoiser $\epsilon_\theta(\cdot)$ is a UNet (Ronneberger et al., 2015) consisting of pairs of down and up-sampling blocks at four resolution levels, as well as a middle bottleneck block. Each network block consists of ResNet (He et al., 2015), spatial self-attention, and cross-attention layers, with the latter introducing the conditioning information.

The cross-attention layers condition the denoising process by computing relationships between intermediate image features during denoising and a set of given conditioning embeddings, usually text. When $F \in \mathbb{R}^{h\times w\times c}$ represents the intermediate image features during denoising (reshaped to $hw \times c$ for attention computation) and $e \in \mathbb{R}^{n\times d}$ are the n conditioning embeddings, the cross-attention mechanism first projects features into queries and embeddings into keys and values as

$$Q = FW_Q,\ K = eW_K,\ V = eW_V$$
$$Q \in \mathbb{R}^{hw\times d_k},\ K \in \mathbb{R}^{n\times d_k},\ V \in \mathbb{R}^{n\times d_v} \tag{1}$$

where $W_Q \in \mathbb{R}^{c\times d_k}$, $W_K \in \mathbb{R}^{d\times d_k}$, and $W_V \in \mathbb{R}^{d\times d_v}$ are learned projection matrices. The cross-attention output is then computed as:

$$\text{CrossAttention}(F, e) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

This mechanism allows each spatial location in the image (rows in Q) to attend to relevant parts of the conditioning (rows in K), with the attention weights determining how much the information in each conditioning embedding contributes to the denoising process at each spatial location.

2.2 Self-supervised image encoders

DINOv2 (Caron et al., 2021; Oquab et al., 2024) is a self-supervised vision transformer trained for hierarchical visual representation learning without manual annotations. Using multiple self-supervised objectives, including a contrastive loss that causes image features that appear together to have similar embeddings and a reconstruction loss that induces patches to redundantly encode information about their surrounding context, DINOv2 represents both local visual details and higher-level semantics. These properties make it an excellent tool to study fixation-by-fixation human scene understanding. Adeli et al. (2023; 2025) have shown how self-supervised encoders were capable of capturing object-centric representations without labels as well as providing a backbone capable of predicting high-level neural activity in the brain.

2.3 Adapting latent diffusion models to new conditions

In text-to-image LDMs (e.g., Stable Diffusion), cross-attention layers condition image features on text embeddings. An efficient approach for incorporating *additional* conditioning types, without retraining the model from scratch, can be achieved through adapter-based frameworks (Mou et al., 2023). These adapters re-use the learned text conditioning pathways in the LDM to introduce other modalities of conditioning. This is done by introducing trainable components that transform and project new condition signals into a format compatible with the UNet's existing cross-attention mechanisms. This approach has proven particularly effective for incorporating visual conditioning into text-to-image models (Ye et al., 2023; Wang & Shi, 2023; Ye et al., 2025).

3 Perceptually-informed conditioning

3.1 Representing foveal & peripheral visual features

Given an image and a set of fixation locations, potentially made by a human during free-viewing, we first aim to extract the foveal information from the fixation locations and the peripheral information regarding the overall image context. We employ a DINOv2-Base model (with registers) as the feature extractor to obtain these two sources of information. In Appendix Section A.10 we validate the choice of DINOv2 as the feature extractor for ***MetamerGen*** by showing its superiority to CLIP.

DINOv2 processes 448×448 images with a patch size of 14×14, yielding 1024 tokens (32×32 grid), each embedded in 768 dimensions (along with a CLS token representing the entire image, and

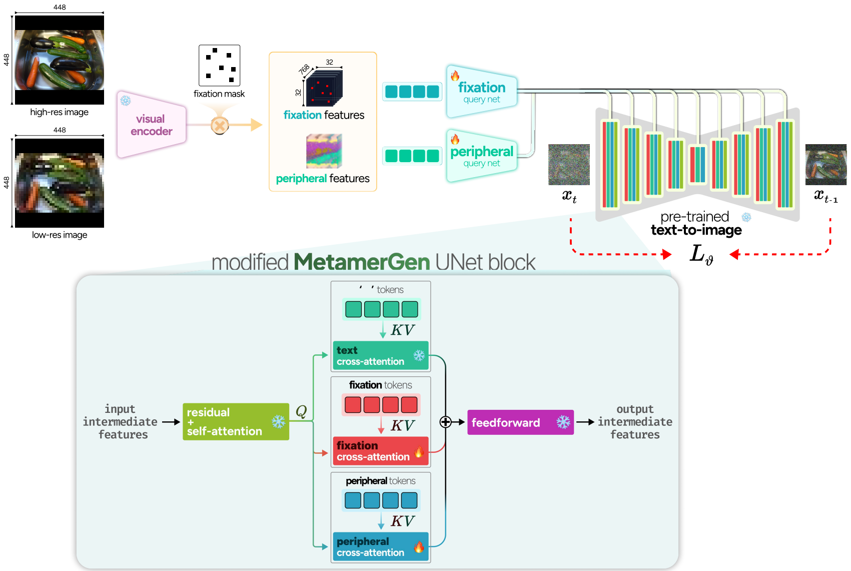

Figure 1: ***MetamerGen* model architecture.** High-resolution and blurred low-resolution images are processed through DINOv2-Base to extract patch tokens each. Foveal features are obtained by applying binary masks to high-resolution patch tokens, retaining only fixated regions. Both foveal and peripheral patch tokens are processed through separate Perceiver-based query networks that compress features into conditioning tokens compatible with Stable Diffusion's cross-attention mechanism. The resulting dual conditioning streams are integrated into the pretrained UNet for guided image denoising and generation.

four register tokens encoding general information about the image). The patch token at a specific location encodes detailed visual and semantic information about that location, analogous to the high-resolution information sampled by the fovea during a fixation. It also encodes limited information about the location's visual context, analogous to low-resolution parafoveal information (Adeli et al., 2023; 2025). To model the information gathered during a series of fixations, we apply a binary mask M_{fixation} to the patch tokens extracted from a scene image I, corresponding to the image locations fixated by humans, zeroing out all non-fixated image patches.

For peripheral visual features, we simulate the inherent uncertainty in peripheral vision by downsampling the input image, and then upsampling it back to 448×448. The blurry image, $I_{\text{peripheral}}$, is also processed with DINOv2, but now retaining all output patch tokens without masking. These peripheral tokens encode uncertain visual representations across the entire scene, capturing the noisy information available in peripheral vision that requires validation through targeted foveal fixations (Srikantharajah & Ellard, 2022; Michel & Geisler, 2011).

3.2 Foveal & peripheral conditioning adapters

We develop foveal and peripheral conditioning adapters to integrate visual information as additional conditioning signals in Stable Diffusion. Similar to IP-adapters (Ye et al., 2023), which integrate CLIP image embeddings into Stable Diffusion, we learn how to incorporate DINOv2 patch embeddings into the cross-attention mechanism of the text-to-image Stable Diffusion model.

Both foveal and peripheral DINOv2 embeddings are first processed through separate Perceiver-based resampler networks $R(\cdot)$ (Alayrac et al., 2022; Jaegle et al., 2021) that compress the 1024 DINOv2 embeddings into 32 conditioning tokens compatible with the pre-trained UNet's cross-attention. (For more information, please refer to Appendix A.3).

$$e_{\text{foveal}} = R_{\text{foveal}}(\text{DINOv2}(I_{\text{original}}) \odot M_{\text{fixation}}), \; e_{\text{peripheral}} = R_{\text{peripheral}}(\text{DINOv2}(I_{\text{downsample}})) \quad (3)$$

The conditions are then integrated through separate cross-attention mechanisms. For each conditioning source (text, foveal, peripheral) we project separately into keys and values

$$K_c = e_c W_K^c, \; V_c = e_c W_V^c,$$
$$K_c \in \mathbb{R}^{n_c \times d_k}, \; V_c \in \mathbb{R}^{n_c \times d_k}, \; c = \{\text{text}, \; \text{foveal}, \; \text{peripheral}\} \quad (4)$$

which we then combine additively into the denoising through cross-attention.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK_{\text{text}}^T}{\sqrt{d_k}}\right)V_{\text{text}} + \lambda_{\text{foveal}} \cdot \text{softmax}\left(\frac{QK_{\text{foveal}}^T}{\sqrt{d_k}}\right)V_{\text{foveal}}$$
$$+ \lambda_{\text{peripheral}} \cdot \text{softmax}\left(\frac{QK_{\text{peripheral}}^T}{\sqrt{d_k}}\right)V_{\text{peripheral}} \qquad (5)$$

λ_{foveal} and $\lambda_{\text{peripheral}}$ are scaling factors that control the contribution of of either foveal or peripheral visual features to the generation process. In practice we “freeze” the text conditioning, by setting the text caption for all images to an empty string “ ”.

3.3 Training and Inference

We start from a pre-trained Stable Diffusion 1.5 network (Rombach et al., 2022). The trainable components of ***MetamerGen*** are the foveal and peripheral resampler networks and their associated key-value projection matrices. Training is conducted on the complete MS-COCO training set (Lin et al., 2015) of approximately $118,000$ images. For foveal conditioning, we apply binary masks that randomly retain $\{1, 2, 3, 5, 10\}$ DINOv2 patch tokens while zeroing all others. This sampling strategy ensures compatibility with our free-viewing behavioral experiments, which constrain scene viewing to a maximum of 10 fixations. For peripheral conditioning, we blur the images by downsampling to $\{0.0625\times, 0.125\times, 0.25\times, 0.5\times, 1\times\}$ of the original resolution.

To enable robust conditioning during inference, we randomly drop conditions with probabilities $p_{\text{foveal}} = 0.05$ and $p_{\text{peripheral}} = 0.10$. The higher peripheral dropout rate prevents over-reliance on peripheral features, which despite blurred image features retain substantial visual information compared to the sparse foveal features. We employ the DDIM sampler (Song et al., 2022) for 50 timesteps, with CFG++ (Chung et al., 2025). We set $\lambda_{\text{foveal}} = 1.2$ and $\lambda_{\text{peripheral}} = 0.7$ to balance detail generation with scene plausibility.

