Seen2Scene: a generative model of fixation-by-fixation scene understanding

Ritik Raina[1] (ritik.raina@stonybrook.edu), **Abe Leite[1,2] (abrahamjleite@gmail.com)**,
Alexandros Graikos[2] (agraikos@cs.stonybrook.edu), **Seoyoung Ahn[3,4] (ahnseoyoung@gmail.com)**,
Gregory J. Zelinsky[1,2] (gregory.zelinsky@stonybrook.edu)
[1]Department of Psychology, [2]Department of Computer Science, Stony Brook University, NY, USA
[3]Department of Molecular and Cell Biology, [4]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA

Abstract

Human scene understanding dynamically evolves over the course of sequential viewing fixations from a gist-level understanding to a more detailed comprehension of the scene. Each fixation provides rich visual information about objects and their spatial relationships. To model this incremental process, we introduce Seen2Scene, a framework for modeling human scene understanding by controlling the inputs used to generate a visual hypothesis of the scene. Seen2Scene uses a self-supervised encoder to extract features from fixated scene regions, which guide a pre-trained text-to-image latent diffusion model through a modular adapter framework. As fixations accumulate, the model iteratively refines its visual hypotheses, filling in unseen areas with contextually plausible content. We evaluated Seen2Scene on COCO-FreeView using two experimental conditions: fixation-only conditioning to isolate the contribution of foveal information, and fixation+gist conditioning to examine how non-fixated scene information integrates with foveal details. Results show that initial fixations drive the greatest gains in semantic and perceptual fidelity and that the fixation+gist condition reached high-fidelity scene understanding with the fewest fixations, thus demonstrating the importance of integrating peripheral gist information with visual details collected foveally.

Keywords: generative modeling, human scene understanding

Introduction

Our understanding of a visual scene evolves as new information is sampled with each fixation during viewing. The information accumulated incrementally refines a visual hypothesis of what exists in non-fixated regions of the scene (Malcolm et al., 2014). Scene understanding begins with the first fixation, which extracts sufficient information to create a *gist* level representation (Potter, 1975) that includes the scene's category and spatial layout. This gist representation is largely obtained from peripheral vision (Stewart et al., 2020) but remains imprecise and can correspond to many more specific scene interpretations.

Our work focuses on the post-gist level of scene understanding that incrementally evolves as people sequentially make fixations while viewing a scene. This aim requires distinguishing between the high-resolution information encoded from each fixation and the gist-enabling information from peripheral vision. To visually model this dynamic evolution of human scene understanding, we leverage recent advances in latent denoising diffusion models (Rombach et al., 2022). We introduce **Seen2Scene**, a novel modular latent diffusion framework for modeling scene understanding fixation-by-fixation (Figure 1A). **Seen2Scene** generates complete scenes from a variable number of fixation features. Its modular architecture allows conditioning on any visual information, including foveal features, peripheral gist, or other visual representations. We treat these generations as visualized hypotheses for the scene understanding gleaned by a human making these fixations.

We evaluated **Seen2Scene** in computational experiments and made a fixation-by-fixation assessment of the fidelity of the model's generated images to the original scenes using a behavioral dataset of free-viewing fixations. In one experiment, we use only foveal information from fixated samples. In a second, we add peripheral gist information to investigate their combined contribution to scene understanding.

The key contributions of this work are as follows:

- We used DINOv2 embeddings to quantify the information that humans encode when viewing an image.
- Leveraging **Seen2Scene**, we generate visual hypotheses of scenes from this information.

Methods

Seen2Scene models incremental scene understanding by leveraging DINOv2's spatially-grounded visual representations within a latent diffusion image generation framework. DINOv2 provides multiple types of embeddings: patch tokens that capture local spatial information in a grid covering the input image and global tokens (CLS and register tokens) that capture broader contextual features (Darcet et al., 2024). This multi-scale structure aligns with human vision; patch tokens correspond to detailed foveal information and global tokens capture gist information from peripheral vision.

