FOVEATED DIFFUSION: EFFICIENT SPATIALLY ADAPTIVE IMAGE AND VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow matching models have unlocked unprecedented capabilities for creative content creation with photorealistic image and 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 can specify the region of interest. 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 natively 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.

1 INTRODUCTION

Interactive image and streaming video generation place strict computational demands on emerging diffusion and flow matching models (Bruce et al., 2024; Alonso et al., 2024; Che et al., 2024; Decart et al., 2024; Feng et al., 2024; Parker-Holder & Fruchter, 2025; Song et al., 2025; Valevski et al., 2024; Weng et al., 2024; Yu et al., 2025; Henschel et al., 2025; Wu et al., 2025c; Po et al., 2025; Zhang et al., 2025d; Yin et al., 2025). At the same time, image resolutions and video frame or context lengths keep growing. How can we generate an ever-increasing number of tokens when attention and MLP operations in modern diffusion transformers (DiTs) (Peebles & Xie, 2023) scales with sequence length?

Our insight answers this question: ultimately, a human observes the generated content, so why not exploit the characteristics of the human visual system to generate it in a computationally efficient, perceptually motivated manner? Specifically, we build on the concept of foveation — humans perceive very high-resolution visual information in a small region around their gaze location (the foveal region) but acuity degrades rapidly in the visual periphery (Anstis, 1974; Weymouth, 1958).

With this work, we introduce *Foveated Diffusion* and develop a framework for post-training existing image or video generation models into *foveated* generation models. Our key idea is eccentricity-dependent token allocation: given a foveation mask that defines the high-acuity foveal region over the image or video frame we wish to generate, we allocate higher token density near the fovea and progressively fewer tokens toward the periphery. Note that this is not equivalent to blurring the periphery of a full-resolution image: peripheral content is never generated at high resolution in the first place, which is what converts eccentricity-dependent acuity into an actual reduction in denoising compute rather than a post-processing step. We then follow standard diffusion or flow-matching procedures to generate an image from Gaussian noise and conditioning text prompts (see Fig. 1);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

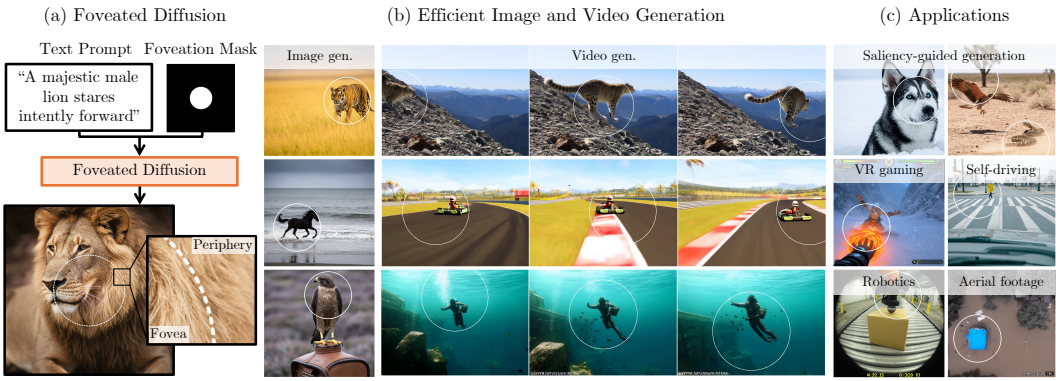

Figure 1: **Foveated Diffusion.** (a) Given user-specified masks and text prompts as input, our method generates foveated content using fewer tokens than full high-resolution generation, resulting in faster inference while maintaining comparable perceptual quality. (b, c) Foveated Diffusion is well suited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

unlike conventional methods, we operate with a significantly reduced set of tokens at every denoising step, achieving substantial computational savings. We develop a mixed-resolution tokenization scheme, accompanied by a modification of Rotary Positional Embeddings (RoPE) (Su et al., 2024; Wu et al., 2025b), along with a post-training strategy that transforms high-resolution pretrained models into foveated generative models. Together, these contributions preserve cross-resolution content consistency while achieving significant speedup.

Our approach is inspired by foveated rendering (Geisler & Perry, 1998; Guenter et al., 2012; Patney et al., 2016), a standard technique in computer graphics. Foveated Diffusion and rendering share the idea of leveraging 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.

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 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 over $2\times$ and $3\times$ for image and video generation, respectively, while maintaining comparable 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 (Curcio & Allen, 1990; Rovamo & Virsu, 1979; Watson, 2014; Geisler & Perry, 1998). Real-time rendering systems exploit this 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 enable interactive rendering at a fraction of the full-resolution cost while maintaining comparable perceptual quality (Guenter et al., 2012; Patney et al., 2016; Levoy & Whitaker, 1990; Reddy, 2001; Stengel et al., 2016; Sun et al., 2017; Kaplanyan et al., 2019; Weier et al., 2016; Tariq et al., 2022). More recently, foveation has been applied to neural rendering and

novel view synthesis (Shi et al., 2024; Franke et al., 2025; Deng et al., 2022) to accelerate the rendering of Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) and Gaussian splats (Kerbl et al., 2023) for immersive displays. In computer vision and robotics, foveation has been used to improve efficiency in neural network architectures or perception tasks (Minut et al., 2000; Bandera & Scott, 1989; Killick et al., 2023), such as mixed-resolution tokenization of vision transformers (Jonnalagadda et al., 2021; Ronen et al., 2023; Schmidt & Newcombe, 2025; Havtorn et al., 2023) and robot policy learning (Kerr et al., 2025; Chuang et al., 2025). However, while foveation has been extensively explored across rendering and perception pipelines, it has not yet been realized in generative modeling, despite the increasing capabilities in immersive and interactive visual generation.

Efficient Visual Generation. Diffusion Transformers (DiTs) have emerged as the dominant paradigm for scalable, high-fidelity generation (Black Forest Labs, 2026; Wan Team et al., 2025; Kong et al., 2024; Esser et al., 2024; Peebles & Xie, 2023). However, the computational cost of DiTs is quadratic with respect to the input token count due to the self- and cross-attention mechanisms (Vaswani et al., 2017), restricting context length and leading either to degraded visual consistency under fixed compute and memory budgets or prohibitive 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 (Li et al., 2025; Xia et al., 2025; Zhan et al., 2025; Katharopoulos et al., 2020; Wang et al., 2020; Choromanski et al., 2021; Beltagy et al., 2020), hardware-aware optimization (Zhang et al., 2025c;e;f; Xi et al., 2025; Dao et al., 2022), KV-caching (Kwon et al., 2023; Shazeer, 2019), etc. Another orthogonal axis of research aims to simply reduce the effective token count while maintaining high image quality. Token merging methods (Bolya et al., 2023; Lee et al., 2024; Chen et al., 2026; Lu et al., 2025; Haurum et al., 2024; Bolya & Hoffman, 2023; Kim et al., 2024; Wu et al., 2025a; Lee et al., 2025; Fang et al., 2025) identify redundant tokens at each layer of a Diffusion Transformer (DiT) and merge similar tokens according to a predefined heuristic or importance metric. Recent training-free mixed-resolution denoising methods (Jeong et al., 2026; Wu et al., 2025b; Tian et al., 2025) 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.

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. 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, concentrating computation in high-acuity regions and sparsifying the periphery.

3 METHOD

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 (Bandera & Scott, 1989; Geisler & Perry, 1998; Guenter et al., 2012; Weier et al., 2016; Patney et al., 2016; Kaplanyan et al., 2019).

Formally, we define a binary foveation mask $M \in \{0,1\}^{H\times W}$, 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 (Ho et al., 2020; Song et al., 2021) 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 (Rombach et al., 2022). Given an image or a video, a variational autoencoder (VAE) (Kingma & Welling, 2019), 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 (Lipman et al., 2023) 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 (Peebles & Xie, 2023). The computational efficiency of such generative models is therefore quadratically related to the number of tokens processed due to the expensive attention (Vaswani et al., 2017) 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

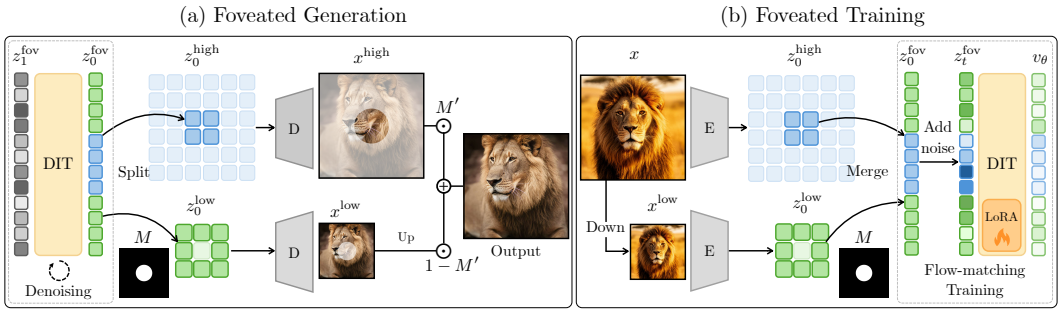

Figure 2: **The Foveated Diffusion Pipeline.** In Foveated Generation (a), we iteratively denoise a foveated token sequence of reduced length instead of the full high-resolution sequence. The resulting tokens z_0^{fov} are split into high- and low-resolution grids, decoded by the VAE, and blended using a user-specified foveation mask. We employ Foveated Training (b) to adapt pretrained DiTs to foveated token sequences using low-rank adaptation (LoRA) (Hu et al., 2022). 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.

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 $M_z \in \{0\ 1\}^{h \times w}$ 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* $\mathbb{R}^{c \times L}$ 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_z. 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:

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

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 upsampled and blended with the high-resolution decoding to form x^{fov}, following Eq. 1 with the pixel-space mask recovered as $M = \text{Up}(M_z)$. The full generation pipeline is illustrated in Fig. 2-(a).

Mixed-Resolution RoPE. Standard Rotary Positional Embedding (RoPE) (Su et al., 2024) 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. (2025b) 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 subsample 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. (2025b) for more details.

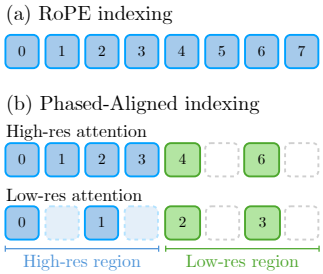

Figure 3: **Mixed-Res RoPE.**

Failure of Naïve Mixed-Resolution Denoising. A pretrained 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.



Figure 4: **Failure of Naïve mixed-res 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:

$$z_0^{\text{high}} = E(x) \in \mathbb{R}^{c\times(h\cdot w)} \tag{7}$$

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

$$z_0^{\text{low}} = E(\text{Down}(x)) \in \mathbb{R}^{c\times(\frac{h}{2}\cdot\frac{w}{2})} \tag{8}$$

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:

$$z_0^{\text{fov}} = \text{Merge}(z_0^{\text{high}}\ z_0^{\text{low}}\ M_z) \in \mathbb{R}^{c\times L} \tag{9}$$

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} corresponds to the foveated image x^{fov} through Eq. 5 to 1.

We train our Foveated Diffusion model using the standard flow matching objective (Lipman et al., 2023). 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_z 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, and may be derived from gaze, saliency, bounding boxes, or task priors. 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. We train a separate Foveated Diffusion model for each of these mask types; all main qualitative and quantitative results are evaluated on the *randomized mask* model.

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

We fine-tune pretrained Diffusion Transformers (DiTs) with our Foveated Diffusion framework. For images, we fine-tune FLUX.2 Klein 4B (Black Forest Labs, 2026) on Aesthetic-Train-V2 (Zhang et al., 2025a;b), randomly sampling 90k images for training and reserving 10k prompt–image pairs

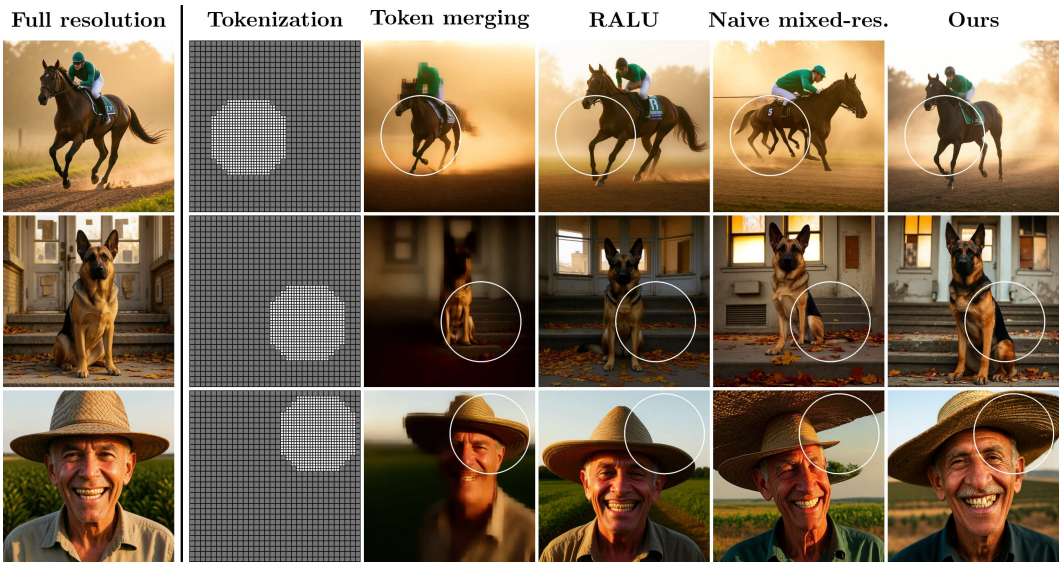

Figure 5: **Qualitative comparison for image generation.** Our method yields perceptually indistinguishable results from full high-resolution synthesis with zero hyperparameter tuning, unlike RALU (Jeong et al., 2026). Token merging (Bolya & Hoffman, 2023) results in overly-blurry peripheral regions, while our naïve baseline introduces scale inconsistencies and structural artifacts across mixed-resolution regions. The high-resolution regions (fovea) are delineated with white borders.

Method	Image-Reward ↑	HPS v2.1 ↑	HPS v3 ↑	CLIP Score ↑	LAION Aes. ↑	FID ↓	LPIPS ↓	Time / Speedup
Full high-res	0.964	0.280	10.32	0.852	5.574	11.56	0.730	41.6s / 1.00×
ToMe (Bolya & Hoffman, 2023)	−0 698	0.181	1.34	0.784	5.014	40.34	0.799	18.2s / 2.28×
RALU (Jeong et al., 2026)	*0.854*	0.267	9.40	*0.861*	*5.526*	12.50	*0.744*	18.9s / 2.19×
Naïve mixed-res	0.823	*0.270*	*10.41*	0.858	5.470	**10.98**	0.753	18.6s / 2.24×
Ours	**0.996**	**0.279**	**12.10**	**0.862**	**5.586**	*11.51*	**0.727**	18.5s / 2.24×

Table 1: **Quantitative comparison for image generation.** We compare against mixed-resolution baselines on 32% token ratio and time-matched performance, highlighting the best result for each in bold and second best in italic. Foveated Diffusion outperforms every baseline on six of seven metrics, and matches or exceeds full high-resolution generation (top row) with 2 24× speedup.

for evaluation; for video, we fine-tune Wan2.1 1.3B (Wan Team et al., 2025) on Vchitect-T2V-Dataverse (Fan et al., 2025; Si et al., 2025), excluding 200 prompts for evaluation. During training, we randomly sample a circular foveation mask for each image or a foveation path for each video. All models are fine-tuned with LoRA (Hu et al., 2022) of rank 32 on NVIDIA H100 GPUs. Further details are in Appendix A.

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.