We point out that although ***MetamerGen*** is conditioned on dense DINOv2 representations of an image (periphery and fixation DINOv2 patch embeddings), the model does not simply reconstruct input images verbatim. We attribute this to the lossiness introduced by the DINOv2 embeddings, as well as stochasticity in the sampling process. We demonstrate this further in Appendix A.7 and A.8.

4 *MetamerGen* for Image Generation

We first evaluate the image quality of samples from our model using Fréchet Inception Distance (FID; Heusel et al., 2017) between images generated from ***MetamerGen*** and COCO-10k-test. Figure 2 shows the results using a single central fixation. Green: we fix the blur level to $0.25\times$, matching our behavioral paradigm, and evaluate how peripheral context affects generation quality by varying the peripheral scale. As peripheral scale increases, FID scores improve showing that the model is able to better integrate the context coming from the peripheral DINOv2 representations. Red: we evaluated the effect of the blur level, showing that our model can consistently generate plausible scenes for all levels of blur. We include a text-to-image baseline (Blue) using SD-1.5 with 10k random captions from the COCO training set. ***MetamerGen***, fine-tuned on the COCO images, consistently outperforms the text-to-image model, proving that we have successfully integrated images of variable resolution into the conditioning mechanism of Stable Diffusion.

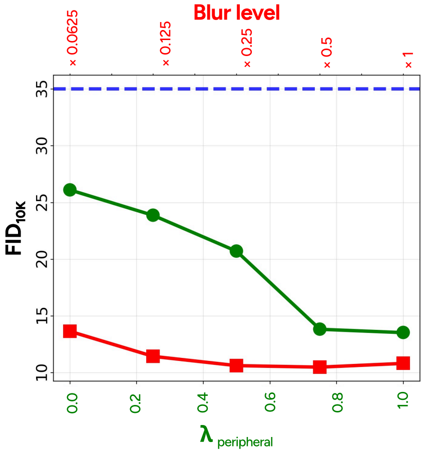

Figure 2: FID values for different input parameters of ***MetamerGen***. Lower FID values indicate closer alignment with real images and better quality.

5 Behaviorally-conditioned scene metamers

5.1 Probing latent scene representations through metamer judgments

Initially established in color science to reveal trichromatic vision, metamers have since been applied to texture perception and visual crowding to infer the underlying structure of human perceptual systems. In the context of scene understanding, metamers offer a unique opportunity to probe what the visual system extracts and retains from complex natural scenes.

Scene perception requires the extraction of meaningful structures from complex inputs. This includes identifying spatial layout, object relations, and global context (Oliva & Torralba, 2006), and is shaped by what the visual system extracts rather than the stimulus itself. When a person views scene A and forms internal representation$_{\mathrm{A}}$, then later sees a different scene B and forms representation$_{\mathrm{B}}$, we can test whether these representations are *perceptually aligned*. If scene B serves as a metamer to scene A, aligning these internal representations reveals what information the brain has perceived and retained from the original scene. By using metamerism as a proxy for the content of scene representations, we can systematically investigate the structure of human scene understanding.

5.2 Real-time behavioral paradigm

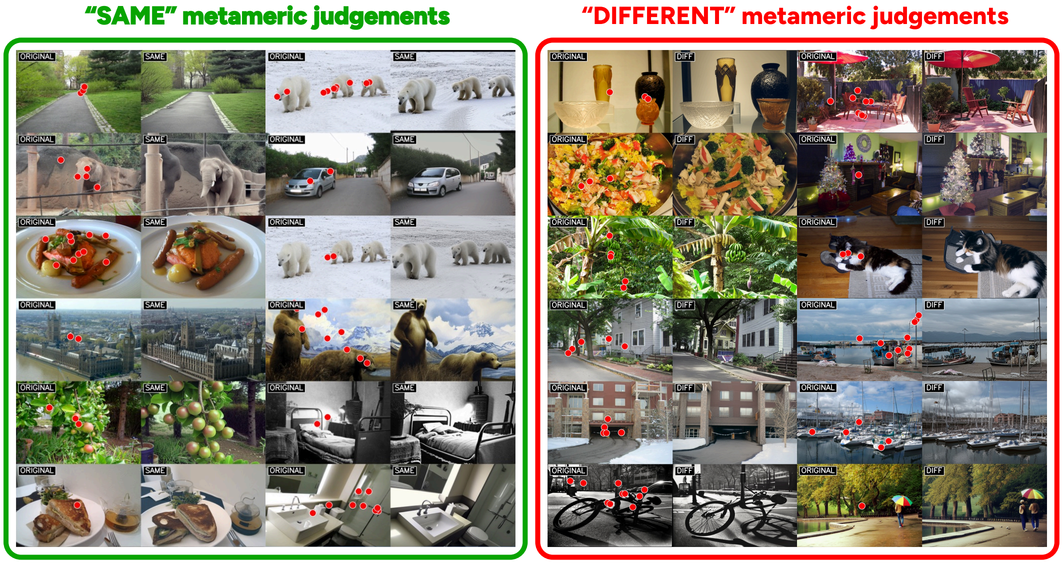

Figure 3: **Metameric vs. non-metameric judgments.** (Left) Original images with human fixations overlaid in red and corresponding generated images judged as ”same” by participants. (Right) Original images with fixations and generated images judged as ”different” by participants. More examples based off of both human-fixation and random-fixation guided generations can be seen in Appendix A.6

We developed a real-time same-different behavioral paradigm to evaluate whether ***MetamerGen*** generates perceptually convincing scene metamers. This paradigm directly tests whether images reconstructed from sparse fixational sampling can achieve perceptual equivalence with the original, thereby revealing the sufficiency of fixated information for scene representation.

Experimental Design We employed 45 participants in a naturalistic free-viewing same-different paradigm. Each trial followed a structured sequence (Appendix A.1.2): participants first completed a drift check, fixated on a central cross, then freely viewed a natural scene image until reaching a predetermined fixation count $\{1, 2, 3, 5, 10\}$, after which the image automatically disappeared. Critically, participants chose their own fixation locations (see Appendix A.1.1). We systematically varied information availability by manipulating fixation count, testing how additional visual information influenced the generation quality.

During a subsequent 5-second interval, participants maintained central fixation while our system processed their actual fixation coordinates and the original image in real-time. ***MetamerGen***'s generated image then appeared briefly for 200 milliseconds—too brief to allow eye movements but sufficient for perceptual comparison (Broderick et al., 2023; Wallis et al., 2019). Participants used a gamepad to indicate whether this second image matched their initial percept.

Participants encountered two primary experimental conditions: metamers generated from their own fixations, and identical original images reshown as controls. As a third comparison condition, 12 participants additionally saw metamers generated from randomly-sampled coordinates instead of their actual fixations. While random metamers often fooled participants, they varied so much that they could not be used to explain human behavior. Generating metamers based on human-fixated locations resulted in better-controlled variability, centered around our best estimate of the participant's scene understanding.

Stimulus Selection Our stimulus set comprised 300 images from the Visual Genome dataset (Krishna et al., 2017), specifically sourced from the YFCC100M subset (Thomee et al., 2016) to avoid overlap with COCO training data used in ***MetamerGen*** training. We employed DreamSim (Fu et al., 2023) to cluster images in semantic representational space and selected one representative image per cluster to maximize visual diversity. Images were filtered to exclude challenging elements for current diffusion models: human hands, faces, and bodies, as well as clocks, text, and numbers.

6 Multiple levels of visual features drive metamer judgments

MetamerGen is conditioned on actual human fixation sequences, providing a richer and more dynamic model of scene understanding. Because it can generate plausible hypotheses for naturalistic images from both peripheral and foveal information (Figure 3), it also enables analysis of which visual features—ranging from low to high levels—shape metameric judgments. In our first analysis, we compared visual similarity from neurally grounded CNN features with human same–different responses. Although images generated from human fixation and random fixation sequences fooled participants about equally often (29.4% and 27.7% of the time, respectively, $p = 0.24$), we found a stark contrast in interpretability between these conditions. For human-fixation-based metamers, higher similarity to the original predicted more "same" judgments. For random-fixation-based metamers, however, high similarity often increased "different" judgments, suggesting that realistic details in non-fixated regions may expose inconsistencies with the viewer's internal scene representation. We confirmed the same pattern across explicitly defined, interpretable feature hierarchies: features at all levels contributed to explaining human metameric judgments, with the fixation-based effect becoming especially pronounced for high-level semantic features (e.g., DreamSim, CLIP).

6.1 Neurally-grounded feature maps

We compared human judgments to a model whose internal representations systematically correspond to human visual processing. We employed a blur-trained AlexNet architecture (Jang & Tong, 2024), which has been specifically trained to be robust to image blur and whose internal representations exhibit strong correlations with human neural responses across visual areas from V1 to inferotemporal cortex (IT). This neurally-grounded model allowed us to isolate contributions from different stages of the visual hierarchy to metameric perception. As illustrated in Figure 4, our analysis pipeline treats early, mid, and late layers as proxies for different stages of visual processing. For each layer, we extracted feature maps from both original and generated images and computed cosine similarity to quantify alignment across the visual hierarchy. We found that as feature similarity increased at any processing level, the proportion of participants judging images as metameric also increased. This relationship held consistently across all layers of the network, from early visual features through high-level representations. The results demonstrate that metamerism spans the entire visual hierarchy rather than being confined to a single processing stage, suggesting that successful scene metamers must maintain representational alignment across multiple levels of visual processing.

