User Study And Evaluation Of Saliency-guided Image Generation Model

Nan Wu
wunan@stanford.edu

Abstract—Diffusion models offer unprecedented image generation capabilities given just a text prompt. While emerging control mechanisms have enabled users to specify the desired spatial arrangements of the generated content [1] [2] [3], they cannot predict or control where viewers will pay more attention due to the complexity of human vision. Recognizing the critical necessity of attention controllable image generation in practical applications, Zhang et al. (2024) proposed a saliency-conditioned generative model designed to direct viewers' attention to specific regions within images. This project employ eye-tracking devices in a user study to evaluate the model's effectiveness. Participants examined images generated by the model alongside baseline models. Their eye gaze patterns were recorded to compute empirical saliency maps. Captured saliency maps were compared with the condition saliency maps. The results demonstrate the model's significant improvement over baselines across various saliency similarity metrics. This suggests the model successfully aligns viewers' focus with the intended saliency regions, validates the model's potential for applications requiring controlled visual attention in generated content.

1 Introduction

The emergence of generative artificial intelligence (AI) marks a paradigm shift for computer graphics. Diffusion models, in particular, enable the generation and editing of photorealistic and stylized images, videos, or 3D objects with little more than a text prompt or high-level user guidance as input [4]. In many applications, including graphic design or advertisement, it is desirable to generate visual content that guides a viewer's attention to the areas of interest. Popular approaches to controlled image [1] generation include lightweight adaptation modules built around a foundation model. The adapter networks are usually conditioned by depth maps, semantic segmentation masks, body poses, or bounding boxes [1], [2], [3], [5] to control the spatial layout of an image or video. Meanwhile, an essential design factor for context creation is to direct viewers' attention to the regions of interest, such as buttons on web pages [6], products being advertised [7], or the storytelling events in a film [8]. Unlike the layout of an image, human attention is selective in nature [9], [10], concurrently influenced by high-level semantics, mid-level layouts, and low-level visual features [11], [12], [13], [14], as well as spatial and temporal characteristics [15], [16]. To capture these complex factors influencing human saliency, and adequately control a viewer's visual spatial attention, Zhang et al. (2024) develop a framework and trained a saliency-conditioned model that guide viewers' visual attention toward specific regions of interest.

To evaluate the model's effectiveness in guiding real users' spatial attention, we conduct a user study with human observers naturally examining the generated images with 3,000 eye-tracked trials. Participants were instructed to browse through the generated images one by one, and their eye gaze patterns were recorded to compute the empirical saliency maps. A series of objective evaluations demonstrate that, compared to the unconditioned and bounding-box-conditioned baseline models, the saliency-guided model achieved significantly better alignment between intended and empirical saliency maps and directed participants' attention to desired image regions at a very high success rate.

2 Related Work

Human Visual Attention Models and Saliency Prediction: Due to the complexity of cognitive visual attention [9], modeling the saliency while perceiving images or videos has been an open challenge. Researchers have attempted to develop saliency models in a bottom-up fashion from image space statistical features [11], [12], [17], [18]. However, these low-level features by themselves are insufficient to account for top-down influences, e.g., our familiarity with different objects [19]. To measure these compounded influences, large-scale eye-tracked studies have been conducted, attempting to establish a paired image–video dataset with human-exhibited gaze fixations. Examples include MIT1003 [17], CAT2000 [20], SALICON [21]. These large-scale datasets catalyzed various deep neural network based saliency metrics for RGB images (e.g., DeepGaze models [13], [22], [23], EMLNet [14], SalGAN [24]). Saliency models are further extended to predict temporal fixation durations [25] and scanpaths [26].

3 Method

The aim of the model is to generate visual content that directs viewer attention in specific ways. This is achieved by incorporating the data priors of human visual attention into the generation process. To systematically analyze the attention-directing properties of the generated images, this project aims at conducting a user study with eye trackers to record participants' eye gaze patterns while they browse through a sequence of generated image samples.

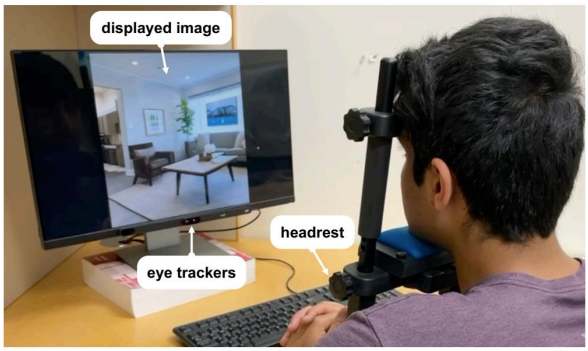

Fig. 1. User study setup. The eye-gaze positions of study participants were recorded while they watched through a sequence of generated images.