Seen2Scene builds on Stable Diffusion by replacing its text conditioning with DINOv2 visual embeddings through the UNet's cross-attention mechanism (H. Ye et al., 2023; Z. Ye et al., 2025). We utilized this architecture to build two variants of our model corresponding to our experiments. The first variant uses a single cross-attention mechanism that conditions on patch tokens from fixated regions (fixation-only) for scene completion. The second variant employs two separate cross-attention mechanisms: one for foveal information (patch tokens) and another for gist information (CLS and register tokens), which we call fixation+gist. This dual-condition design allows us to investigate how foveal and peripheral gist information jointly contribute to scene understanding. During inference, both variants generate complete scenes by denoising random noise conditioned on the visual embeddings from fixated regions, with the number of available fixations incrementally increasing to model progressive scene understanding.

To evaluate **Seen2Scene** we used COCO-FreeView's validation set (Yang et al., 2023), which contains approximately 82K fixations from a 5-second free-viewing task where participants viewed images with memory instructions. Fixations are accumulated sequentially, with each new fixation adding a new token to the set to model progressive scene understanding. We evaluate generation fidelity using CLIP image score (Ge et al., 2024), which assesses text-based semantic alignment, and DreamSim (Fu et al., 2023), which is a powerful method that estimates human similarity judgments gathered on an ABX task using fine-tuned CLIP, OpenCLIP (Ilharco et al., 2021), and DINOv1 (Caron et al., 2021) features.

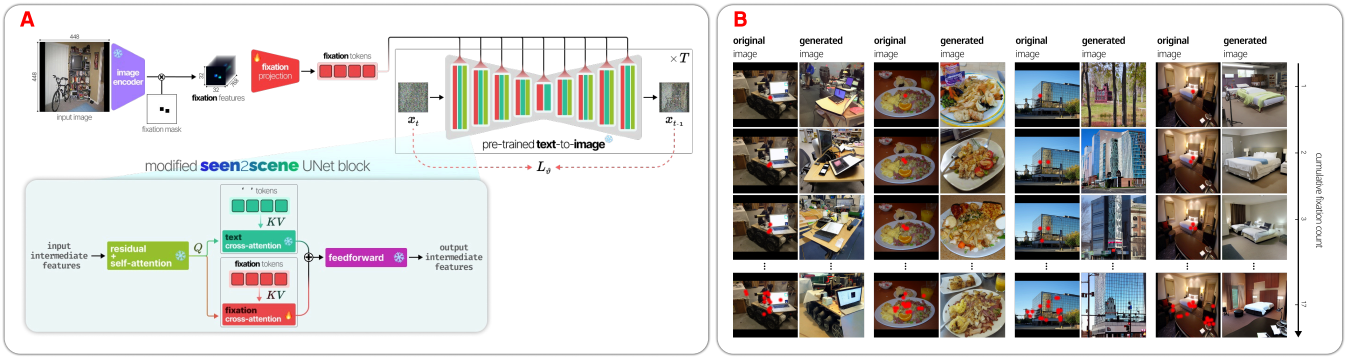

Figure 1: **A: Seen2Scene** architecture. **B:** Fixation-only outputs with fixations as red dots on COCO-FreeView images.

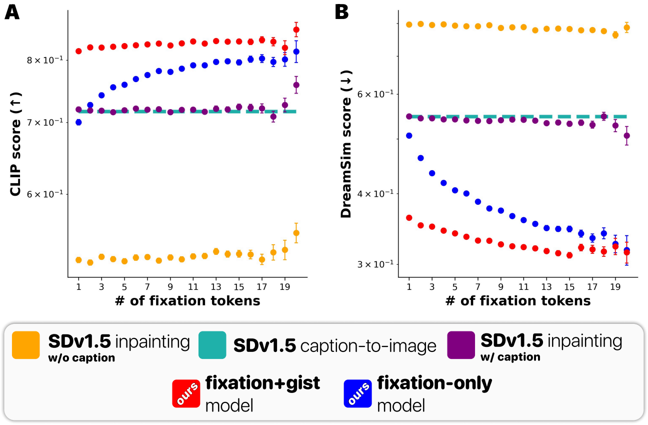

Figure 2: Image generation fidelity on the COCO-FreeView validation set improves with fixation count. The fixation-only model (blue dots) uses only fixation tokens (Exp 1), while the fixation+gist model (red dots) uses both fixation and gist tokens (Exp 2). We report CLIP and DreamSim scores. (**A**) CLIP similarity is positively correlated with fixation count for the fixation-only model ($R = 0.28,\ p < 0.05$). Fixation+gist shows weaker correlation ($R = 0.06,\ p < 0.05$) but achieves near-ceiling performance. (**B**) DreamSim distance shows negative correlations for both fixation-only ($R = -0.43$, $p < 0.05$) and fixation+gist ($R = -0.19$, $p < 0.05$). Both **Seen2Scene** models outperform the SDv1.5 variants, inpainting (with and without COCO captions) and caption-to-image generation.

