GazeFusion: Saliency-guided Image Generation

YUNXIANG ZHANG, New York University, USA
NAN WU, Stanford University, USA
CONNOR Z. LIN, Stanford University, USA
GORDON WETZSTEIN, Stanford University, USA
QI SUN, New York University, USA

Diffusion models offer unprecedented image generation power given just a text prompt. While emerging approaches for controlling diffusion models have enabled users to specify the desired spatial layouts of the generated content, they cannot predict or control where viewers will pay more attention due to the complexity of human vision. Recognizing the significance of attention-controllable image generation in practical applications, we present a saliency-guided framework to incorporate the data priors of human visual attention mechanisms into the generation process. Given a user-specified viewer attention distribution, our control module conditions a diffusion model to generate images that attract viewers' attention toward the desired regions. To assess the efficacy of our approach, we performed an eye-tracked user study and a large-scale model-based saliency analysis. The results evidence that both the cross-user eye gaze distributions and the saliency models' predictions align with the desired attention distributions. Lastly, we outline several applications, including interactive design of saliency guidance, attention suppression in unwanted regions, and adaptive generation for varied display/viewing conditions.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Perception**.

Additional Key Words and Phrases: Human Visual Attention, Perceptual Computer Graphics, Controllable Image Generation

ACM Reference Format:
Yunxiang Zhang, Nan Wu, Connor Z. Lin, Gordon Wetzstein, and Qi Sun. 2024. GazeFusion: Saliency-guided Image Generation. *ACM Trans. Appl. Percept.* 1, 1, Article 1 (January 2024), 20 pages.

1 INTRODUCTION

The recent emergence of generative artificial intelligence marks a paradigm shift for computer graphics. Diffusion models, in particular, enable the generation and editing of both photorealistic and stylized images, videos, and 3D objects with little more than a text prompt or high-level user guidance as inputs [Po et al. 2023]. In many applications, including graphic design and advertisement, it is desirable to generate visual content that attracts a viewer's attention toward specific areas of interest. Such a human-centric control mechanism for the generation process, however, is not supported by current image diffusion models.

Popular approaches to controlled image [Zhang et al. 2023b] and video [Guo et al. 2023] 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 [Li et al. 2023b; Ye et al. 2023; Zhang et al. 2023b; Zhao et al. 2023] to control the spatial layout of an image or video. Meanwhile, an essential

Authors' addresses: Yunxiang Zhang, yunxiang.zhang@nyu.edu, New York University, USA; Nan Wu, wunan@stanford.edu, Stanford University, USA; Connor Z. Lin, connorzl@stanford.edu, Stanford University, USA; Gordon Wetzstein, gordon.wetzstein@stanford.edu, Stanford University, USA; Qi Sun, qisun@nyu.edu, New York University, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Association for Computing Machinery.
1544-3558/2024/1-ART1 $15.00
https://doi.org/

(a) saliency-guided image generation for triggering desired viewer attention (b) variations of a fixed layout

Fig. 1. *Attention-controllable image generation with saliency guidance.* Given a text prompt and saliency map pair, GazeFusion generates images that not only contain the content as described by the text prompt but also attract viewers' attention toward the desired image regions as emphasized by the saliency map. As illustrated in (a), GazeFusion understands and exploits a diversity of factors inducing visual saliency, including low-level image features, such as color, contrast, frequency, orientation, and layout, and high-level semantic information, such as objects, texts, and faces; (b) demonstrates that GazeFusion can flexibly manipulate viewers' visual attention within the generated images by purely adjusting the color and contrast of specific image content while precisely following the user-specified layout.

design factor for context creation is to direct viewers' attention to the regions of interest, such as buttons on web pages [Pang et al. 2016], products being advertised [Bakar et al. 2015], or the storytelling events in a film [Shimamura et al. 2015]. Unlike the layout of an image, human attention is selective in nature [Kastner and Ungerleider 2000; Kümmerer et al. 2015], concurrently influenced by high-level semantics, mid-level layouts, and low-level visual features [Harel et al. 2006; Itti et al. 1998; Jia and Bruce 2020; Kümmerer et al. 2014] (see examples in Figure 1), as well as spatial and temporal characteristics [Droste et al. 2020; Wang et al. 2018]. Therefore, existing conditioning mechanisms do not adequately control a viewer's visual spatial attention.

Our work aims to generate images and videos that guide viewers' visual attention toward specific regions of interest. To this end, we first analyze the spatial visual saliency [Jia and Bruce 2020] of a large-scale image dataset [Jiang et al. 2015]. The resulting image-saliency pairs are then leveraged to train a custom adapter network conditioned on saliency maps for text-to-image diffusion models. We further extend our saliency-aware conditioning from static images to temporally consistent video generation.

To evaluate the model's effectiveness in guiding real users' visual attention, we conducted a user study with human observers naturally examining the generated images with 3,000 eye-tracked trials. A series of objective evaluations demonstrate that our method consistently outperforms alternative control mechanisms in guiding user attention when viewing its generated visual content. We also showcase how the model can be applied in practical applications, including interactive designing saliency guidance, suppressing viewer attention in unwanted regions, and adapting generated content to various display and viewing conditions. This research serves as an important step toward human-centric control over generative models, focusing on the integration of

typically implicit human design intentions within the content produced by these models. Our source code and pre-trained models are released at https://github.com/NYU-ICL/saliency-guided-image-generation.

To summarize, our main contributions include:

- introduce a saliency-guided image and video generation framework that incorporates the data priors of human visual attention mechanisms into the generation process;
- systematically study and validate the attention-directing performance of the proposed framework through an eye-tracked user study and a large-scale model-based saliency analysis;
- demonstrate the availability and flexibility of the proposed framework in practical applications, including interactive design of saliency guidance, viewer attention suppression in unwanted regions, and adaptive visual content generation for varied display and viewing conditions.

