Foveated Diffusion:
Efficient Spatially Adaptive Image and Video Generation

Brian Chao*, Lior Yariv*, Howard Xiao[1], and Gordon Wetzstein[1]

Stanford University, USA

Abstract. Diffusion and flow matching models have unlocked unprecedented capabilities for creative content creation, such as interactive image and streaming video generation. The growing demand for higher resolutions, frame rates, and context lengths, however, makes efficient generation increasingly challenging, as computational complexity grows quadratically with the number of generated tokens. Our work seeks to optimize the efficiency of the generation process in settings where the user's gaze location is known or can be estimated, for example, by using eye tracking. In these settings, we leverage the eccentricity-dependent acuity of human vision: while a user perceives very high-resolution visual information in a small region around their gaze location (the foveal region), the ability to resolve detail quickly degrades in the periphery of the visual field. Our approach starts with a mask modeling the foveated resolution to allocate tokens non-uniformly, assigning higher token density to foveal regions and lower density to peripheral regions. An image or video is generated in a mixed-resolution token setting, yielding results perceptually indistinguishable from full-resolution generation, while drastically reducing the token count and generation time. To this end, we develop a principled mechanism for constructing mixed-resolution tokens directly from high-resolution data, allowing a foveated diffusion model to be post-trained from an existing base model while maintaining content consistency across resolutions. We validate our approach through extensive analysis and a carefully designed user study, demonstrating the efficacy of foveation as a practical and scalable axis for efficient generation. Project website at https://bchao1.github.io/foveated-diffusion/.

Keywords: Diffusion models · Foveation · Efficient Visual Generation

1 Introduction

Interactive image and streaming video generation place strict demands on the frame rates of emerging diffusion and flow matching models used for this purpose [1,7,8,14,21,27,34,44,62,72,79,85,89,93,94,99]. At the same time, demands on image resolutions and video frame or context lengths are also growing. How can we generate an ever-increasing number of tokens at fast frame rates when the computational complexity of the attention mechanism in modern diffusion transformers (DiTs) [60] grows quadratically with the token sequence length?

* Denotes equal contribution.

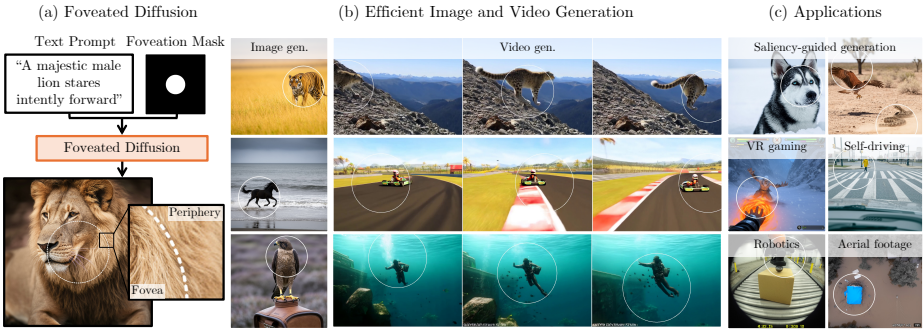

Fig. 1: Foveated Diffusion. (a) Given user-specified masks and text prompts as input, our method generates foveated content using fewer tokens than full high-resolution generation, resulting in faster inference while maintaining comparable perceptual quality. (b, c) Foveated Diffusion is well suited for tasks where salient regions require high-resolution synthesis, while peripheral regions can be generated at a lower resolution.

Our work builds on an intuitive insight that answers this question: ultimately, a human observes the generated content, so why not exploit the unique characteristics of the human visual system to generate the content in a computationally efficient, perceptually motivated manner? Specifically, we build on the concept of foveation — humans are able to perceive very high-resolution visual information in a small region around their gaze location (the foveal region) but their ability to resolve detail rapidly degrades in the visual periphery [2, 86].

With this work, we introduce the concept of *Foveated Diffusion* and develop a practical framework for post-training existing image or video generation models for foveated visual generation. Our framework starts with a foveation mask that guides the spatial layout of non-uniformly distributed tokens over the image or video frame that we wish to generate. Our key idea is eccentricity-dependent token allocation: given a foveation mask that defines the high-acuity foveal region, we allocate higher token density near the fovea and progressively fewer tokens toward the periphery, enabling spatially adaptive computation aligned with human perceptual sensitivity. Using a foveated token layout, we follow standard diffusion or flow-matching procedures to generate an image from Gaussian noise and conditioning text prompts (see Fig. 1); the key difference between Foveated Diffusion and conventional methods is that we operate with a significantly reduced set of tokens during denoising at all times, achieving substantial computational savings. We develop a simple yet highly effective mixed-resolution tokenization scheme, accompanied by a suitable modification of Rotary Positional Embeddings (RoPE) [75, 88], along with a post-training strategy that transforms high-resolution pretrained models into foveated generative models. Together, these contributions establish a principled framework that preserves cross-resolution content consistency while achieving significant speedup.

Our approach is inspired by foveated rendering [23, 24, 59], a standard technique widely used in traditional computer graphics. Foveated Diffusion and

rendering share the idea of leveraging the user's gaze location and a model of eccentricity-dependent acuity to reduce computation in the visual generation process. The key difference is that foveated rendering accelerates modules of the traditional graphics pipeline, such as the geometry and shading engines, whereas our approach seeks to achieve similar benefits for modern DiT-based diffusion and flow-matching models. While similar in spirit, these two approaches to foveated content creation differ substantially in their methods.

In summary, we propose a perceptually motivated framework for computationally efficient and spatially adaptive image and video generation. Our approach is backed by an extensive set of evaluations, including a detailed analysis of the compute–quality trade-off in various settings as well as a user study. Our key contributions are as follows:

- We introduce of the concept of *Foveated Diffusion*: a perceptually motivated, mixed-resolution diffusion algorithm for efficient image and video generation.
- We present a principled approach for tokenization, training, and inference of DiT-based generative models using spatially adaptive mixed-resolution tokens, providing cross-resolution content consistency by design.
- We demonstrate significant speedups of up to 2× and 4× for image and video generation, respectively, while preserving perceptual quality, validated through a carefully designed user study and visual quality metrics.

2 Related Work

Foveation for Computer Vision and Rendering. Decades of vision research have shown that visual acuity decreases rapidly with retinal eccentricity, i.e. distance from the fovea, where spatial resolution is highest [12, 23, 67, 83]. As a result, the human visual system processes central vision at significantly higher spatial precision than the periphery.

Real-time rendering systems exploit this eccentricity-dependent resolution of human vision by allocating higher spatial resolution near the gaze location and lower resolution in peripheral regions. When combined with real-time eye tracking, such foveated rendering systems reduce bandwidth and compute substantially, enabling interactive rendering at a fraction of the full-resolution cost while maintaining comparable perceptual quality [24, 38, 52, 59, 64, 74, 76, 77, 84]. More recently, foveation has been applied to neural rendering and novel view synthesis [15, 22, 70] to accelerate the rendering of Neural Radiance Fields (NeRFs) [57] and Gaussian splats [40] for immersive displays. In computer vision and robotics, foveation has also been used to improve efficiency in neural network architectures or perception tasks [3, 42, 58], such as mixed-resolution tokenization of vision transformers [26, 36, 66, 68] and robot policy learning [11, 41].

However, while foveation has been extensively explored across rendering and perception pipelines, it has not yet been realized in generative modeling, despite their increasing capabilities in immersive and interactive visual generation. This gap motivates the need for a generative framework that can allocate capacity according to visual eccentricity.

Efficient Visual Generation. Diffusion models have fundamentally reshaped visual generative modeling, setting new standards in photorealism, diversity, and controllability for both images and videos. While early diffusion models leverage U-Net backbones [65], Diffusion Transformer (DiT)-based architectures have emerged as the dominant paradigm for scalable, high-fidelity generation [18, 45, 49, 60, 81]. However, the computational cost of DiTs is quadratic with respect to the input token count due to the expensive self- and cross-attention mechanisms [80]. This fundamental limitation of transformer architectures severely constrains context length, leading either to degraded visual consistency under fixed compute and memory budgets or to prohibitive computational costs for immersive, high-fidelity, long-form generation.

There have been significant efforts in improving the computational efficiency of the attention mechanism, including various attention variants that reduce the algorithmic complexity [4, 10, 39, 53, 82, 92, 95], hardware-aware optimization [13, 91, 98, 100, 101], KV-caching [47, 69], etc. Another orthogonal axis of research aims to simply reduce the effective token count while maintaining high image quality. Token merging methods [5, 9, 51] identify redundant tokens at each layer of a Diffusion Transformer (DiT) and merge similar tokens according to a predefined heuristic or importance metric. While originally developed for vision transformers in recognition and perception tasks, they have recently been shown to effectively reduce token counts for generative models as well [6, 20, 25, 43, 50, 56, 87]. Recent training-free mixed-resolution denoising methods [33, 78, 88] downsample or upsample tokens during the diffusion process using fixed importance metrics such as entropy or saliency to reduce token counts for efficient generation. However, directly applying standard denoising to mixed-resolution tokens requires carefully tuned noise schedules and re-noising strategies to preserve diffusion noise statistics and maintain global content structure. These procedures are brittle; without them, mixed-resolution generation leads to structural inconsistencies and cross-scale artifacts, as we demonstrate in Sec. 4. In addition, existing approaches rely on multi-stage pipelines in which a low-resolution image first establishes global layout, followed by progressive token upsampling. Such designs complicate the diffusion trajectory and hinder compatibility with real-time generation and model distillation.

Although all these methods significantly improve efficiency in visual generation, they ignore a key characteristic of visual perception: human visual acuity decreases sharply with eccentricity. These methods focus on reconstructing high-resolution imagery everywhere and treat all spatial regions uniformly. However, because generated images are intended for human observers, generation should be optimized for perceptual relevance rather than uniform pixel fidelity. In contrast, our Foveated Diffusion pipeline leverages this principle by embedding spatially adaptive token allocation directly into the diffusion process given a predetermined foveation mask. By concentrating computation in high-acuity regions and sparsifying peripheral regions, we depart from uniform-resolution synthesis and achieve perceptually aligned, computationally efficient generation.

3 Method

In this section, we first review the basic concepts of foveated rendering in traditional graphics, as well as standard diffusion and flow-matching models in Sec. 3.1. We then introduce our Foveated Diffusion framework (Fig. 2) and explain its tokenization, inference, and training pipelines in Sec. 3.2.

3.1 Preliminaries

Foveated Rendering. Foveated rendering refers to the spatially adaptive computation where computational resources are allocated unevenly across the image according to a specified user gaze location. Modern real-time graphics systems leverage eye-tracking to render high-resolution imagery in the foveal regions while aggressively reducing shading, rasterization, or sampling rates in the peripheral regions [3,23,24,38,59,84].

Formally, we define a binary foveation mask $M \in \{0,1\}^{H\times W}$ constructed from visual eccentricity, where $M(i,j) = 1$ denotes high-resolution (HR) regions near the fovea and $M(i,j) = 0$ denotes low-resolution (LR) peripheral regions. The rendering quality is concentrated in the HR region near the fixation point (center of foveation), and progressively reduced toward the periphery. Most importantly, in foveated rendering, the scene content is unknown a priori and the foveation mask is known via gaze.

We denote $x^{\text{high}} \in \mathbb{R}^{3\times H\times W}$ as the underlying high-resolution content, and $x^{\text{low}} \in \mathbb{R}^{3\times (H/d)\times (W/d)}$ as the underlying low-resolution content, where d is the spatial downsampling factor. In this paper, we define $d = 2$, allowing $4\times$ computational gain in the periphery. During foveated rendering, only pixels (i,j) with $M(i,j) = 1$ are synthesized at high resolution from x^{high} (the foveal region), while pixels with $M(i,j) = 0$ are synthesized from x^{low} (the peripheral region). Thus, computation is performed exclusively on the masked regions at their respective resolutions, rather than producing full high- and low-resolution renderings. Composing the final foveated image in pixel space is simply achieved by blending, that is:

$$x_{\text{fov}} = M \odot x^{\text{high}} + (1 - M) \odot \text{Up}(x^{\text{low}}), \tag{1}$$

where $\text{Up}(\cdot)$ denotes the spatial upsampling operator, and $\odot$ denotes elementwise multiplication.

Diffusion Models. Diffusion models [29,30,73] define a generative process that gradually transforms samples from an easy-to-sample distribution (i.e. Gaussian) into data samples via a learned reverse-time process. Modern large-scale diffusion models operate in a compressed latent space to improve computational efficiency [60,65]. Given an image or a video, a variational autoencoder (VAE) [17], consisting of an encoder E and a decoder D, maps it into a latent representation $z_0 \in \mathbb{R}^{c\times (h\cdot w)}$ (images) or $z_0 \in \mathbb{R}^{c\times (f\cdot h\cdot w)}$ (videos, with additional frame dimension f), where the diffusion process is defined.

Flow Matching. Flow matching [55] reformulates diffusion as a continuous-time optimal transport problem between the data distribution and a simple prior, usually a Gaussian distribution $\mathcal{N}(0, I)$. Instead of learning to predict noise or score functions, flow matching learns a velocity field that deterministically transports samples along straight-line paths in latent space.

Specifically, given a data sample z_0 and a noise sample $z_1 \sim \mathcal{N}(0, I)$, the noise-to-data path is defined via a linear interpolation:

$$z_t = (1 - t)z_0 + tz_1. \tag{2}$$

A neural network $v_\theta(z_t, t)$ is optimized to predict its corresponding velocity field:

$$\frac{d}{dt}z_t = z_1 - z_0, \tag{3}$$

Therefore, the training objective is to minimize

$$\mathbb{E}_{z_0, z_1, t}\left[\|v_\theta(z_t, t) - (z_1 - z_0)\|_2^2\right]. \tag{4}$$

At inference time, data samples are generated through sampling z_1 and solving the flow ODE: $\frac{d}{dt}z_t = v_\theta(z_t, t)$. Flow matching yields faster convergence and more stable training compared to score-based diffusion models.

Almost all modern diffusion and flow matching models are built on top of the Diffusion Transformer (DiT) architecture [60]. The computational efficiency of such generative models is therefore quadratically related to the number of tokens processed due to the expensive attention [80] and MLP operations in the DiT. This motivates our method, which generates images and videos using a reduced set of tokens where the low-resolution tokens are specified by spatial or spatiotemporal foveation masks, while preserving perceptual image quality.

3.2 Foveated Diffusion

To achieve true computational savings in foveated visual generation, we introduce *Foveated Diffusion*, a principled training and generation framework that enables diffusion or flow-matching models to directly generate spatially foveated images and videos with reduced token complexity. We describe our pipeline using latent-space image generation models here, but this concept applies equally to video generation models, as can be seen in Sec. 4.