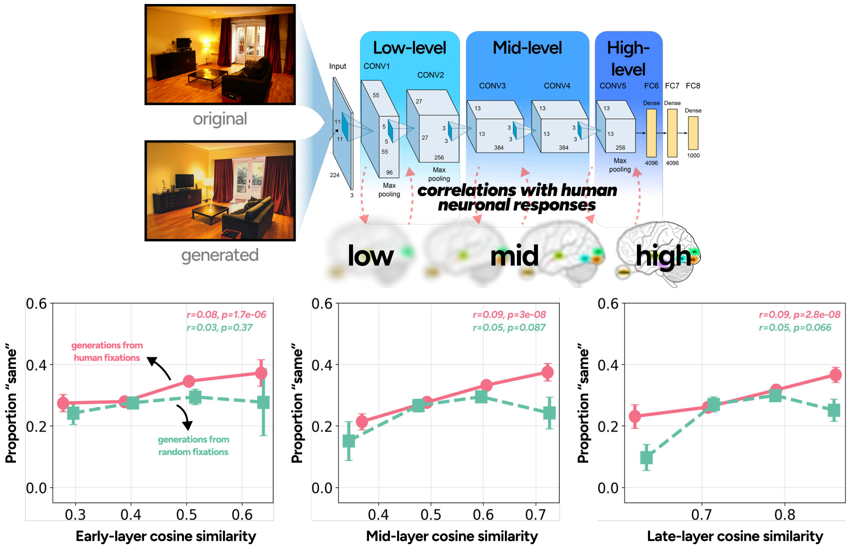

Figure 4: **Multi-level feature analysis pipeline using neurally-grounded model:** (Top) Early, mid, and late network layers serve as proxies for different stages of visual processing from V1 to IT. (Bottom) Results show that as feature similarity increased at different processing stages, the proportion of participants judging generated images as metameric also increased. These effects were clearer when metamers were generated based on fixated locations (salmon) than on randomly-sampled locations (turquoise).

Importantly, we observed distinct patterns when comparing images generated based on the random-sampling human-fixation sampling conditions. While early-layer feature similarities showed little difference between the two conditions, mid- and late-layer similarities revealed divergent trends: that human-fixation sampling maintained a consistent linear relationship between high-level feature alignment and "same" metameric judgments, while random-sampling produced an inverted-U trend for late-layer feature similarities.

6.2 Interpretable visual feature analysis

Having demonstrated that neurally grounded feature similarity aligns with human metameric judgments, we turned next to explicitly defined, interpretable visual features. To capture contributions across different levels of the visual hierarchy, we analyzed a diverse set of features: low-level (e.g., edges, Gabor filters, color), mid-level (e.g., depth cues, proto-object structure), and high-level (e.g., object, semantics, overall perceptual similarity). Because many of these features are correlated, we applied a forward stepwise regression model to identify the most predictive subset ($R^2 = 0.039$), which we focus on in the main text. Detailed contributions of each feature to the regression are provided in Appendix A.5.

6.2.1 Low-level visual features

We compared human "same" judgments as a function of (i) Gabor filter intensities and (ii) Sobel edge density response differences between the generated and original images. By comparing normalized Gabor filter responses, of four orientations (0°, 45°, 90°, 135°), between the original and generated images, we assessed how low-level texture detection affects scene similarity judgments. Surprisingly, we found that positive differences in Gabor filter responses—where generated images showed stronger texture responses than originals—correlated with more "same" judgments. This suggests that enhanced texture definition, which makes boundaries more distinctive, increases the perceived realism of generated images, even when they differ substantially from the originals (Ho et al., 2012). We also found that greater Sobel edge density responses (Kanopoulos et al., 1988) led to greater "same" judgments, though this effect was redundant with the Gabor filter effect (see A.4).

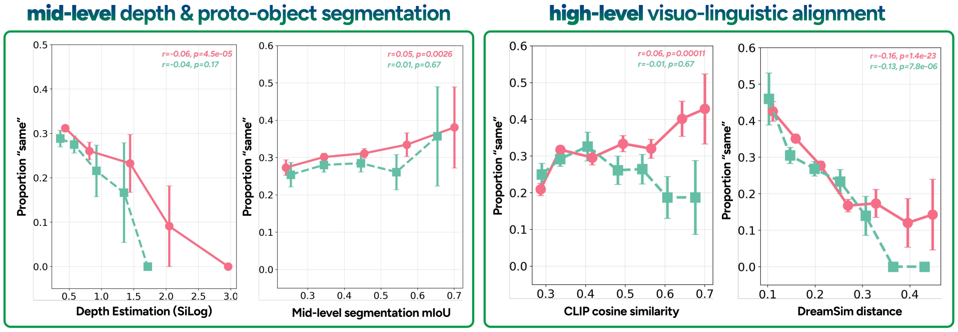

Figure 5: **(Left) Mid-level visual features driving metameric judgments:** For metamers generated based on human-fixated locations (salmon), the preservation of monocular depth estimates in scene structure was an indicator of how more depth discrepancies yielded a decrease in metameric judgments. Additionally, when it came towards the mid-level organizational structure as seen from proto-object candidates, greater mIoU scores correlated with greater proportions of "same" metameric judgments. **(Right) High-level visual features driving metameric judgments:** Semantic similarity strongly predicts metameric perception, with larger DreamSim distances corresponding to reduced perceptual alignment. This result is shared with the CLIP similarity trends as well. However, these trends are less apparent when metamers were generated based on randomly-sampled locations (turquoise).

6.2.2 MID-LEVEL VISUAL FEATURES

We tested two different mid-level visual features, representing local scene layout information available prior to full scene segmentation: (i) relative depth and (ii) proto-object segmentation. **Depth information** proved crucial for metameric perception.

We utilized the Depth Anything model (Yang et al., 2024a) to obtain depth maps from both original and generated images, then compared them using the Scale-Invariant Logarithmic (SiLog) error metric (Lee et al., 2021; Eigen et al., 2014). As discrepancies between depth maps increased, the proportion of "same" metameric judgments systematically decreased (Figure 5). This finding highlights how fundamental depth perception is to mid-level scene understanding and spatial layout representation (Verhoef et al., 2016).

We also extracted **proto-object segmentations** to analyze mid-level grouping structures using the `conv3` layer of the blur-trained AlexNet model (Jang & Tong, 2024). These mid-layer representations are crucial for forming robust "proto-object" identities (Finkel & Sajda, 1992; Yu et al., 2014), which are initial, structured percepts that represent candidate objects by integrating visual parts and features before final semantic recognition. Greater proto-object segmentation similarity (mIoU) predicted more "same" judgments (Figure 5). This demonstrates that proto-object structures—the intermediate groupings that bridge low-level features and high-level object recognition—play a role in scene representation.

6.2.3 HIGH-LEVEL VISUAL FEATURES

For high-level semantic comparison, we employed both (i) CLIP (Radford et al., 2021) and (ii) DreamSim (Fu et al., 2023) as learned semantic similarity models.

For metamers generated using human-guided fixations, DreamSim served as the strongest predictor of metameric perception among all features tested. DreamSim was specifically trained on human judgments using a two-alternative forced-choice paradigm to capture human-like notions of visual similarity, with smaller DreamSim discrepancies predicting more 'same' responses (Figure 5, right). DreamSim's superior explanatory power likely stems from its ability to capture mid-to-high visual features that cannot be described in language, making it a more comprehensive measure of perceptual similarity than models focused solely on semantic content.

Similarly, as semantic alignment increased between generated and original scenes, measured by CLIP representation similarity, participants were more likely to judge them as metameric (Figure 5 Right). This effect was specific to human-generated fixations and absent for randomly generated fixations, where higher CLIP similarity did not translate into more "same" metameric judgments. We suggest that this discrepancy reflects the fact that random fixations often fall on contextually irrelevant regions, exposing semantic details misaligned with participants' internal scene representations. Together, these results indicate that metamers generated from human fixation sequences produced scenes that are better aligned with participants' internal representations, particularly at the level of high-level semantics. Additional object-level visual feature analyses can be seen in Appendix A.4.

6.3 Foveal and peripheral features both contribute to metameric judgments

We ran an ablation experiment to isolate the contributions of foveal and peripheral conditioning in ***MetamerGen***. We recruited 10 additional participants for a same-different task similar to the primary experiment reported above, with four second-image conditions that systematically assessed the impact of conditioning: identical original images (actual "same" images), generated images using both foveal and peripheral conditioning (as in the primary experiment), generated images using peripheral-only conditioning, and generated images using foveal-only conditioning. We found that, whereas both foveal and peripheral conditioning played a role in whether a generation becomes a metamer, the role played by peripheral conditioning was greater. As expected, the full model had the highest fool rate of 54.5%, compared to the second highest fool rate of 45.8% in the peripheral-only generation condition. Because the model learned to rely on peripheral conditioning for generating scene structure and foveal conditioning for generating the fine-grained visual information at fixated locations, images generated using only foveal conditioning tended to be easily distinguishable from original images (8.4% "same" judgments). Nevertheless, this shows that conditioning from foveal inputs contribute visual and semantic information that produces generations that are better aligned with human scene understanding, beyond what peripheral-only conditioning alone can achieve.

We replicated the preceding multi-level visual feature analysis under each condition of this ablation experiment, and found that (1) under equivalent levels of feature similarity, the full model is the most likely to fool participants; and (2) in general, feature similarity predicted participant judgments the best amongst full model generations, more poorly amongst peripheral-only generations, and very little amongst foveal-only generations. For more detailed results and figures, see Appendix A.9.

7 Limitations

While ***MetamerGen*** is effective at reconstructing semantically coherent scenes from sparse visual inputs, it inherits limitations from the pre-trained Stable Diffusion model on which it is built. In our work, we identified two main limitations in the generated images: (1) difficulties producing fine-grained facial details and accurate limb articulations (Narasimhaswamy et al., 2024; Wang et al., 2025), and (2) generations of text were often unreadable (Yang et al., 2024b) even when directly fixated. To mitigate the effects of these model weaknesses on our behavioral experiment, we excluded images containing such problematic elements as inputs to ***MetamerGen***. Including these elements would have caused participants to respond "different" due to Stable Diffusion artifacts rather than differences in their own scene representations.