2 RELATED WORK

2.1 Controllable Diffusion Models

Recent advancements in diffusion-based generative models have found great success in image [Rombach et al. 2022] and video [Blattmann et al. 2023] content generation, as surveyed by Po et al. [2023] and Yang et al. [2023]. These foundation models, however, demand non-trivial prompt engineering to achieve desired user control over the generation process. To overcome this limitation, model customization approaches [Hu et al. 2021; Ye et al. 2023] as well as lightweight adapter networks [Li et al. 2023b; Zhang et al. 2023b; Zhao et al. 2023] have been established as the primary mechanisms for adding control over of the generated content. Current control strategies, however, use image-space annotations, including body pose, depth maps, or bounding boxes, to guide the *spatial layout* of a generated image. While visual attention is impacted by spatial layout arrangements, it also largely depends on the interplay among low-level visual features (e.g., contrast, frequency, and color) and high-level semantic information (e.g., objects, texts, and faces). Therefore, spatial attention exhibits selective and sometimes individually inconsistent patterns [Kümmerer et al. 2015]. We use model-predicted saliency maps to condition a customized conditional diffusion model. This model is then used to steer the generation of images and videos to align with specific attention patterns based on the design intentions.

2.2 Human Visual Attention Modeling and Prediction

Due to the complexity of cognitive visual attention [Kastner and Ungerleider 2000], 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 [Bruce and Tsotsos 2005, 2007; Harel et al. 2006; Itti and Koch 2001; Itti et al. 1998; Judd et al. 2009]. However, these low-level features by themselves are insufficient to account for top-down influences, e.g., our familiarity with different objects [Elazary and Itti 2008]. 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 [Judd et al. 2009], CAT2000 [Borji and Itti 2015], SALICON [Huang et al. 2015; Jiang et al. 2015] for images, VR saliency for 360 videos [Sitzmann et al. 2018], and DHF1K for videos [Wang et al. 2019]. These large-scale datasets catalyzed various deep neural network based saliency metrics for RGB images (e.g., DeepGaze models [Kümmerer et al. 2014; Kummerer et al. 2017; Linardos et al. 2021], EMLNet [Jia and Bruce 2020], SalGAN [Pan et al. 2017]), RGB-D frames [Ren et al. 2015; Sun et al. 2021; Zhang et al. 2021], videos ([Droste et al. 2020; Jiang et al. 2018; Min and Corso 2019; Wang et al. 2018]), and panoramas ([Zhang et al. 2018]). Saliency models are further extended to predict temporal fixation durations [Fosco et al. 2020] and scanpaths [Martin et al. 2022].

2.3 Attention-Aware Computer Graphics

Leveraging the selective attention distributions predicted by computational models has facilitated practical applications in enhancing the end-user experience or improving system efficiency. For instance, an existing image may be automatically enhanced by identifying and removing distractions without losing the fidelity [Aberman et al. 2022; Jiang et al. 2021; McDonnell et al. 2009; Mejjati et al. 2020; Miangoleh et al. 2023]. High salient regions may also be prioritized for image loading [Valliappan et al. 2020]. Beyond images, character animation may also be enhanced with identified salient body parts [McDonnell et al. 2009]. So far, saliency-based optimization has been used to guide broad editing tasks given existing content. This research aims to introduce the building block of interactively creating desired images and videos from only simple user interference and text prompts.

3 METHOD

3.1 Saliency-guided Diffusion Model for Image Generation

Diffusion models define a Markov chain that iteratively adds Gaussian noise to samples from an empirical data distribution and gradually converts them into noisy samples from a standard Gaussian distribution. They then learn the reverse diffusion (or denoising) process to iteratively remove noise and generate new data samples from randomly sampled Gaussian noise. State-of-the-art image diffusion models are commonly trained on large-scale image datasets and are capable of synthesizing visually appealing and content-diverse images [Po et al. 2023].

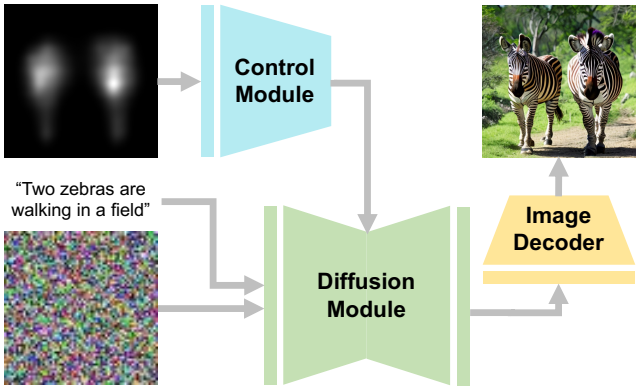

Fig. 2. *Saliency-guided image generation with GazeFusion.* Given randomly sampled noisy images as inputs, GazeFusion conditions the denoising process on user-specified saliency maps and text prompts such that the image features and semantic content in the generated images can trigger similarly distributed viewer attention.

Recent efforts toward controlling these large models, such as ControlNet [Zhang et al. 2023b] and GLIGEN [Li et al. 2023b], have shown that it is possible to incorporate a variety of multimodal conditions into the generation process. This integration allows for the manipulation of semantic information and spatial layout in the content produced by these large models. Such conditional generation is typically achieved by first augmenting pre-trained image diffusion models with an adaptation module and then fine-tuning on a considerably smaller set of condition-image pairs.

To achieve attention-aware image generation, we first curate a dataset of saliency-image pairs. The scale of existing image datasets with paired eye-tracked human saliency data is commonly too limited ($< 10k$ images) to support generative learning. Therefore, we leverage the captioned MSCOCO dataset [Lin et al. 2014] for images and a learning-based saliency model, EMLNet [Jia and Bruce 2020], to predict their corresponding saliency maps. As visualized in Figure 2, we attach a ControlNet module to the encoder and middle blocks of a pre-trained Stable Diffusion (SD2.1) model [Rombach et al. 2022] to inject saliency map conditions through zero convolutions and perform saliency-to-image translation. Particularly, the train, test, and unlabeled splits of MSCOCO 2017 (a total of 282k images) were taken to construct our training set $\mathcal{D}$. The 5k validation images of MSCOCO 2017 were held out for evaluation. GazeFusion was initialized with the pre-trained SD2.1 model checkpoint and finetuned using an Adam optimizer (constant learning rate $1e^{-5}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$) [Kingma and Ba 2015] for $5e^5$ steps. Mathematically, our saliency-conditioned training process can be summarized as