Foveated Tokenization. Let the latent space of a high-resolution image be $\mathbb{R}^{c\times(h\cdot w)}$. The VAE encodes and patchifies each image $x \in \mathbb{R}^{3\times H\times W}$ into a sequence of $h \times w$ tokens with feature dimension c. Standard DiTs perform training and generation directly on this full set of $h \cdot w$ tokens. In Foveated Diffusion, we are given a foveation mask $M \in \{0, 1\}^{h\times w}$ that specifies the spatial locations where high-resolution tokens are retained; meanwhile, peripheral regions are represented with fewer tokens to reduce the sequence length and computational complexity. Consequently, we operate entirely in the *foveated token space*

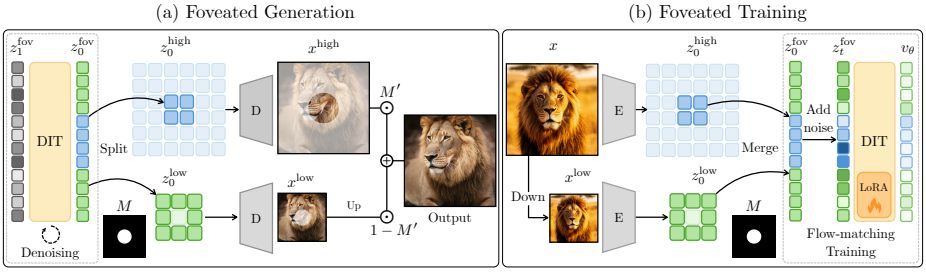

Fig. 2: The Foveated Diffusion Pipeline. In Foveated Generation (a), we iteratively denoise a foveated token sequence of reduced length instead of the full high-resolution sequence. The resulting tokens z_0^{fov} are split into high- and low-resolution grids, decoded by the VAE, and blended using a user-specified foveation mask. We employ Foveated Training (b) to adapt pretrained DiTs to foveated token sequences using low-rank adaptation (LoRA) [31]. The image and its downsampled version are independently encoded by the VAE encoder and merged into a clean foveated token sequence for flow-matching training.

$\mathbb{R}^{c\times L}$ where the token sequence has a variable length $L \ll h \cdot w$. In our setting, a single low-resolution token represents the spatial area of a 2×2 block of high-resolution tokens. This results in a total sequence length of $L = m+(h\cdot w-m)/4$, where m is the number of effective tokens in the mask M. This approach directly parallels foveated rendering in traditional graphics (see Sec. 3.1), where shading and rasterization are computed asymmetrically based on a user-specified mask to achieve computational savings.

Foveated Generation. Foveated generation is performed by sampling Gaussian noise $z_1^{\text{fov}} \sim \mathcal{N}(0, I)$ in the foveated token space $\mathbb{R}^{c\times L}$ at a reduced sequence length L. The foveated token sequence is then iteratively denoised from $t = 1$ to $t = 0$, producing a clean foveated token sequence $z_0^{\text{fov}} \in \mathbb{R}^{c\times L}$. This procedure can be done in a completely training-free setting using a pretrained generative model, which we refer to as *Naïve Mixed-Resolution Denoising*.

To obtain a full-resolution image, we first partition the clean foveated token sequence z_0^{fov} into high-resolution and low-resolution components:

$$(z_0^{\text{high}}, z_0^{\text{low}}) = \text{Split}(z_0^{\text{fov}}, M). \tag{5}$$

We then decode each subset separately with the VAE decoder D:

$$x^{\text{high}} = D(z_0^{\text{high}}), \qquad x^{\text{low}} = D(z_0^{\text{low}}). \tag{6}$$

The decoded low-resolution image is spatially upsampled to the original spatial resolution and blended with the high-resolution decoding to form the final image using the upsampled latent foveation mask $M' = \text{Up}(M) \in \mathbb{R}^{H\times W}$:

$$x^{\text{fov}} = M' \odot x^{\text{high}} + (1 - M') \odot \text{Up}(x^{\text{low}}), \tag{7}$$

where $\text{Up}(\cdot)$ denotes spatial upsampling and $\odot$ denotes elementwise multiplication. The full generation pipeline is illustrated in Fig. 2-(a).

Mixed-Resolution RoPE. Standard Rotary Positional Embedding (RoPE) [75] typically assumes a uniform grid with fixed-phase spacing (Fig. 3(a)). However, because Foveated Diffusion introduces mixed-resolution tokenization, we must modify the RoPE indexing accordingly. Therefore, we follow Wu et al. [88] and align the key RoPE phases with query RoPE phases based on their corresponding token resolutions. Specifically, when computing attention with high-resolution query tokens, we sub-sample low-resolution key tokens from the full-resolution tokens. For attention with low-resolution query tokens, we subsample high-resolution key tokens and normalize their RoPE indices to the low-resolution grid. Please see illustration in Fig. 3(b) or refer to Wu et al. [88] for more details.

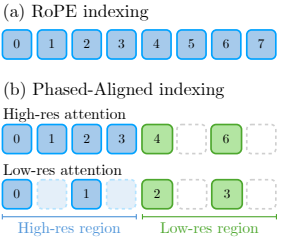

Fig. 3: Adapting RoPE for mixed-resolution attention [88].

Failure of Naïve Mixed-Resolution Denoising. A pre-trained DiT is not, by default, compatible with a mixed-resolution, or foveated, token layout since tokens and positional embeddings are defined to be on a uniform grid. Even after adapting RoPE to handle mixed-resolution tokens, the low- and high-resolution regions still exhibit significant scale and structural inconsistencies, frequently resulting in duplicated objects or multiple entities fused together unnaturally (Fig. 4, and Fig. 5). These findings suggest that high-quality foveated generation cannot be achieved with training-free mixed-resolution denoising, and a more principled approach is required to achieve our objective.

Fig. 4: Failure of Naïve mixed-resolution denoising.

Foveated Training. To address the aforementioned failure, we design an effective post-training procedure in the foveated token space by constructing foveated training targets that are mixed-resolution-consistent by design, as shown in Fig. 2-(b). Given a high-resolution training image x, we first form two latent token sequences using the VAE encoder E. The high resolution token sequence is:

$$z_0^{\text{high}} = E(x) \in \mathbb{R}^{c\times(h\cdot w)} \tag{8}$$

The low-resolution token sequence is obtained by bicubically downsampling the image and encoding it:

$$z_0^{\text{low}} = E(\text{Down}(x)) \in \mathbb{R}^{c\times(\frac{h}{2}\cdot\frac{w}{2})} \tag{9}$$

We then construct a clean foveated target token sequence by merging tokens from z_0^{high} and z_0^{low} according to the foveation mask:

$$z_0^{\text{fov}} = \text{Merge}(z_0^{\text{high}}, z_0^{\text{low}}, M) \in \mathbb{R}^{c\times L}. \tag{10}$$

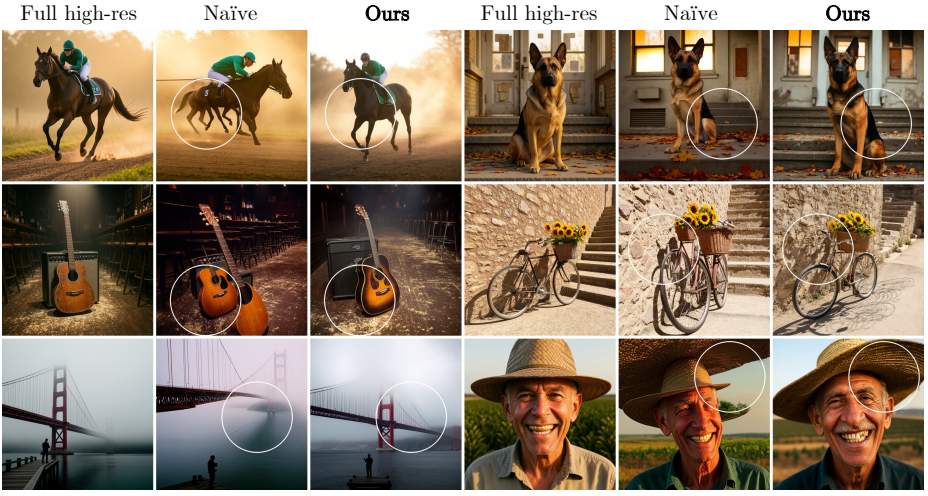

Fig. 5: Qualitative comparison for image generation. Our method yields perceptually indistinguishable results from full high-resolution synthesis, whereas the naïve baseline introduces scale inconsistencies and structural artifacts across mixed-resolution regions. The high-resolution regions (fovea) are delineated with white borders.

By construction, both z_0^{high} and z_0^{low} are derived from the same underlying image content, and thus z_0^{fov} defines a single coherent target token sequence for mixed-resolution denoising. This clean foveated token sequence z_0^{fov} exactly corresponds to the foveated image x^{fov} through Equations 5 to 7.

We train our Foveated Diffusion model using the standard flow matching objective [55]. Concretely, for a sampled timestep t and noise $z_1^{\text{fov}} \sim \mathcal{N}(0, I)$, we generate a noisy foveated token sequence z_t^{fov} from z_0^{fov} following the flow-matching parameterization (Eq. 2) and optimize the model to predict the corresponding target velocity (Eq. 4). Most importantly, all computations are performed on the variable-length L foveated token sequence.

Foveation Masking Strategies. Our Foveated Training procedure is agnostic to how the foveation masks M are designed, as it is simply a user-specified binary mask indicating the locations of high-resolution tokens. The form of the masks for training is therefore flexible and completely task-dependent. Specifically, we present two variants in the main paper: *randomized masks*, which produce a generative model agnostic to the specified foveation, allowing the gaze center to be shifted to arbitrary image regions; and *saliency-guided masks*, which encourage the foveal region to encompass salient objects in the scene, reflecting the regions where viewer attention is most likely to be directed. We additionally show *bounding-box masks* results in the supplementary materials.

Notably, training with different masking strategies does not require any modification to the model architecture or to the training objective; it only involves changing the foveation masks.

4 Experiments

4.1 Implementation Details

For image and video generation, we adopt pretrained Diffusion Transformers (DiTs) as base models and fine-tune them using our Foveated Diffusion framework. For image generation, we fine-tune FLUX.2 Klein 4B [49] on the Aesthetic-Train-V2 dataset [96,97]. We randomly sample 90k images for training and reserve 10k prompt–image pairs for evaluation. For video generation, we fine-tune Wan2.1 1.3B [81] on Vchitect-T2V-Dataverse [19,71], excluding 200 prompts to serve as test samples. During training, we randomly sample a circular foveation mask for each image or a random foveation path for video to simulate diverse foveation patterns. For a fair comparison, the full-resolution and naïve mixed-resolution baselines use the same base models and training data, but without Foveated Diffusion training. All models are fine-tuned using Low-Rank Adaptation (LoRA) [31] with rank 32. Experiments are conducted on NVIDIA H100 GPUs. Please refer to the supplementary materials for more details.

4.2 Results

Runtime Comparison. Foveated Diffusion greatly reduces computational complexity via foveated tokenization, significantly accelerating visual generation. This computational efficiency is determined by the foveation mask (see Sec. 3.2). By expanding the low-resolution periphery, we drastically reduce the effective sequence length to L tokens, compared to $h{\cdot}w$ for images and $f{\cdot}h{\cdot}w$ for videos. We define Token Ratio as the proportion of the reduced sequence relative to the full sequence and measure the resulting computational savings compared to full high-resolution generation as Speedup. As shown in Fig. 6, using 25% of the tokens yields remarkably over 2× and over 4× speedup for image and video generation, respectively. Video models achieve higher speedup due to the higher computational cost of spatiotemporal (3D) attention operations.

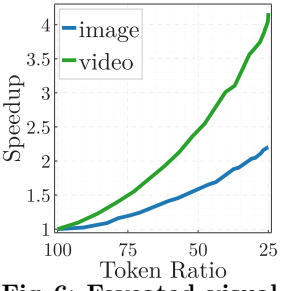

Fig. 6: Foveated visual generation speedup.

Image Generation. For foveated image generation, we report standard generative metrics including FID [28], Precision [48], and a human preference metric HPSv2.1 [90], and measure prompt alignment using CLIP score [63]. For the naïve mixed-resolution baseline and our Foveated Diffusion pipeline, we fix the foveation mask to be a centered rectangular mask with varying token ratios.

As shown in Tab. 1 and Fig. 5, Foveated Diffusion consistently achieves substantially better performance than the naïve mixed-resolution baseline across all foveation sizes and across all metrics aside from FID, while maintaining performance similar to full high-resolution generation. Importantly, our method significantly surpasses the naïve baseline on a human preference–aligned metric

Method	Token Ratio	HPSv2.1 ↑	FID ↓	Precision ↑	CLIP ↑	Runtime ↓ (Speedup ↑)
Full high-res	100%	0.280	11.38	0.792	0.292	10.45s
Naïve mixed-res	43%	0.268	**10.99**	0.769	0.292	6.53s (1.61×)
Ours		**0.279**	11.38	**0.777**	**0.294**	
Naïve mixed-res	30%	0.270	**11.70**	0.762	0.292	5.27s (1.98×)
Ours		**0.279**	11.91	**0.789**	**0.293**	
Naïve mixed-res	26%	0.275	12.83	0.775	0.292	5.02s (2.08×)
Ours		**0.280**	**12.62**	**0.792**	**0.293**	

Table 1: Quantitative comparison for image generation. We compare against the naïve mixed-resolution baseline across various token count ratios, highlighting the best result for each. Foveated Diffusion surpasses the baseline, excluding FID which we find less reliable for our task. Our method matches the image quality of full high-resolution generation (top row) while achieving up to a 2× speedup.

HPSv2.1 [90], reinforcing the perceptual, human-centered focus of our approach. We observe that FID may not be a reliable metric for evaluating foveated visual generation, as the naïve baseline, despite exhibiting clear structural artifacts (see Fig. 5), even outperforms full high-resolution generation. As the foveation size decreases, the peripheral low-resolution area increases, leading to a significant reduction in generation time.

Perceptual User Study. Standard generative metrics weigh all pixels uniformly and ignore eccentricity-dependent visual acuity, leading to trends that conflict with human preference (e.g., the HPSv2.1–FID discrepancy in Tab. 1). This is fundamentally misaligned with the perceptually-driven and gaze-contingent motivation of Foveated Diffusion. We therefore perform a Two-Alternative Forced Choice (2AFC) user study under a pseudo-eye-tracked protocol. Participants fixate on a red point before a pair of images randomly drawn from two of the three methods (our method, full high-resolution generation, and the naïve mixed-resolution baseline) are sequentially displayed for 1 second to discourage eye movements. Participants then select the image with higher overall visual quality, avoiding bias from actively searching for distortions.

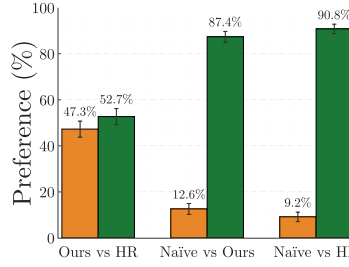

Fig. 7: User study results.

In Fig. 7, we show that our method achieves near perceptual parity with full high-resolution generation and is strongly preferred over the naïve baseline. These results confirm that Foveated Diffusion preserves perceptual quality under gaze-contingent viewing while substantially reducing latency (user study images are generated with a 1.85× speedup), establishing its practicality for real-time, wide-field-of-view applications such as gaming and immersive video. We include a detailed description of our user study design, procedure, analysis, and results in the supplementary materials.