8 Discussion

In this paper, we introduced ***MetamerGen***, a latent diffusion model that generates image metamers aligned with human scene representations by combining peripheral gist with fixation-based information. While ***MetamerGen*** was trained to predict images from randomly sampled locations, we found that the scientific value of the model is maximized when conditioned on human fixations. For some applications, such as large-scale or crowdsourced experiments, random conditioning offers practical flexibility by removing the need for eye-tracking, and we found that it is indeed capable of fooling viewers. However, fixation-based conditioning better reflects human perceptual processes, reducing noise in behavioral judgments and yielding stronger correlations across all feature hierarchies (Figs. 4, 5). ***MetamerGen*** advances generative modeling by producing semantically coherent and diverse scenes from sparsely sampled inputs. We also believe it offers a powerful tool for cognitive scientists studying scene perception, enabling testing fixation-specific hypotheses on how scene representations unfold dynamically.

Ethics Statement

This behavioral experiment presented in this work was conducted in accordance with ethical guidelines for a human subjects research. The study protocol was reviewed and approved by the Institutional Review Board.

All participants provided informed consent before participating in the behavioral experiment. Participants were fully informed about study procedures during trials. Participation was entirely voluntary, with participants retaining the right to withdraw at any time without penalty. The data we collected involved non-invasive eye-tracking using the EyeLink 1000 eye-tracker. All data were de-identified (codified) and cannot be linked back to individual participants.

References

Hossein Adeli, Seoyoung Ahn, Nikolaus Kriegeskorte, and Gregory Zelinsky. Affinity-based attention in self-supervised transformers predicts dynamics of object grouping in humans. *arXiv preprint arXiv:2306.00294*, 2023.

Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *arXiv preprint arXiv:2505.17329*, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 9(12):13.1–13.18, 2009.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22035, 2025.

Michael F. Bonner and Russell A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12:4081, 2021. doi: 10.1038/s41467-021-24368-2.

William F. Broderick, Gizem Rufo, Jonathan Winawer, and Eero P. Simoncelli. Foveated metamers of the early visual system. *eLife*, 12:RP90554, 2023. doi: 10.7554/eLife.90554.1. URL https://doi.org/10.7554/eLife.90554.1.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

S. Eberhardt, C. Zetzsche, and K. Schill. Peripheral pooling is tuned to the localization task. *Journal of Vision*, 16(2):14, 2016. doi: 10.1167/16.2.14. URL https://doi.org/10.1167/16.2.14.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. URL https://arxiv.org/abs/1406.2283.

Russell A. Epstein and Chris I. Baker. Scene perception in the human brain. *Annual Review of Vision Science*, 5:373–397, 2019. doi: 10.1146/annurev-vision-091718-014809.

L. H. Finkel and P. Sajda. Proto-objects: an intermediate-level visual representation. In *Optical Society of America Annual Meeting, Technical Digest Series*. Optica Publishing Group, 1992. paper FO1.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14 (9):1195–1201, 2011.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

M. R. Greene and A. Oliva. The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472, 2009. doi: 10.1111/j.1467-9280.2009.02316.x.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Tiffany C. Ho, Scott Brown, Newton A. Abuyo, Eun-Hae J. Ku, and John T. Serences. Perceptual consequences of feature-based attentional enhancement and suppression. *Journal of Vision*, 12 (8):15–15, 08 2012. ISSN 1534-7362. doi: 10.1167/12.8.15. URL https://doi.org/10.1167/12.8.15.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021. URL https://arxiv.org/abs/2103.03206.

Hyodong Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1989), 2024. doi: 10.1038/s41467-024-45679-0. URL https://doi.org/10.1038/s41467-024-45679-0.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL https://github.com/ultralytics/ultralytics.

Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

A. M. Larson and L. C. Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):1–16, 2009. doi: 10.1167/9.10.6. URL https://doi.org/10.1167/9.10.6.

A. M. Larson, T. E. Freeman, R. V. Ringer, and L. C. Loschky. The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2):471, 2014. doi: 10.1037/a0034986. URL https://doi.org/10.1037/a0034986.

Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation, 2021. URL https://arxiv.org/abs/1907.10326.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P. Kording. Does object binding naturally emerge in large pretrained vision transformers?, 2025. URL https://arxiv.org/abs/2510.24709.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

George L. Malcolm, IIA Groen, and Chris I. Baker. Making sense of real-world scenes. *Trends in Cognitive Sciences*, 20(11):843–856, 2016. doi: 10.1016/j.tics.2016.09.003.

Melchi Michel and Wilson S. Geisler. Intrinsic position uncertainty explains detection and localization performance in peripheral vision. *Journal of Vision*, 11(1):18, 2011. doi: 10.1167/11.1.18. URL https://doi.org/10.1167/11.1.18.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2468–2479. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.00239. URL http://dx.doi.org/10.1109/CVPR52733.2024.00239.

Aude Oliva and Antonio Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Mary C. Potter. Meaning in visual search. *Science*, 187:965–966, 1975. doi: 10.1126/science.1145183.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Ruth Rosenholtz. Demystifying visual awareness: Peripheral encoding plus limited decision complexity resolve the paradox of rich visual experience and curious perceptual failures. *Attention, Perception, & Psychophysics*, 82(3):901–925, 2020.

Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J Balas, and Livia Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, 12(4):14, 2012.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL `https://arxiv.org/abs/2010.02502`.

SR Research Ltd. Eyelink 1000. [Apparatus and software], 2006. URL `https://web.archive.org/web/20060615055704/http://www.sr-research.com/fixed_main.php`.

Jatheesh Srikantharajah and Colin Ellard. How central and peripheral vision influence focal and ambient processing during scene viewing. *Journal of Vision*, 22(12):4, 2022. doi: 10.1167/jov.22.12.4. URL `https://doi.org/10.1167/jov.22.12.4`.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL `https://doi.org/10.1145/2812802`.

Bram-Ernst Verhoef, Rufin Vogels, and Peter Janssen. Binocular depth processing in the ventral visual pathway. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1697):20150259, 2016. doi: 10.1098/rstb.2015.0259. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2015.0259`.

Melissa Le-Hoa Võ. The meaning and structure of scenes. *Vision Research*, 181:10–20, 2021. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2020.11.003. URL `https://www.sciencedirect.com/science/article/pii/S0042698920301796`.

Thomas SA Wallis, Christina M Funke, Alexander S Ecker, Leon A Gatys, Felix A Wichmann, and Matthias Bethge. Image content is more important than bouma's law for scene metamers. *eLife*, 8:e42512, 2019. doi: 10.7554/eLife.42512. URL `https://doi.org/10.7554/eLife.42512`.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference, 2024. URL `https://arxiv.org/abs/2312.01597`.

Kaihong Wang, Lingzhi Zhang, and Jianming Zhang. Detecting human artifacts from text-to-image models, 2025. URL `https://arxiv.org/abs/2411.13842`.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation, 2023. URL `https://arxiv.org/abs/2312.02201`.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Zhipeng Ye, Feng Jiang, Qiufeng Wang, Kaizhu Huang, and Jiaqi Huang. Idea: Image description enhanced clip-adapter, 2025. URL `https://arxiv.org/abs/2501.08816`.

Chen-Ping Yu, Dimitris Samaras, and Gregory J. Zelinsky. Modeling visual clutter perception using proto-object segmentation. *Journal of Vision*, 14(7):4, 2014. doi: 10.1167/14.7.4. URL `https://doi.org/10.1167/14.7.4`.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL `https://arxiv.org/abs/2302.05543`.

A Appendix

A.1 Eye-tracking behavioral paradigm

A.1.1 Eye-tracking methodology and fixation coordinate extraction

Eye movements were recorded using an EyeLink 1000 eye-tracker (SR Research Ltd., 2006) configured with the Tower Mount setup. This configuration positions the infrared camera above the participant via a mirror, providing an unobstructed view while enabling monocular tracking across $55°$ horizontally and $45°$ vertically. Participants viewed stimuli on a 27 inch 2560×1440 resolution 240Hz OLED monitor positioned 24 inches from their eyes (subtending approximately $55° \times 30°$ visual angle). Prior to each experimental session, a standard 13-point calibration procedure was performed to ensure accurate gaze tracking. During free-viewing trials, fixations were detected online using the EyeLink's built-in saccade detection algorithm.

A.1.2 Experimental design overview

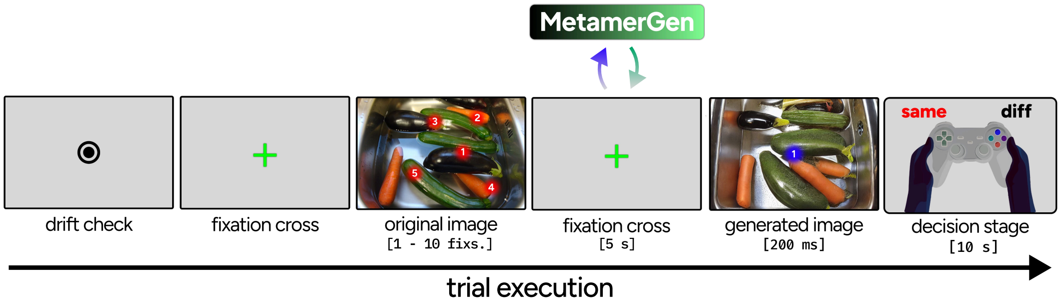

Figure 6: **Real-time metameric judgment paradigm.** Each trial begins with drift correction and central fixation, followed by free viewing of an original scene for a predetermined number of fixations. After image offset, participants maintain central fixation for 5 seconds while fixation coordinates are transmitted via API to ***MetamerGen*** for a real-time image generation. The generated image (or original as control) is then presented for 200ms, followed by a same-different judgment using a gamepad within a 10-second response window.