3.1 Participants

Twenty adults participated in the study (ages 23–57, 9 female). All of them have normal or corrected-to-normal vision, no history of visual deficiency, and no color blindness. None of them were aware of the hypothesis, the research, or the number of conditions. The research protocol was approved by the Institutional Review Board (IRB) at the host institution, and all subjects gave informed consent prior to the study.

3.2 Setup and procedure

During the study, subjects remained seated in a well-lit room and viewed a 24-inch Dell monitor (Model No. S2415H, resolution 1920×1080, luminance 250 $\mathrm{cd/m^2}$) binocularly from an SR Research headrest positioned 60 cm away. The effective field of view and resolution were $46° \times 26.8°$ and 40 pixels per degree of visual angle. A Tobii Pro Spark eye tracker was mounted to the bottom of the monitor to record their eye gaze at 60 FPS. A 5-point eye-tracking calibration was performed before each session began. Figure **??** shows the experimental setup of our user study.

3.3 Stimuli

We first sampled 50 images from the held-out validation set of MSCOCO 2017, where half of them have humans/animals as the main content and the rest show close-up shots of objects or nature/city scenes. These selected images were annotated using the BLIP-2 Image2Text model [27] for text prompt conditioning. Image saliency maps were extracted using the EML-Net saliency model [14] for visual saliency conditioning. We then fed the obtained paired text prompts and saliency maps to our saliency-guided model to generate 50 images of 512×512 resolution as the visual stimuli for the user study. The hypothesis is that these generated images should direct viewers' attention toward the intended regions as depicted by the saliency maps while observing the text prompts and maintaining non-degraded image quality/diversity.

3.4 Conditions

We also included two baseline conditions, the Stable Diffusion v2.1 (SD2.1) model [28] (UNCOND) and the GLIGEN BBox-guided model [2] (BBOX), to compare with the saliency-condition model (SMAP) in terms of the accuracy and robustness of manipulating human visual attention. All three conditions share the same input text prompts to control what visual content is generated. Additionally, takes in text-annotated bounding boxes predicted by the Grounding DINO model [**?**], and takes in saliency maps predicted by the EML-Net saliency model. Similar to , 50 images were generated for and , respectively.

3.5 Task and duration

The total of 150 images generated for the three conditions was shuffled in random order and sequentially displayed to each subject, with a 5-second duration for each image and a 1-second pause between consecutive images. The complete study, including hardware setup/calibration, pre-study instructions, and breaks, took about 30 minutes per subject. Throughout the study, all subjects were instructed to keep their head stationary on the headrest and freely explore the displayed images by shifting their eye gaze. Their eye gaze patterns on each image were recorded to compute the corresponding empirical saliency map.

4 Results

4.1 Metrics

To quantitatively evaluate each model's performance in directing users' attention to intended image regions, we adopted five saliency similarity metrics from the MIT/Tuebingen Saliency Benchmark [29]: Area Under ROC Curve (AUC) [10], Normalized Scanpath Saliency (NSS) [30], Correlation Coefficient (CC), Kullback–Leibler Divergence (KL), and Histogram Intersection (SIM). Notably, AUC and NSS take a saliency map and a sequence of eye fixations as inputs, while computing CC, KL, and SIM requires two saliency maps.

AUC: This metric assesses the performance of the saliency map by using eye fixations. Thresholds are determined using the combined saliency values of all fixations and nonfixations. It essentially measures how well the saliency map predicts a given sequence of fixations by computing the area under the Receiver Operating Characteristic (ROC) curve.

NSS: This metric computes the mean saliency of fixations after normalizing the saliency map to have zero mean and unit variance. It provides a scale-invariant measure of saliency efficiency by evaluating how salient the actual human fixation points are relative to a given saliency map on a normalized scale.

CC: Fixations are convolved with a Gaussian kernel to compute empirical saliency maps. The correlation coefficient between the empirical saliency maps and the condition saliency maps provides a measure of accuracy, specifically indicating how well the shapes of the two saliency distributions match.

KL: Saliency maps are first converted to probability distribution vectors, then compute the KL divergence between the condition saliency distribution and the empirical distribution. Lower values indicate better performance, as there is less divergence between the two distributions.

TABLE 1
Eye-tracked user study. The saliency-conditioned model largely outperforms the two baselines in directing viewers' attention toward the specified image regions. ↑/↓ indicates that higher/lower score is better.