Method	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Image Quality ↑
Full high-res	0.9407	0.9363	0.9943	0.263	0.5434	0.653
Naïve mixed-res	0.9072	0.9239	0.9899	**0.465**	0.4795	0.522
Ours	**0.9446**	**0.9393**	**0.9946**	0.265	**0.5432**	**0.587**

Table 2: Quantitative comparison for video generation (VBench). Foveated Diffusion outperforms the naïve mixed-resolution baseline across key metrics, achieving performance comparable to full-resolution generation. Notably, our framework maintains high subject consistency while providing a 3.5× speedup.

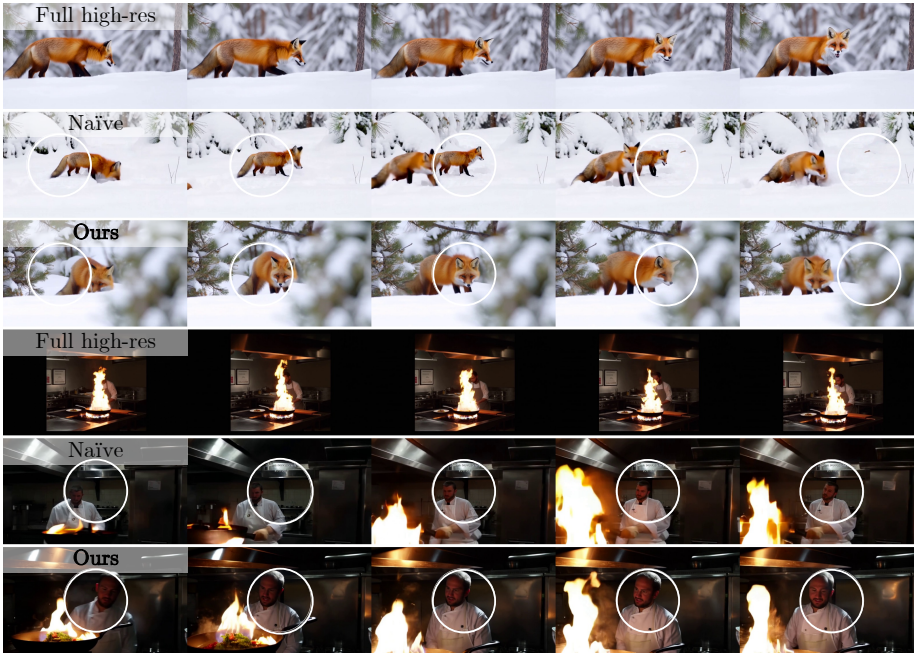

Fig. 8: Qualitative comparison for video generation. Foveated Diffusion outperforms the naïve mixed-resolution baseline in video generation, which exhibits scale mismatches or duplicate entities near the low–high resolution boundary (white outline).

Video Generation. For foveated video generation, we report the standard generative video metric VBench [32]. Similar to our image generation experiments, we fix the foveation mask across all frames as a centered circular mask with a token ratio of 38% relative to the original token length of $(f \cdot h \cdot w)$. Foveated Diffusion surpasses the naïve mixed-resolution baseline while achieving results comparable to full high-resolution generation with a 3.5× speedup, as clearly shown in Tab. 2. This parity across quality and consistency metrics highlights our framework's ability to maintain a coherent global structure. Fig. 8 provides supporting qualitative results, where the naïve baseline exhibits severe structural and scale mismatches.

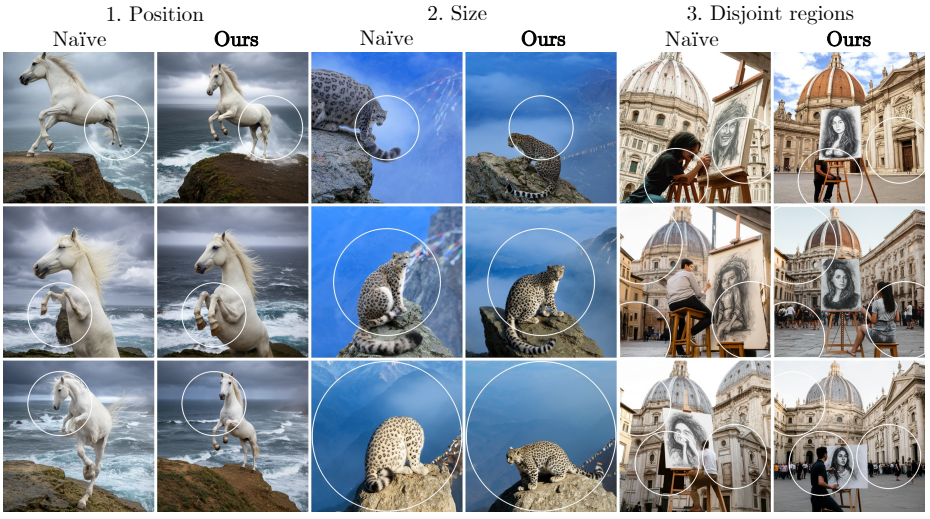

Fig. 9: Image generation with different foveation patterns. We generate images using the same prompt and noise seed while varying the foveation pattern in shape, position, and size. High-resolution regions are delineated with white borders. Our method maintains content consistency across resolution regions, while the naïve mixed-resolution baseline exhibits inconsistencies of scale and structure.

Foveation Mask Strategies. As discussed in Sec. 3.2, the strategy used to construct foveation masks during training significantly affects the behavior of the resulting foveated generative model. In Fig. 9, we present foveated image generation results using various masks while keeping the text prompt and noise seed constant. We vary the masks in size, position, and shape, including non-contiguous masks with multiple disjoint high-resolution regions. Foveated Diffusion generates coherent content under unseen foveation masks at inference, which is uniquely enabled by our *randomized mask* training protocol.

Furthermore, Foveated Diffusion shows great potential for *saliency-guided* generation. Specifically, we construct foveation masks by binarizing image and video saliency maps predicted by DeepGaze [54]. As evident in Fig. 10, we observe that salient objects align with specified foveal regions, demonstrating the generalization of our Foveated Diffusion framework beyond random mask placements. This is potentially useful for generative VR gaming or generative robotics simulation scenarios where only salient objects in view are required to be rendered in high resolution [41].

5 Conclusion

In this work, we introduce Foveated Diffusion, a perceptually motivated framework for efficient visual generation. By leveraging the eccentricity-dependent nature of the human visual system, we achieve significant computational savings while maintaining the perceived quality of the generated content.

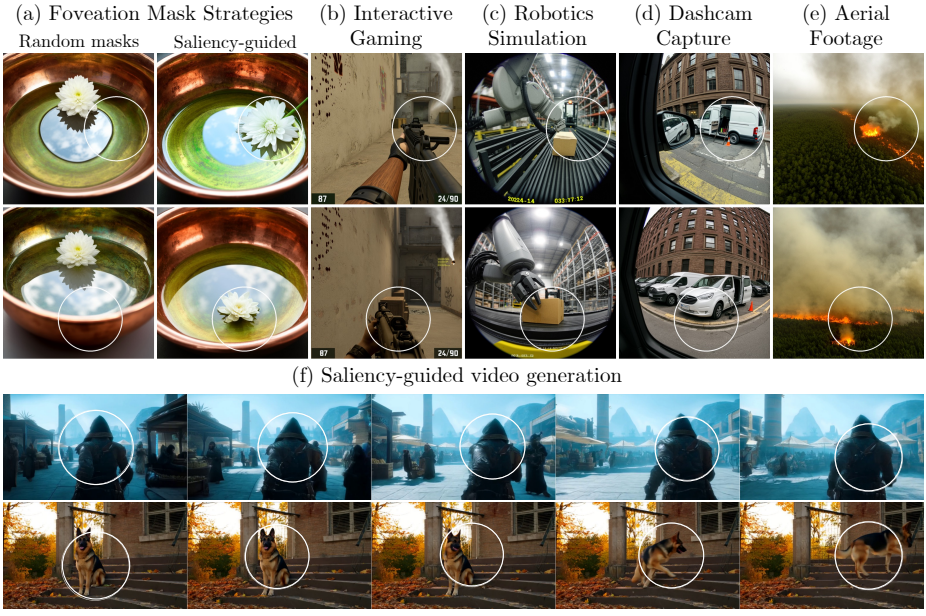

Fig. 10: Towards saliency-guided visual generation. We show that the *saliency-guided* Foveated Diffusion models enable saliency-guided image (a-e) and video generation (f), where salient objects align with the fovea. The *randomized masks* model do not exhibit such behavior (a). This is potentially useful for generative simulation applications such as VR gaming or robotics simulations (b-e), where only the salient objects have to be generated at the highest resolution, i.e. the machine gun in (b), the robotics arm and box in (c), and the dog in (f).

Our method yields promising results for foveated generation, generating coherent content across low- and high-resolution regions. Nevertheless, we observe occasional color artifacts near the foveation boundary (see the supplementary materials). This is primarily due to the blending of the VAE-decoded low- and high-resolution content (see Eq. 6 and 7). A promising direction for future work is redesigning the VAE to directly encode and decode mixed-resolution tokens. Additionally, we present two levels of foveation with a spatial reduction factor of 2×2, whereas traditional foveated rendering can employ even coarser peripheral resolutions. Our general framework naturally extends to multiple levels of foveation, and such an extension calls for the corresponding multi-level adaptation of phase-aligned RoPE [88]. Finally, we believe that our method is most impactful when deployed on a streaming autoregressive video generation system equipped with an eye tracker. While we are the first to establish the foundations of such a system, integrating Foveated Diffusion into a real-time video world model remains a compelling direction for future work.

In conclusion, Foveated Diffusion offers a new paradigm and opens a new avenue for scaling generative models: aligning model computation with human visual perception, complementary to advances in hardware and algorithmic efficiency.

Acknowledgments

We thank Ryan Po, Hansheng Chen, and Tong Wu for fruitful discussions. Brian Chao and Howard Xiao are supported by Stanford Graduate Fellowships (SGF). Brian Chao is also supported by the NSF Graduate Research Fellowship Program (GRFP). Compute resources were provided by the Marlowe cluster at Stanford University [37].

References

1. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024)
2. Anstis, S.M.: A chart demonstrating variations in acuity with retinal eccentricity. Vision Research **14**(7), 589–592 (1974)
3. Bandera, C., Scott, P.D.: Foveal machine vision systems. In: Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics. pp. 596–599. IEEE (1989)
4. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
5. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: ICLR (2023)
6. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: CVPR. pp. 4599–4603 (2023)
7. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
8. Che, H., He, X., Liu, Q., Jin, C., Chen, H.: Gamegen-x: Interactive open-world game video generation. arXiv preprint arXiv:2411.00769 (2024)
9. Chen, Y., Qiu, Y., Li, R., Agha, A., Omidshafiei, S., Patrikar, J., Scherer, S.: Come: Confidence-guided token merging for visual geometric transformers (2025), https://arxiv.org/abs/2511.14751
10. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. In: International Conference on Learning Representations (ICLR) (2021)
11. Chuang, I., Lee, A., Gao, D., Zou, J., Soltani, I.: Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers (2025), https://arxiv.org/abs/2507.15833
12. Curcio, C.A., Allen, K.A.: Topography of ganglion cells in human retina. Journal of comparative Neurology **300**(1), 5–25 (1990)
13. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
14. Decart, E., McIntyre, Q., Campbell, S., Chen, X., Wachen, R.: Oasis: A universe in a transformer. URL: https://oasis-model. github. io **2**(3), 6 (2024)

15. Deng, N., He, Z., Ye, J., Duinkharjav, B., Chakravarthula, P., Yang, X., Sun, Q.: Fov-nerf: Foveated neural radiance fields for virtual reality. IEEE TVCG **28**(11), 3854–3864 (2022)
16. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
17. Diederik, P.K., Max, W.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (Nov 2019). https://doi.org/10.1561/2200000056, http://dx.doi.org/10.1561/2200000056
18. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
19. Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al.: Vchitect-2.0: Parallel transformer for scaling up video diffusion models. arXiv preprint arXiv:2501.08453 (2025)
20. Fang, H., Tang, S., Cao, J., Zhang, E., Tang, F., Lee, T.Y.: Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In: CVPR. pp. 18083–18092 (2025)
21. Feng, R., Zhang, H., Yang, Z., Xiao, J., Shu, Z., Liu, Z., Zheng, A., Huang, Y., Liu, Y., Zhang, H.: The matrix: Infinite-horizon world generation with real-time moving control. arXiv preprint arXiv:2412.03568 (2024)
22. Franke, L., Fink, L., Stamminger, M.: Vr-splatting: Foveated radiance field rendering via 3d gaussian splatting and neural points. Proceedings of the ACM on Computer Graphics and Interactive Techniques **8**(1), 1–21 (2025)
23. Geisler, W.S., Perry, J.S.: Real-time foveated multiresolution system for low-bandwidth video communication. In: Human vision and electronic imaging III. vol. 3299, pp. 294–305. SPIE (1998)
24. Guenter, B., Finch, M., Drucker, S., Tan, D., Snyder, J.: Foveated 3d graphics. ACM TOG **31**(6), 1–10 (2012)
25. Haurum, J.B., Escalera, S., Taylor, G.W., Moeslund, T.B.: Agglomerative token clustering. In: ECCV. pp. 200–218. Springer (2024)
26. Havtorn, J.D., Royer, A., Blankevoort, T., Bejnordi, B.E.: Msvit: Dynamic mixed-scale tokenization for vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 838–848 (2023)
27. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
28. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
29. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
31. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022), https://arxiv.org/abs/2106.05615
32. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench:

Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
33. Jeong, W., Lee, K., Seo, H., Chun, S.Y.: Upsample what matters: Region-adaptive latent sampling for accelerated diffusion transformers. arXiv preprint arXiv:2507.08422 (2025)
34. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
35. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics yolo (Jan 2023), https://ultralytics.com, software release, January 10, 2023
36. Jonnalagadda, A., Wang, W.Y., Manjunath, B., Eckstein, M.P.: Foveater: Foveated transformer for image classification. arXiv preprint arXiv:2105.14173 (2021)
37. Kapfer, C., Stine, K., Narasimhan, B., Mentzel, C., Candès, E.: Marlowe: Stanford's GPU-based computational instrument (2025). https://doi.org/10.5281/zenodo.14751899, https://doi.org/10.5281/zenodo.14751899
38. Kaplanyan, A.S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., Rufo, G.: Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. ACM TOG **38**(6), 1–13 (2019)
39. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning (ICML). pp. 5156–5165. PMLR (2020)
40. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
41. Kerr, J., Hari, K., Weber, E., Kim, C.M., Yi, B., Bonnen, T., Goldberg, K., Kanazawa, A.: Eye, robot: Learning to look to act with a bc-rl perception-action loop. In: 9th Annual Conference on Robot Learning (2025), http://arxiv.org/abs/2506.10968
42. Killick, G., Henderson, P., Siebert, P., Aragon-Camarasa, G.: Foveation in the era of deep learning. arXiv preprint arXiv:2312.01450 (2023)
43. Kim, M., Gao, S., Hsu, Y.C., Shen, Y., Jin, H.: Token fusion: Bridging the gap between token pruning and token merging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1383–1392 (2024)
44. Kodaira, A., Hou, T., Hou, J., Tomizuka, M., Zhao, Y.: Genie 3: A new frontier for world models (2025), https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, google DeepMind Technical Report
45. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
46. Krajancich, B., Kellnhofer, P., Wetzstein, G.: Towards attention–aware foveated rendering. ACM TOG **42**(4), 1–10 (2023)
47. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, H., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th Symposium on Operating Systems Principles. pp. 611–626 (2023)
48. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019)
49. Labs, B.F.: Flux.2 klein 4b (2025), https://blackforestlabs.ai, available at https://huggingface.co/black-forest-labs