During a given trial, given fixation coordinates (x, y) from eye-tracking data, we map each fixation to the corresponding patch token in DINOv2's 32×32 grid. For 448×448 input images, each patch token represents a 14×14 pixel region (roughly $1.2° \times 1.2°$ visual angle). Fixation coordinates are normalized to this grid space, with the nearest patch token selected and all others zeroed out, forcing the model to reconstruct the entire scene from sparse fixation inputs.

A.2 ***MetamerGen*** training and inference details

We train following the configuration of Stable Diffusion 1.5 (linear scheduler, fixed variance) for $200K$ steps with a batch size of 32, distributed across 4 NVIDIA H100 GPUs, using the AdamW optimizer with a learning rate of 10^{-4} and weight decay of 0.01. Images from the dataset are padded with 0s to preserve aspect ratios. The model generates output RGB images of size 512×512.

A.3 Perceiver-based Resampler Architecture

The Perceiver-based resampler networks $R(\cdot)$ compress variable-length visual embeddings into a fixed number of conditioning tokens suitable for cross-attention in the pre-trained UNet of Stable Diffusion. This architecture is adapted from Alayrac et al. (2022) and Jaegle et al. (2021) Alternative approaches than resamplers like mean pooling or convolutional downsampling would lose spatial relationships and semantic structure in the conditioning tokens (e.g. in our case DINOv2) that are crucial for high-quality image generation.

Perceiver Attention The core component is a cross-attention mechanism that allows a fixed set of learned latent queries to attend to variable-length input sequences (DINOv2 tokens). Given input features $x \in \mathbb{R}^{n \times d}$ and latent queries $\ell \in \mathbb{R}^{m \times d}$, the Perceiver attention computes:

$$Q = \ell W_Q, \quad K, V = \text{concat}(x, \ell) W_{KV} \tag{6}$$

$$\text{PerceiverAttn}(x, \ell) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{7}$$

The key insight is that queries come solely from the learned latents ℓ, while keys and values are computed from both input features x and latents ℓ concatenated together. This allows the latents to attend to relevant information in the input sequence while maintaining their learned structure.

Resampler Architecture The full resampler consists of:

- **Learned latents**: $m = 32$ learned query vectors initialized from $\mathcal{N}(0, d^{-0.5})$
- **Input projection**: Linear layer mapping from DINOv2 embedding dimension (1024) to internal dimension d
- **Attention layers**: $L = 8$ layers of Perceiver attention followed by feedforward networks with residual connections
- **Output projection**: Final linear projection to match UNet's cross-attention dimension

The resampler processes the 1024 DINOv2 patch embeddings (whether it is via high-resolution fixations or low-resolution peripheral images) and outputs exactly 32 conditioning tokens regardless of input length.

A.4 Additional visual feature drivers of metameric judgment

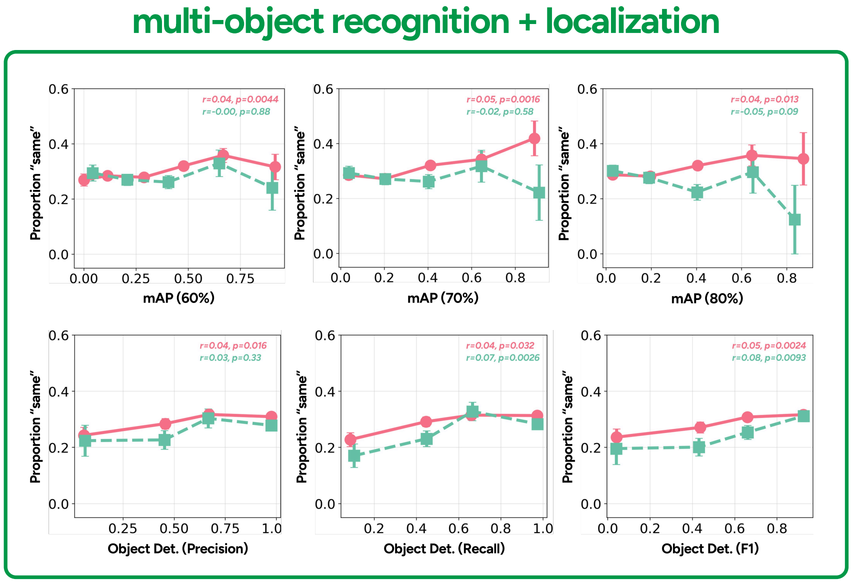

Figure 7: **Object detection errors predict metameric perception:** (Top) mAP scores demonstrate that higher precision accuracies (from mAP 60% to mAP 80%) with better alignment at strict localization boundaries correlate with increased ”same” metameric judgments. (Bottom) Object detection metrics show a positive relationship where improvements in model precision, recall, and F1 scores correspond to increased ”same” metameric judgments.

Multi-object recognition & localization To analyze object-level scene understanding, we employed YOLOv8 (Jocher et al., 2023) to extract object detection bounding boxes and class predictions from both original and generated images. Our pipeline compared object inventories between image pairs, quantifying detection errors across multiple metrics: precision (avoiding extra objects), recall (retaining original objects), and localization accuracy measured by mean Average Precision (mAP) at different IoU thresholds.

Analysis revealed that object-level localization inconsistencies systematically impacted metameric perception (Figure 7 (Left)). Localization accuracy showed consistent relationships with metameric perception. As we required increasingly precise object positioning (shown by increasing mAP thresholds), then the gap between human-guided and random fixation conditions systematically widened. This suggests that extremely precise spatial localization becomes increasingly critical for metameric judgments, and that it can best be exemplified using human fixations.

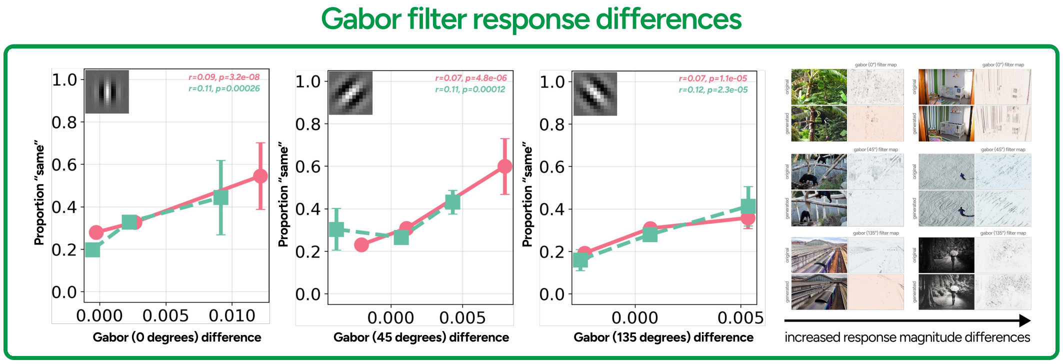

Figure 8: **Stronger Gabor texture responses than originals coincided with greater proportions of metameric judgments.** This suggests that enhanced texture definition, like enhanced edge information, contributes to the perceived realism of generated metamers across multiple spatial frequencies and orientations.

A.5 Stepwise regression model details

We performed a forward stepwise linear regression analysis to measure the extent to which human judgments could be explained by feature differences in our primary behavioral experiment. The resulting linear model had an R^2 value of **0.039**, representing a small but meaningful effect size (in psychological terms). This model incorporated 8 variables, and we evaluated their importance to the model by comparing the full linear model to a model omitting each of them and reporting the change in R^2 for each. In descending importance, these variables were: **DreamSim distance** ($\Delta R^2 = 0.10$), **vertical Gabor intensity** ($\Delta R^2 = 0.006$), **predicted depth map RMSE** ($\Delta R^2 = 0.003$), **D3 (Percentage of pixels with depth error** $< 1.25^3$ **threshold)** ($\Delta R^2 = 0.003$), **mid-level blur-trained CNN feature similarity** ($\Delta R^2 = 0.002$), **CLIP feature similarity** (last hidden layer) ($\Delta R^2 = 0.001$), **CLIP image similarity** (CLS) ($\Delta R^2 = 0.001$), and **D0.25** (Percentage of pixels with depth error $< 1.25^0$.25 threshold) ($\Delta R^2 = 0.001$). These results highlight that human scene similarity judgments depend on independent features distributed across the levels of visual processing, and indeed the three most important features in this regression included low-level, mid-level, and high-level measures.

For comparison, we also ran a stepwise regression on the generations conditioned on random fixations. The resulting linear model had an R^2 value of **0.031**, meaning that in spite of the generations' variability, we were able to begin explaining scene judgments in this case. However, consistent with our earlier findings that these generations differed from the original image in such unpredictable ways that interpretable predictors were no longer significant, this regression only found 2 significant regressors: **DreamSim distance** ($\Delta R^2 = 0.016$) and **135° Gabor intensity** ($\Delta R^2 = 0.014$).