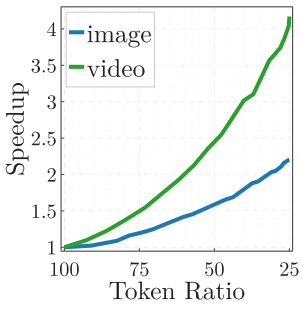

Figure 6: **Foveated visual generation speedup.**

Method	Subject Cons. ↑	Background Cons. ↑	Motion Smooth. ↑	Dynamic Degree	Aesthetic Quality ↑	Imaging Quality ↑	Time / Speedup
Full high-res	0.9319	0.9298	0.9943	0.2400	0.5492	0.6411	86.2s / 1.00×
ToMe (Bolya & Hoffman, 2023)	0.8048	0.8938	0.9666	**0.5000**	0.4099	0.4804	32.2s / 2.68×
Naïve mixed-res	*0.8985*	*0.9180*	*0.9917*	*0.2350*	*0.4832*	*0.5210*	27.4s / 3.15×
Ours	**0.9377**	**0.9349**	**0.9946**	0.1750	**0.5479**	**0.5906**	27.4s / 3.15×

Table 2: **Quantitative comparison for video generation (VBench).** Foveated Diffusion outperforms both token-reduction baselines on all quality metrics, while achieving performance comparable to full-resolution generation. Dynamic Degree measures the amount of motion rather than its quality, as can be seen in Fig. 8. Notably, our framework maintains high subject consistency while providing a 3 15× speedup. Full high-resolution generation is the reference baseline; bold marks the best mixed-resolution result and italics the second best.

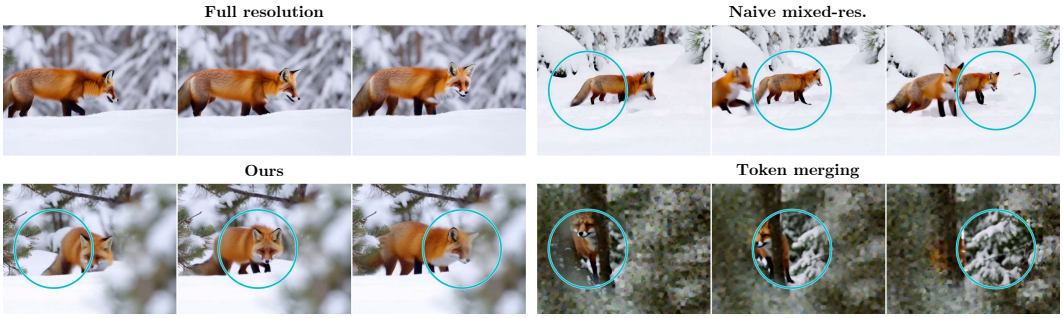

Figure 8: **Qualitative comparison for video generation.** Foveated Diffusion outperforms the naïve mixed-resolution baseline (duplicate objects and scale mismatch) and ToMe Bolya et al. (2023) (blurry periphery) in video generation.

Image Generation. For foveated image generation, we compare with three other mixed-resolution diffusion baselines: ToMe (Bolya & Hoffman, 2023), RALU (Jeong et al., 2026), and the naïve mixed-resolution diffusion introduced in this paper, along with the full high-resolution baseline. We report image quality across seven metrics: ImageReward (Xu et al., 2023), HPSv2.1 (Wu et al., 2023), HPSv3 (Ma et al., 2025), CLIP score (Radford et al., 2021), LAION aesthetic predictor (Schuhmann et al., 2022), FID (Heusel et al., 2017) and LPIPS (Zhang et al., 2018). We fix the foveation mask to be a circular mask with randomly sampled centers and varying radii that is shared across baselines. Please refer to the supplementary materials for additional image generation results with varying token ratios (Appendix C.1) and compatibility with distilled models (Appendix C.3). 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. Setting FID aside, Foveated Diffusion surpasses all mixed-resolution baselines on all remaining metrics with a 2 24× end-to-end speedup as shown in Table 1.

Perceptual User Study. Standard generative metrics weigh all pixels uniformly and ignore eccentricity-dependent visual acuity, so their trends can conflict with human preference (e.g., the HPSv2.1–FID discrepancy in Tab. 1). We therefore run a two-alternative forced choice (2AFC) study under a pseudo-eye-tracked protocol: participants fixate a red point, view a pair of images drawn from two of the three methods (high-resolution, naïve mixed-resolution, and ours) for 1 second each to discourage eye movements, then select the one with higher overall quality. Fig. 7 shows near perceptual parity with full high-resolution generation and a strong preference over the naïve baseline; further analysis is in the supplement.

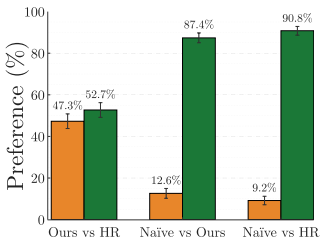

Figure 7: **User study results.**

Video Generation. For foveated video generation, we report the standard generative video metric VBench (Huang et al., 2024). Here the foveation mask follows a random path across frames at a token ratio of 38% relative to the original token length of $(f \cdot h \cdot w)$, so that the foveal region moves during generation. We additionally compare against token merging (Bolya & Hoffman, 2023), driven by the same path. Foveated Diffusion surpasses both mixed-resolution baselines while achieving results comparable to full high-resolution generation with a 3 15× speedup, as clearly shown in

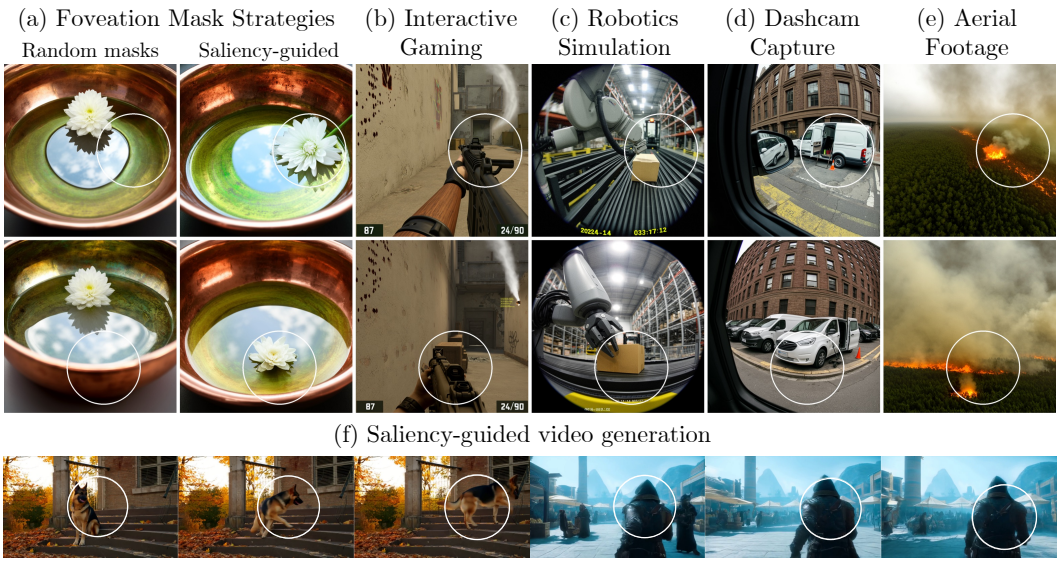

Figure 9: **Towards saliency-guided visual generation.** We show that the *saliency-guided* Foveated Diffusion models enable saliency-guided image (a-e) and video generation (f), where salient objects align with the fovea. The *randomized masks* model do not exhibit such behavior (a). This is potentially 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.

Tab. 2. Fig. 8 provides supporting qualitative results, where ToMe (Bolya & Hoffman, 2023) shows high noise in the periphery and the naïve baseline exhibits severe structural and scale mismatches. This explains ToMe's high Dynamic Degree, which measures the amount of motion rather than its quality; peripheral noise registers as motion, making the metric unreliable in this setting.

Foveation Mask Strategies. As discussed in Sec. 3.2, our *randomized mask* training protocol uniquely enables one model to generate coherent content under foveation masks unseen at training, varying in size, position, and shape, including non-contiguous masks with multiple disjoint high-resolution regions (Appendix C.4). More importantly, replacing random masks with image and video saliency maps binarized from DeepGaze (Linardos et al., 2021) further enables *saliency-guided* generation. In Fig. 9, salient objects align with the specified foveal regions, which is useful for generative VR gaming, robotics simulation, or self-driving, where only salient objects in view need high-resolution rendering (Kerr et al., 2025).

5 Conclusion

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

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 1). 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 (Wu et al., 2025b). 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.

AI USE STATEMENT

We used generative AI tools to assist with implementing and debugging methods. We also used these tools for manuscript editing and making visualization code. Reported numerical results are computed by evaluation code rather than supplied by a language model. The authors take responsibility for the final text, claims, code, and experimental results.

ETHICS STATEMENT

Foveated Diffusion builds on pretrained image and video generation models and inherits their underlying biases and potential societal risks, including stereotypical representations and the generation of misleading or harmful content. Our changes to tokenization and mixed-resolution token allocation do not explicitly address these risks, and fine-tuning may alter the models' behavior. Reducing the computational cost of generation can broaden access to creative tools and research, but can also lower the cost of producing deceptive media at scale. Potential misuse includes impersonation and non-consensual imagery, with associated privacy and intellectual-property concerns. Responsible deployment should therefore include appropriate content safeguards, disclosure of synthetic content, respect for consent and applicable usage restrictions, and application-specific bias and safety evaluation. Our quality and efficiency evaluations do not establish the safety or fairness of downstream deployments.

REPRODUCIBILITY STATEMENT

Detailed explanations of the method is in 3. Implementation details are provided in Section 4 and A, user study details are described in Section B, and all baseline comparison and ablation study settings are described in Sections 4, C, and D. We will release code and model weights upon paper acceptance.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Stuart M Anstis. A chart demonstrating variations in acuity with retinal position. *Vision Research*, 14(7):589–592, 1974.

Cesar Bandera and Peter D Scott. Foveal machine vision systems. In *Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics*, pp. 596–599. IEEE, 1989.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence, 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence. Model release, January 15, 2026.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPRW*, pp. 4599–4603, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

Yutian Chen, Yuheng Qiu, Ruogu Li, Jay Patrikar, and Sebastian Scherer. Co-me: Confidence guided token merging for visual geometric transformers. In *CVPR*, pp. 14590–14599, 2026.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021.

Ian Chuang, Andrew Lee, Dechen Gao, Jinyu Zou, and Iman Soltani. Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers, 2025. URL https://arxiv.org/abs/2507.15833v1.

Christine A Curcio and Kimberly A Allen. Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1):5–25, 1990.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. Project website, 2024. URL https://oasis-model.github.io/.

Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE TVCG*, 28 (11):3854–3864, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *CVPR*, pp. 18083–18092, 2025.

Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

Linus Franke, Laura Fink, and Marc Stamminger. 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.

Wilson S Geisler and Jeffrey S Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Human vision and electronic imaging III*, volume 3299, pp. 294–305. SPIE, 1998.

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM TOG*, 31(6):1–10, 2012.

Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Agglomerative token clustering. In *ECCV*, pp. 200–218. Springer, 2024.

Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Msvit: Dynamic mixed-scale tokenization for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 838–848, 2023.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2568–2577, 2025.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Wongi Jeong, Kyungryeol Lee, Hoigi Seo, and Se Young Chun. Training-free mixed-resolution latent upsampling for spatially accelerated diffusion transformers. In *CVPR*, 2026. Highlight.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, January 2023. URL https://ultralytics.com. Software release, January 10, 2023.

Aditya Jonnalagadda, William Yang Wang, BS Manjunath, and Miguel P Eckstein. FoveaTer: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM TOG*, 38(6):1–13, 2019.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, pp. 5156–5165. PMLR, 2020.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. In *9th Annual Conference on Robot Learning*, 2025. URL http://arxiv.org/abs/2506.10968.

George Killick, Paul Henderson, Paul Siebert, and Gerardo Aragon-Camarasa. Foveation in the era of deep learning. *arXiv preprint arXiv:2312.01450*, 2023.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 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.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, November 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. Towards attention–aware foveated rendering. *ACM TOG*, 42(4):1–10, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Min-Jeong Lee, Hee-Dong Kim, and Seong-Whan Lee. Local representative token guided merging for text-to-image generation. *arXiv preprint arXiv:2507.12771*, 2025.

Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long video understanding. *NeurIPS*, 37:13851–13871, 2024.

Marc Levoy and Ross Whitaker. Gaze-directed volume rendering. In *Proceedings of the 1990 symposium on interactive 3d graphics*, pp. 217–223, 1990.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, 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.

Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 12919–12928, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. ToMA: Token merge with attention for diffusion models. In *ICML*, 2025. URL https://openreview.net/forum?id=51l8tvuIxo.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Silviu Minut, Sridhar Mahadevan, John M Henderson, and Fred C Dyer. Face recognition using foveal vision. In *International Workshop on Biologically Motivated Computer Vision*, pp. 424–433. Springer, 2000.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/. Google DeepMind blog post, August 5, 2025.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM TOG*, 35(6):1–12, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8733–8744, 2025.

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

Martin Reddy. Perceptually optimized 3d graphics. *IEEE computer Graphics and Applications*, 21 (5):68–75, 2001.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *CVPRW*, pp. 4613–4622, 2023.

Jyrki Rovamo and Veijo Virsu. An estimation and application of the human cortical magnification factor. *Experimental brain research*, 37(3):495–510, 1979.

Tanner Schmidt and Richard Newcombe. Segment this thing: Foveated tokenization for efficient point-prompted segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29428–29437, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Xuehuai Shi, Lili Wang, Xinda Liu, Jian Wu, and Zhiwen Shao. Scene-aware foveated neural radiance fields. *IEEE TVCG*, 2024.

Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation. *arXiv preprint arXiv:2501.08994*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *ICML*, pp. 56242–56280, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor. Adaptive image-space sampling for gaze-contingent real-time rendering. *Comput. Graph. Forum*, 35(4):129–139, 2016.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Qi Sun, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman. Perceptually-guided foveation for light field displays. *ACM TOG*, 36(6):1–13, 2017.

Taimoor Tariq, Cara Tursun, and Piotr Didyk. Noise-based enhancement for foveated rendering. *ACM TOG*, 41(4):1–14, 2022.

Ye Tian, Xin Xia, Yuxi Ren, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Yunhai Tong, Ling Yang, and Bin Cui. Training-free diffusion acceleration with bottleneck sampling. *arXiv preprint arXiv:2503.18940*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Andrew B Watson. 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.

Martin Weier, Thorsten Roth, Ernst Kruijff, André Hinkenjann, Arsène Pérard-Gayot, Philipp Slusallek, and Yongmin Li. Foveated real-time ray tracing for head-mounted displays. *Comput. Graph. Forum*, 35(7):289–298, 2016.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *CVPRW*, pp. 7395–7405, 2024.

Frank W Weymouth. Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, 46(1):102–113, 1958.

Haoyu Wu, Jingyi Xu, Hieu Le, and Dimitris Samaras. Importance-based token merging for efficient image and video generation. In *ICCV*, pp. 4983–4995, 2025a.

Haoyu Wu, Jingyi Xu, Qiaomu Miao, Dimitris Samaras, and Hieu Le. One attention, one scale: Phase-aligned rotary positional embeddings for mixed-resolution diffusion transformer. *arXiv preprint arXiv:2511.19778*, 2025b. URL https://arxiv.org/abs/2511.19778v1.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *NeurIPS*, 2025c.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *International Conference on Machine Learning (ICML)*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2023.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11590–11599, 2025.