Results

Experiment 1: Scene generations conditioned on fixations. Figure 1B shows images generated incrementally (top to bottom) as fixations are added to **Seen2Scene**. For each of four scenes, the original is shown on the left with behavioral fixations superimposed (red dots) and the generated scene is on the right. Note that increasing the number of fixations yields generations that are more perceptually and semantically aligned with the original. As shown in Figure 2, a fixation-only version of **Seen2Scene** (blue) shows steepest gains in the first 3–4 fixations, with performance plateauing thereafter. We observed this logistic relationship between fidelity gains and the number of fixation tokens in both CLIP (Fig. 2A) and DreamSim (Fig. 2B) similarity scores.

Experiment 2: Scene generations conditioned on fixation+gist information. The fixation+gist version of **Seen2Scene** (Fig. 2, red), aimed at better isolating the role of peripheral vision in scene understanding, showed weaker correlations with increasing numbers of fixations. However, this was largely due to the model achieving near-ceiling performance with information from only the single central fixation, a pattern aligned with the perspective that gist information provides a strong foundational understanding of scene structure that requires fewer additional fixations to refine.

Comparisons to SDv1.5 baseline model variants. Both **Seen2Scene** variants outperformed all SDv1.5 baselines, with inpainting+captions and caption-to-image being the closest competitors. Inpainting alone performed very poorly. This stratification of performance demonstrates the benefit of combining both visual and language information, which is particularly evidenced by the performance gap between inpainting with and without captions. Notably, **Seen2Scene** achieves its superior semantic understanding using only visual fixation inputs devoid of textual descriptions.

Discussion & Future Work

Seen2Scene demonstrates that visual features alone can drive a complex scene understanding, which we quantify through scene generation. Our fixation-only model successfully completed non-fixated regions with plausible objects over the first few fixations of viewing, and the fixation+gist variant showed that this already-rapid scene understanding can be further accelerated with the addition of global context. The fixation+gist model's near-ceiling performance with minimal fixations reinforces the belief that peripheral gist provides strong foundational understanding, reducing the burden of reasoning about non-fixated content. In contrast, the fixation-only model must incrementally obtain foveal samples until it is possible to infer plausible completions from sparse local information.

In future work, we will conduct behavioral evaluations investigating the confusability between generated scenes with originals using gaze-contingent change detection tasks and same/different memory tests. We will do this to identify scene generations that are perceptual and memory metamers for the latent representations of scenes by humans.

Acknowledgments

We would like to thank the National Science Foundation for supporting this work through awards 2123920 and 2444540 to GJZ, and to the National Institutes of Health through their award R01EY030669, also to GJZ. AL is supported by NSF-GRFP award 2234683 and RR is supported in part by NSF award 2125295. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency.

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9650–9660).

Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2024). *Vision transformers need registers.* Retrieved from https://arxiv.org/abs/2309.16588

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. In *Advances in neural information processing systems* (Vol. 36, pp. 50742–50768).

Ge, Y., Zeng, X., Huffman, J. S., Lin, T.-Y., Liu, M.-Y., & Cui, Y. (2024). Visual fact checker: Enabling high-fidelity detailed caption generation. In *Ieee conference on computer vision and pattern recognition (CVPR).*

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., … Schmidt, L. (2021, July). *OpenCLIP.* doi: 10.5281/zenodo.5143773

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: strategic and incremental information accumulation for scene categorization. *Psychological Science*, *25*(5), 1087–1097. doi: 10.1177/0956797614522816

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *CVPR*.

Stewart, E. E. M., Valsecchi, M., & Schütz, A. C. (2020, 11). A review of interactions between peripheral and foveal vision. *Journal of Vision*, *20*(12), 2-2. Retrieved from https://doi.org/10.1167/jov.20.12.2 doi: 10.1167/jov.20.12.2

Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., & Samaras, D. (2023). Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*.

Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.

Ye, Z., Jiang, F., Wang, Q., Huang, K., & Huang, J. (2025). *Idea: Image description enhanced clip-adapter.* Retrieved from https://arxiv.org/abs/2501.08816