A.6 Additional generation visualizations based on fixated inputs

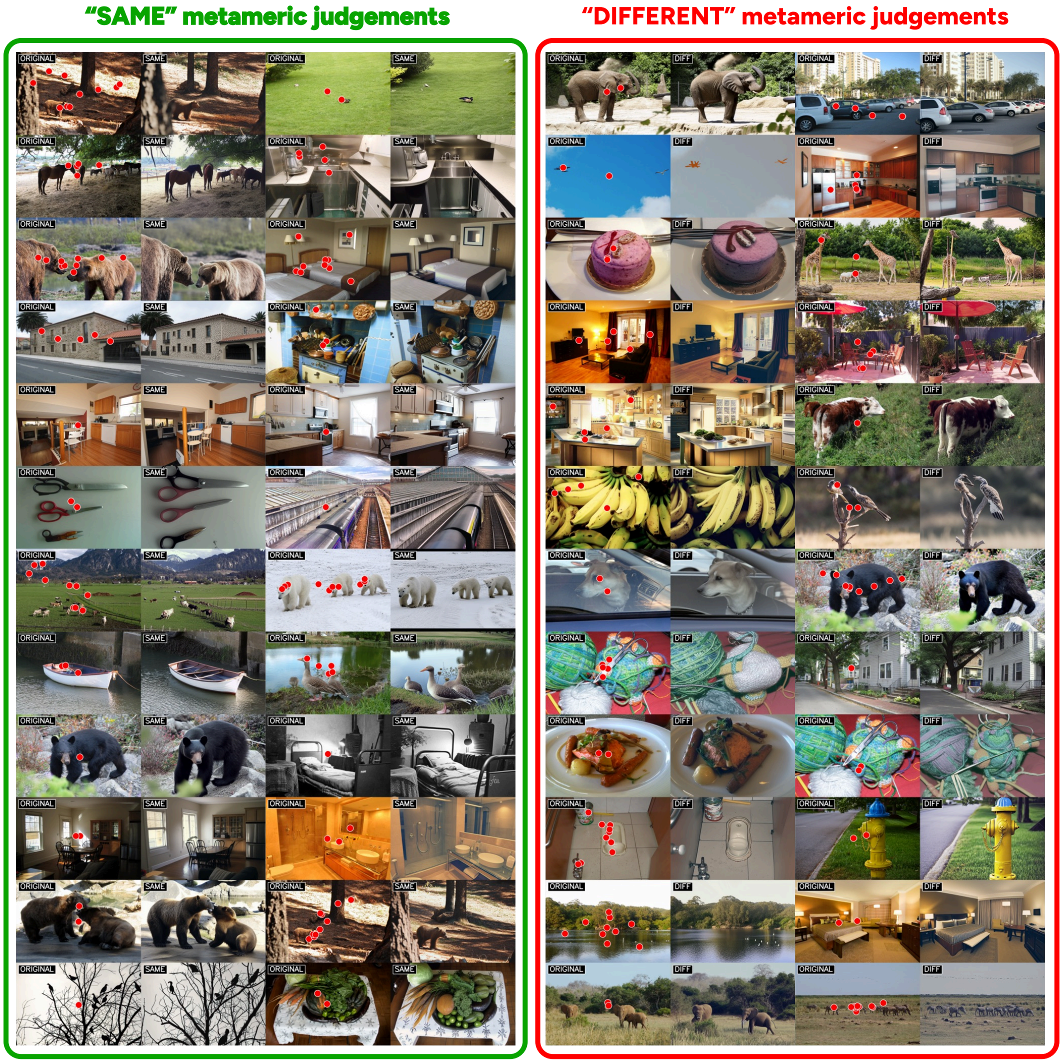

Figure 9: **Additional metameric vs. non-metameric judgment example images based on human fixations.** (Left) Original images with human fixations overlaid in red and corresponding generated images judged as ”same” by participants. (Right) Original images with fixations and generated images judged as ”different” by participants.

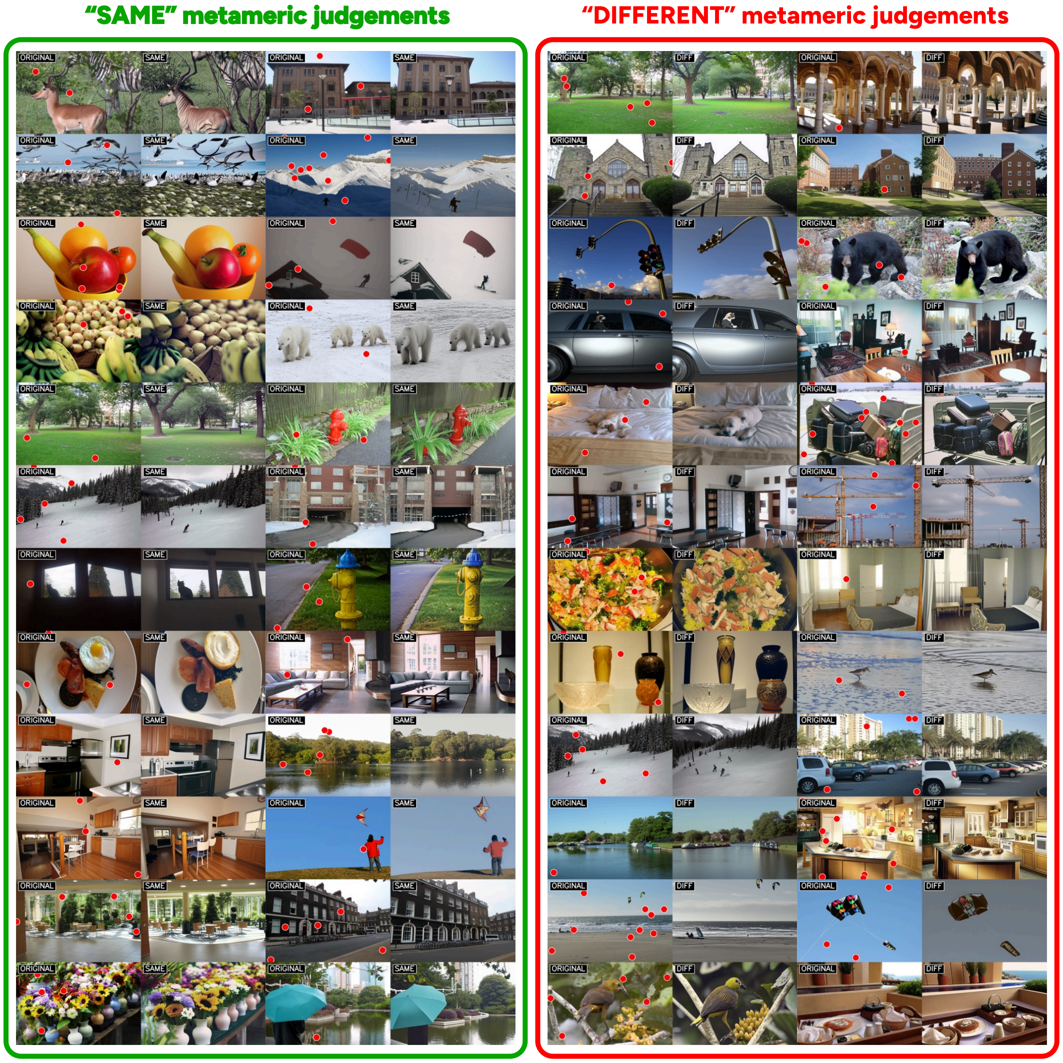

Figure 10: **Additional metameric vs. non-metameric judgment example images based on randomly-sampled fixations.** (Left) Original images with randomly-sampled fixations overlaid in red and corresponding generated images judged as "same" by participants. (Right) Original images with fixations and generated images judged as "different" by participants.

A.7 Effect of peripheral blur and foveal tokens on image generation quality

We ran two computational ablation studies to measure how image generation quality on the COCO-10k-test set is affected by (1) peripheral blur level and (2) foveal token count.

For the study of blur levels (Figure 11, Top), we provided peripheral features as the sole input (at varying levels of downsampling) and masked the foveal tokens. We found that greater downsampling blur yields lower CLIP similarities and higher DreamSim distances, as well as higher (worse) FID, though downsampling 0.25x to 112 × 112 seems to have little effect on quality compared to keeping the original 448 × 448 image, perhaps due to the limited capacity of the DINOv2 embedding and the stochasticity in sampling. Empirically, we find that this introduced just enough uncertainty in the reconstruction that it still resembles the original (CLIP similarity $\sim$0.88) without maintaining fine details. This is the blur level we choose in our human experiments.

For the study of foveal token count (Figure 11, Bottom), we instead varied the number of randomly-positioned foveal tokens, and provided them as the sole input to ***MetamerGen***, masking the peripheral tokens entirely. We found that increasing number of foveal tokens improves CLIP, DreamSim, and FID scores, which nevertheless remain worse than the scores of reconstructions attained by peripheral features only. This highlights the importance of the peripheral conditioning, which we behaviorally confirmed in Section 6.2.3 and Appendix A.9.

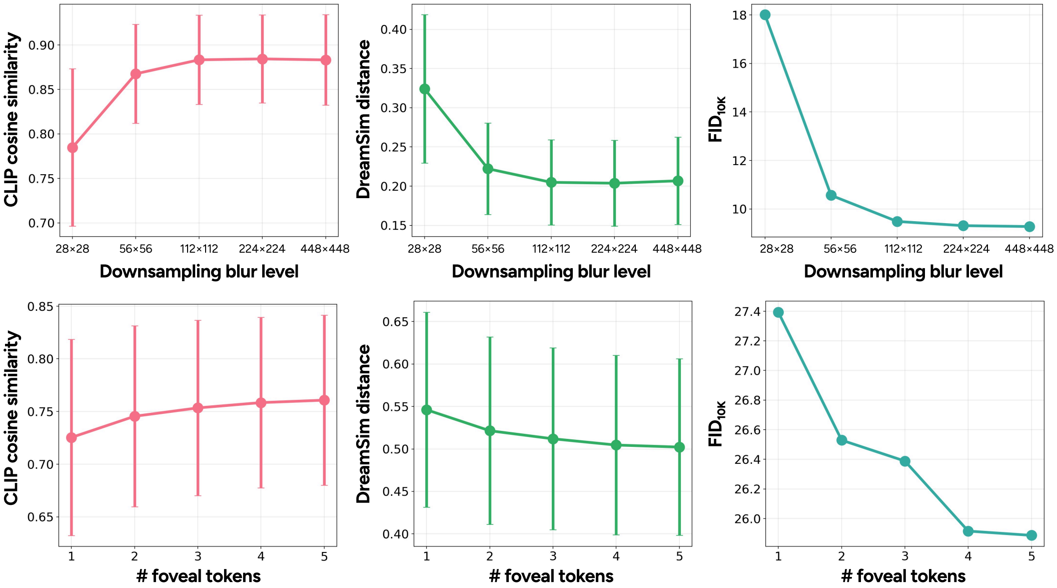

Figure 11: **Influence of blur-level and foveal token count on image generation quality.** (Top Row) Image generation quality decreases as greater blur degrades the base image. (Bottom Row) Image generation quality increases as a function of increasing foveal tokens.