Chenlu Zhan, Wen Li, Chuyu Shen, Jun Zhang, Suhui Wu, and Hao Zhang. Bidirectional sparse attention for faster video diffusion training. *arXiv preprint arXiv:2509.01085*, 2025.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Ultra-high-resolution image synthesis: Data, method and evaluation, 2025b. URL https://arxiv.org/abs/2506.01331.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In *International Conference on Machine Learning (ICML)*, 2025c.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025d.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric P Xing, and Hao Zhang. Faster video diffusion with trainable sparse attention. In *NeurIPS*, 2025e.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025f.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

APPENDIX

A. **Additional Implementation Details** 18
A.1 Image Generation 18
A.2 Video Generation 18
B. **Additional User Study Details** 19
B.1 Study Design and Participants 19
B.2 Study Setup and Images 19
B.3 Procedure 19
B.4 Statistical Analysis 20
C. **Additional Image Generation Results** 22
C.1 Image Generation with Varying Token Ratios 22
C.2 Extended Image Generation Baseline Comparisons 23
C.3 Extended Image Generation Baseline Comparisons on Distilled Models 32
C.4 Image Generation with Different Foveation Patterns 39
C.5 Towards Saliency-guided Image Generation 43
C.6 Towards Bounding-box-guided Image Generation 49
D. **Additional Video Generation Results** 53
D.1 Extended Video Generation Baseline Comparisons 53
D.2 Video Generation with Different Foveation Patterns 53
D.3 Towards Saliency-guided Video Generation 53
E. **Discussion** 54

A Additional Implementation Details

A.1 Image Generation

For our foveated image generation experiments, we fine-tune the FLUX 2.1 Klein 4B model (Black Forest Labs, 2026) using the Aesthetic-Train-V2 dataset (Zhang et al., 2025a;b). We adopt a Low-Rank Adaptation (LoRA) (Hu et al., 2022) 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 (Zhang et al., 2025a;b). We adopt the standard evaluation protocol in (Dhariwal & Nichol, 2021) and report standard generative metrics including HPSv2.1 (Wu et al., 2023), FID (Heusel et al., 2017), Precision (Kynkäänniemi et al., 2019), and CLIP score (Radford et al., 2021) 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 (Wan Team et al., 2025) using the Vchitect-T2V-Dataverse dataset (Fan et al., 2025; Si et al., 2025). We adopt a Low-Rank Adaptation (LoRA) (Hu et al., 2022) 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 (Huang et al., 2024) 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 (Krajancich et al., 2023). 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 $P = 0\ 5$.

We employ a pseudo–eye-tracked protocol for our user study. 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 (Peirce, 2007), and images were streamed to the display via a wired HDMI connection.

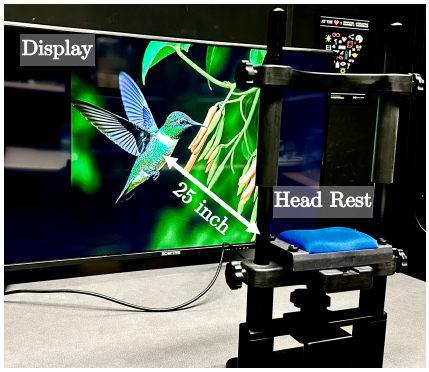

Figure 10: **User study setup.**

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. 10). At this distance, the display subtends approximately 24 pixels per degree (ppd) of visual angle, so each full-screen image spans roughly $45° \times 45°$ of the visual field.

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 (Geisler & Perry, 1998), 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

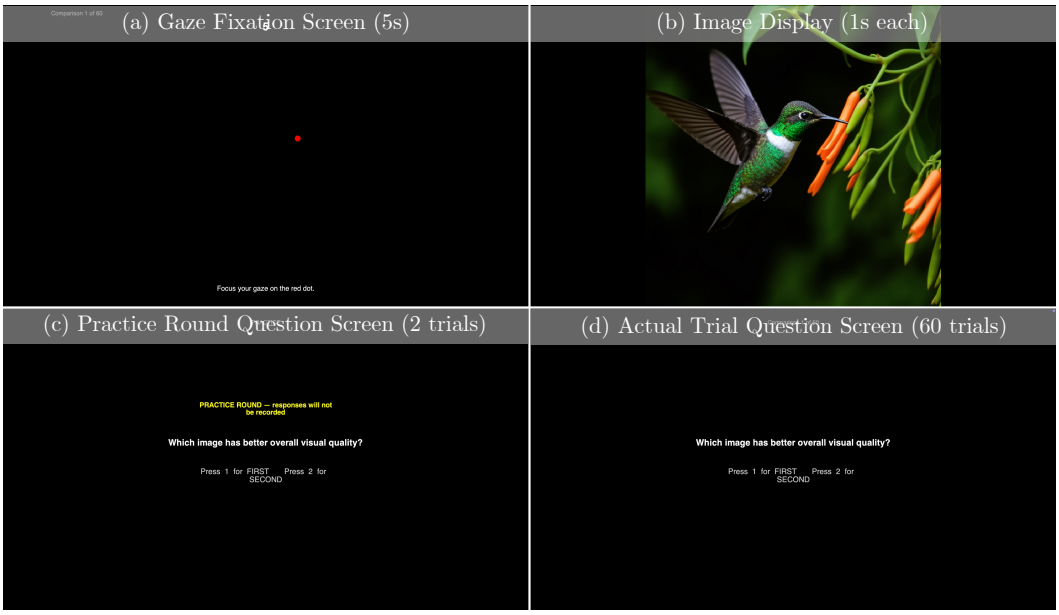

Figure 11: **User study interface.** Each trial began with a red gaze fixation dot displayed for 5 seconds (**a**), followed by a pair of images, each displayed for 1 s (**b**). Participants answered visual quality questions after both images were shown. Participants completed two practice trials (**c**) before beginning 60 data collection trials (**d**).

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 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. 11).

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 comparison, 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 comparison contains 220 data points (votes) in total, and equal preference corresponds to $P = 0\ 5$. We report a 95% Wilson score interval with each rate.

Significance testing. To test whether a pairwise preference rate differs significantly from chance, we apply a two-sided binomial test under the null hypothesis $H_0 \colon P = 0\ 5$ (equal preference). A result is considered statistically significant at the $\alpha = 0\ 05$ level. A p-value above $0\ 05$ for a pair indicates that we detect no significant preference between the two methods; a p-value below $0\ 05$ indicates a statistically significant preference for one method over the other.

Table 3 reports the preference rates, confidence intervals, and p-values for all three pairwise conditions. Foveated Diffusion is preferred over full high-resolution generation in $47\ 3\%$ of trials, with an interval of $[40\ 8\ \ 53\ 9]$ that brackets chance and bounds any residual preference for either method at

Pair	Preference P (%)	95% CI (%)	p-value
Ours vs. Full high-res	47.3	$[40.8,\ 53.9]$	0.4829
Ours vs. Naïve mixed-res	**87.4**	$[\mathbf{82.4},\ \mathbf{91.2}]$	< 0.0001
Full high-res vs. Naïve mixed-res	90.8	$[86.2,\ 94.0]$	< 0.0001

Table 3: **User study statistical analysis.** Preference rate for the first-listed method in each pair (higher = more preferred), over 220 votes per condition from $N = 11$ participants. Intervals are 95% Wilson score intervals; p-values are from a two-sided binomial test against $P = 0\ 5$.

roughly 9 percentage points. Both our method and full high-resolution generation are significantly preferred over the naïve baseline (87 4% [82 4 91 2] and 90 8% [86 2 94 0]), due to visual artifacts in the naïve baseline generations (main paper Fig. 4, Fig. 5).

C Additional Image Generation Results

C.1 Image Generation with Varying Token Ratios

Method	Token Ratio	Image-Reward ↑	HPS v2.1 ↑	HPS v3 ↑	CLIP Score ↑	LAION Aes. ↑	FID ↓	LPIPS ↓	Time / Speedup
Full high-res	100%	0.964	0.280	10.32	0.852	5.574	11.56	0.730	41.6s / 1.00×
ToMe (Bolya & Hoffman, 2023)	54%	0.447	0.242	7.12	0.843	5.485	18.84	0.764	25.9s / 1.61×
RALU (Jeong et al., 2026)		*0.832*	0.264	9.23	*0.858*	*5.508*	12.60	0.746	26.9s / 1.55×
Naïve mixed-res		0.804	*0.270*	*9.64*	0.857	5.501	**10.40**	*0.742*	26.0s / 1.60×
Ours		**0.974**	**0.279**	**10.92**	**0.858**	**5.574**	*11.30*	**0.721**	26.1s / 1.60×
ToMe (Bolya & Hoffman, 2023)	32%	−0.698	0.181	1.34	0.784	5.014	40.34	0.799	18.2s / 2.28×
RALU (Jeong et al., 2026)		*0.854*	0.267	9.40	*0.861*	*5.526*	12.50	*0.744*	18.9s / 2.19×
Naïve mixed-res		0.823	*0.270*	*10.41*	0.858	5.470	**10.98**	0.753	18.6s / 2.24×
Ours		**0.996**	**0.279**	**12.10**	**0.862**	**5.586**	*11.51*	**0.727**	18.5s / 2.24×
ToMe (Bolya & Hoffman, 2023)	26%	−1.172	0.154	−3.46	0.739	4.603	55.84	0.810	16.0s / 2.60×
RALU (Jeong et al., 2026)		0.872	0.269	9.55	*0.859*	*5.535*	12.67	*0.747*	16.8s / 2.47×
Naïve mixed-res		*0.918*	*0.275*	*11.16*	0.859	5.490	**11.73**	0.756	16.5s / 2.53×
Ours		**1.013**	**0.281**	**12.48**	**0.863**	**5.603**	*11.97*	**0.729**	16.5s / 2.52×

Table 4: **Extended Image Generation Baseline Comparisons with Varying Token Ratios.**

Foveated Diffusion outperforms every baseline on six of seven metrics at all three token ratios (the ratio between the number of mixed-resoluton tokens in our generated image and the number of high-resolution tokens in the reference image), and matches or exceeds full high-resolution generation (top row) on human preference, prompt alignment, and aesthetic quality. All methods share the same foveation masks, base model, sampler, guidance scale, prompts, and fixation points, and are scored over $10\,000$ samples with randomized fixation points. The full high-resolution row is the reference baseline and is excluded from the per-token-ratio ranking. Time is the average end-to-end wall-clock per image, covering both the DiT denoising loop and VAE decoding, and speedup is the ratio of that time to the end-to-end time of full high-resolution generation; all baselines are time-matched to ours within each token ratio. Within each token ratio, the best result per metric is shown in **bold** and the second best in *italics*.

We observe that FID may not be a reliable metric for evaluating foveated visual generation, as the naïve baseline, despite exhibiting clear structural artifacts even outperforms full high-resolution generation. Setting FID aside, Foveated Diffusion surpasses both all mixed-resolution baselines on all seven metrics with increasing speedups as token ratio decreases, reaching $2\ 52\times$ speedup with a 26% token ratio.

C.2 Extended Image Generation Baseline Comparisons

In Figures 12-19, 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 $1\ 85\times$ speedup in image generation time.

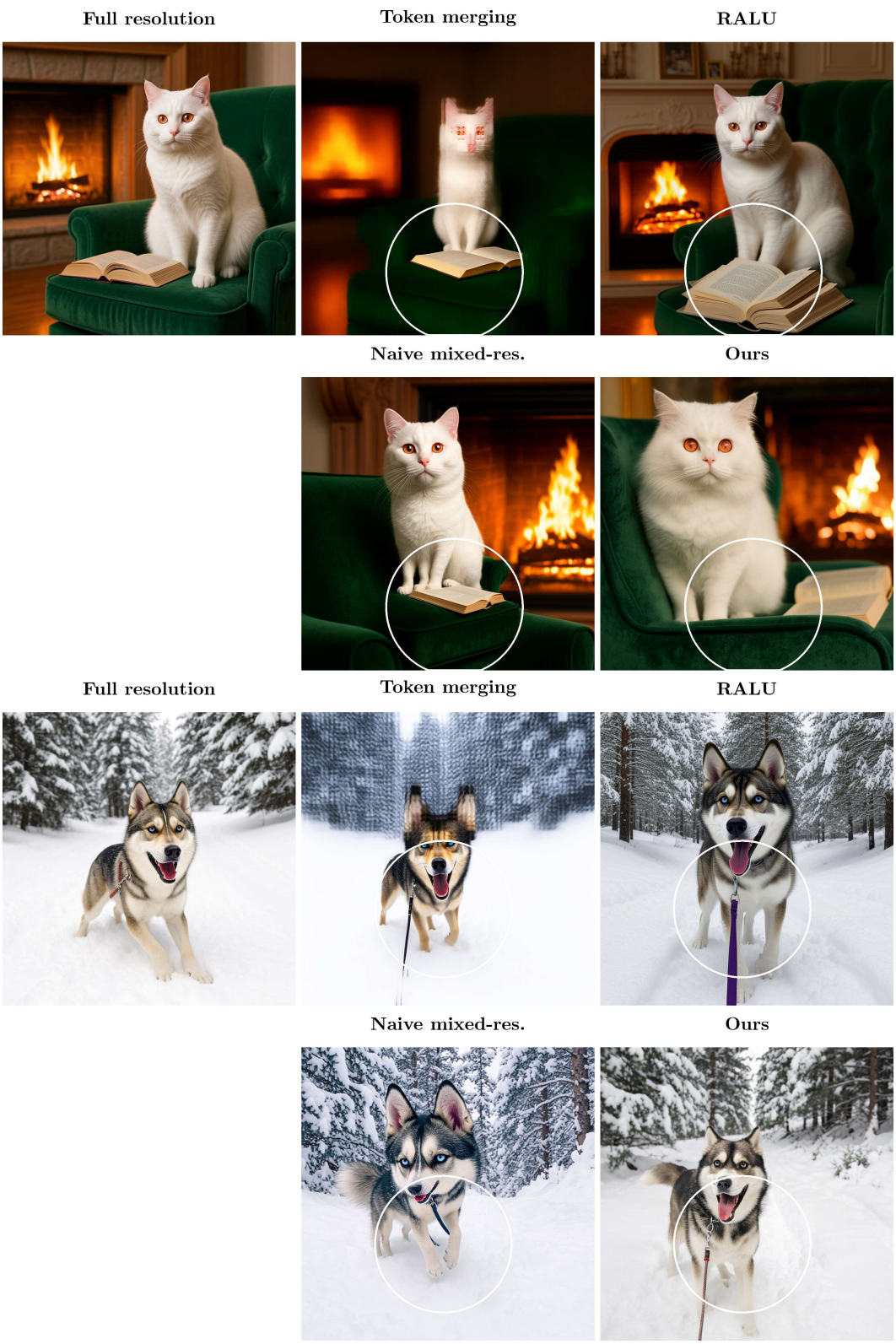

Figure 12: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results. All images are uncompressed in this figure.