50. Lee, M.J., Kim, H.D., Lee, S.W.: Local representative token guided merging for text-to-image generation. arXiv preprint arXiv:2507.12771 (2025)
51. Lee, S.H., Wang, J., Zhang, Z., Fan, D., Li, X.: Video token merging for long video understanding. NeurIPS **37**, 13851–13871 (2024)
52. Levoy, M., Whitaker, R.: Gaze-directed volume rendering. In: Proceedings of the 1990 symposium on interactive 3d graphics. pp. 217–223 (1990)
53. Li, X., Li, M., Cai, T., Xi, H., Yang, S., Lin, Y., Zhang, L., Yang, S., Hu, J., Peng, K., et al.: Radial attention: $\mathcal{O}(n \log n)$ sparse attention with energy decay for long video generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
54. Linardos, A., Kümmerer, M., Press, O., Bethge, M.: 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. pp. 12919–12928 (2021)
55. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
56. Lu, W., Zheng, S., Xia, Y., Wang, S.: ToMA: Token merge with attention for diffusion models. In: ICML (2025), https://openreview.net/forum?id=51l8tvuIxo
57. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
58. Minut, S., Mahadevan, S., Henderson, J.M., Dyer, F.C.: Face recognition using foveal vision. In: International Workshop on Biologically Motivated Computer Vision. pp. 424–433. Springer (2000)
59. Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., Lefohn, A.: Towards foveated rendering for gaze-tracked virtual reality. ACM TOG **35**(6), 1–12 (2016)
60. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
61. Peirce, J.W.: Psychopy—psychophysics software in python. Journal of neuroscience methods **162**(1-2), 8–13 (2007)
62. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8733–8744 (2025)
63. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
64. Reddy, M.: Perceptually optimized 3d graphics. IEEE computer Graphics and Applications **21**(5), 68–75 (2002)
65. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
66. Ronen, T., Levy, O., Golbert, A.: Vision transformers with mixed-resolution tokenization. In: CVPR. pp. 4613–4622 (2023)
67. Rovamo, J., Virsu, V.: An estimation and application of the human cortical magnification factor. Experimental brain research **37**(3), 495–510 (1979)
68. Schmidt, T., Newcombe, R.: Segment this thing: Foveated tokenization for efficient point-prompted segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29428–29437 (2025)

69. Shazeer, N.: Fast transformer decoding: One write-head is all you need. arXiv preprint arXiv:1911.02150 (2019)
70. Shi, X., Wang, L., Liu, X., Wu, J., Shao, Z.: Scene-aware foveated neural radiance fields. IEEE TVCG (2024)
71. Si, C., Fan, W., Lv, Z., Huang, Z., Qiao, Y., Liu, Z.: Repvideo: Rethinking cross-layer representation for video generation. arXiv 2501.08994 (2025)
72. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
73. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
74. Stengel, M., Grogorick, S., Eisemann, M., Magnor, M.: Adaptive image-space sampling for gaze-contingent real-time rendering. In: Comput. Graph. Forum. vol. 35, pp. 129–139. Wiley Online Library (2016)
75. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
76. Sun, Q., Huang, F.C., Kim, J., Wei, L.Y., Luebke, D., Kaufman, A.: Perceptually-guided foveation for light field displays. ACM TOG **36**(6), 1–13 (2017)
77. Tariq, T., Tursun, C., Didyk, P.: Noise-based enhancement for foveated rendering. ACM TOG **41**(4), 1–14 (2022)
78. Tian, Y., Xia, X., Ren, Y., Lin, S., Wang, X., Xiao, X., Tong, Y., Yang, L., Cui, B.: Training-free diffusion acceleration with bottleneck sampling. arXiv preprint arXiv:2503.18940 (2025)
79. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
80. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
81. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
82. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
83. Watson, A.B.: A formula for human retinal ganglion cell receptive field density as a function of visual field location. Journal of vision **14**(7), 15–15 (2014)
84. Weier, M., Roth, T., Kruijff, E., Hinkenjann, A., Pérard-Gayot, A., Slusallek, P., Li, Y.: Foveated real-time ray tracing for head-mounted displays. In: Comput. Graph. Forum. vol. 35, pp. 289–298. Wiley Online Library (2016)
85. Weng, W., Feng, R., Wang, Y., Dai, Q., Wang, C., Yin, D., Zhao, Z., Qiu, K., Bao, J., Yuan, Y., et al.: Art-v: Auto-regressive text-to-video generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7395–7405 (2024)
86. Weymouth, F.W.: Visual sensory units and the minimal angle of resolution. American Journal of Ophthalmology **46**(1), 102–113 (1958)
87. Wu, H., Xu, J., Le, H., Samaras, D.: Importance-based token merging for efficient image and video generation. In: ICCV. pp. 4983–4995 (2025)
88. Wu, H., Xu, J., Miao, Q., Samaras, D., Le, H.: One attention, one scale: Phase-aligned rotary positional embeddings for mixed-resolution diffusion transformer. arXiv preprint arXiv:2511.19778 (2025)
89. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory (2025), https://arxiv.org/abs/2506.05284

90. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
91. Xi, H., Yang, S., Zhao, Y., Xu, C., et al.: Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. In: International Conference on Machine Learning (ICML) (2025)
92. Xia, Y., Ling, S., Fu, F., Wang, Y., Li, H., Xiao, X., Cui, B.: Training-free and adaptive sparse attention for efficient long video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
93. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
94. Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., Liu, X.: Gamefactory: Creating new games with generative interactive videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11590–11599 (2025)
95. Zhan, C., Li, W., Shen, C., Zhang, J., Wu, S., Zhang, H.: Bidirectional sparse attention for faster video diffusion training. arXiv preprint arXiv:2509.01085 (2025)
96. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
97. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Ultra-high-resolution image synthesis: Data, method and evaluation (2025), arXiv:2506.01331
98. Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., Chen, J.: Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In: International Conference on Machine Learning (ICML) (2025)
99. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. arXiv preprint arXiv:2504.12626 (2025)
100. Zhang, P., Chen, Y., Huang, H., Lin, W., Liu, Z., Stoica, I., Xing, E., Zhang, H.: Vsa: Faster video diffusion with trainable sparse attention. arXiv preprint arXiv:2505.13389 (2025)
101. Zhang, P., Chen, Y., Su, R., Ding, H., Stoica, I., Liu, Z., Zhang, H.: Fast video generation with sliding tile attention. arXiv preprint arXiv:2502.04507 (2025)

Supplementary Material

Foveated Diffusion: Efficient Spatially Adaptive Image and Video Generation

Table of Contents

A Additional Implementation Details

A.1 Image Generation

For our foveated image generation experiments, we fine-tune the FLUX 2.1 Klein 4B model [49] using the Aesthetic-Train-V2 dataset [96, 97]. We adopt a Low-Rank Adaptation (LoRA) [31] approach with a rank of 32, training for 10,000 steps. The optimization was conducted on a cluster of eight NVIDIA H100 GPUs with an effective batch size of 8. Our implementation utilizes the DiffSynth-Studio codebase [1] and follows its default hyperparameter settings for LoRA training.

For quantitative results, we generate 10K images, one for each prompt in the reserved test set from the Aesthetic-Train-V2 dataset [96, 97]. We adopt the standard evaluation protocol in [16] and report standard generative metrics including HPSv2.1 [90], FID [28], Precision [48], and CLIP score [63] in Table 1 of the main paper. FID and Precision are computed against real images in the evaluation set, reflecting the data alignment between generated and real images. The CLIP and HPSv2.1 scores are averaged across all generated images, where the CLIP score measures prompt alignment and the HPSv2.1 score captures human preference.

For all image generation experiments, we generate images at 1024×1024 resolution for all methods.

A.2 Video Generation

For our foveated video generation experiments, we fine-tune the Wan2.1 1.3B model [81] using the Vchitect-T2V-Dataverse dataset [19, 71]. We adopt a Low-Rank Adaptation (LoRA) [31] approach with a rank of 32, training for 10,000 steps. The optimization was conducted on a cluster of eight NVIDIA H100 GPUs with an effective batch size of 8. Our implementation utilizes the DiffSynth-Studio codebase and follows its default hyperparameter settings for LoRA training.

For quantitative evaluations, we generate 200 videos using the held-out test prompts and report the standard VBench [32] metrics. We evaluate our approach and all baselines at a consistent 480p resolution.

For qualitative results, we provide samples at the original 480p training resolution and additionally demonstrate generalization to 720p.

[1] https://github.com/modelscope/DiffSynth-Studio/tree/main

B Additional User Study Details

B.1 Study Design and Participants

Study design. We evaluate the perceptual quality of Foveated Diffusion against both full high-resolution generation and the naïve mixed-resolution baseline using a Two-Alternative Forced Choice (2AFC) paradigm, a standard protocol for preference-based perceptual evaluation [46]. In each trial, participants are shown a pair of images sequentially and are asked to select the one they judge to have higher overall visual quality. Forced choice eliminates neutral or indecisive responses and yields a clean preference rate $P \in [0, 1]$, where the null hypothesis of perceptual equivalence corresponds to $P = 0.5$.

Since Foveated Diffusion targets gaze-contingent generation as a primary use case, the study employs a pseudo–eye-tracking protocol. Rather than using physical eye-tracking hardware, each trial begins with a red fixation dot displayed on a black screen at the position corresponding to the center of the foveal region for that trial. Participants fixate on this dot before each full-screen image is shown, ensuring that their gaze is directed toward the foveal center. For foveated and naïve mixed-resolution images, the dot is placed precisely at the center of the high-resolution region. Although full high-resolution images have uniform resolution across the entire image, participants fixate at the same location for consistency. Detailed trial procedures are described in Sec. B.3.

Study participants. A total of 11 participants (8 male and 3 female; ages 21–32) took part in the study. All participants reported normal or corrected-to-normal vision, no history of visual deficiencies, and no color blindness. All participants provided informed consent.

B.2 Study Setup and Images

Due to the need to display high-resolution content, we used a Sceptre X405BV-FSR LED monitor (40-inch, 16:9 aspect ratio) at a native resolution of 1920 × 1080 (Full HD), a 60 Hz refresh rate, and a peak luminance of 250 cd/m^2 for image display. All test images were square and centered on the display to preserve their aspect ratio. The experiment was implemented in Python using the PsychoPy package [61], and images were streamed to the display via a wired HDMI connection.

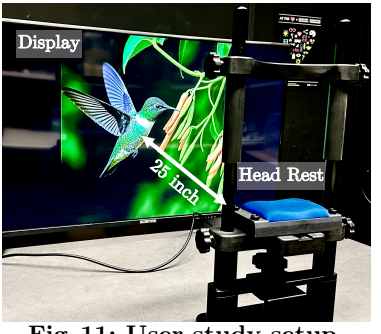

Fig. 11: User study setup.

Participants were seated at a fixed viewing distance of 25 inches (approximately 63.5 cm), maintained by a headrest that also controlled viewing height (Fig. 11). At this distance, the display subtends approximately 24 pixels per degree (ppd) of visual angle, so each full-screen image spans roughly 45° × 45° of the visual field.

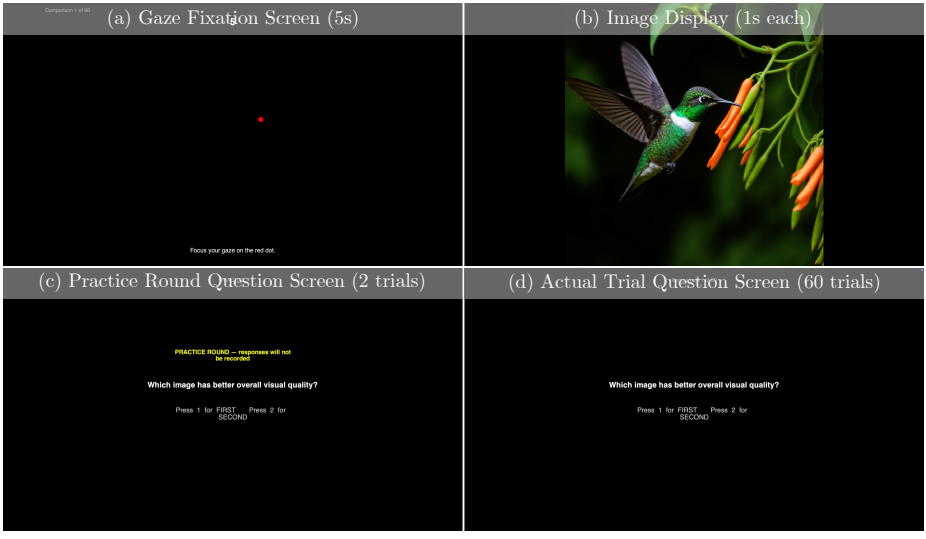

Fig. 12: User study interface. Each trial began with a red gaze fixation dot displayed for 5 seconds (**a**), followed by a pair of images, each displayed for 1 s (**b**). Participants answered visual quality questions after both images were shown. Participants completed two practice trials (**c**) before beginning 60 data collection trials (**d**).

The foveal region diameter was set to one-third of the image width, subtending approximately 15° of visual angle at the prescribed viewing distance. This boundary was chosen based on the known eccentricity-dependent decline in human visual acuity [23], placing the peripheral region beyond the zone of high acuity. Images were drawn from 40 unique prompts in the same test set used for the quantitative evaluations in Table 1 of the main paper, spanning diverse scenes and objects. For each prompt, foveated and naïve mixed-resolution images were generated with matching foveal region locations, which were randomized across trials (see main paper Sec. 3), ensuring that participants could not anticipate the foveal center from one trial to the next.

B.3 Procedure

Each participant completed 60 trials, divided equally across three pairwise comparison conditions: Foveated Diffusion vs. full high-resolution generation, Foveated Diffusion vs. the naïve mixed-resolution baseline, and full high-resolution generation vs. the naïve mixed-resolution baseline (20 trials each). Within each condition, the two presentation orders were counterbalanced equally (10 trials per order). For each trial, the test image was randomly selected from the 40 available, with the comparison condition and presentation order assigned according to the counterbalanced scheme.

Each trial began with a five-second red fixation dot at the center of the foveal region, orienting participants' gaze before the first image was shown. The first

Pair	Preference P (%)	p-value ($H_0 : P = 0.5$)
Ours vs. Full high-res	47.3	0.4829
Ours vs. Naïve mixed-res	**87.4**	< 0.0001
Full high-res vs. Naïve mixed-res	90.8	< 0.0001

Table 3: User study statistical analysis. Preference rate for the first-listed method in each pair (higher = more preferred). p-values are from a two-sided binomial test under $H_0 : P = 0.5$. $p > 0.05$ indicates failure to reject H_0 and implies perceptual indistinguishability.

image was then displayed for one second, followed by a one-second fixation dot to recalibrate gaze, and then the second image for one second. This brief exposure duration limited peripheral exploration, specifically testing whether Foveated Diffusion remained perceptually indistinguishable from full high-resolution generation when participants could perceive little peripheral content. After both images had been shown, participants pressed a keyboard key to indicate which image–the first or the second–had higher overall visual quality (Fig. 12).