A.8 Semantic similarity is more important than physical distance

While ***MetamerGen*** was trained to denoise the original image from sparse visual inputs, we found that it never precisely recreated the original images, even when presented with all patch tokens. More importantly, we found that pixel-level similarities between the generated and original images had no effect on whether an image was judged as a metamer. Instead, this judgment was predominantly driven by high-level semantic similarities between the generated and original images.

This result indicates that observers rely primarily on a conceptual and semantic understanding of the scene rather than on low-level pixel features when making metameric judgments. In Figure 12 A, we found that the pixel-level similarities measured by PSNR did not predict whether an image would be a metamer; in Figure 12B, we establish that in a PSNR–DreamSim plot, low DreamSim distances predict “same” judgments, but high PSNR values do not, with examples of each included. We conclude that if a researcher wishes to titrate the rate of similarity judgments, they should do so by selecting images based on DreamSim scores, not physical stimulus distance.

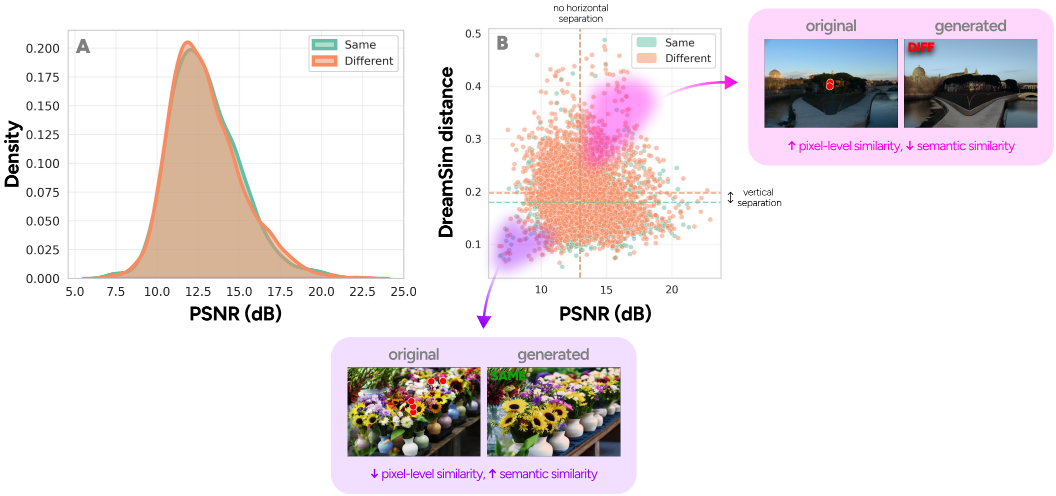

Figure 12: **Comparison of pixel-level (PSNR) and semantic (DreamSim) similarities in metameric judgments.** (A, Left) portrays the histograms of PSNR values for generated images judged as "same" or "different" in the behavioral task, with nearly identical distributions between the two groups. (B, Right) shows the relationship between PSNR and DreamSim distances for all image pairs. There is a clear vertical separation by DreamSim distance that corresponds with metameric judgments, while PSNR values do not discriminate between what is considered metameric.

A.9 Foveal and peripheral token contributions to metameric judgments

We present further statistics and analyses of the behavioral ablation experiment presented in Section 6.2.3.

Figure 13 provides qualitative examples of generations from each treatment that were judged "same" and "different", together with a plot of the fool rate in each condition. The generations in the full-model condition are both qualitatively the highest-quality and quantitatively the most likely to fool participants.

One subtlety that emerged during this experiment was that, despite the same participant population and an identical model, participants were substantially more likely to judge images generated by ***MetamerGen*** as "same" (54.5%) than in the primary experiment (29.4%). Our interpretation of this unexpected pattern of results is that the foveal-only condition, which was generally easy for participants to distinguish from the original image, acted as a low anchor on image similarity, thus lowering participants' threshold for making a same judgment. Because this increases the amount of variability in judgments that can be explained (see below), we see this as a 'feature' rather than a 'bug'.

Figure 14 presents our exhaustive replication of the multi-scale feature correlation analysis from the primary experiment under each condition of the ablation experiment. The main text describes the two conclusions following from this analysis.

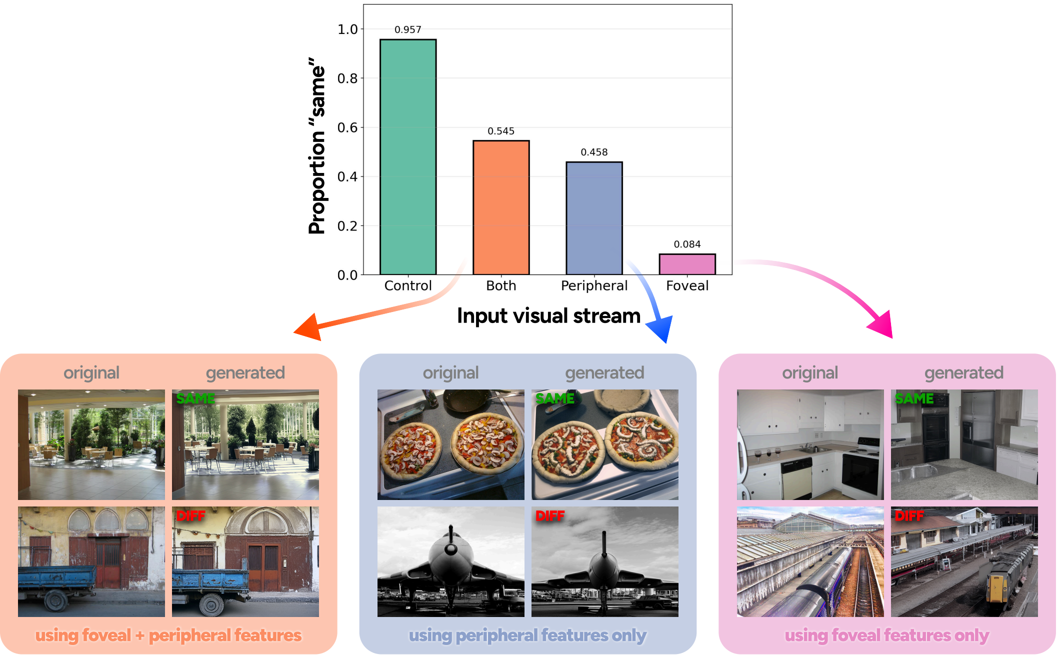

Figure 13: **Higher metameric (same) judgments for images incorporating both peripheral and foveal information.** Using both foveal and peripheral features produced the highest fooling rates. Peripheral-only conditioning yielded the second best results, while foveal-only generations lagged significantly behind. Although the difference between peripheral-only and combined foveal-peripheral conditioning is small, it is meaningful: the additional high-resolution details from fixations lead participants to be more easily fooled.

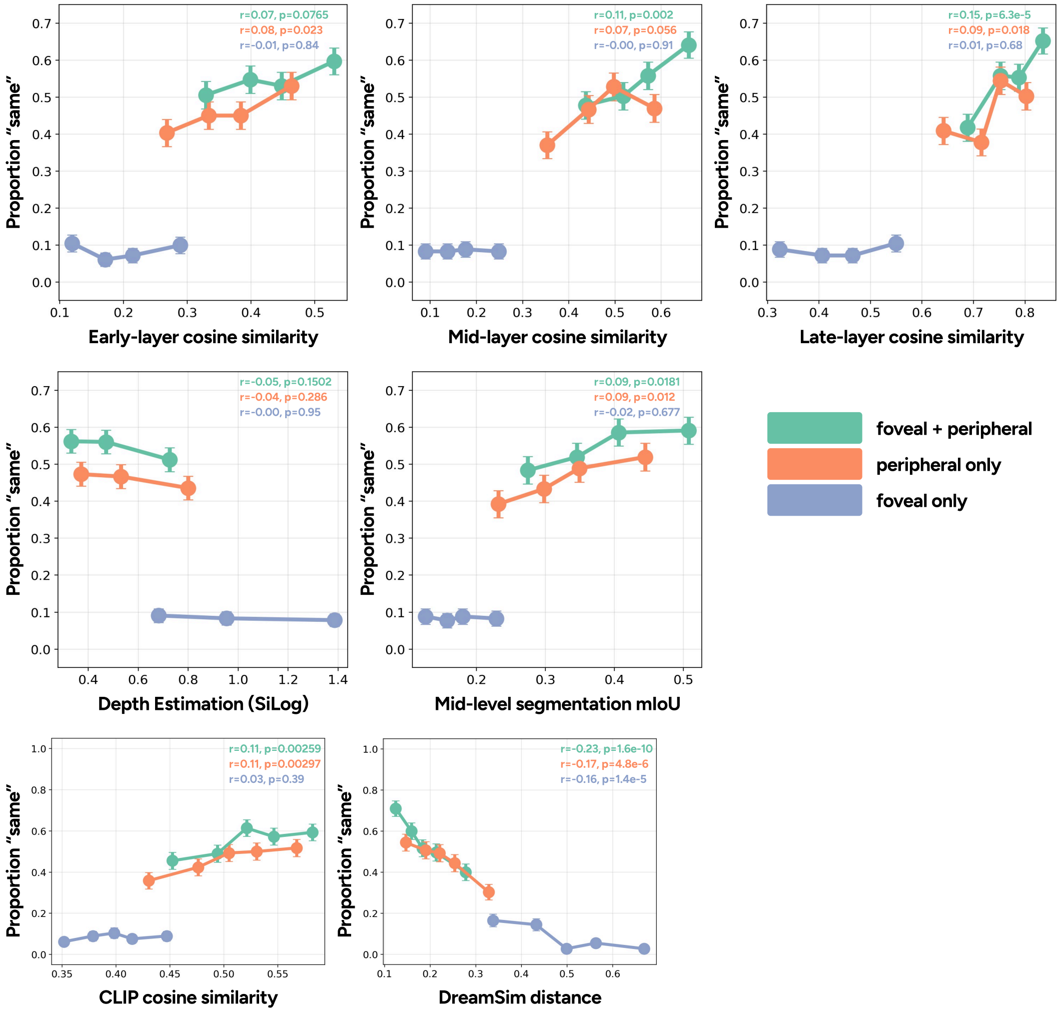

Figure 14: **Impact of input visual streams on hierarchical feature analyses.** (Top Row) Multi-level feature analysis using neurally-grounded model (Jang & Tong, 2024) on driving metameric judgments. (Middle Row) Mid-level visual features driving metameric judgments (mid-level segemention mIoU and SiLog depth estimation). (Bottom Row) High-level visual features driving metameric judgments (CLIP cosine similarity and DreamSim distance).