	AUC ↑	NSS ↑	CC ↑	KL ↓	SIM ↑
UNCOND	0.65	0.71	0.21	4.75	0.34
BBOX	0.78	1.21	0.47	2.67	0.48
SMAP	**0.84**	**1.82**	**0.78**	**0.79**	**0.66**

SIM: Similar to KL, SIM first convert saliency maps to probability vectors, then takes the pixelwise minimum values and sums them. This essentially measures the $L_1 - distance$ between the condition saliency map and the empirical saliency map.

To convert our collected eye-tracking data into empirical saliency maps, we followed the same post-processing procedures described in [31].

4.2 Results

As summarized in table 1, our saliency-guided approach consistently outperforms the two baselines in controlling and directing users' visual attention toward desired image regions across all five metrics. Figure 2 shows seven groups of generated images used in our user study, their corresponding text prompts and empirical saliency maps, as well as the input conditioning saliency maps. As can be observed, the saliency-condition model not only produces the content as depicted by the text prompts but also achieves viewer attention distributions that align well with the intended ones. These results strongly validate the attention-directing capabilities of the saliency-condition model.

5 Discussions and Limitation

Our user study employed a gender-balanced participant pool; however, the limited age range (most participants are 25-30 years old) could introduce bias. A more diverse age distribution in future studies would enhance the generalizability of the findings to a broader population. Additionally, the sample size of 20 participants might limit the statistical power of the results. Increasing the number of participants in future studies would strengthen the reliability of the conclusions drawn.

The StableDiffusion 2.1 model used for image generation exhibited artifacts, particularly in images containing human hands and faces. These artifacts have the potential to influence the empirical saliency of the images, as viewers reported a tendency to fixate on these unusual elements. Exploring alternative diffusion model variants that produce images with fewer artifacts presents a valuable avenue for future research.

Zhang et al. (2024) also proposed a saliency-guided video generation process within the same paper. Unfortunately, incorporating human observers in our study was not feasible due to the substantial sampling requirements needed to acquire eye-tracking data revealing spatiotemporal saliency. To address this challenge, investigating eye-tracker-free attention assessment approaches through crowdsourcing platforms, e.g., [32], offers a promising direction for future research.

6 Conclusion

A user study assessed a new model guiding viewers' attention within images. Participants examined text-prompted images generated by the model and compared them to baselines. Eye-tracking captured their focus, revealing the model's superiority. It significantly outperformed baseline models across several saliency similarity metrics in aligning viewers' gaze with intended regions as described the condition saliency maps. Though limitations exist, the user study validates the model's effectiveness, opening doors for future applications.

References

[1] A. R. Lvmin Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[2] Q. W. F. M. J. Y. J. G. C. L. Yuheng Li, Haotian Liu and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] S. L. X. H. Hu Ye, Jun Zhang and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[4] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. H. Bermano, E. R. Chan, T. Dekel, A. Holynski, A. Kanazawa *et al.*, "State of the art on diffusion models for visual computing," *arXiv preprint arXiv:2310.07204*, 2023.

[5] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *arXiv preprint arXiv:2305.16322*, 2023.

[6] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan, "Directing user attention via visual flow on web designs," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016.

[7] M. H. A. Bakar, M. A. M. Desa, and M. Mustafa, "Attributes for image content that attract consumers' attention to advertisements," *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 309–314, 2015.

[8] A. P. Shimamura, B. I. Cohn-Sheehy, B. L. Pogue, and T. A. Shimamura, "How attention is driven by film edits: A multimodal experience." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 9, no. 4, p. 417, 2015.

[9] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annual review of neuroscience*, vol. 23, no. 1, pp. 315–341, 2000.

[10] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, 2006.

[13] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," in *International Conference on Learning Representations (ICLR 2015)*, 2014, pp. 1–12.

[14] S. Jia and N. D. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," *Image and vision computing 95 (2020)*, 2020.

[15] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 419–435.

[16] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 4894–4903.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2106–2113.