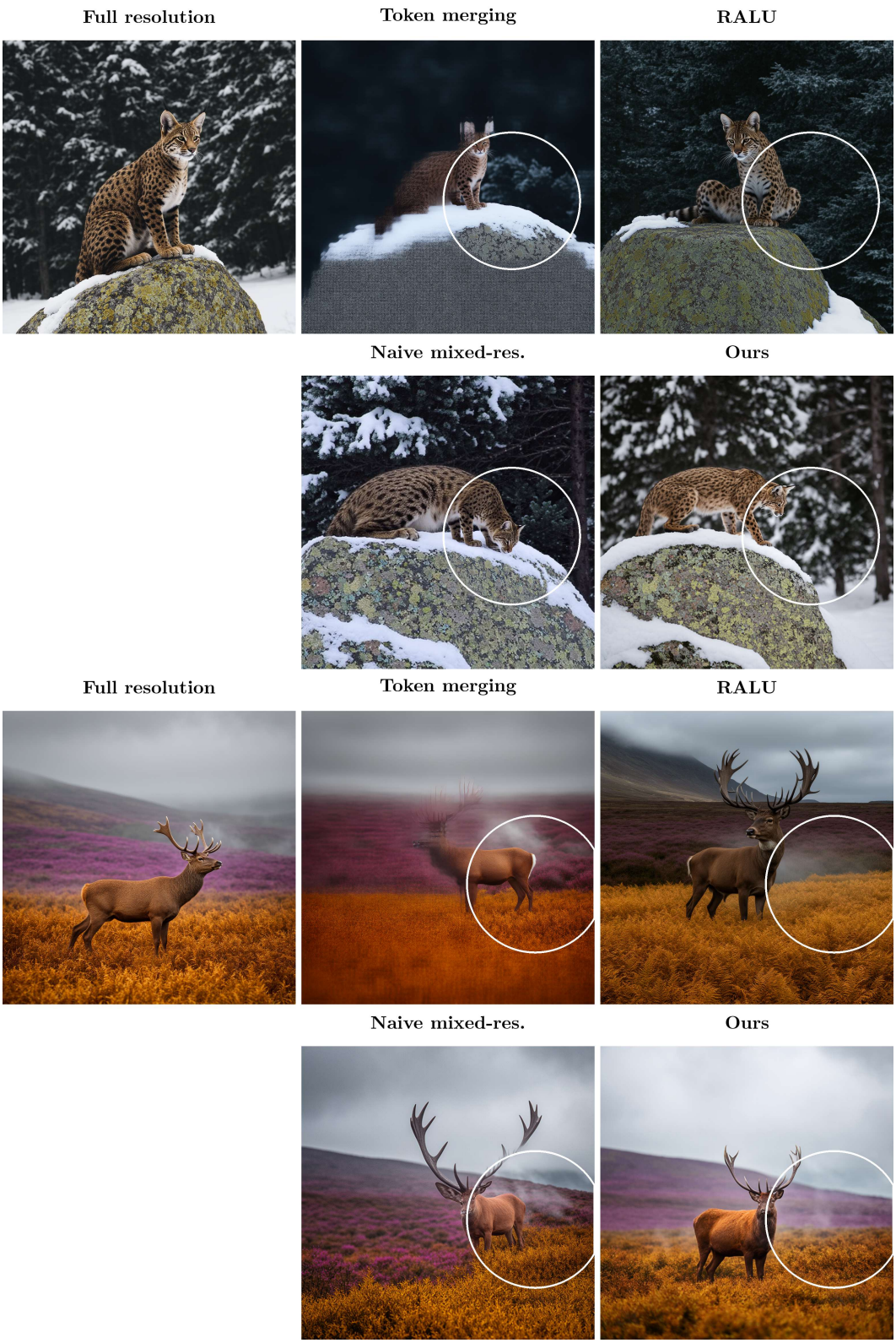

Figure 13: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results. All images are uncompressed in this figure.



Figure 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, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