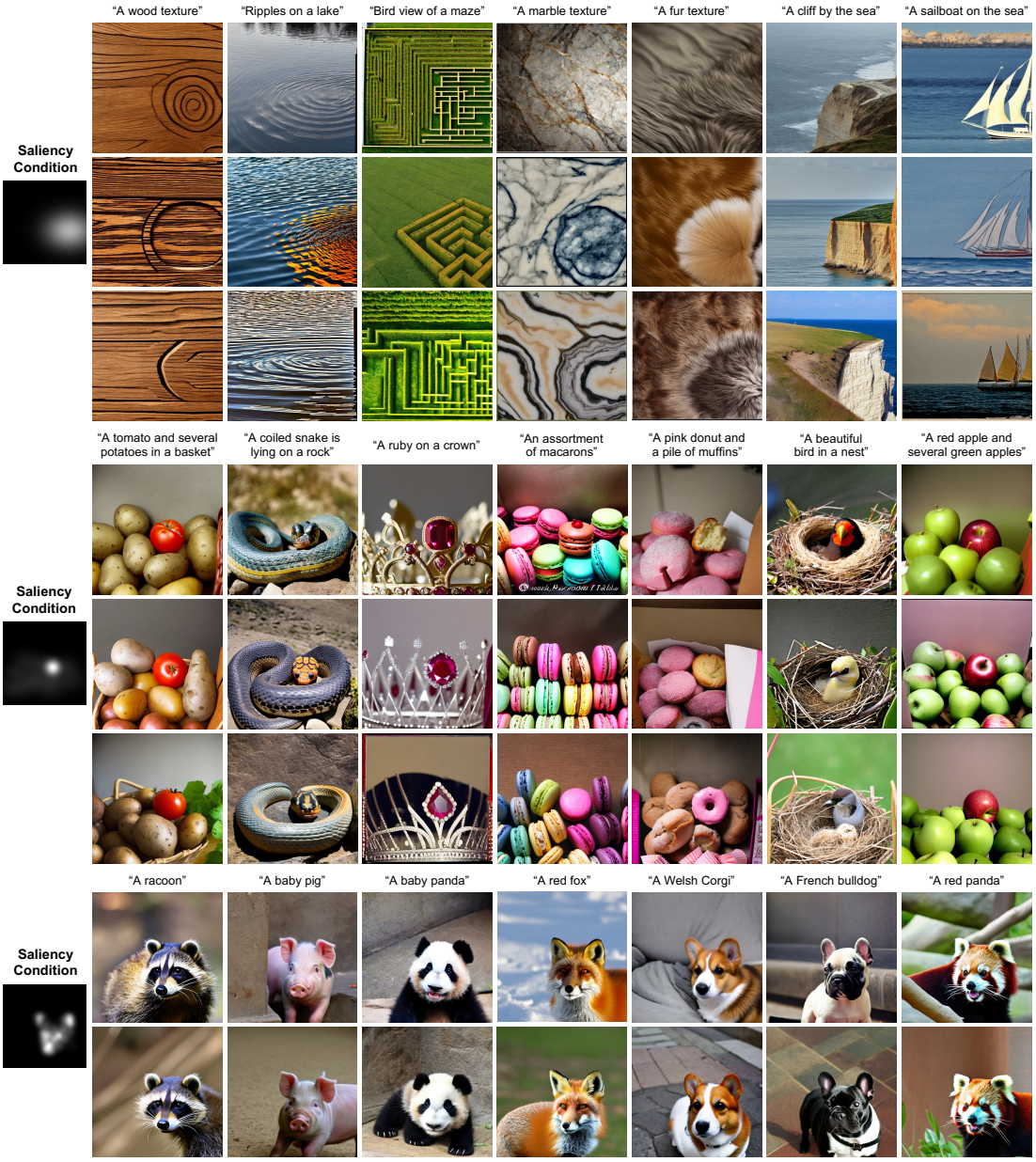

Fig. 3. *Saliency-guided image generation with GazeFusion.* During the generation process, GazeFusion leverages a variety of factors that affect visual saliency, such as color (e.g., the tomato and the apple scenes), frequency (e.g., the maze and the lake scenes), contrast (e.g., the fur and the marble texture scenes), orientation (e.g., the wood texture and the macaron scenes), layout (e.g., the last 3 rows), and high-level semantic information (e.g., the snake and the bird scenes).

Fig. 4. *Saliency-guided video generation with GazeFusion.* By leveraging an off-the-shelf zero-shot video generation pipeline, GazeFusion can be extended to generate temporally consistent video clips with spatial-temporal saliency guidance.

optimizing the denoising network in SD2.1, denoted as ϵ_θ, to predict the Gaussian noise added at each time step. The loss function driving the optimization can be formulated as follows:

$$\mathcal{L} = \mathbb{E}_{z,t,c_t,c_s,\epsilon\sim\mathcal{N}(0,1)} \|\epsilon_\theta(z_t, t, c_t, c_s) - \epsilon\|_2^2. \tag{1}$$

where z, z_t, c_t, and c_s denote a sample from the latent image distribution, the corresponding noisy sample after adding Gaussian noise for t steps, the text prompt, and the conditioning saliency map, respectively. As sampled cases in Figures 1 and 3 (and Figures 10 and 13 in the appendix) evidence, GazeFusion automatically incorporates a range of factors that lead to visual saliency, such as low-level image features (e.g. frequency, color, contrast, orientation, and layout) and high-level semantic information (e.g. objects, texts, and faces).

3.2 Extension to Saliency-guided Video Generation

We further explore the possibility of applying GazeFusion to saliency-guided video generation. Existing zero-shot (i.e. training-free) video generation pipelines extend the generation of individual frames to video clips by enforcing the temporal consistency across adjacent frames [Guo et al. 2023; Khachatryan et al. 2023]. Therefore, they are also adaptable to controlling modules with conditions such as body poses and edges.

However, a unique challenge for human perception is the domain gap between viewing individual frames with extended duration (spatial-only saliency, $\mathbf{S} = S_{\{1,2,\ldots,T\}}$, as used in Section 3.1) vs. watching them temporally composed as a video sequence (spatio-temporal saliency $\mathbf{V} = V_{\{1,2,\ldots,T\}}$) [Droste et al. 2020]. Here, S and V indicate the saliency maps on individual frames and videos of the same image sequence. This is due to various temporally induced factors – such as camera and object motions – influencing selective attention. Therefore, although it is possible to directly apply our control module to perform per-frame saliency-guided image generation $\mathbf{S}$ to compose a temporally consistent video, the resulting user attention will differ from the target $\mathbf{V}$. Concerning the domain gap, we adopt a video saliency prediction model, TASED-Net [Min and Corso 2019], to approximate $\mathbf{V}$ for temporally consistent saliency sequences. Figure 4 shows two sequences of GazeFusion-generated video frames.