First, under nearly all metrics, full-model (foveal + peripheral) generations fool participants at a greater rate than peripheral-only generation, even at the same metric scores. This is evidenced by the clear separation between the green and orange lines in most Fig 14 panels. The only features where this gap was not apparent were late-layer cosine similarity and DreamSim distance, suggesting that these metrics may capture a large proportion of the factors that caused humans to judge full-model generations as "same" more than peripheral-only generations.

Second, participant judgments on full-model generations are more explainable than judgments on peripheral-only generations. Along nearly all feature axes (apart from early-layer cosine similarity and mid-level proto-object segmentation mIoU), judgments were more strongly correlated with features under the foveal + peripheral condition. Foveal-only generations were even more poorly explained by these metrics. In fact, the only metric which could explain judgments of foveal-only generations to statistical significance was DreamSim, indicating that these generations, which lacked the gross scene structure and layout provided from the periphery, were so far from the original image that ordinarily important feature axes did not influence judgments.

A.10 DINOv2 versus CLIP as the vision encoder of ***MetamerGen***

Previous adapter-based approaches, like IP-Adapter Ye et al. (2023), have utilized CLIP embeddings as image conditioning inputs for Stable Diffusion. We choose DINOv2 as the visual encoder for foveal and peripheral feature extraction because DINOv2 patch tokens have been shown to better encode both local and contextual information. This contextual encoding emerges from DINOv2's self-supervised training objectives: its reconstruction loss encourages patches to redundantly encode information about their surroundings, while its contrastive loss causes semantically related patches to have similar embeddings (via object and scene structure) (Barsellotti et al., 2025; Adeli et al., 2023). This means a single DINOv2 patch token naturally captures both foveal detail (local information) and parafoveal context (relationships to nearby regions) – precisely the type of representation needed for modeling human fixations.

On the other hand, vision-language models like CLIP optimize for global image-text alignment, which limits their patch-level spatial selectivity and the spatial relationships modeled in their deep layers (Wang et al., 2024; Li et al., 2025). (CNN-based encoders lack emergent representations of patch context – ie, parafoveal information – entirely.)

To empirically validate our choice of DINOv2, we retrained ***MetamerGen***, using a CLIP vision encoder, for 100K steps. Using the CLIP-conditioned model, we conducted two image generation ablations using COCO-10K-test images: (1) varying the blur levels for a peripheral-only image generation, and (2) varying the number of foveal tokens for a foveal-only image generation. This ablation is directly comparable to the DINOv2 ablation presented in Appendix A.7.

We observed that CLIP features perform significantly worse at encoding peripheral information, reflected by the higher FID values obtained in Figure 15 (left), 16. In contrast to DINOv2 tokens, the FID value does not show a significant decrease when we reduce the blur level, meaning that CLIP encodes similar (impoverished) information in its patch tokens for the blurred and non-blurred images, focusing on global scene category, at the expense of more fine-grained scene structure. This is consistent with previous observations made regarding CLIP's relatively low ability to encode contextual information with its patch tokens (Li et al., 2025). For foveal information, increasing the number of tokens has a similar effect for both encoders, though a single DINOv2 token seems to encode more information than a single CLIP token. Overall, our choice of DINOv2 is mainly motivated by its ability to accurately encode both the peripheral and foveal information present in the image.

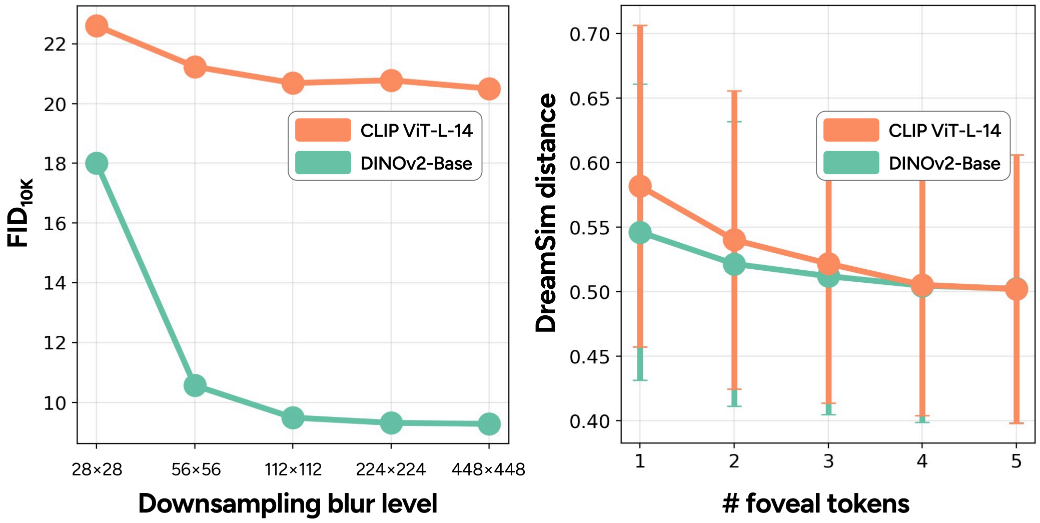

Figure 15: **FID and DreamSim evaluations based on DINOv2 and CLIP as vision encoders for foveal and peripheral feature extraction.** (Left) The image generation quality (FID) for DINOv2-based peripheral generations is consistently better than CLIP patch embeddings. For DINOv2, we observe a sharp drop when decreasing the blur level, showing how decreasing blur results in the model encoding different, more accurate image features. This is not true for CLIP patch tokens, which seem to encode the same limited information across all blur levels. (Right) With increasing numbers of foveal token inputs, the DreamSim distance for both DINOv2 and CLIP-based embeddings decreases. However, DINOv2-based generations yield greater semantic similarities with the original images, especially at low token counts.

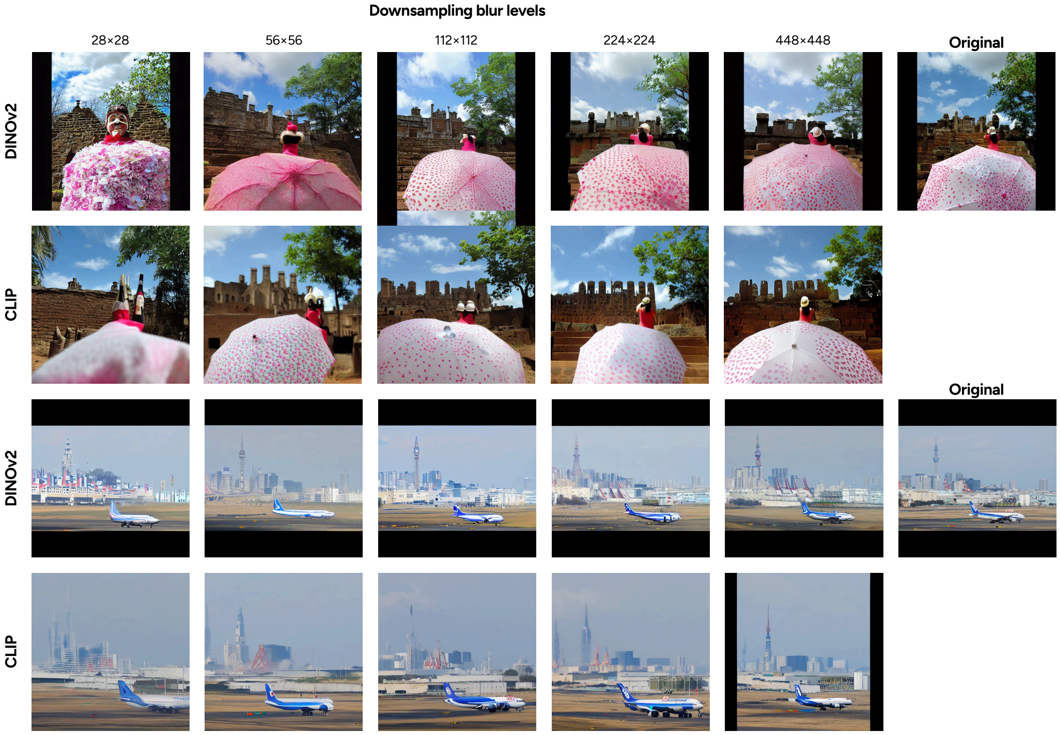

Figure 16: **Image generation examples across blur levels using DINOv2 and CLIP as vision encoders.** DINOv2-based peripheral generations resemble the original images more than CLIP-based generations, even at low blur levels. As the rate of downsampling decreases ($28 \times 28 \rightarrow 448 \times 448$), DINOv2-based generations continue to show substantial improvements while CLIP-based generations exhibit minimal improvements. For the bottom two row pairs, DINOv2-based generations are able to keep the size of the plane (as well as its spatial position) intact irregardless of blur level input. However, that is not the case for the CLIP-based generations.