Before data collection, participants completed two practice trials to familiarize themselves with the interface and task. All trials were conducted without breaks; the total completion time was approximately 10 minutes.

B.4 Statistical Analysis

Preference rate estimation. We group all participants' votes for each pairwise condition, and the preference rate P for each pairwise condition is calculated as the fraction of votes for the target method. Across all participants, each pairwise condition contains 220 data points (votes) in total.

Significance testing. To test whether a pairwise preference rate differs significantly from chance, we apply a two-sided binomial test under the null hypothesis $H_0 : P = 0.5$ (equal preference). A result is considered statistically significant at the $\alpha = 0.05$ level. A p-value above 0.05 for a pair indicates failure to reject H_0, i.e., the two methods are perceptually indistinguishable; a p-value below 0.05 indicates a statistically significant preference for one method over the other. Table 3 reports the preference rates and p-values for all three pairwise conditions.

Based on the p-values in Table 3, our user study confirms that Foveated Diffusion is perceptually indistinguishable from full high-resolution generation under gaze-contingent viewing conditions. Both our method and full high-resolution generation are significantly preferred over the naïve baseline, due to visual artifacts in the naïve baseline generations (main paper Fig. 4, Fig. 5).

C Additional Image Qualitative Results

C.1 Extended Image Generation Baseline Comparisons

In Figures 13-16, we provide extended baseline comparisons against full high-resolution generation and naïve mixed-resolution generation using our *randomized mask model*. The high-resolution region defined by the foveation mask is a circle with radius 0.5 relative to the image diagonal and a randomly placed center.

Our method consistently outperforms the naïve mixed-resolution baseline, generating coherent and consistent content with consistent structure and scale, whereas the mixed-resolution baseline exhibits significant distortions. Most importantly, our method achieves perceptually indistinguishable quality from the full high-resolution baseline while using approximately 57% fewer tokens, resulting in a 1.85× speedup in image generation time.

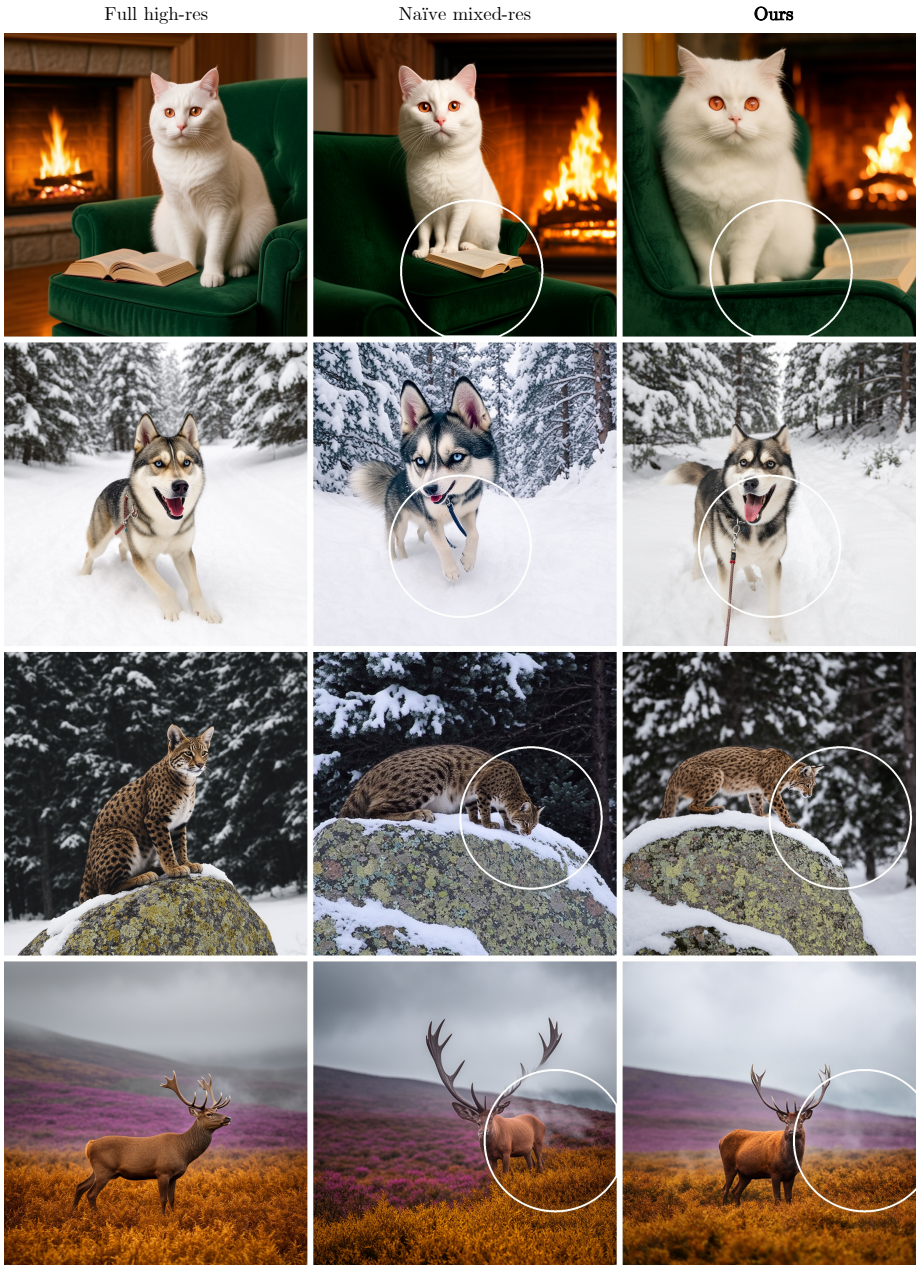

Fig. 13: Extended baseline comparisons. Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions. All images are uncompressed in this figure.

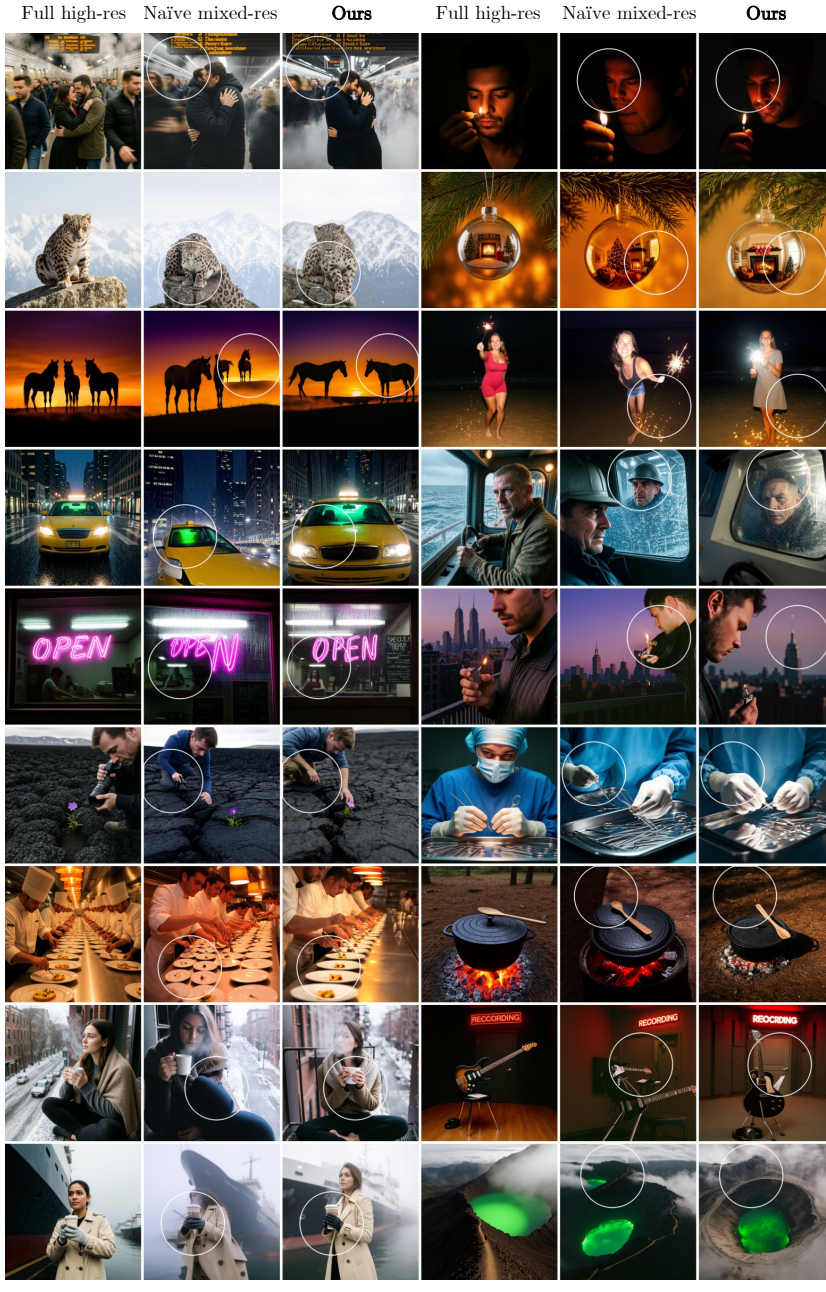

Fig. 14: Extended baseline comparisons. Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions.

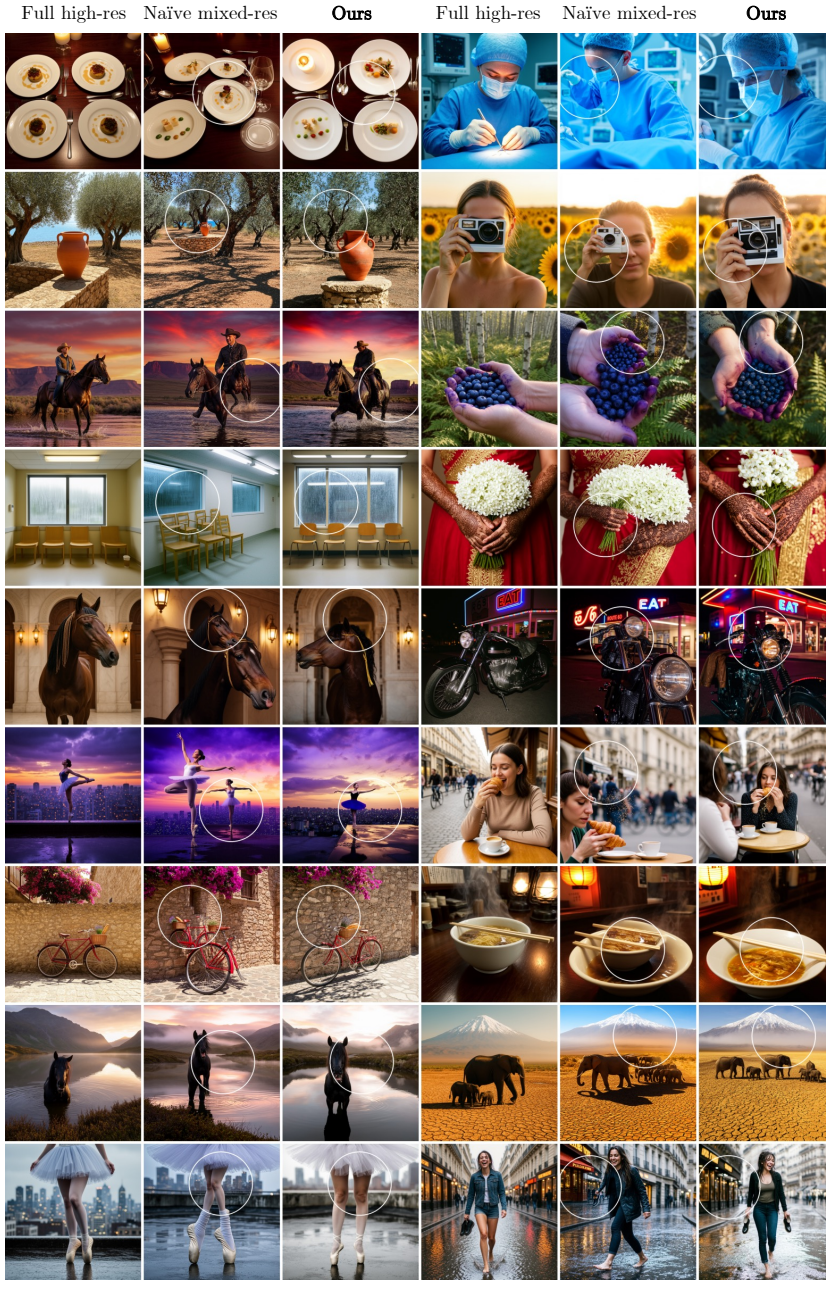

Fig. 15: Extended baseline comparisons. Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions.

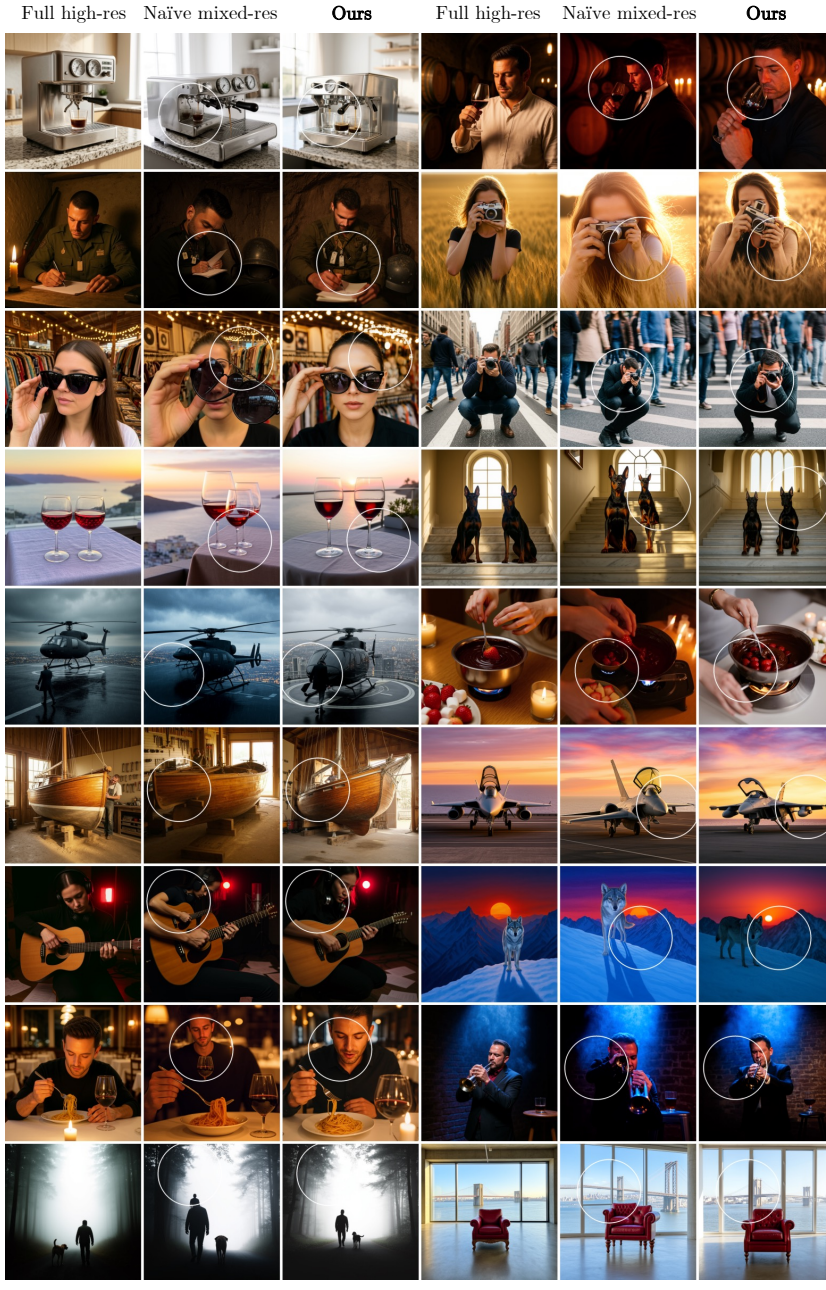

Fig. 16: Extended baseline comparisons. Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions.