4 EVALUATION

To quantitatively evaluate our saliency-guided approach to visual content generation, we conducted 1) an eye-tracked user study on GazeFusion-generated images to analyze its attention-directing performance (Section 4.1); 2) a large-scale objective evaluation on GazeFusion-generated images and videos using off-the-shelf saliency models (Section 4.2); 3) an ablation study on the quality and diversity of GazeFusion-generated images.

4.1 User Study: Tracking and Analyzing Viewers' Eye Gaze over Generated Images

The aim of our research 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, we conducted a user study with eye trackers to record participants' eye gaze patterns while they browse through a sequence of generated image samples.

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.

Setup and procedures. 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 cd/m^2) binocularly from an SR Research headrest positioned 60 cm away. The effective field of view and resolution were $46^\circ \times 26.8^\circ$ 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 5 shows the experimental setup of our user study.

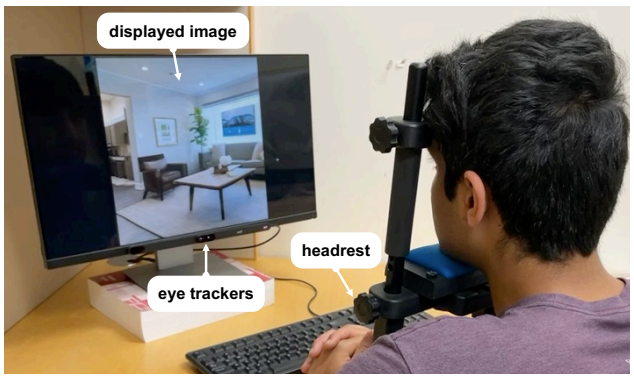

Fig. 5. *User study setup.* We used a Tobii Pro Spark eye tracker to record study participants' eye-gaze directions while they browsed through a sequence of generated images.

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 [Li et al. 2023a] for text prompt conditioning. Image saliency maps were extracted using the EML-Net saliency model [Jia and Bruce 2020] 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.

Conditions. For attention-directing performance comparison, we included two baseline conditions, the text-only model (**TEXT**), Stable Diffusion v2.1 [Rombach et al. 2022], and the bounding-box-guided model (**BBOX**), GLIGEN [Li et al. 2023b]. All three conditions share the same input text prompts to control what visual content should be generated. Additionally, **BBOX** takes in text-annotated bounding boxes predicted by the Grounding DINO model [Liu et al. 2023], and **OURS** (i.e. GazeFusion) takes in saliency maps predicted by the EML-Net saliency model. Similar to **OURS**, 50 images were generated for **TEXT** and **BBOX**, respectively.

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

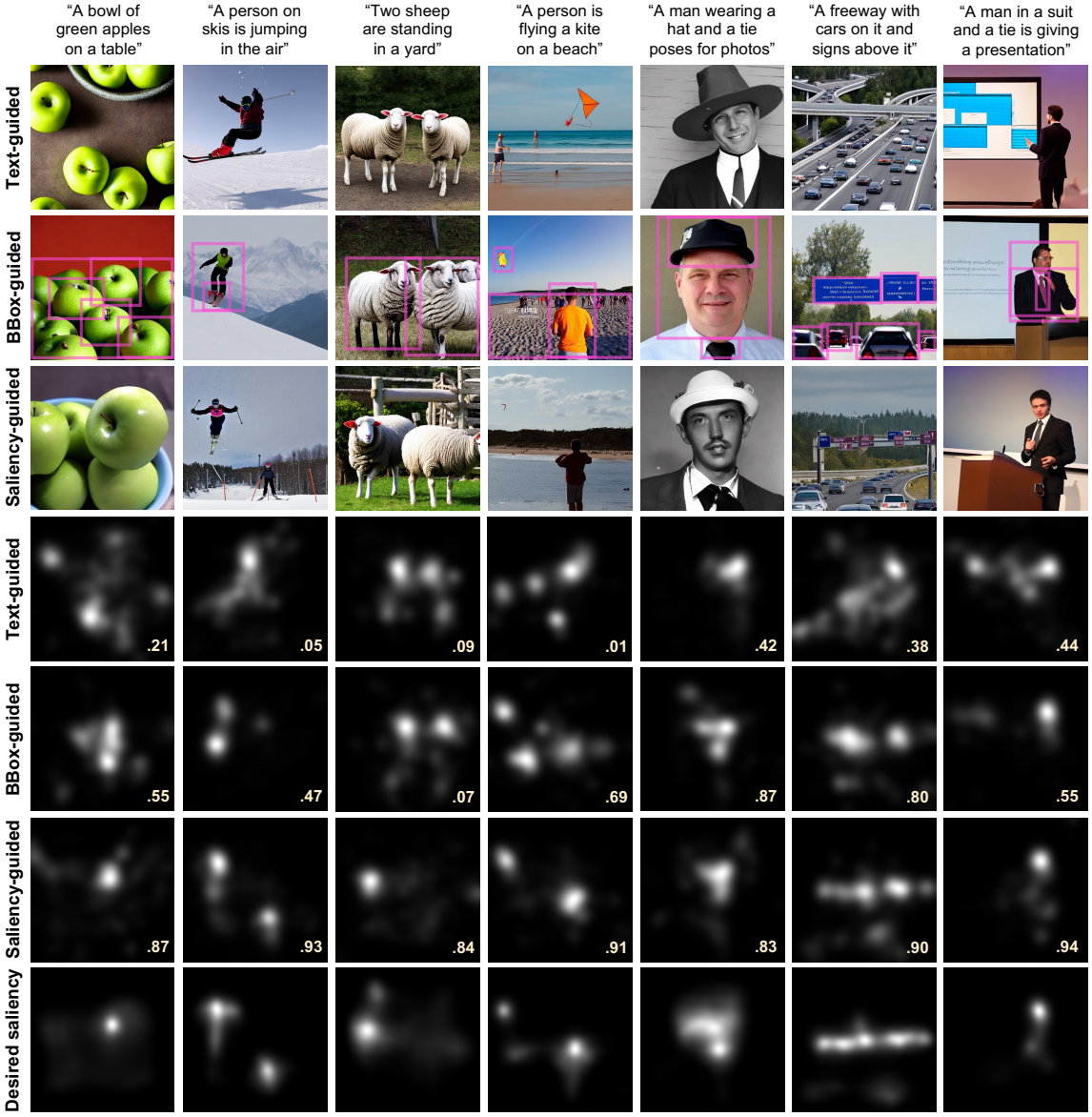

Fig. 6. *Eye-tracked user study.* Rows 1–3 show the generated images, rows 4–6 show the empirical saliency maps obtained by aggregating 20 users' eye gaze data, and row 7 shows the input conditioning saliency maps (i.e., the desired viewer attention distributions). The conditioning bounding boxes for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

Table 1. *Eye-tracked user study.* GazeFusion outperforms the two baseline methods in directing viewers' visual attention toward the user-intended image regions. ↑/↓ indicates that higher/lower score is better.