[18] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

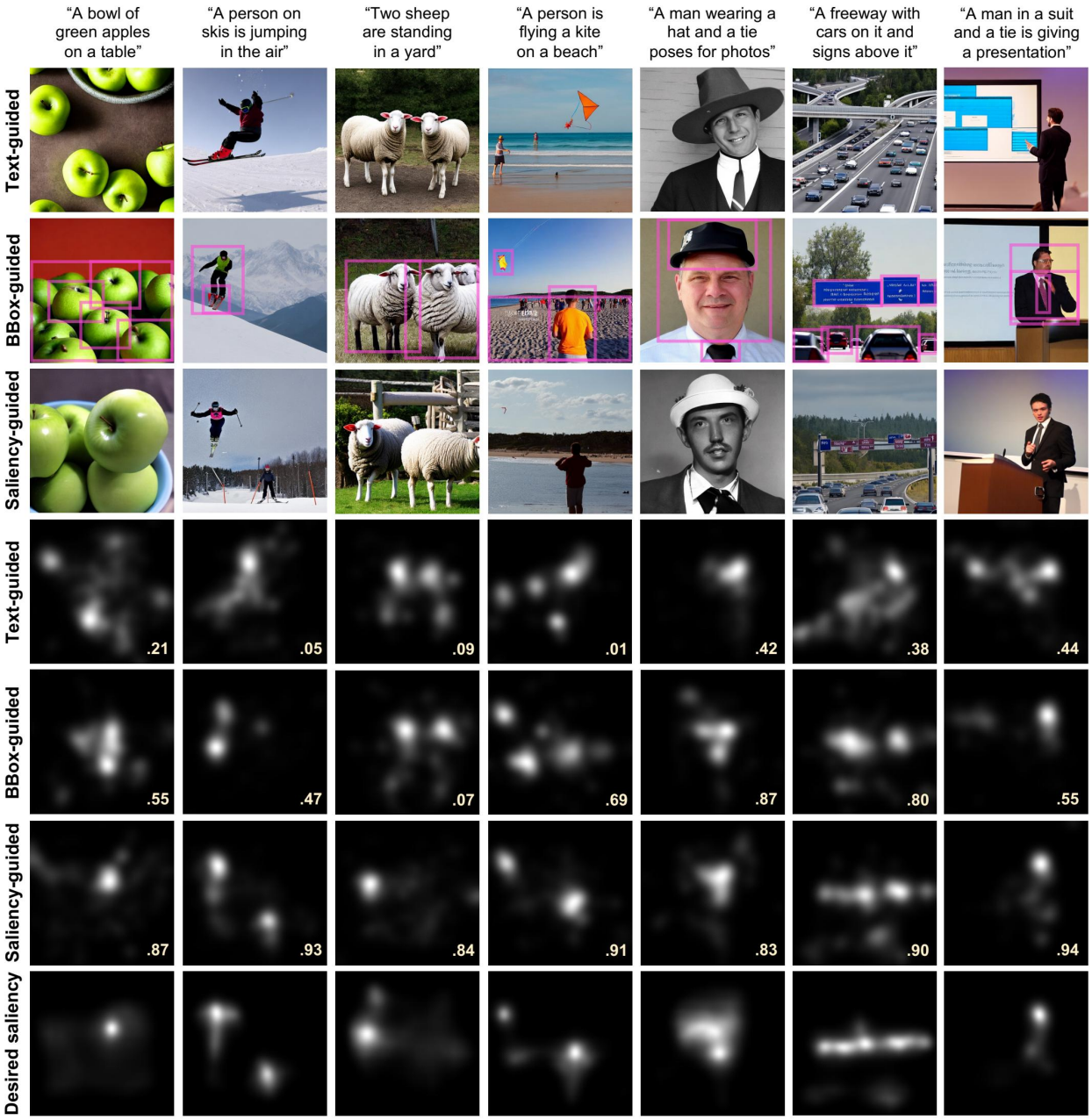

Fig. 2. Eye-tracked user study. Rows 1–3 show the generated images, rows 4–6 show the empirical saliency maps obtained by aggregating 20 users' eye gaze data, and row 7 shows the input conditioning saliency maps (i.e., the desired saliency distribution). The conditioning bounding boxes for BBOX are visualized as overlays. The number associated with each empirical saliency map shows its correlation with the desired saliency distribution. Compared to the two baseline methods, the images generated by the saliency-conditioned model not only contain the exact content as described by the text prompts but also trigger viewer attention that aligns with the saliency conditioning.

[19] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of vision*, vol. 8, no. 3, pp. 3–3, 2008.

[20] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[21] X. B. Xun Huang, Chengyao Shen and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," *In Proceedings of the IEEE international conference on computer vision*, 2015.

[22] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4789–4798.

[23] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 919–12 928.

[24] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[25] C. Fosco, A. Newman, P. Sukhum, Y. B. Zhang, N. Zhao, A. Oliva, and Z. Bylinskii, "How much time do you have? modeling multi-

duration saliency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4473–4482.

[26] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360 images," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2003–2013, 2022.

[27] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[29] M. Kummerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 770–787.

[30] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[31] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[32] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister, "Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 24, no. 5, pp. 1–40, 2017.