C.2 Image Generation with Different Foveation Patterns

We present additional results using the *randomized-mask* model, where the high-resolution region varies in shape, size, and position, including foveation masks that contain multiple disjoint regions (Figs. 17–19). Given the same prompt and noise seed, Foveated Diffusion generates coherent and consistent content independent of the foveation mask geometry; the mask only determines which regions are synthesized at high resolution.

Specifically, images generated with foveation masks containing multiple disjoint regions (Fig. 19) suggest that Foveated Diffusion could support multi-viewer scenarios with multiple gaze locations.

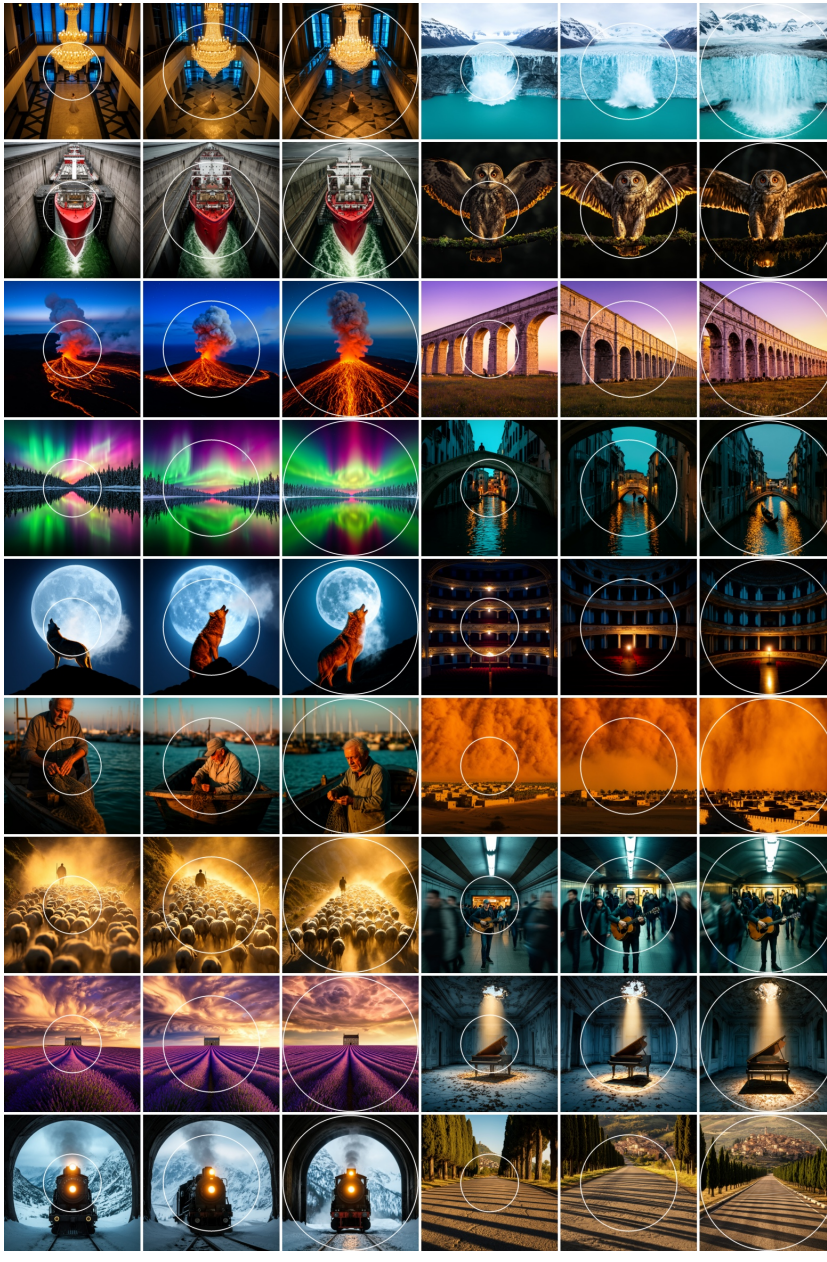

Fig. 17: Varying foveation mask radius. Foveated Diffusion generates coherent content across varying foveation mask radii. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

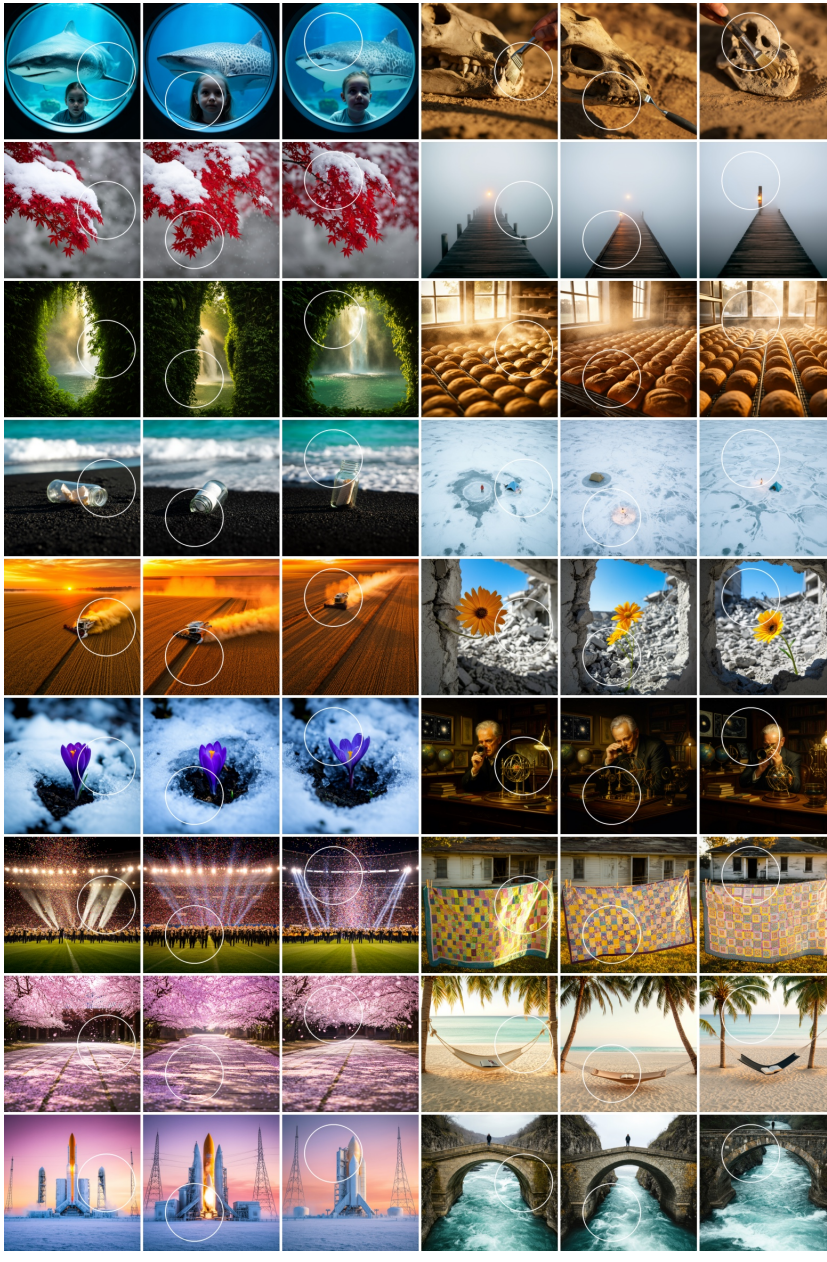

Fig. 18: Varying foveation mask position. Foveated Diffusion generates coherent content across varying foveation mask positions. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

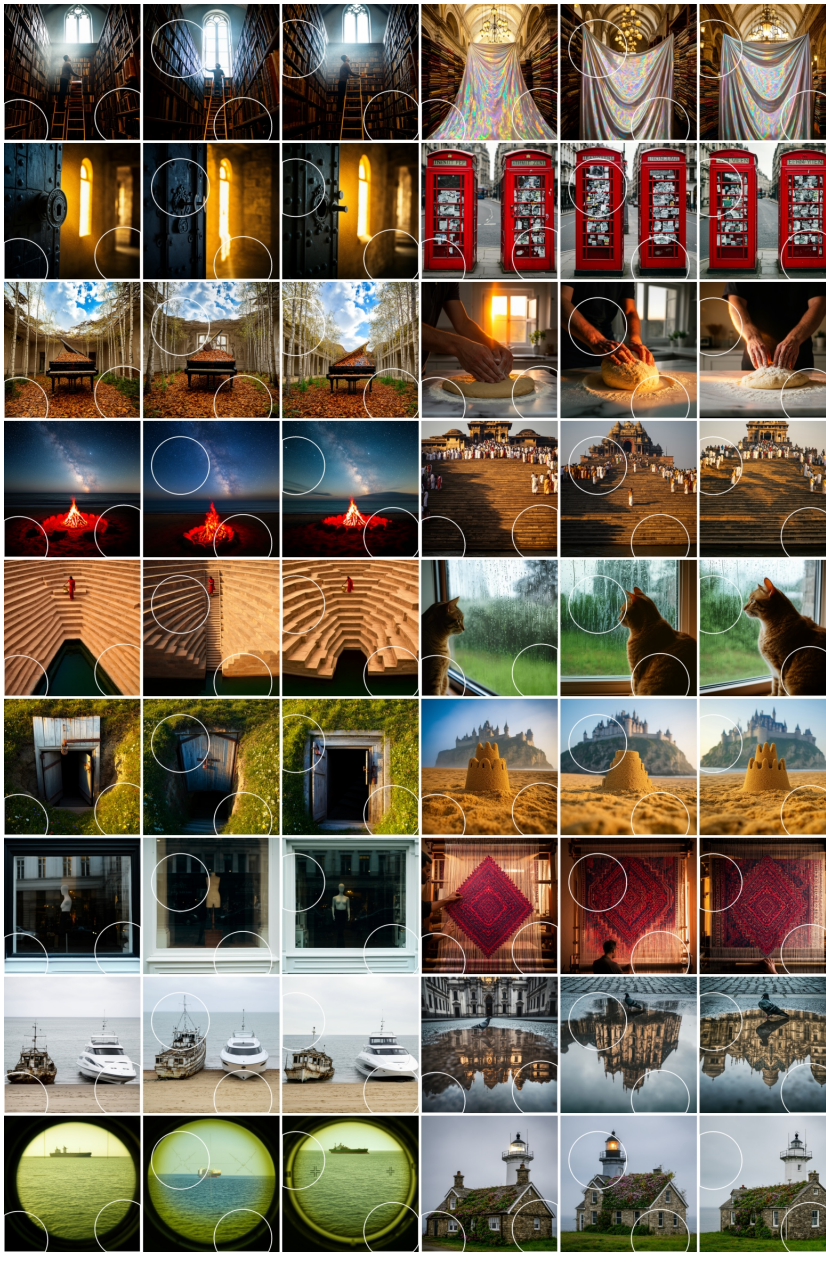

Fig. 19: Foveation mask containing multiple disjoint regions. Foveated Diffusion generates coherent content with foveation masks containing multiple disjoint regions. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

C.3 Towards Saliency-guided Image Generation

We show additional Foveated Diffusion additional results using the *saliency-guided* model in Figures 20 to 24. Compared to the randomized-mask model, the saliency-guided model generates images in which salient objects are aligned with the high-resolution regions defined by the foveation mask. Furthermore, the saliency-guided model also natively supports controllable multi-object generation when the foveation mask contains multiple disjoint high-resolution regions.

In Figures 22–24, we illustrate potential applications of saliency-guided Foveated Diffusion, including immersive VR gaming (Fig. 22), generative robotics (Fig. 23), and autonomous driving simulation for robotics policy learning (Fig. 24). Foveated Diffusion is particularly well suited for these scenarios because only the most salient objects need to be rendered in high resolution (e.g., wielded objects in VR games, robot arms and manipulated objects, and pedestrians or vehicles in dashcam scenes), while the remaining regions can be rendered at lower resolution.

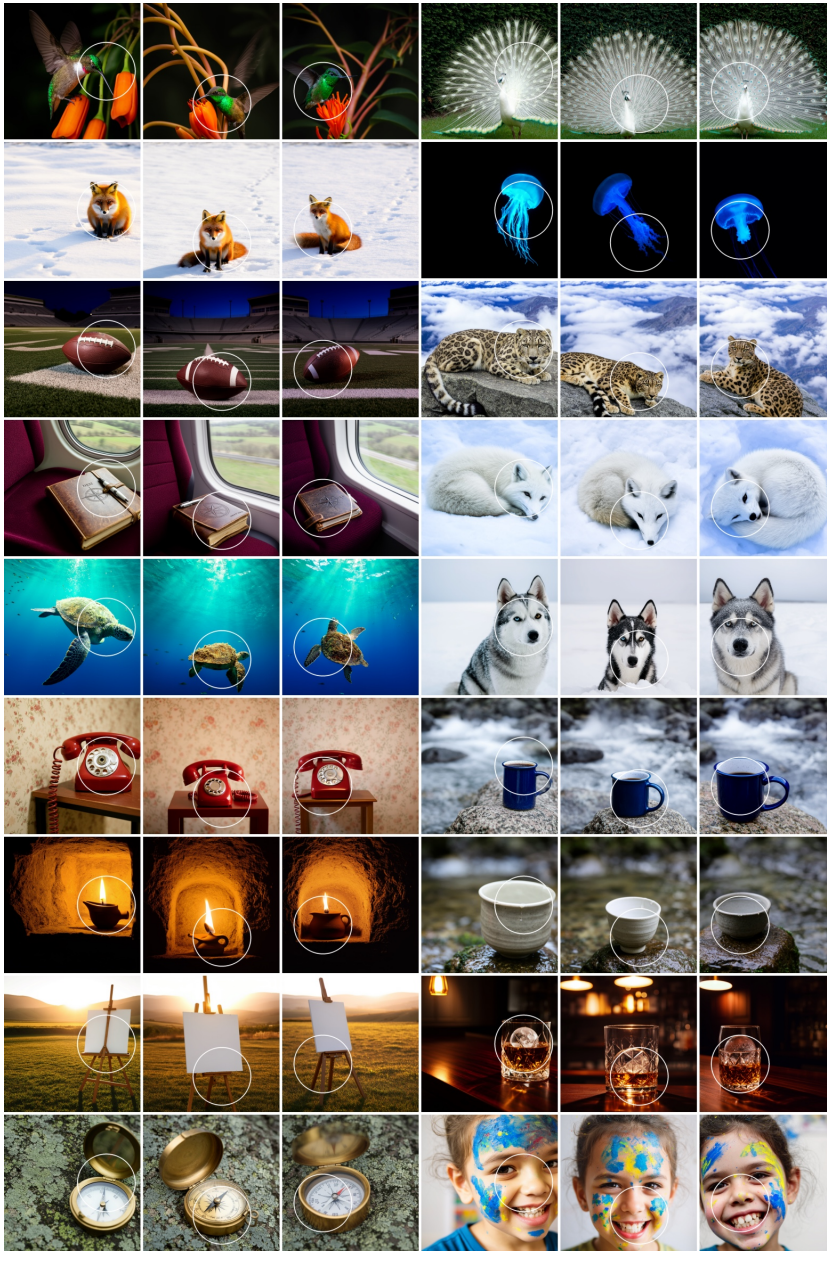

Fig. 20: Single-object saliency-guided generation. Foveated Diffusion with saliency-guided training enables coarse, controllable single-object generation, where the salient object is approximately aligned with the center of the foveation mask.