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

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.

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 [Kummerer et al. 2018]: Area Under ROC Curve (AUC) [Kümmerer et al. 2015], Normalized Scanpath Saliency (NSS) [Peters et al. 2005], 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. To convert our collected eye-tracking data into empirical saliency maps, we followed the same post-processing procedures described in [Sitzmann et al. 2018].

Results and discussion. As shown in Table 1, our saliency-guided approach consistently outperforms the two baselines in controlling and directing users' visual attention toward intended image regions across all 5 metrics. Figure 6 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, GazeFusion 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 capability of GazeFusion.

4.2 Model-based Saliency Analysis

In addition to the eye-tracked user study, we further performed a large-scale objective analysis of GazeFusion-generated images and videos to more comprehensively understand its capabilities. The two previously introduced baseline methods, **TEXT** and **BBOX**, are again adopted for comparisons.

Image. Similar to the data preparation procedures in Section 4.1, we first computed the BLIP-2 captions, spatial saliency maps, and text-annotated bounding boxes on the held-out MSCOCO data (5K images) to condition the image generation process. Next, we applied GazeFusion to generate 5K images for **OURS** using the 5K paired captions and saliency maps. Using their respective model and input conditions, 5K images were also generated for **BBOX** and **TEXT**. Finally, we measured the discrepancy between the saliency maps used as input conditions (i.e., the desired attention distributions) and the saliency maps of the generated images per the EMG-Net image saliency predictor (i.e., the achieved attention distributions).

Video. To generate a large set of video clips with GazeFusion, we first sampled 1K videos from the WebVid-10M dataset [Bain et al. 2021], trimmed them down to 4-second clips, and uniformized the frame rate to 8 FPS. Next, we took the TASED-Net video saliency predictor to extract the spatial-temporal saliency map sequence from each processed video clip. Leveraging the Text2Video-Zero pipeline [Khachatryan et al. 2023], we then applied our GazeFusion model to generate 1K video clips of 4 seconds and 8 FPS based on the previously extracted saliency map sequences. Finally, similar to generated images, we measured the frame-wise discrepancy between the desired and achieved attention distributions for all generated video clips.

Table 2. *Model-based saliency analysis.* The images and videos generated by GazeFusion achieve more aligned visual attention distributions with the input saliency maps according to EML-Net and TASED-Net. Note that the video results of **BBOX** are unavailable due to GLIGEN's incompatibility with the Text2Video-Zero pipeline.

	Image			Video		
	CC ↑	KL ↓	SIM ↑	CC ↑	KL ↓	SIM ↑
TEXT	0.22	7.27	0.35	0.21	5.68	0.34
BBOX	0.54	3.97	0.54	N/A	N/A	N/A
OURS	**0.84**	**0.75**	**0.75**	**0.79**	**0.85**	**0.68**

Results and discussion. In this large-scale evaluation, since we take advantage of ML-based saliency models to directly extract saliency maps from generated images/videos as ground truth and do not have access to eye fixation data, only CC, KL, and SIM are feasible and thus adopted. As shown in Table 2, GazeFusion significantly outperformed the two baselines across all three metrics on both image and video generation. The analysis above not only validates the extendability of GazeFusion to saliency-guided video generation but also demonstrates its robustness and generality on a wide range of text prompts and attention distributions. These results laid the foundations for the practical applications that we discuss in Section 5.

4.3 Ablation Study on Image Quality and Diversity

To verify that our saliency conditioning does not result in any performance drop to the base SD2.1 model, we adopted two image quality/diversity metrics, inception score (IS, higher is better) and Fréchet inception distance (FID, lower is better), to evaluate how do GazeFusion-generated images compare with those generated by the base model. To this end, we reused the two sets of images generated for the conditions **TEXT** (the base model with text guidance only) and **OURS** (our GazeFusion model). Note that both sets of images were generated using the same text prompts and are thus fair to be compared. Notably, GazeFusion achieved FID = 16.43 and IS = 36.05, while SD2.1 achieved FID = 22.29 and IS = 34.81, indicating that our saliency-guided generation even improves upon the base model in terms of image quality and diversity. Such improvements are likely due to the similar image distribution of our training and evaluation sets, which were both taken from MSCOCO 2017. Our GazeFusion model was carefully fine-tuned on these photo-based realistic images to learn saliency guidance while SD2.1 was not specifically optimized to generate such images. These results strongly validate that we can effectively incorporate the data priors of human attentional behaviors into the diffusion process and generate saliency-guided images while not degrading the image generation performance of the base diffusion models.

5 APPLICATIONS

5.1 Interactive Design of Saliency Guidance

Unlike other non-semantic controlling inputs, such as body poses, line sketches, and depth maps, prior research has observed a strong correlation between the content of an image and its incurred visual saliency [Borji et al. 2015]. Consequently, when the text prompt is incompatible with the saliency guidance provided by a designer, GazeFusion may generate unnaturally looking images with artifacts (e.g., distorted objects and human bodies, see the top row of Figure 7), as we have repeatedly observed in our experiments.