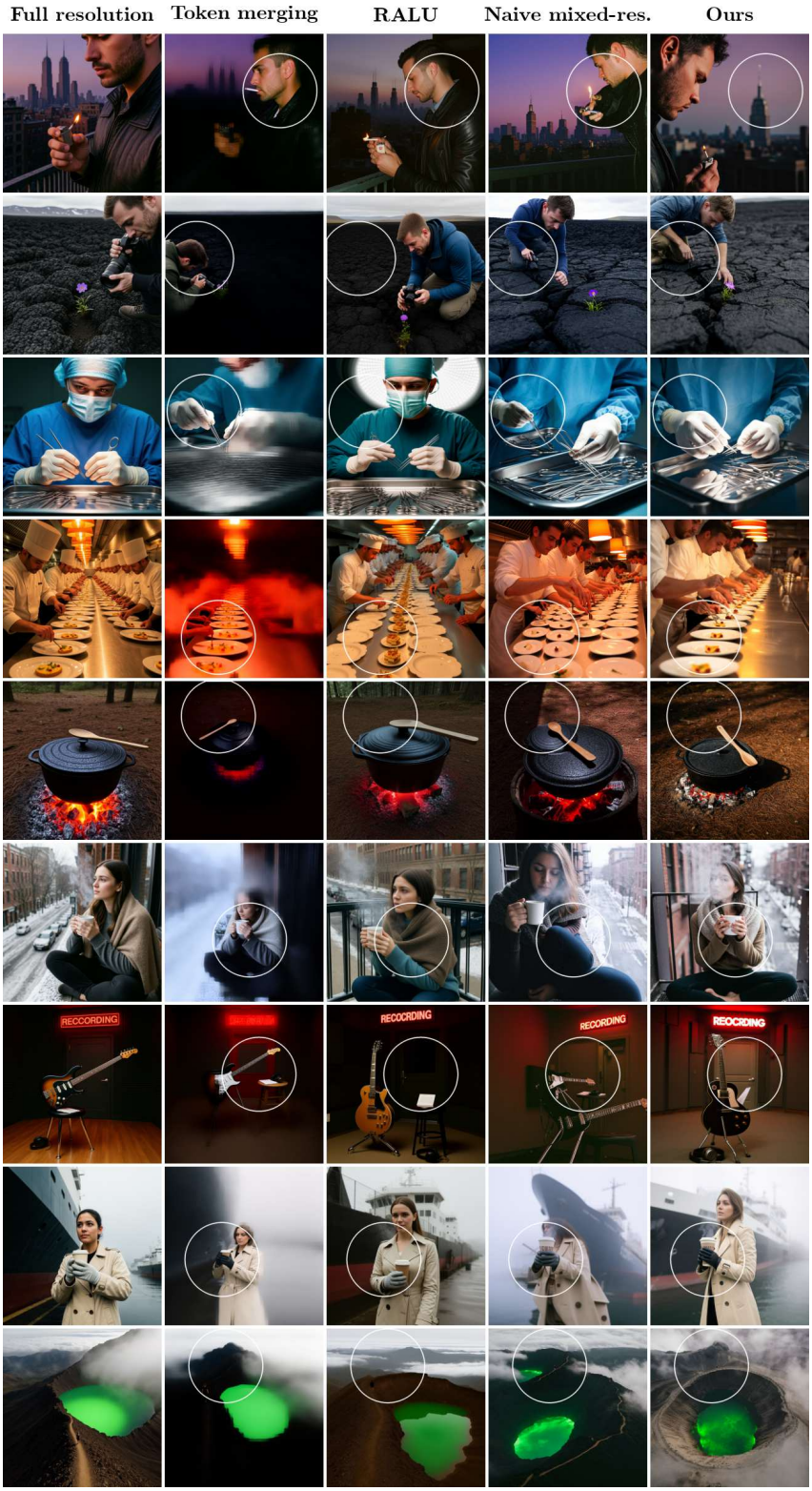

Figure 15: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

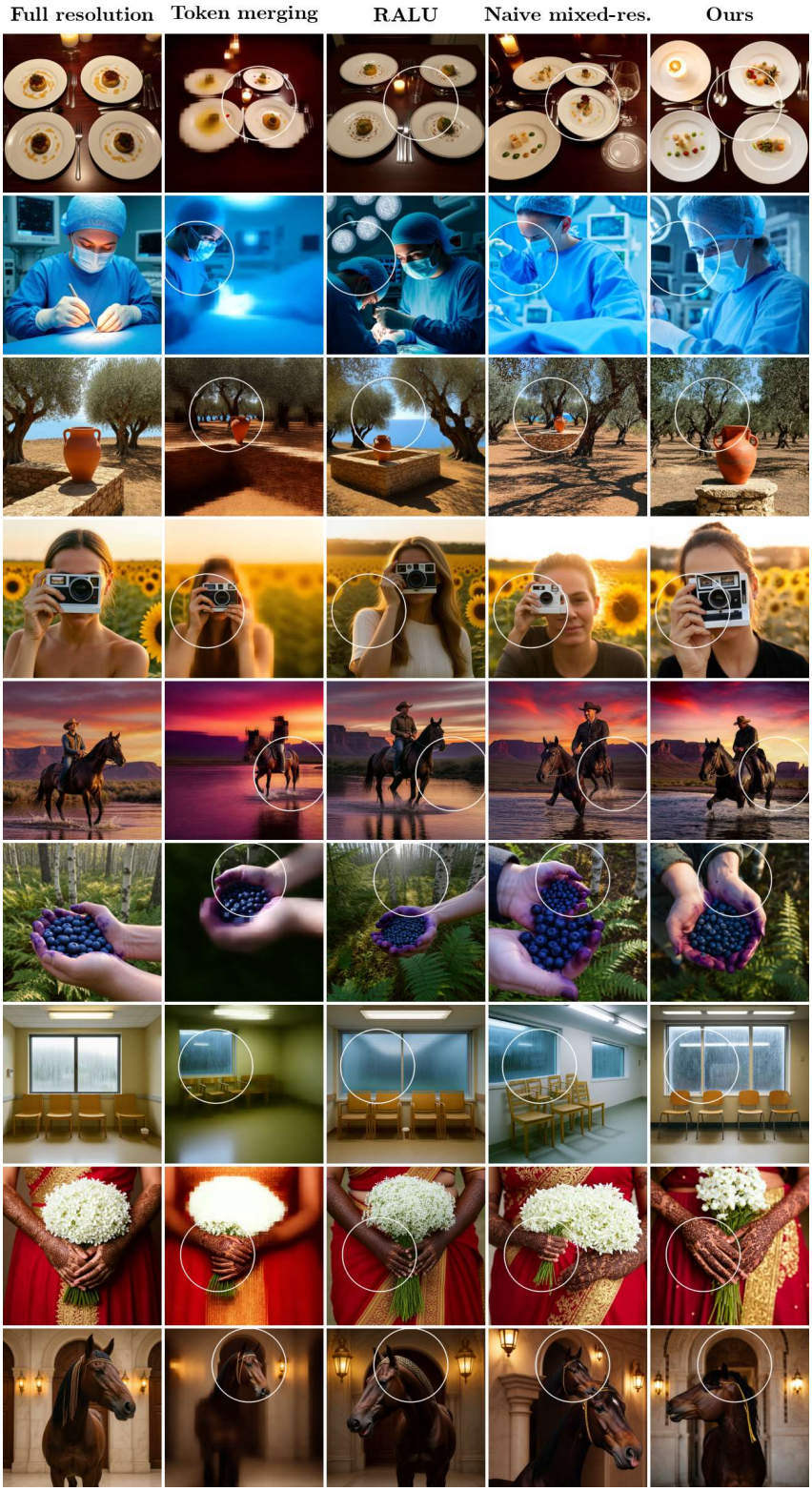

Figure 16: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

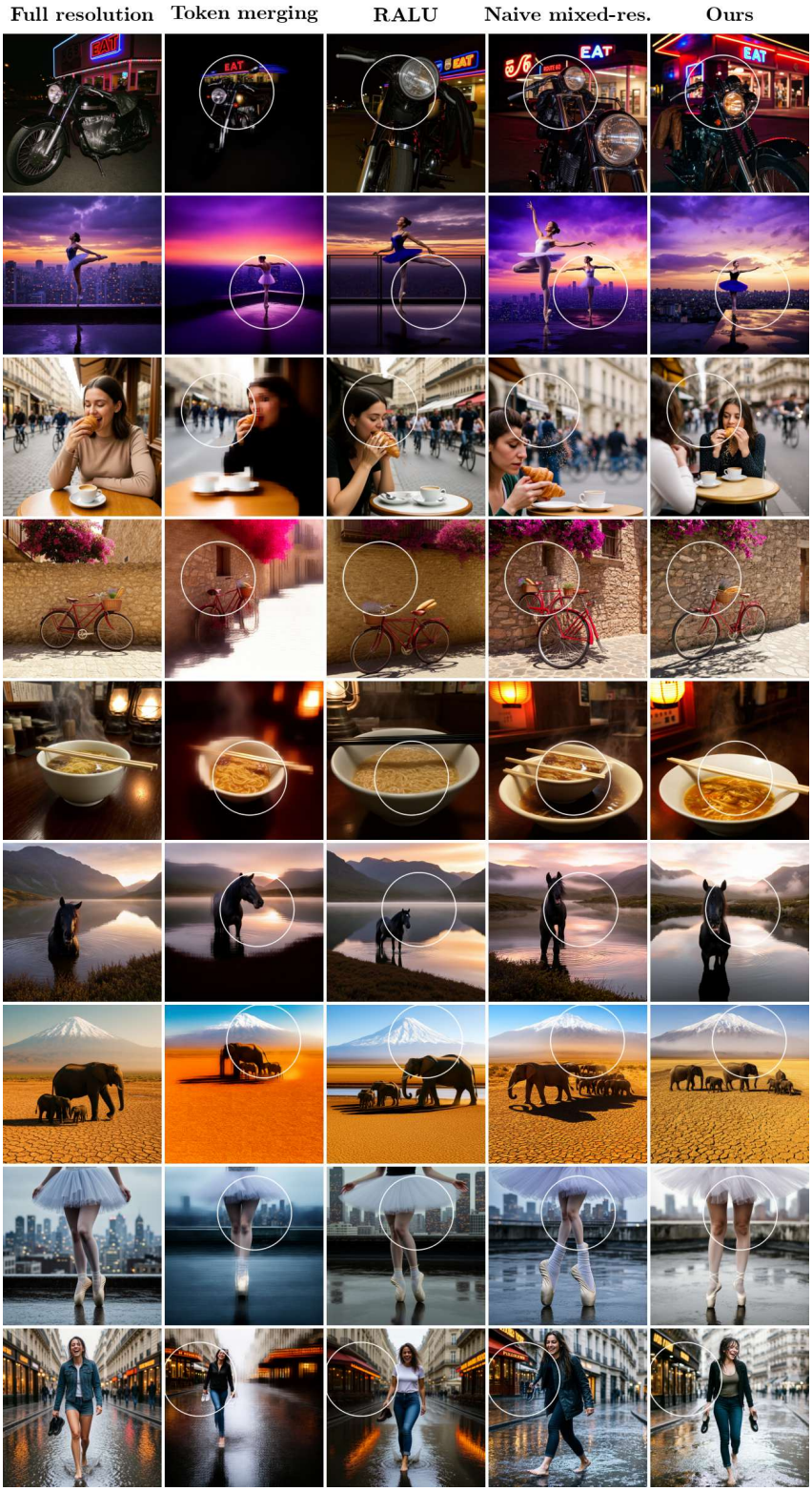

Figure 17: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

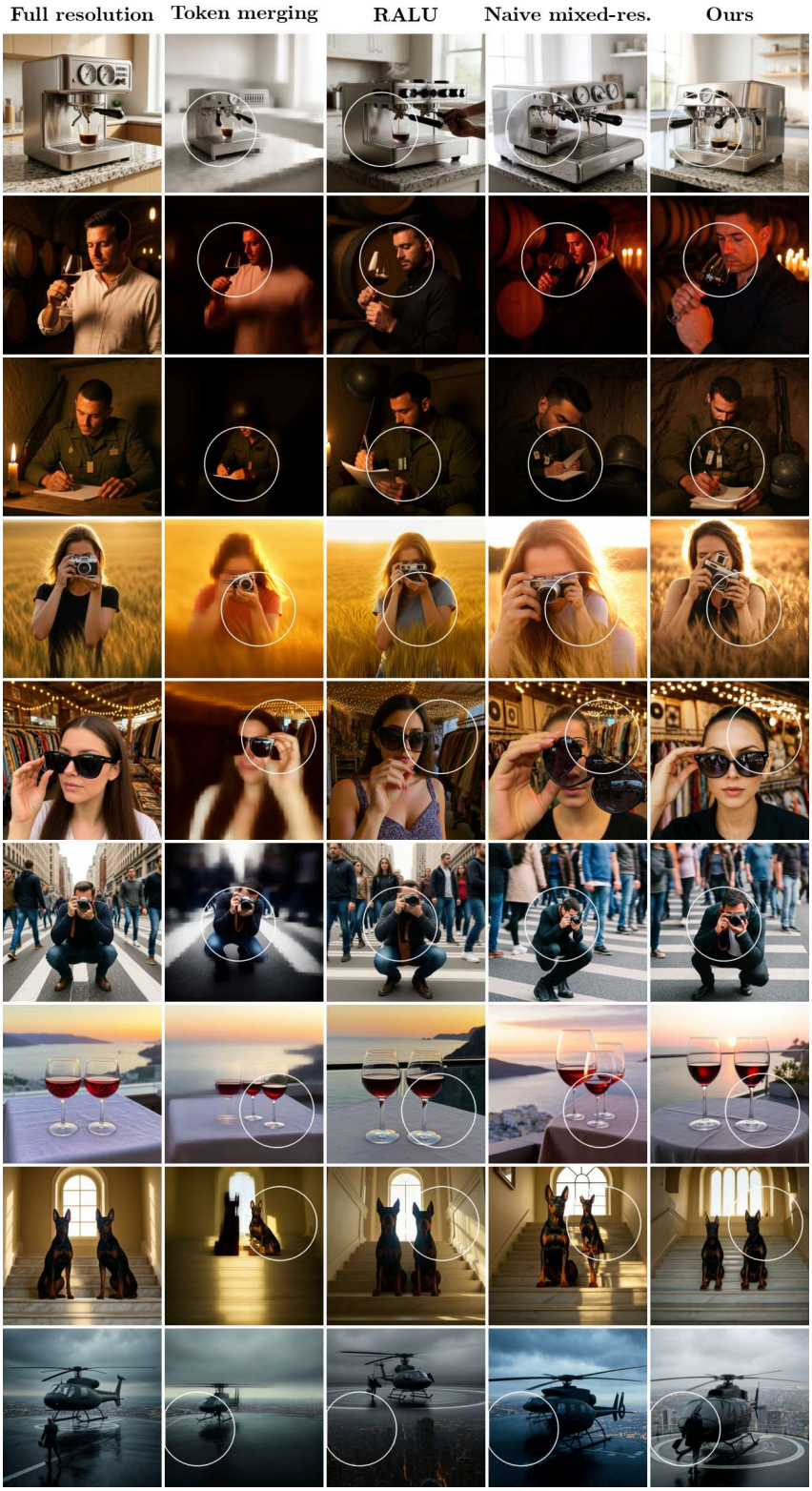

Figure 18: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

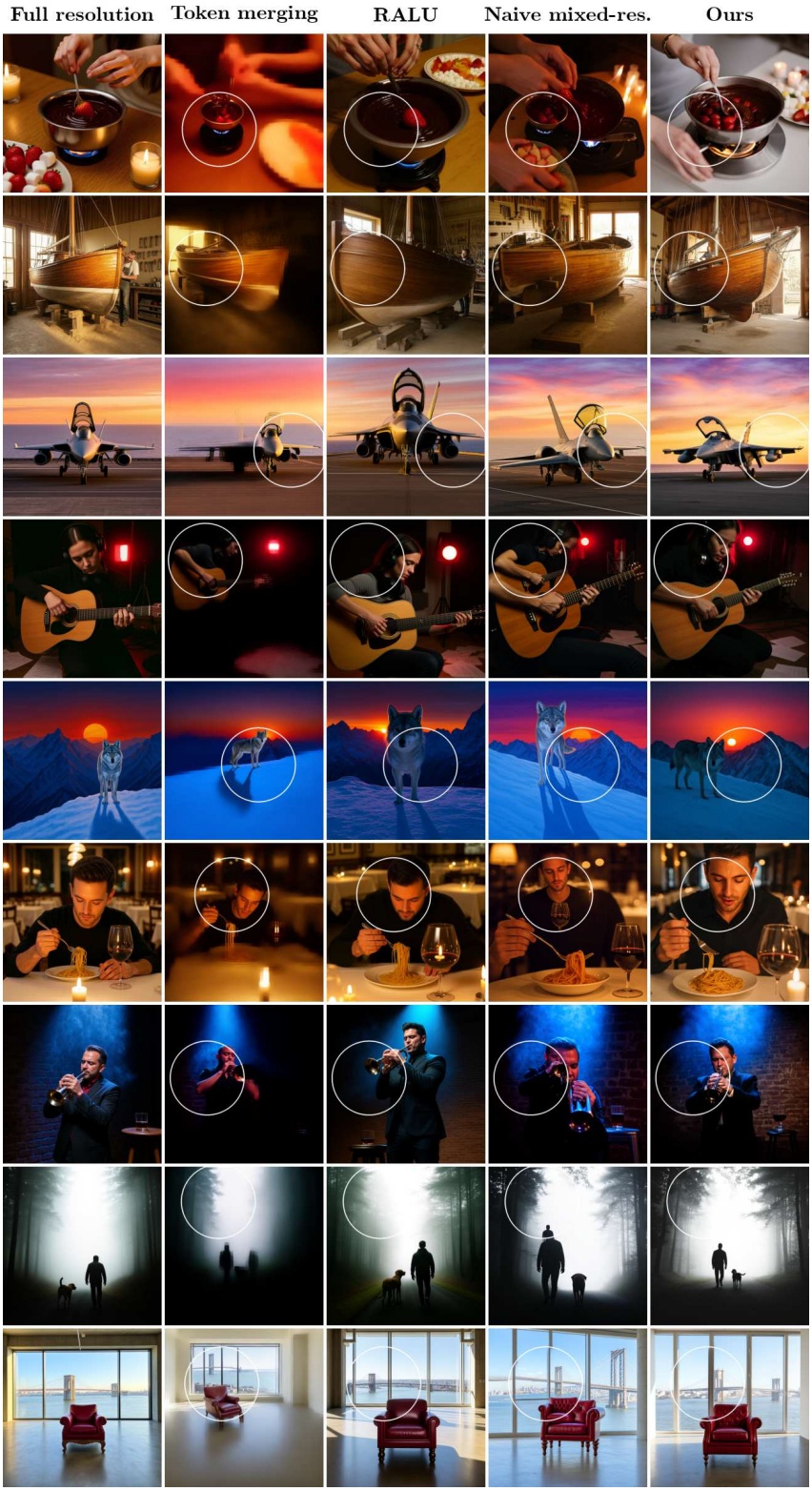

Figure 19: **Extended baseline comparisons.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation, whereas the naïve mixed-resolution baseline exhibits severe scale mismatches and structural inconsistencies across resolutions, and token merging leads to overly blurry results in the periphery. RALU generates plausible results.

C.3 Extended Image Generation Baseline Comparisons on Distilled Models

Method	Token Ratio	Image-Reward ↑	HPS v2.1 ↑	HPS v3 ↑	CLIP Score ↑	LAION Aes. ↑	FID ↓	LPIPS ↓	Time / Speedup
Full high-res	100%	1.063	0.288	11.46	0.846	5.547	12.34	0.706	1.83s / 1.00×
ToMe (Bolya & Hoffman, 2023)	54%	0.801	0.265	8.84	*0.855*	5.451	13.76	0.723	1.24s / 1.48×
RALU (Jeong et al., 2026)		*0.983*	0.278	10.54	**0.856**	*5.541*	12.84	**0.710**	1.22s / 1.51×
Naïve mixed-res		0.912	*0.280*	*10.71*	0.853	5.491	**11.96**	0.725	1.30s / 1.41×
Ours		**1.027**	**0.286**	**11.53**	0.853	**5.556**	*12.52*	*0.717*	1.30s / 1.41×
ToMe (Bolya & Hoffman, 2023)	32%	0.376	0.230	5.41	0.846	5.165	19.85	0.733	0.93s / 1.97×
RALU (Jeong et al., 2026)		*1.010*	0.281	10.69	**0.858**	*5.550*	*12.87*	**0.709**	1.06s / 1.73×
Naïve mixed-res		0.967	*0.281*	*11.80*	0.855	5.484	**12.56**	0.736	1.00s / 1.83×
Ours		**1.055**	**0.286**	**12.80**	*0.856*	**5.561**	13.21	*0.723*	1.00s / 1.83×
ToMe (Bolya & Hoffman, 2023)	26%	0.181	0.220	3.55	0.837	5.002	24.57	0.734	0.84s / 2.17×
RALU (Jeong et al., 2026)		1.021	0.282	10.82	0.856	*5.562*	**12.99**	**0.710**	1.02s / 1.80×
Naïve mixed-res		*1.052*	*0.286*	*12.52*	**0.857**	5.512	*13.49*	0.739	0.92s / 1.99×
Ours		**1.074**	**0.287**	**13.12**	*0.857*	**5.577**	13.82	*0.725*	0.92s / 1.99×

Table 5: **Extended Image Generation Baseline Comparisons on Distilled Models.**

The table above reports quantitative results on the step- and guidance-distilled FLUX.2-klein-4B model, where all five arms run at 4 denoising steps with CFG $= 1\ 0$, at three token ratios. All methods share the same foveation masks, base model, sampler, guidance scale, prompts, and fixation points, and are scored over $10\ 000$ samples with randomized fixation points; within each token ratio the best result per metric is shown in **bold** and the second best in *italics*, and the full high-resolution row is the reference baseline and is excluded from the ranking. Time is the average end-to-end wall-clock per image, covering both the DiT denoising loop and VAE decoding, and speedup is the ratio of that time to the end-to-end time of full high-resolution generation; RALU is run in its time-matched $N = [1\ 2\ 1]$ configuration, which at the two smaller token ratios is still slower than ours and therefore receives more compute.

Foveated Diffusion is preferred by both human-preference metrics, ImageReward and HPSv2.1, at all three token ratios, also leads HPSv3 and aesthetic quality, and at 32% and 26% exceeds full high-resolution generation on human preference. RALU leads CLIPScore and LPIPS, and FID favours the naïve mixed-resolution baseline: these metrics score the full frame, so a method that restores a globally high-resolution output is not penalized for a periphery that ours renders at low resolution by design. In Figure 20 to 25, we provide extended qualitative comparisons on the same model. The high-resolution region defined by the foveation mask is a circle with radius $0\ 33$ relative to the image diagonal and a randomly placed center. Our method consistently outperforms the naïve mixed-resolution baseline and token merging (Bolya & Hoffman, 2023), while RALUJeong et al. (2026) generates plausible results. These artifacts stem from the spatial resampling itself rather than from the length of the denoising trajectory, so shortening the trajectory does not remove them.

Most importantly, our method achieves perceptually indistinguishable quality from the full high-resolution distilled baseline at a 32% token ratio, resulting in an additional $1\ 83\times$ speedup. Notably, our foveation LoRA is trained on the undistilled model and is plugged into the distilled model untouched, with no retraining and no degradation relative to the undistilled setting. Step distillation shortens the denoising trajectory while foveation reduces the tokens processed at each step, which is an orthogonal axis whose speedups compose multiplicatively, so Foveated Diffusion delivers a further $1\ 83\times$ on top of an already heavily accelerated 4-step, guidance-free model ($41\ 6\times$ total speedup compared to to un-distilled, full-res model in Table 4) at no additional training cost.



Figure 20: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.

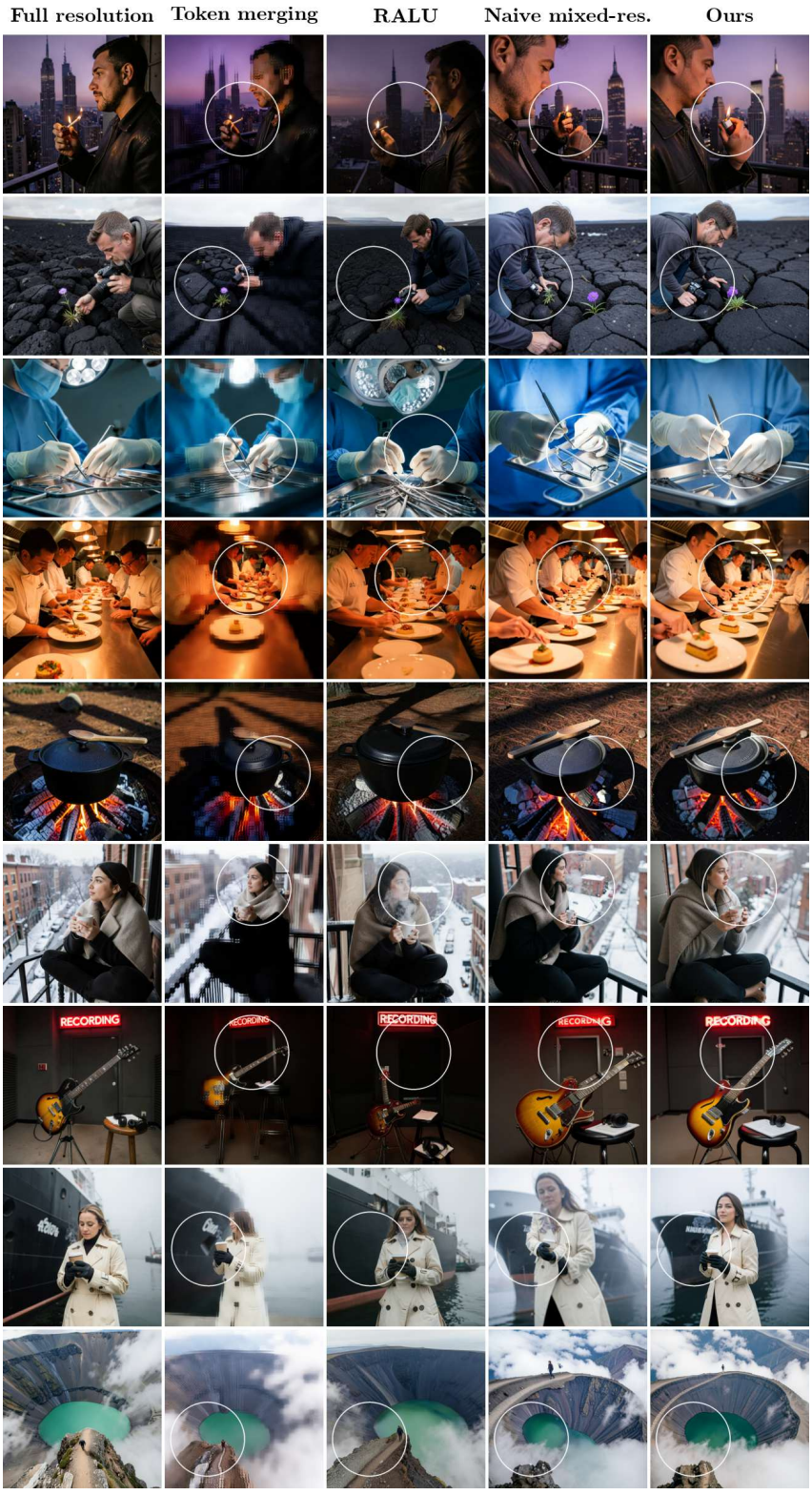

Figure 21: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.

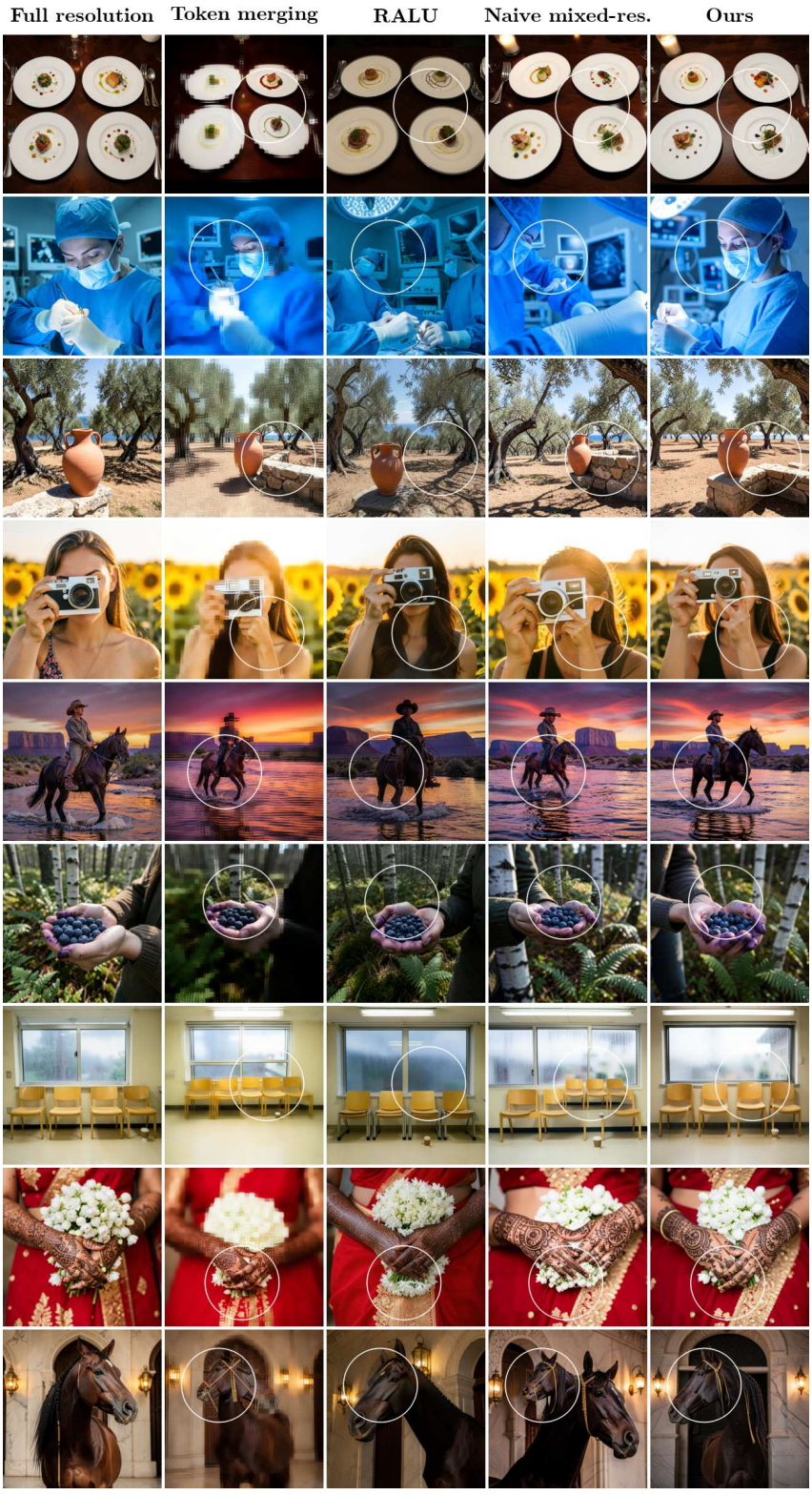

Figure 22: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.