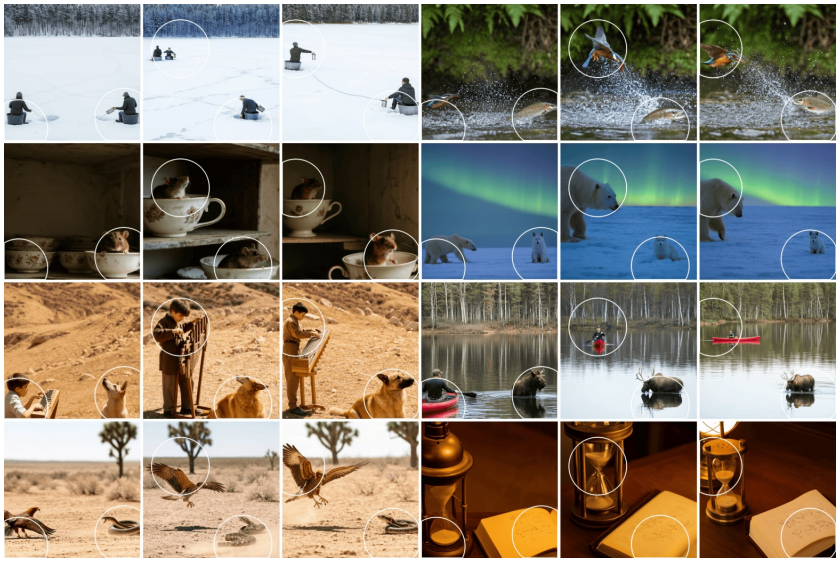

Fig. 21: Multi-object saliency-guided generation. Foveated Diffusion with saliency-guided training also enables coarse, controllable multi-object generation, where salient objects approximately align with the centers of the disjoint regions in the foveation mask.

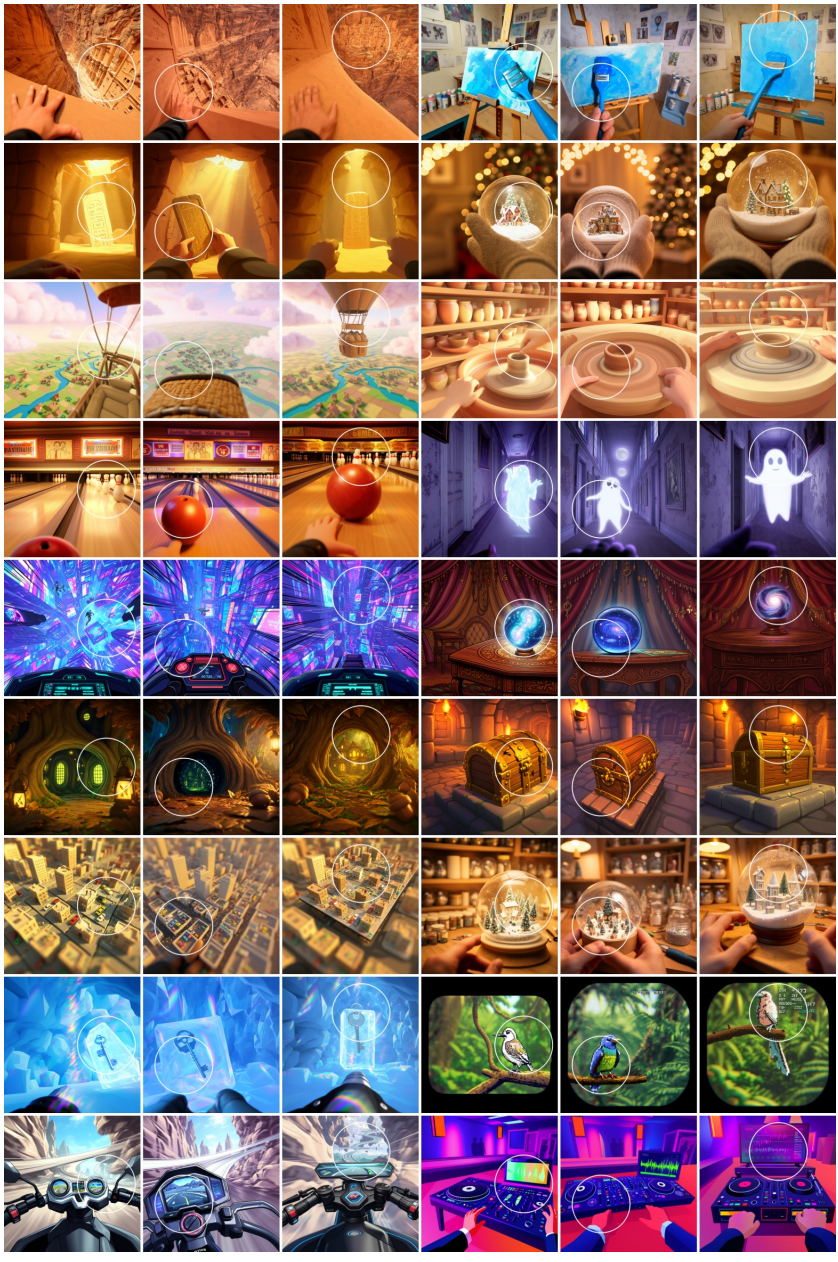

Fig. 22: Saliency-guided generation for immersive gaming. Foveated Diffusion is well suited for immersive first-person generative gaming applications, where salient objects can be generated near the gaze-tracked location (fovea) and rendered in high resolution, while the remaining regions are rendered at lower resolution.

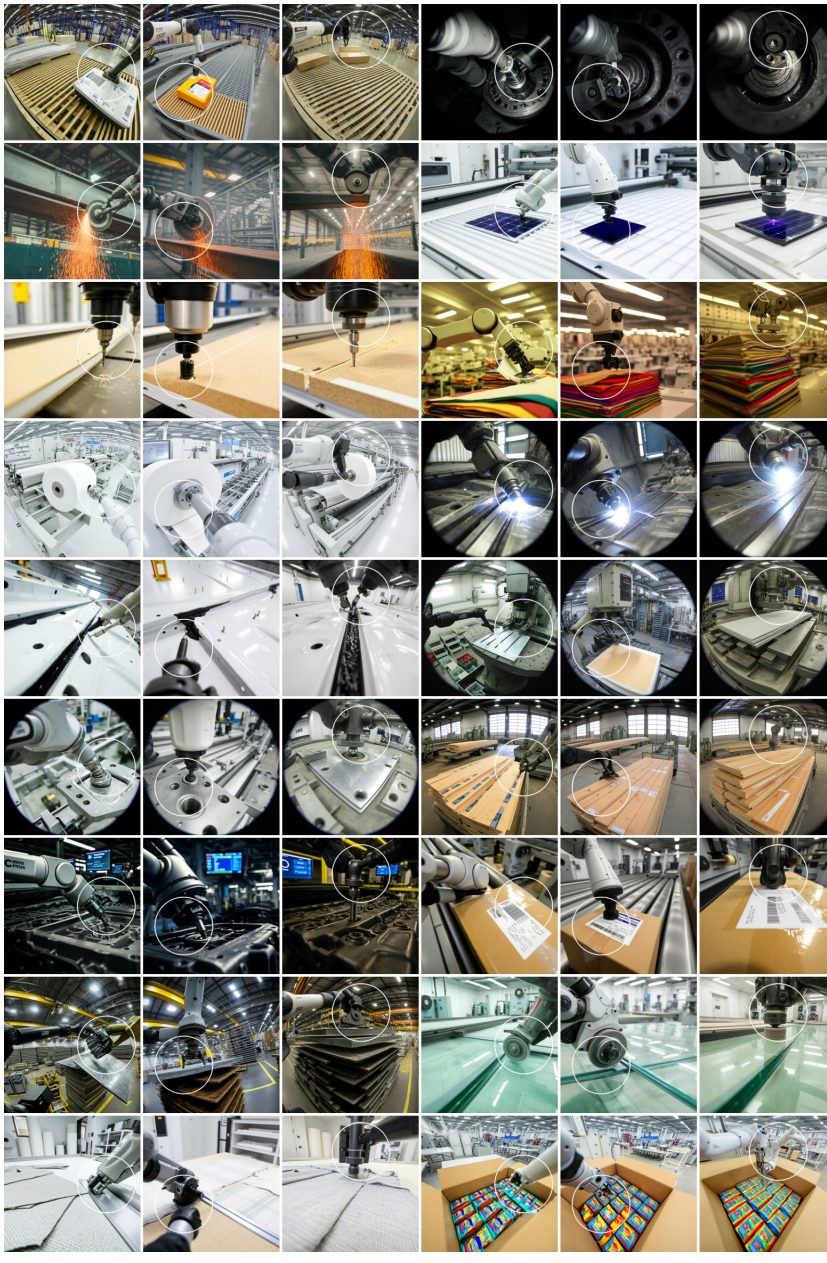

Fig. 23: Saliency-guided generation for robotics simulation. Foveated Diffusion is well suited for generative robotics simulation, where foveated imagery can be used for robotics policy learning. In this setting, only robot arms and manipulated objects are generated at high resolution, while the background is rendered at lower resolution to provide global context.

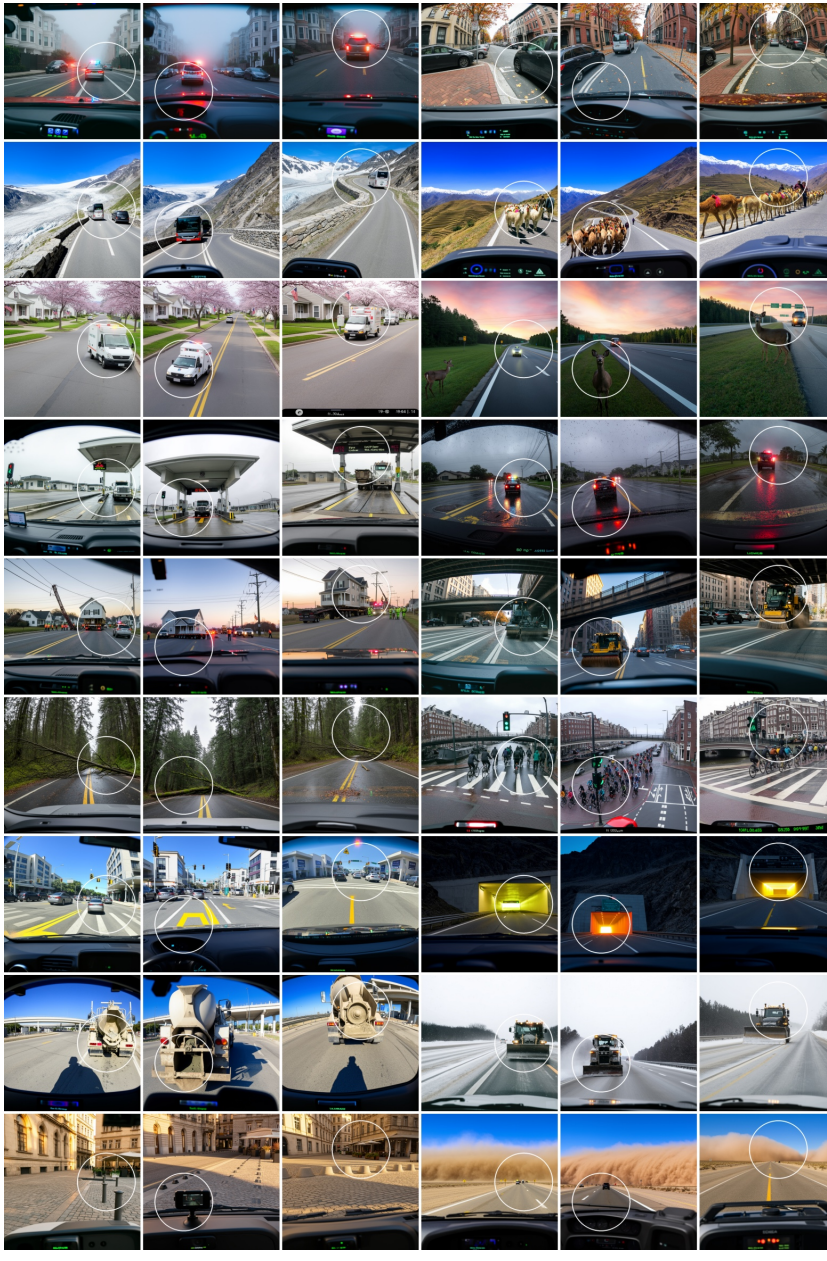

Fig. 24: Saliency-guided generation for autonomous vehicles. Foveated Diffusion is also well suited for generative autonomous driving simulation, where foveated imagery can be used for policy learning in self-driving systems. In this setting, only important objects in the scene (e.g., pedestrians, other vehicles, roadblocks) are generated at high resolution, while the background is rendered at lower resolution to provide global context.

C.4 Towards Bounding-box-guided Image Generation

Similar to saliency-guided visual generation, we adapt Foveated Diffusion for *bounding-box-guided* visual generation. We use the Ultralytics software library [35], which integrates multiple YOLO models, for bounding box detection.

As shown in Figures 25 to 27, the bounding-box-guided model successfully generates objects within the foveation boundary. Similar to the saliency-guided model, the bounding-box-guided model inherently enables controllable multi-object generation when the foveation mask comprises multiple disjoint high-resolution regions.

The difference between the saliency-guided and bounding-box-guided models is subtle but informative. Bounding boxes explicitly delineate object contours, encouraging the model to generate entire objects within, or closely aligned to, the foveal region (Fig. 27). In contrast, the saliency-guided model aligns only the most salient portions of objects with the fovea, rather than enforcing full-object containment. Notably, this behavior is not imposed by any architectural modification or specialized algorithm. This behavior arises purely from data construction, namely how foveation masks structure the interaction between high- and low-resolution tokens during training, highlighting the generality of our Foveated Diffusion framework.