To this end, we developed an interactive saliency authoring system to ease designers' trial-and-error efforts in crafting prompt-compatible saliency guidance. Specifically, a user provides a text prompt p_{in} and generates a tentative saliency map by clicking over a black canvas. Each click creates a bivariate Gaussian $G_i(w_i, \mu_i, \Sigma_i)$ that can be further adjusted to compose the desired saliency guidance. Here, w_i, μ_i, and Σ_i denote the Gaussian's

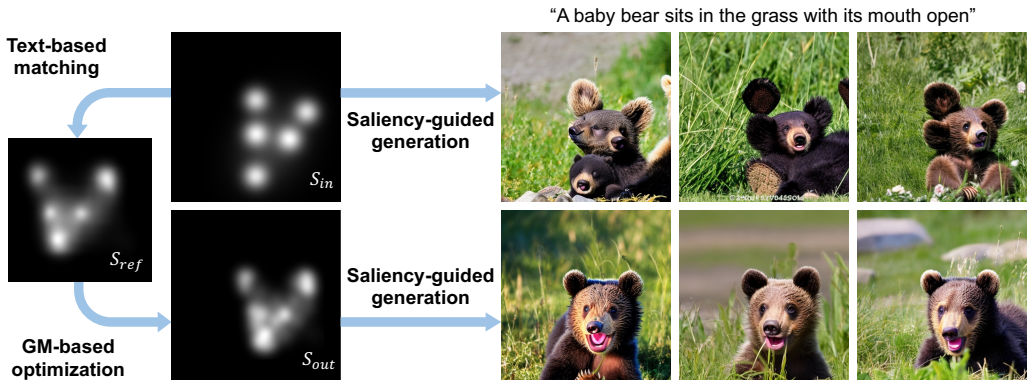

Fig. 7. *Interactive design of saliency guidance.* Our interactive saliency design system eases users' burden of crafting high-quality, prompt-compatible saliency maps by proposing corrections when necessary. The top row shows a tentative user-created saliency map and its corresponding GazeFusion-generated images; the bottom row shows the artifact-free generation results obtained after optimizing the saliency guidance.

weight, mean, and covariance matrix. The resulting saliency map is represented as a Gaussian mixture (GM):

$$S_{\text{in}} := \sum_{i=1}^{N_{\text{in}}} G_i\left(w_i, \mu_i, \Sigma_i\right). \tag{2}$$

Next, from our 282k text-saliency paired dataset $\mathcal{D}$ (Section 3.1), we search for the text prompt closest to the user-provided one within the CLIP-embedded language space [Radford et al. 2021], and retrieve its corresponding saliency map (i.e. the reference saliency map) as a starting point for optimization:

$$p_{\text{ref}}, S_{\text{ref}} = \underset{\{p,S\}\in\mathcal{D}}{\arg\min} \left\| \mathcal{E}(p) - \mathcal{E}(p_{\text{in}}) \right\|_2. \tag{3}$$

Here, p_{ref} and S_{ref} denote the retrieved reference text-saliency pair; $\mathcal{E}$ denotes the CLIP text encoder. Finally, we convert S_{ref} into a Gaussian mixture $\hat{S}_{\text{ref}}$ and optimize an image-space transformation matrix $T \in \mathbb{R}^{2\times 2}$, which is composed of a translation vector $t \in \mathbb{R}^2$, a rotation angle $r \in [0, \pi]$, and a scaling vector $s \in \mathbb{R}^2_+$, to approximate the user-intended saliency guidance S_{in} as follows:

$$\hat{S}_{\text{ref}} := \sum_{j=1}^{N_{\text{out}}} G_j\left(w_j, \mu_j, \Sigma_j\right), \tag{4}$$

$$\mathcal{L} = \left\| T \cdot \hat{S}_{\text{ref}} - S_{\text{in}} \right\|_2, \quad T_{\text{out}} = \underset{T\in\mathbb{R}^{2\times 2}}{\arg\min}\, \mathcal{L}, \quad S_{\text{out}} = T \cdot \hat{S}_{\text{ref}}. \tag{5}$$

Here, $\mathcal{L}$, T_{out}, and S_{out} denote the objective function, optimized transformation matrix, and corrected saliency map, respectively. Figure 7 shows an example where the initial user-specified saliency map S_{in} is incompatible with the accompanying text prompt due to inappropriate saliency intensity and layout. In the corresponding generated images, the baby bear exhibits a severely disfigured face and incorrect body anatomy. In comparison, the images generated based on the system-corrected saliency map S_{out} not only present a realistically looking baby bear without artifacts but also maintain the user-intended saliency layout. Notably, the whole generation process only involves a few mouse clicks from the users with the correction automatically performed. Using this interactive design interface, users can easily create complicated custom saliency maps to perform attention-controlled image generation (refer to Figures 1 and 3 for a few examples).

Fig. 8. *Demonstration of attention suppression.* GazeFusion is capable of completely shifting viewers' attention away from specific image regions (absolute suppression) or reducing viewers' attention in less important regions (relative suppression). Brighter/darker colors indicate higher/lower saliency (or desired attention levels).

5.2 Attention Suppression

Now that we have demonstrated GazeFusion's capabilities in generating high-quality images and videos that attract viewers' attention toward the intended regions, we also explore the opposite effects of suppressing viewers' attention in unwanted regions. In particular, we differentiate between two types of attention suppression using GazeFusion: 1) absolute suppression, where viewer attention is completely shifted away from the target regions; 2) relative suppression, where viewer attention is adaptively reduced in the less important regions compared to the more important ones, though not entirely removed.

As demonstrated in Figure 8, we crafted two pairs of saliency maps and fed them to GazeFusion. In each pair, one saliency map is designed to demonstrate a type of attention suppression (**TEST**), and the other is crafted to be the same except in the suppressed regions (**CONTROL**). For 1) *absolute suppression*: **TEST** is strongly salient everywhere except for an oval-shaped region. The **TEST**-guided image features several arrays of donuts except for the suppressed region, while the **CONTROL**-guided image shows donuts everywhere. For 2) *relative suppression*, the text prompt describes a two-person chatting scenario. **CONTROL** shows two separate bright regions that are equally salient, and **TEST** shows the same two bright regions but with one of them being considerably brighter than the other. While the **CONTROL**-guided image depicts two persons at the two bright regions, the **TEST**-guided image, interestingly, not only positions the two persons to the two bright regions but also makes the person at the brighter one face the camera and the other person turn his back to the camera and out of focus. More attention suppression results can be found in Figures 11 and 12 in the appendix.