Figure 23: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.



Figure 24: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.

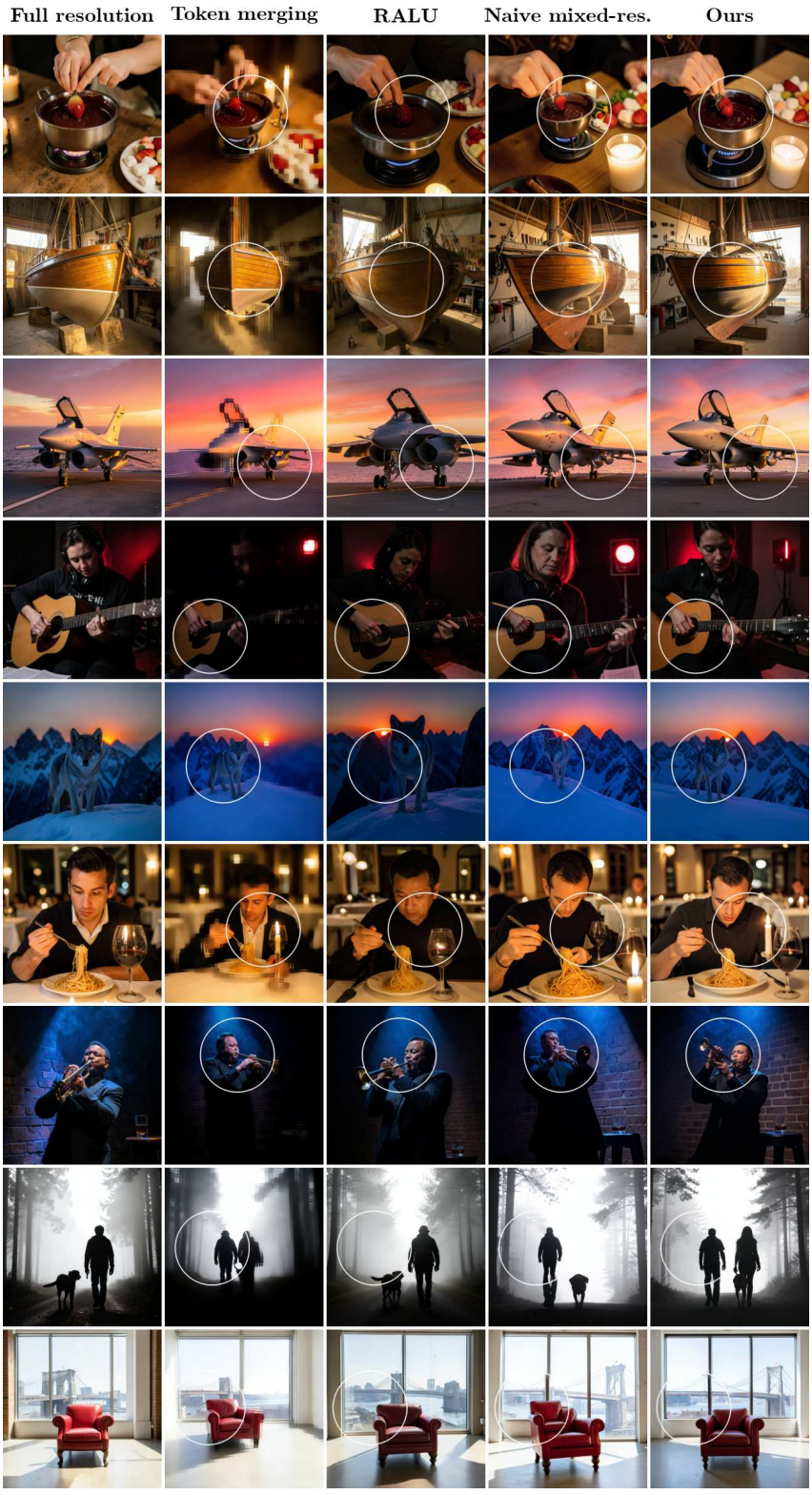

Figure 25: **Extended baseline comparisons on distilled models.** Foveated Diffusion (ours) produces perceptually similar quality to full high-resolution generation while transferring to distilled models, while other baselines exhibit similar artifacts as their un-distilled version.

C.4 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. 26–28). 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. 28) suggest that Foveated Diffusion could support multi-viewer scenarios with multiple gaze locations.

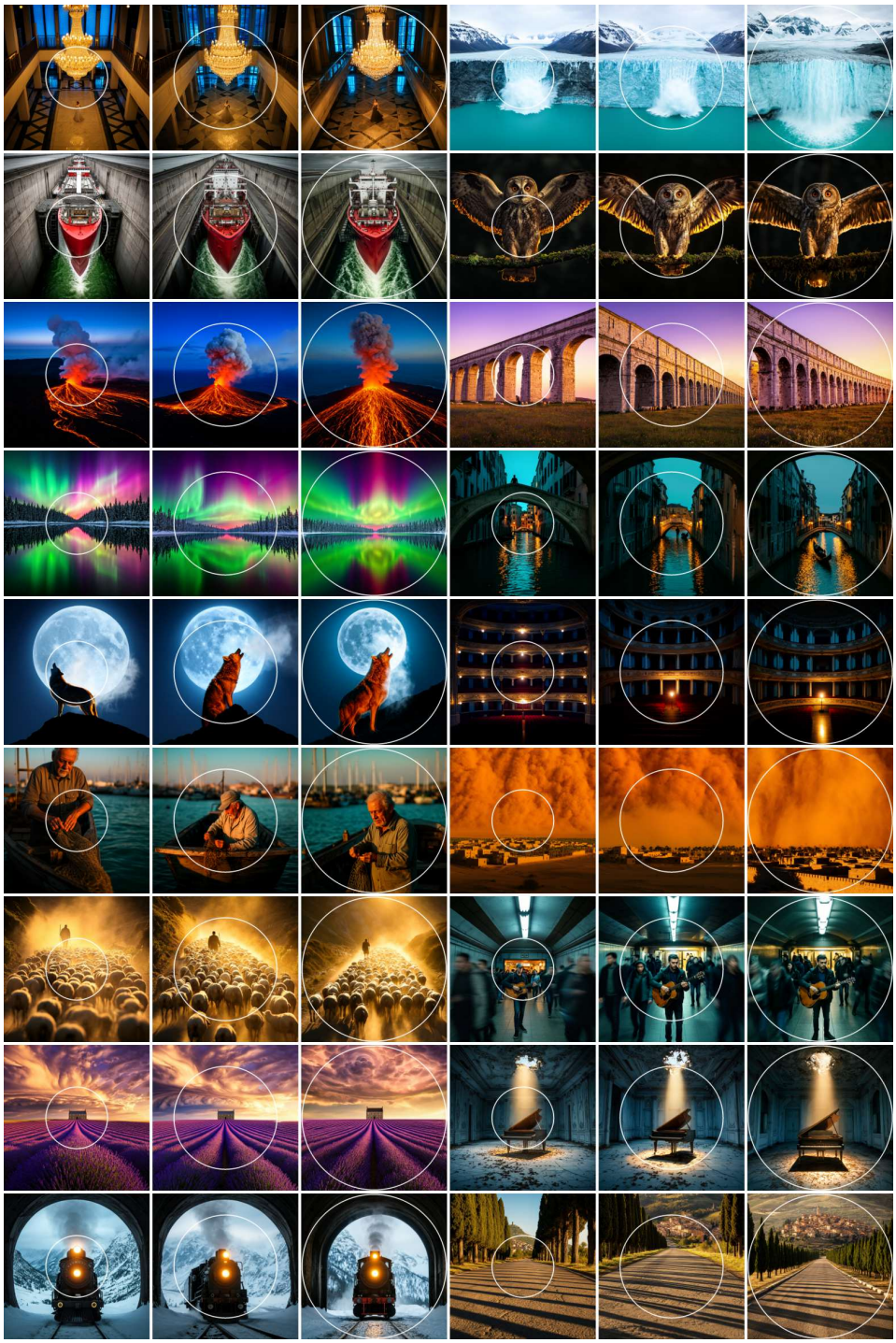

Figure 26: **Varying foveation mask radius.** Foveated Diffusion generates coherent content across varying foveation mask radii. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

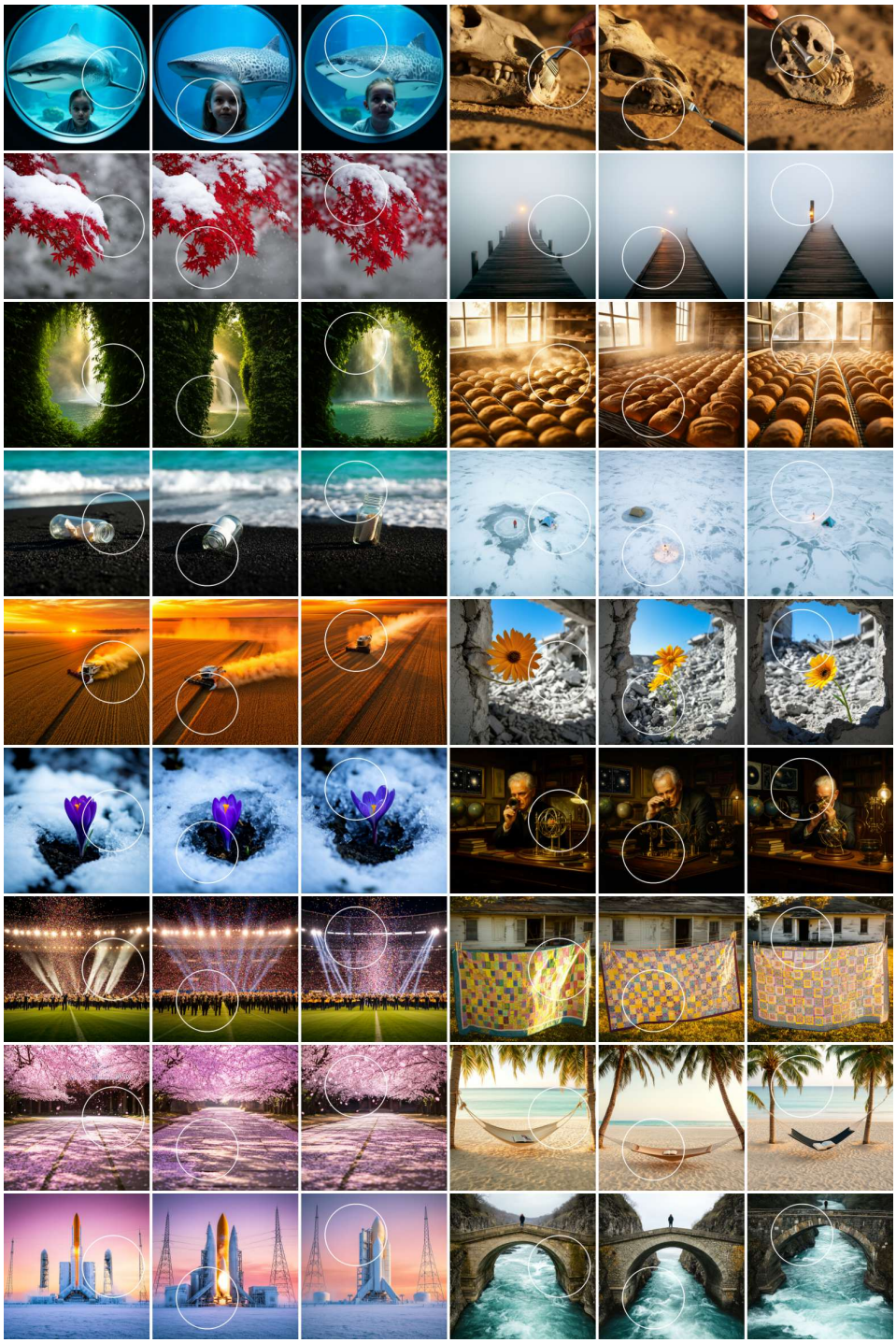

Figure 27: **Varying foveation mask position.** Foveated Diffusion generates coherent content across varying foveation mask positions. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

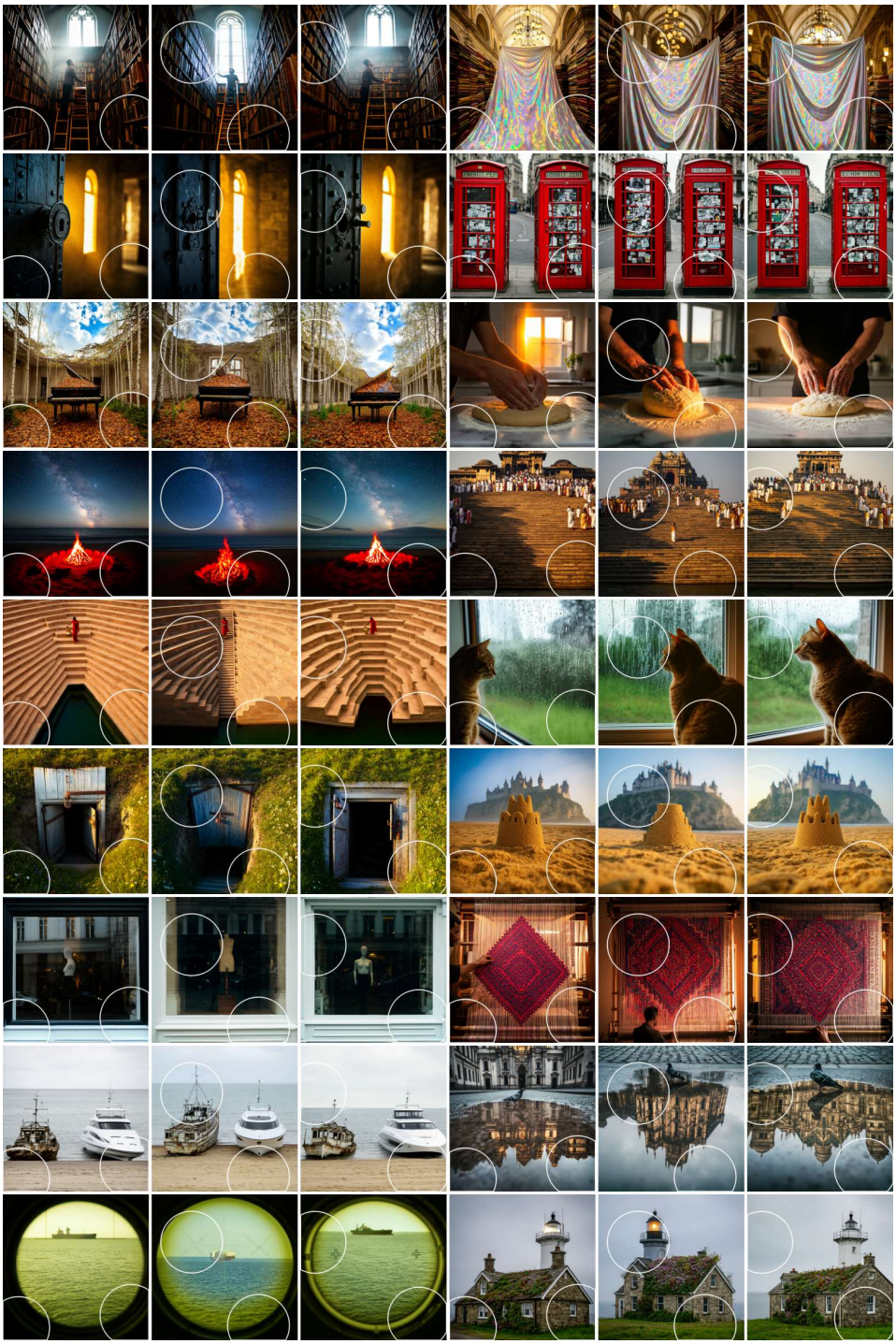

Figure 28: **Foveation mask containing multiple disjoint regions.** Foveated Diffusion generates coherent content with foveation masks containing multiple disjoint regions. Given the same prompt and noise seed, the generated images remain consistent and adhere to the prompt.