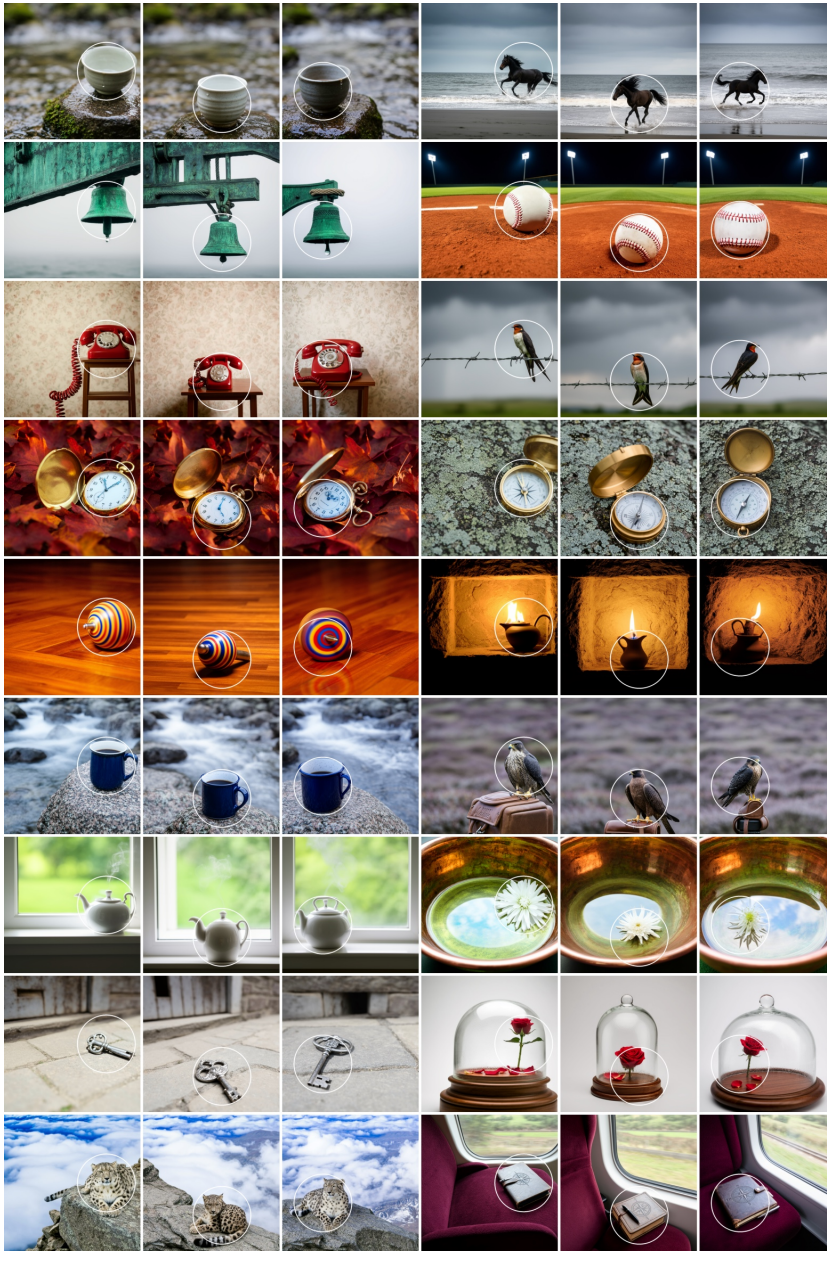

Fig. 25: Single-object bounding-box-guided generation. Foveated Diffusion with bounding-box-guided training enables coarse, controllable single-object generation, where the salient object is approximately constrained within the foveation mask.

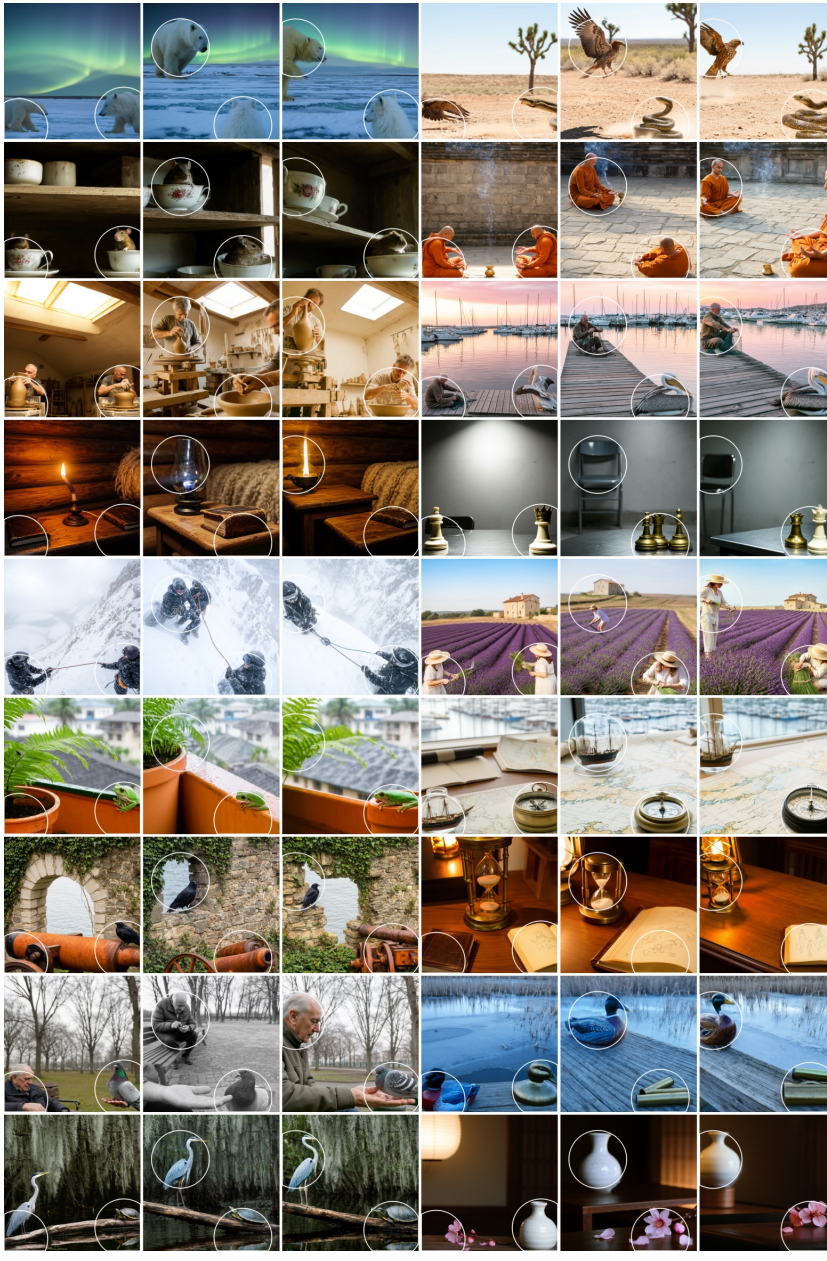

Fig. 26: Multi-object bounding-box-guided generation. Foveated Diffusion with bounding-box-guided training enables coarse, controllable multi-object generation, where salient objects are approximately constrained within the disjoint regions of the foveation mask.

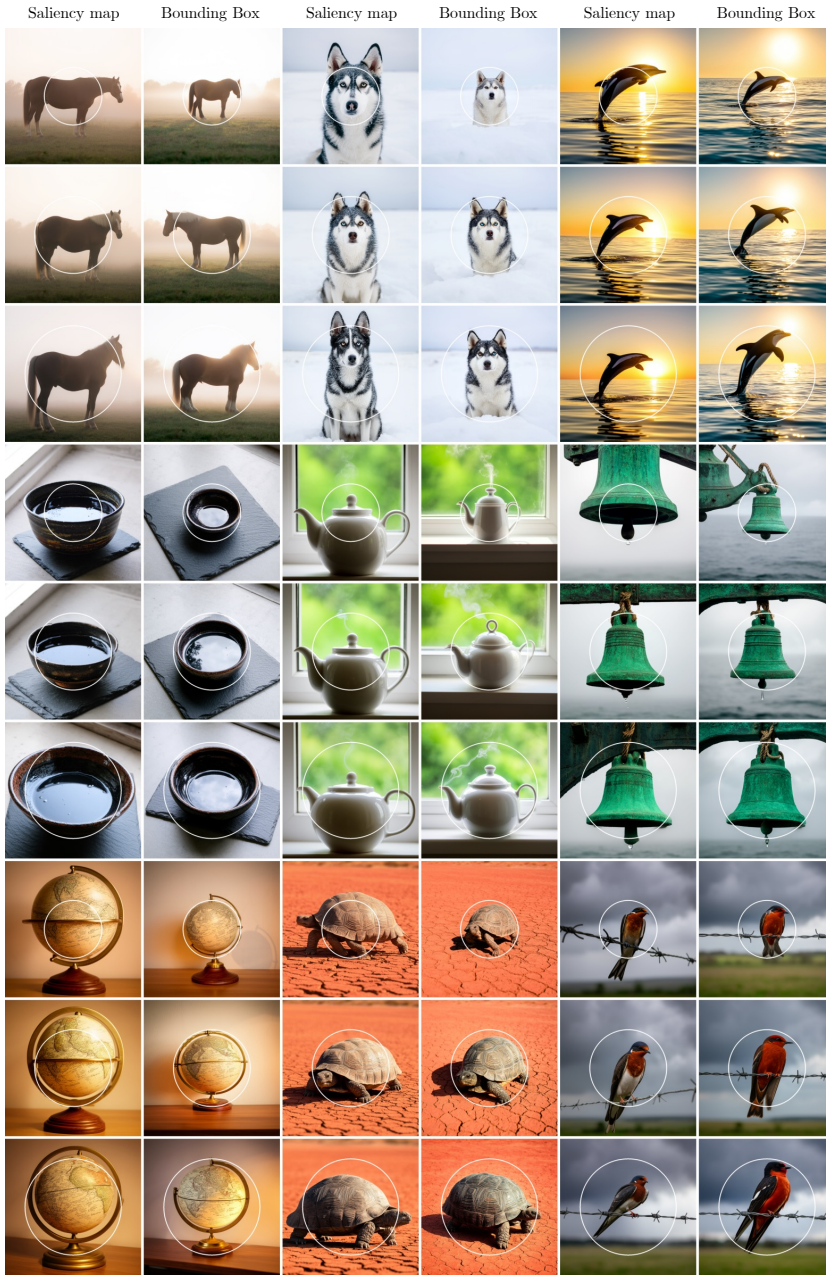

Fig. 27: Bounding-box-guided generation and saliency-guided generation. We compare bounding-box-guided and saliency-guided generation. Because bounding-box-derived masks precisely delineate object contours, the bounding-box-guided model generates entire objects within the high-resolution region. In contrast, the saliency-guided model aligns only the most salient parts of objects with the center of the high-resolution region.

D Additional Video Qualitative Results

We include selected video generation results on our project website [2].

D.1 Extended Video Generation Baseline Comparisons

We provide extended baseline comparisons against full high-resolution video generation and naïve mixed-resolution video generation using our *randomized mask model*. While the foveation mask for image generation is defined as a circle with a randomized center and radius, video generation requires temporal coherence. To achieve this, we sample three key control points with randomized spatial coordinates and radii across the video sequence. We then apply cubic spline interpolation to these points to generate a smooth, continuous foveation trajectory for the duration of the video, ensuring the high-resolution window moves fluidly across frames. We show both 480p and 720p generation results to show the generality of our model.

Our method consistently outperforms the naïve mixed-resolution baseline, generating coherent and consistent content with consistent structure and scale without color distortions, whereas the mixed-resolution baseline exhibits significant artifacts.

D.2 Video Generation with Different Foveation Patterns

We present additional video generation results using the *randomized-mask* model, where the high-resolution region follows different randomized spline trajectories that vary in position and size across frames. Given the same prompt and noise seed, Foveated Diffusion generates coherent and consistent content independent of the foveation mask trajectory; the mask only determines which regions are synthesized at high resolution.

D.3 Towards Saliency-guided Video Generation

We show additional Foveated Diffusion additional results using the *saliency-guided* model. The saliency-guided model generates videos in which salient objects are aligned with the high-resolution regions defined by the foveation mask trajectory.

We illustrate potential applications of saliency-guided Foveated Diffusion, including immersive VR gaming and generative robotics and autonomous driving simulation for robotics policy learning. Foveated Diffusion is particularly well suited for these scenarios because only the most salient objects need to be rendered in high resolution (e.g., wielded objects in VR games, robot arms and manipulated objects, and pedestrians or vehicles in dashcam scenes), while the remaining regions can be rendered at lower resolution.

[2] https://bchao1.github.io/foveated-diffusion/

E Discussion

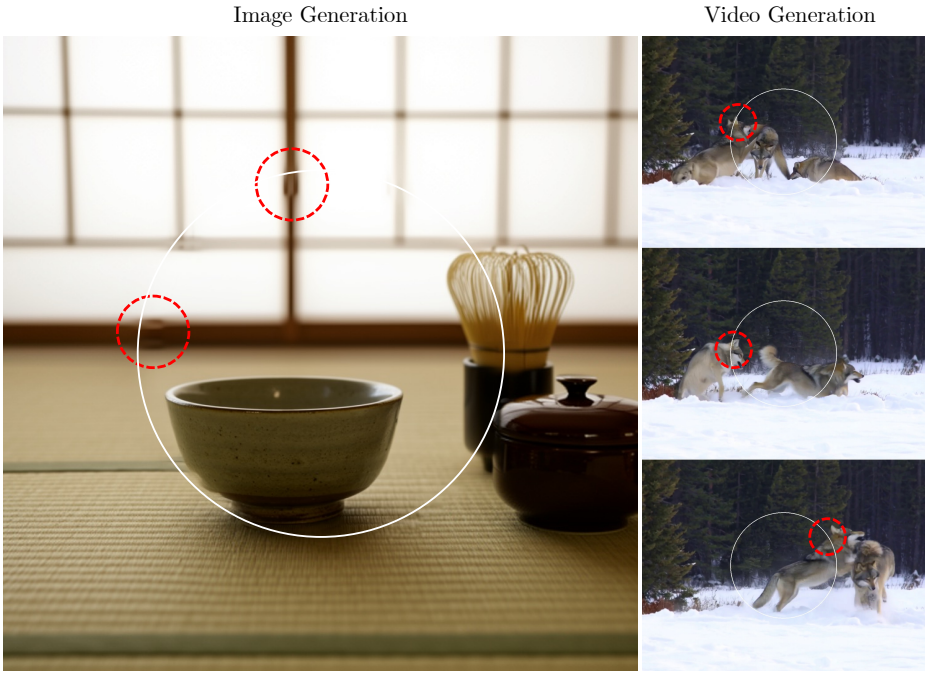

Fig. 28: Foveated Diffusion artifacts. We delineate the foveation border with a white circular outline. The red dashed lines indicate regions with blending artifacts.

Foveated Diffusion occasionally exhibits color or discontinuity artifacts near foveation boundaries, as shown in Fig. 28. We attribute these artifacts to the final VAE decoding and alpha-blending step between low- and high-resolution regions. We believe these artifacts could be mitigated by adapting the VAE to directly decode mixed-resolution tokens, thereby avoiding separate decoding and blending of low- and high-resolution regions.