5.3 Display-Adaptive Generation

The increasing size of the latest displays has widened the field of view (FoV) they deliver. However, most visual content is created without considering the exact environment in which it will be displayed. In practice, this may lead to unpredictable and compromised viewing experiences. For instance, when viewing an image on a 13-inch MacBook, most of the image content will be located within the central part of the viewers' visual field; when viewing the same image on a 34-inch curved monitor with a 21:9 aspect ratio, however, a non-trivial part of the

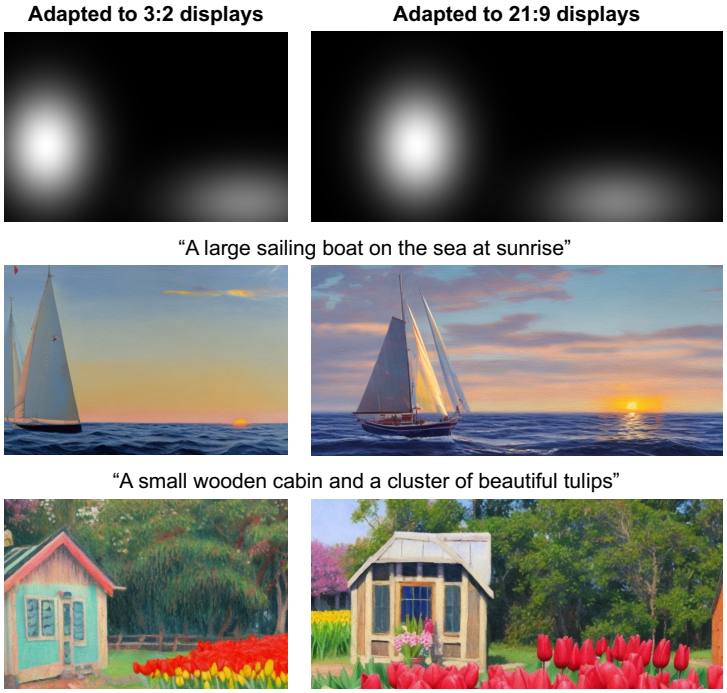

Fig. 9. *Demonstration of display-adaptive image generation.* The left/right column shows the target saliency distribution and the corresponding generated images for a narrow/wide field display. Notably, the spatial guidance for the wide-field display is designed to ensure that the main image content remains in a comfortable FoV for viewers. This can help improve content visibility and reduce excessive eye/head movements, countering the low visual acuity of peripheral vision.

image will be displayed in the viewers' low-acuity peripheral visual field [Eriksen and Yeh 1985; Reeves et al. 1999] and not be properly consumed. In addition, viewers also have to shift their eye gaze or rotate their necks more frequently, which may lead to ergonomic problems [Gallagher et al. 2021; Zhang et al. 2023a].

Leveraging GazeFusion, artists and designers can readily customize their desired saliency guidance and generate graphics that are perceptually optimized to varying displays (aspect ratio and size) and viewing conditions (viewing distance and angle). This feature can help, for instance, keep image content of particular interests and importance within a reasonable FoV for viewers to enjoy an uncompromised viewing experience when switching from one display to another. As demonstrated in Figure 9, GazeFusion-based adaptive generation shifts important image content to the optimal viewing angle for the display in use through saliency conditioning, effectively improving content visibility and reducing the necessity of frequent eye/head movements due to peripheral viewing.

6 LIMITATIONS AND FUTURE WORK

Our saliency-guided controlling module was established based on an end-to-end computational saliency model optimized on human gaze data [Jia and Bruce 2020]. However, the factors inducing human visual attention are diverse and multi-dimensional, including low-level image features, mid-level local structures, and high-level semantics [Hayes and Henderson 2021]. Currently, GazeFusion model does not differentiate between these underlying causes. Integrating various saliency models targeting different levels of saliency-triggering factors

under a probabilistic framework [Kümmerer et al. 2015] may shed light on more fine-grained saliency-based control, such as specifying local color- and contrast-induced saliency vs. semantic labels in the saliency map.

The extension to saliency-guided video generation, while showing numerical saliency alignments between the input control and the computed prediction with TASED-Net [Min and Corso 2019], has not been measured with human observers. This is due to the significantly large sampling requirement to obtain an eye-tracking-revealed spatial-temporal saliency [Wang et al. 2018]. To this end, we plan to investigate eye-tracking-free human attention assessment approaches via crowdsourcing platforms, e.g., [Kim et al. 2017]. In addition, the current approach treats the saliency frames in the control sequence separately to generate temporally consistent videos frame-by-frame with [Khachatryan et al. 2023]. Introducing an explicit temporal module to exploit the temporally induced factors, such as motions between adjacent frames, may further improve the controlling effectiveness.

The proposed interactive design tool automatically corrects users' arbitrarily specified saliency input through a text-saliency compatibility matching and adjustment, as shown in Figure 7. Recognizing that image artifacts may also influence saliency [Yang et al. 2021], we plan to investigate the three-fold cross effects among characterized saliency maps, text feature spaces, and their generated image quality assessment [Golestaneh et al. 2022; Yang et al. 2019; Zhang et al. 2014]. A quantifiable correlation in the continuous domain may provide guidance for quality-predictable and saliency-aware generation.

7 CONCLUSION

In this paper, we present a saliency-guided generative model that guides users' viewing attention. It may match designers' specifications with a simple click-and-run user interface. An eye-tracked user study evidences the real-world effectiveness. With various demonstrated applications, such as inversely suppressing attention and adapting the generations to various viewer-display conditions, we hope the research will initiate the first step of viewer-perception-aware generative models.

ACKNOWLEDGMENTS

This research is partially supported by the National Science Foundation grant #2232817 and a gift from Google.

REFERENCES

Kfir Aberman, Junfeng He, Yossi Gandelsman, Inbar Mosseri, David E Jacobs, Kai Kohlhoff, Yael Pritch, and Michael Rubinstein. 2022. Deep saliency prior for reducing visual distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19851–19860.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.

Muhammad Helmi Abu Bakar, Mohd Asyiek Mat Desa, and Muhizam Mustafa. 2015. Attributes for image content that attract consumers' attention to advertisements. *Procedia-Social and Behavioral Sciences* 195 (2015), 309–314.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing* 24, 12 (2015), 5706–5722.

Ali Borji and Laurent Itti. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581* (2015).

Neil Bruce and John Tsotsos. 2005. Saliency based on information maximization. *Advances in neural information processing systems* 18 (2005).

Neil Bruce and John Tsotsos. 2007. Attention based on information maximization. *Journal of Vision* 7, 9 (2007), 950–950.

Richard Droste, Jianbo Jiao, and J Alison Noble. 2020. 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, 419–435.

Lior Elazary and Laurent Itti. 2008. Interesting objects are visually salient. *Journal of vision* 8, 3 (2008), 3–3.

Charles W Eriksen and Yei-yu Yeh. 1985. Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance* 11, 5 (1985), 583.

Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. 2020. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4473–4482.

Kaitlin M Gallagher, Laura Cameron, Diana De Carvalho, and Madison Boule. 2021. Does using multiple computer monitors for office tasks affect user experience? a systematic review. *Human Factors* 63, 3 (2021), 433–449.