C.5 TOWARDS SALIENCY-GUIDED IMAGE GENERATION

We show additional Foveated Diffusion additional results using the *saliency-guided* model in Figures 29 to 33. 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 31–33, we illustrate potential applications of saliency-guided Foveated Diffusion, including immersive VR gaming (Fig. 31), generative robotics (Fig. 32), and autonomous driving simulation for robotics policy learning (Fig. 33). 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.

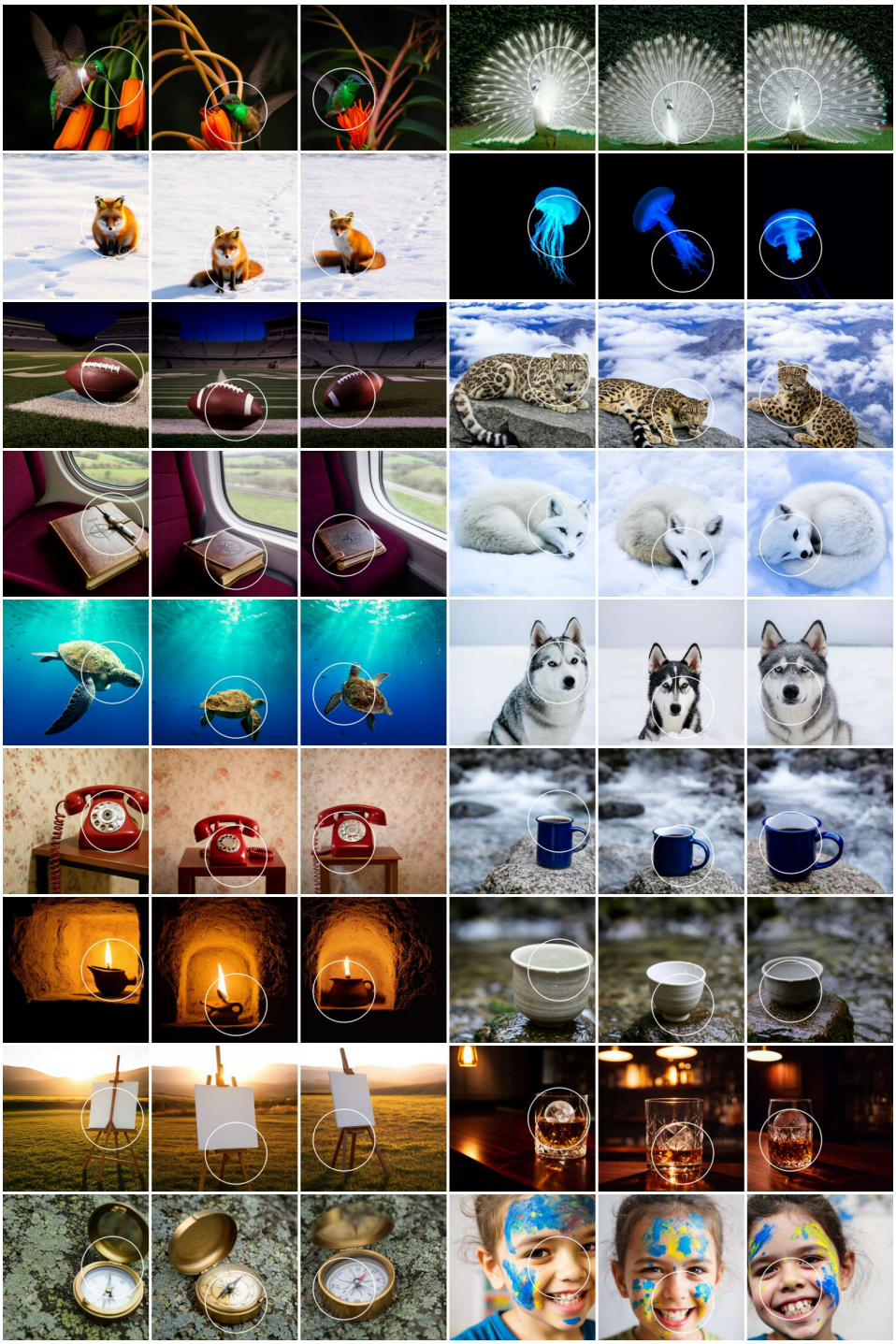

Figure 29: **Single-object saliency-guided generation.** Foveated Diffusion with saliency-guided training enables coarse, controllable single-object generation, where the salient object is approximately aligned with the center of the foveation mask.

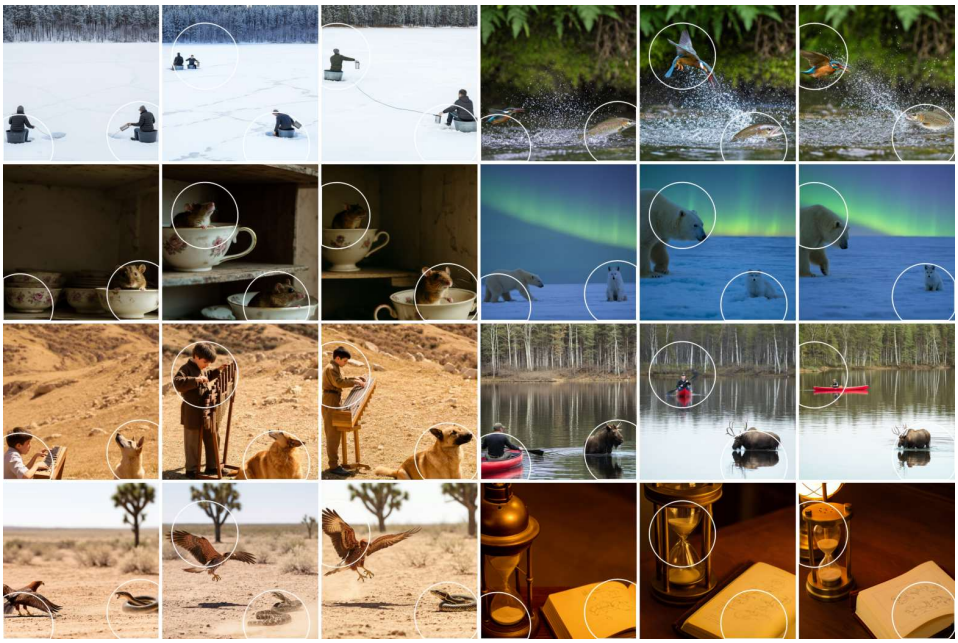

Figure 30: **Multi-object saliency-guided generation.** Foveated Diffusion with saliency-guided training also enables coarse, controllable multi-object generation, where salient objects approximately align with the centers of the disjoint regions in the foveation mask.

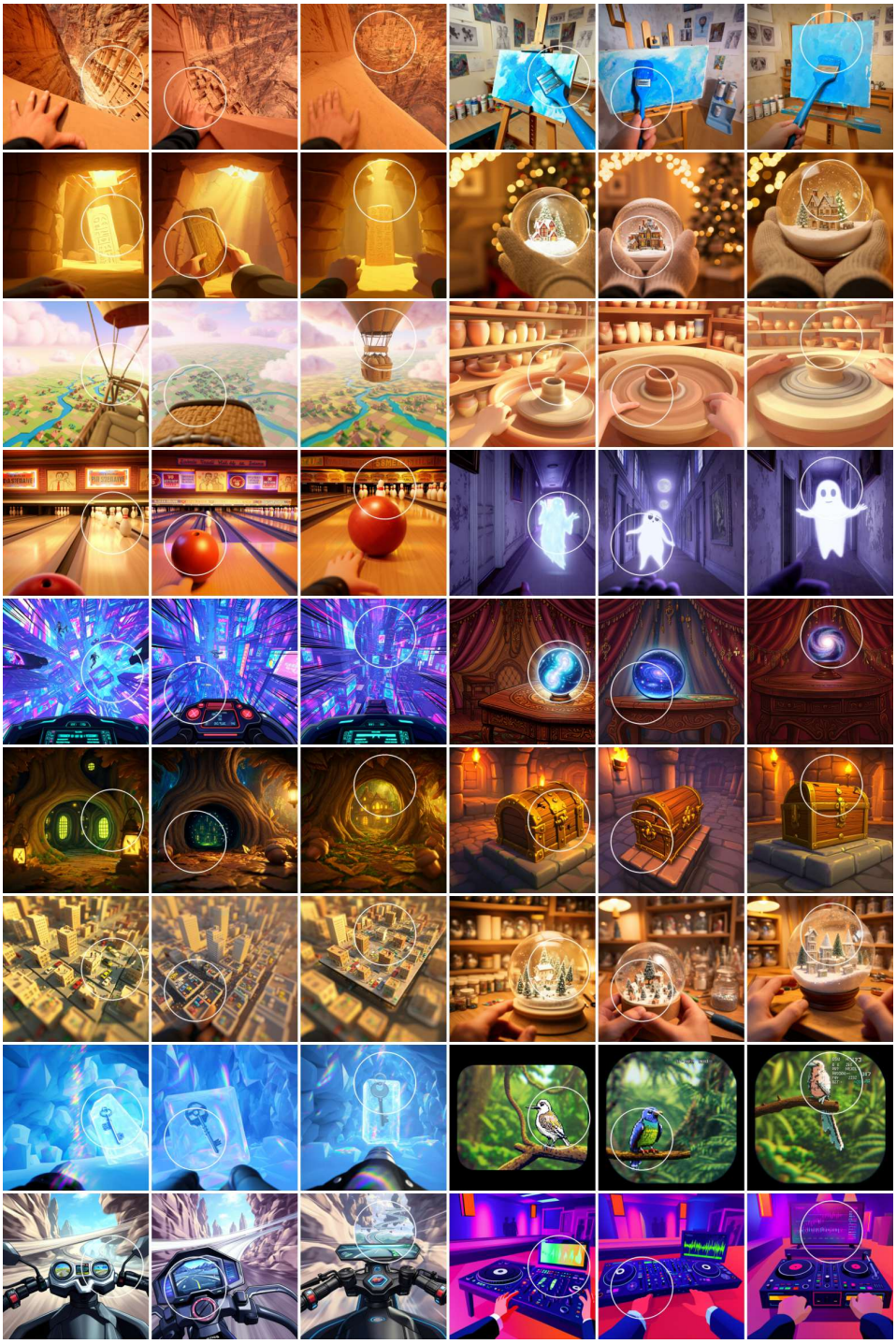

Figure 31: **Saliency-guided generation for immersive gaming.** Foveated Diffusion is well suited for immersive first-person generative gaming applications, where salient objects can be generated near the gaze-tracked location (fovea) and rendered in high resolution, while the remaining regions are rendered at lower resolution.

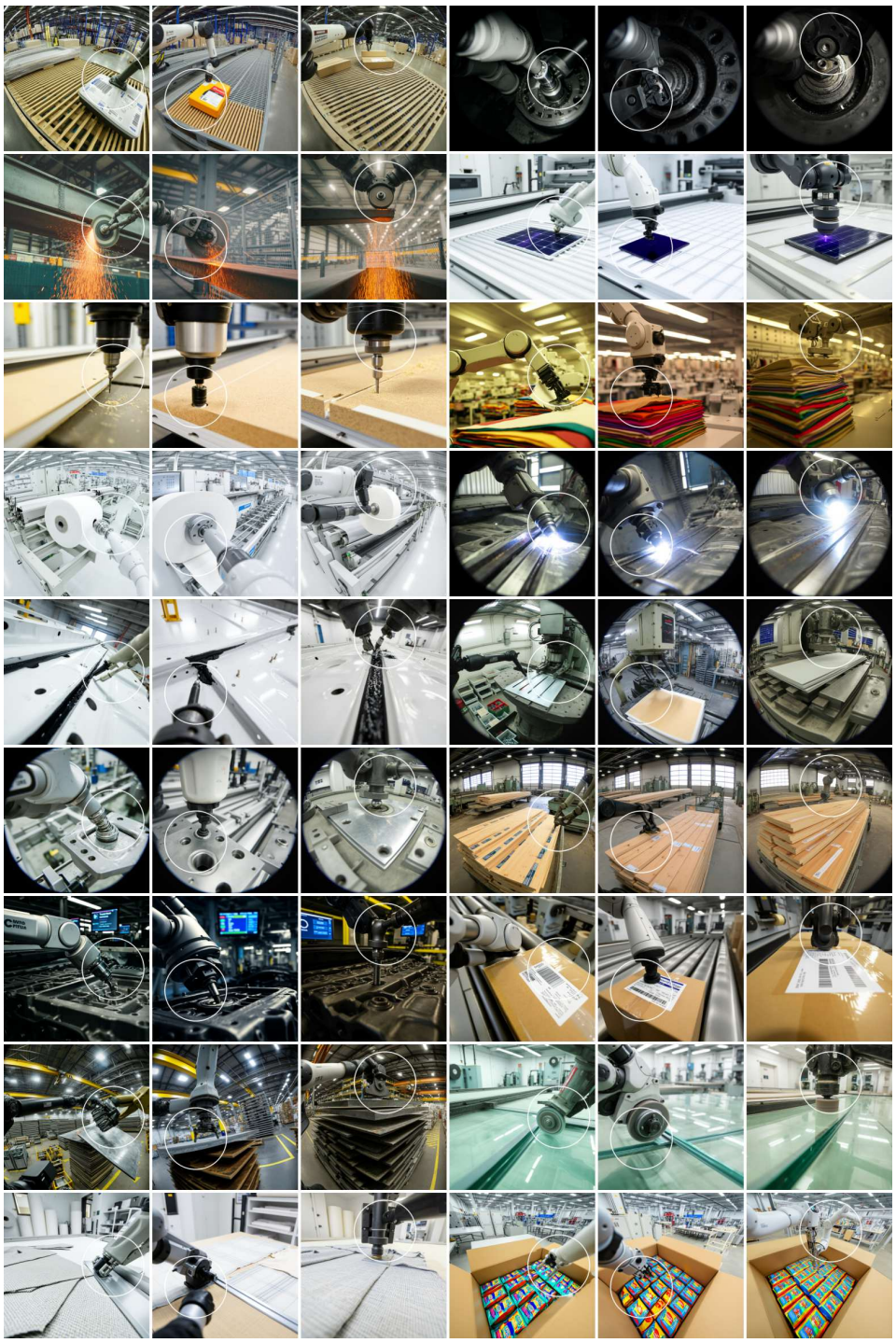

Figure 32: **Saliency-guided generation for robotics simulation.** Foveated Diffusion is well suited for generative robotics simulation, where foveated imagery can be used for robotics policy learning. In this setting, only robot arms and manipulated objects are generated at high resolution, while the background is rendered at lower resolution to provide global context.