S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1220–1230.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. *Advances in neural information processing systems* 19 (2006).

Taylor R Hayes and John M Henderson. 2021. Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific reports* 11, 1 (2021), 18434.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*. 262–270.

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience* 2, 3 (2001), 194–203.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (1998), 1254–1259.

Sen Jia and Neil DB Bruce. 2020. Eml-net: An expandable multi-layer network for saliency prediction. *Image and vision computing* 95 (2020), 103887.

Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. 2018. Deepvs: A deep learning based video saliency prediction approach. In *Proceedings of the european conference on computer vision (eccv)*. 602–617.

Lai Jiang, Mai Xu, Xiaofei Wang, and Leonid Sigal. 2021. Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16509–16518.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2106–2113.

Sabine Kastner and Leslie G. Ungerleider. 2000. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience* 23, 1 (2000), 315–341.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023).

Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–40.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

M Kümmerer, L Theis, and M Bethge. 2014. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. In *International Conference on Learning Representations (ICLR 2015)*. 1–12.

Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. 2015. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences* 112, 52 (2015), 16054–16059.

Matthias Kummerer, Thomas SA Wallis, and Matthias Bethge. 2018. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 770–787.

Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. 2017. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE international conference on computer vision*. 4789–4798.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. 2021. 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*. 12919–12928.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

Daniel Martin, Ana Serrano, Alexander W Bergman, Gordon Wetzstein, and Belen Masia. 2022. Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2003–2013.

Rachel McDonnell, Michéal Larkin, Benjamín Hernández, Isaac Rudomin, and Carol O'Sullivan. 2009. Eye-catching crowds: saliency based selective variation. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–10.

Youssef A Mejjati, Celso F Gomez, Kwang In Kim, Eli Shechtman, and Zoya Bylinskii. 2020. Look Here! A Parametric Learning Based Approach to Redirect Visual Attention. In *European Conference on Computer Vision*. 343–361.

S Mahdi H Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, and Yağiz Aksoy. 2023. Realistic Saliency Guided Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 186–194.

Kyle Min and Jason J Corso. 2019. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2394–2403.

Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. 2017. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081* (2017).

Xufang Pang, Ying Cao, Rynson WH Lau, and Antoni B Chan. 2016. Directing user attention via visual flow on web designs. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–11.

Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision research* 45, 18 (2005), 2397–2416.

Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit H Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. 2023. State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204* (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Byron Reeves, Annie Lang, Eun Young Kim, and Deborah Tatar. 1999. The effects of screen size and message content on attention and arousal. *Media psychology* 1, 1 (1999), 49–67.

Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. 2015. Exploiting global priors for RGB-D saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 25–32.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Arthur P Shimamura, Brendan I Cohn-Sheehy, Brianna L Pogue, and Thomas A Shimamura. 2015. How attention is driven by film edits: A multimodal experience. *Psychology of Aesthetics, Creativity, and the Arts* 9, 4 (2015), 417.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1633–1642.

Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. 2021. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1407–1417.

Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. 2020. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications* 11, 1 (2020), 4553.

Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. 2018. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 4894–4903.

Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. 2019. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 220–237.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.

Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. 2019. SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the 27th ACM international conference on multimedia*. 1383–1391.

Xiaohan Yang, Fan Li, and Hantao Liu. 2021. A measurement for distortion induced saliency variation in natural images. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–14.

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

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. 2021. Uncertainty inspired RGB-D saliency detection. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 5761–5779.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* 23, 10 (2014), 4270–4281.

Yunxiang Zhang, Kenneth Chen, and Qi Sun. 2023a. Toward Optimized VR/AR Ergonomics: Modeling and Predicting User Neck Muscle Contraction. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–12.

Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. 2018. Saliency detection in 360 videos. In *Proceedings of the European conference on computer vision (ECCV)*. 488–503.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. 2023. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.16322* (2023).

Fig. 10. *Ablation study on subtle changes in saliency guidance.* Visually less significant components in the conditioning saliency map (highlighted regions) may have significant impacts on the generation results.

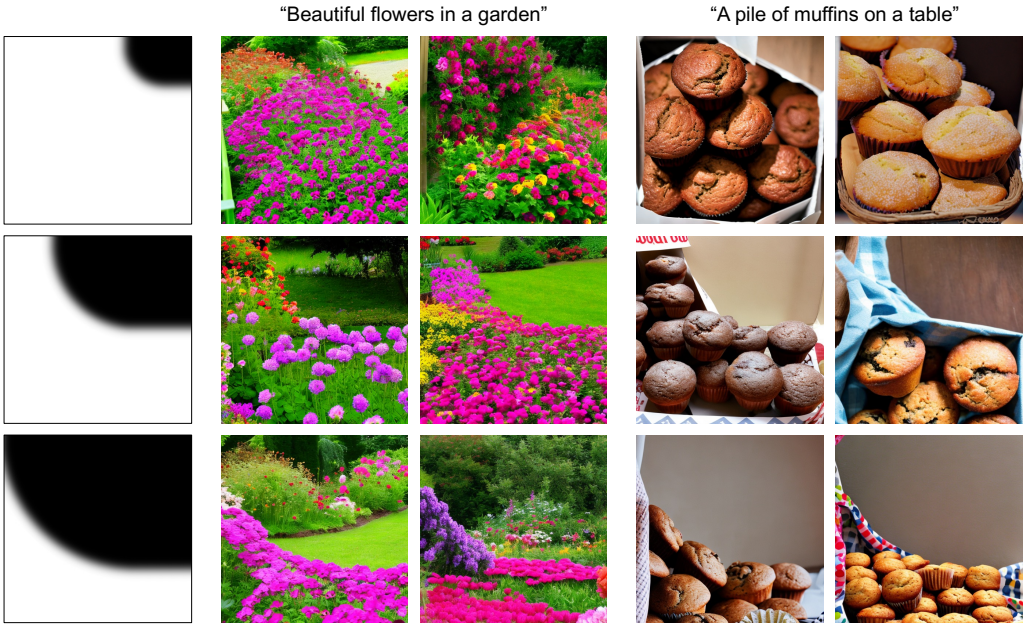

Fig. 11. *Absolute attention suppression with GazeFusion.*

Fig. 12. *Relative attention suppression with GazeFusion.*

Fig. 13. *Saliency-guided image generation with GazeFusion.*