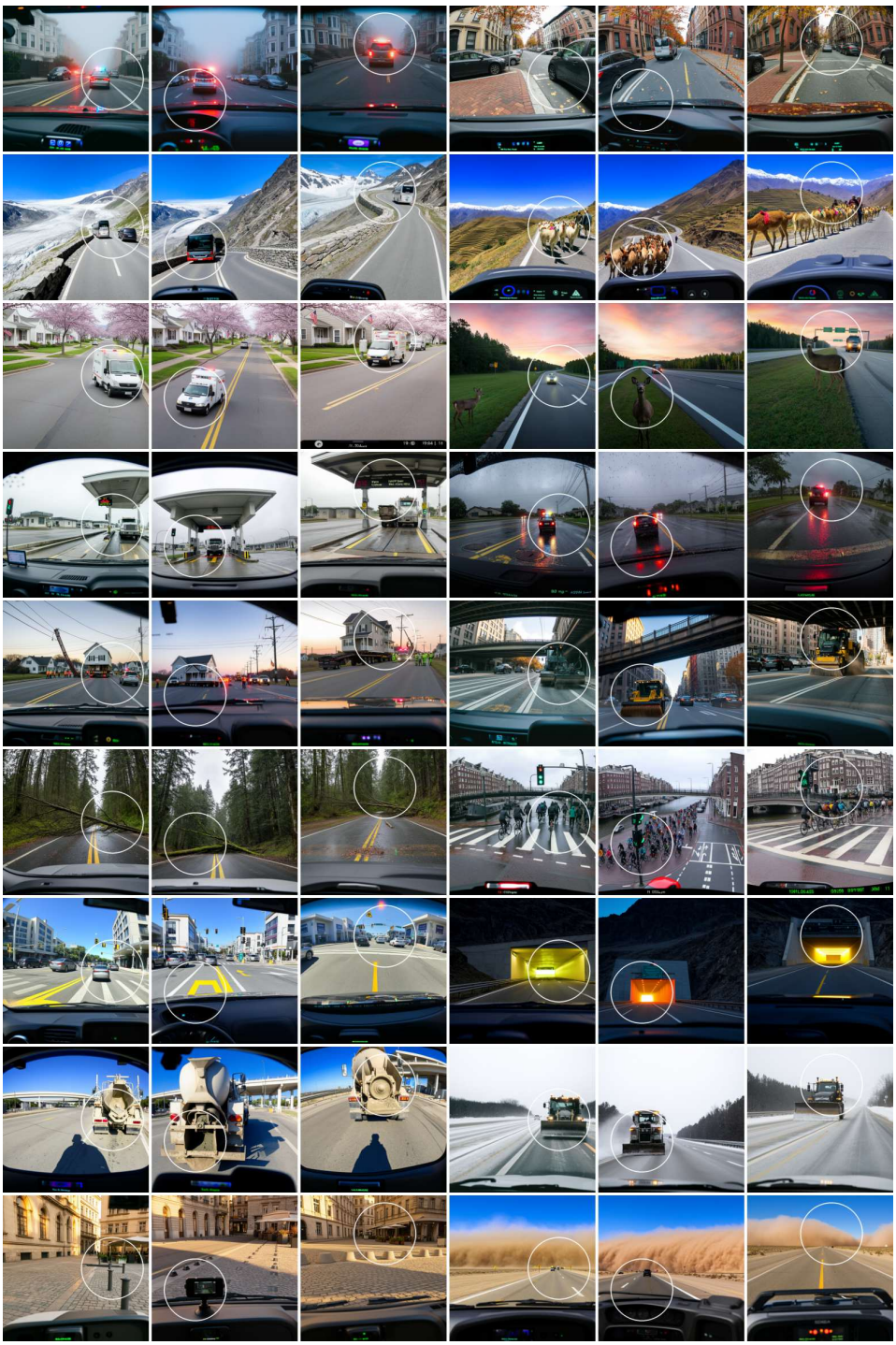

Figure 33: **Saliency-guided generation for autonomous vehicles.** Foveated Diffusion is also well suited for generative autonomous driving simulation, where foveated imagery can be used for policy learning in self-driving systems. In this setting, only important objects in the scene (e.g., pedestrians, other vehicles, roadblocks) are generated at high resolution, while the background is rendered at lower resolution to provide global context.

C.6 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 (Jocher et al., 2023), which integrates multiple YOLO models, for bounding box detection.

As shown in Figures 34 to 36, 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. 36). 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.



Figure 34: **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.

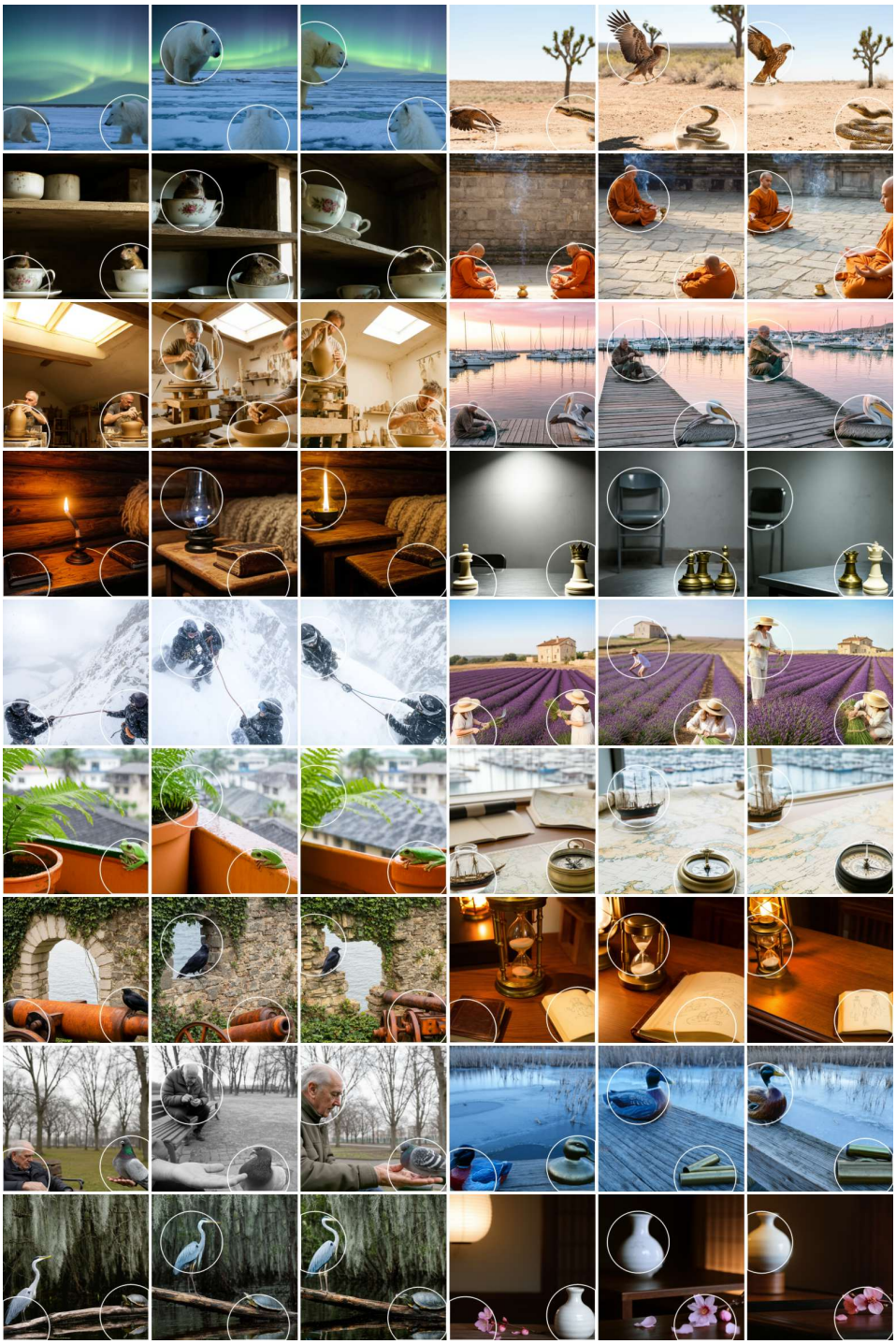

Figure 35: **Multi-object bounding-box-guided generation.** Foveated Diffusion with bounding-box-guided training enables coarse, controllable multi-object generation, where salient objects are approximately constrained within the disjoint regions of the foveation mask.

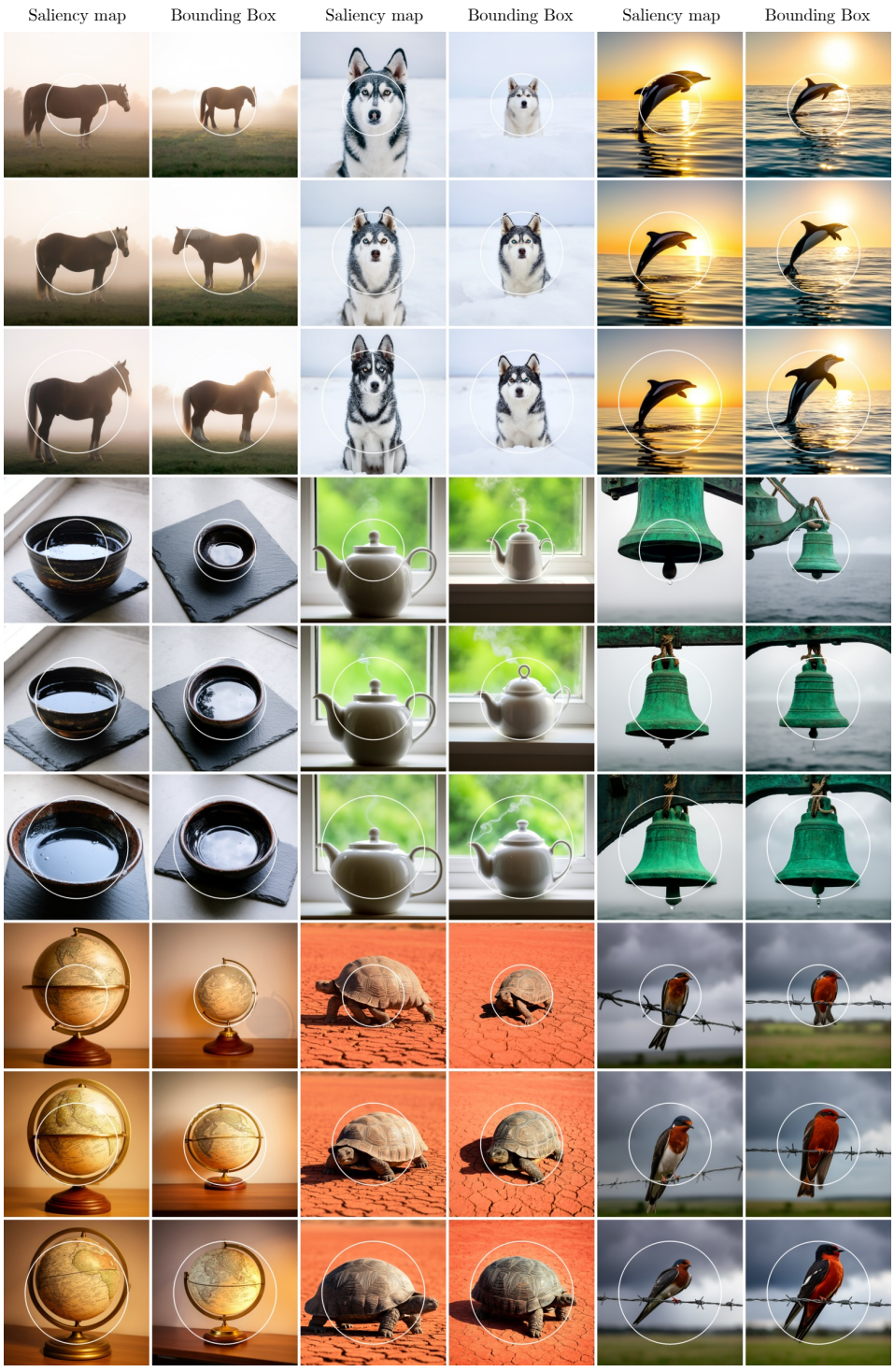

Figure 36: **Bounding-box-guided generation and saliency-guided generation.** We compare bounding-box-guided and saliency-guided generation. Because bounding-box-derived masks precisely delineate object contours, the bounding-box-guided model generates entire objects within the high-resolution region. 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 Generation Results

We include all video generation results on a supplementary website, attached to the supplemental document as a self-contained website package containing `.html` and `.js` files. Please open the `index.html` file in the `website/` folder locally to click through all the video results.

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 and token-merging Bolya et al. (2023), generating coherent and consistent content with consistent structure and scale without color distortions, whereas the mixed-resolution baseline exhibits significant artifacts and token-merging results in overly blurry periphery.

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.

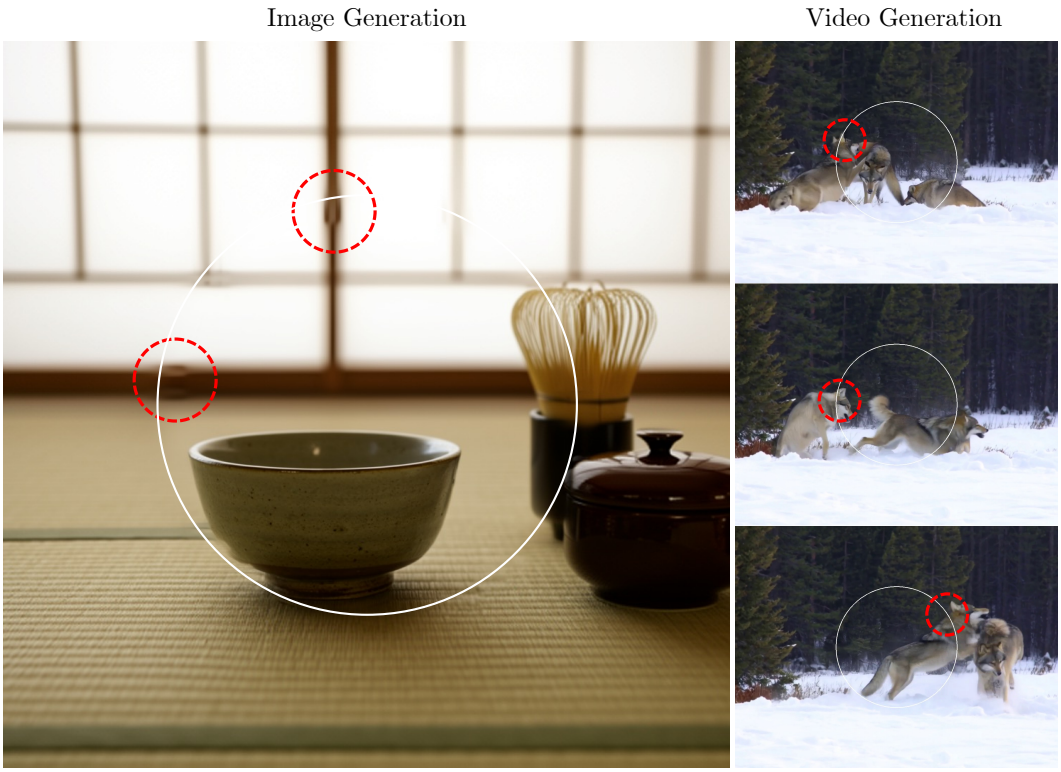

Figure 37: **Foveated Diffusion artifacts.** We delineate the foveation border with a white circular outline. The red dashed lines indicate regions with blending artifacts.

E Discussion

Foveated Diffusion occasionally exhibits color or discontinuity artifacts near foveation boundaries, as shown in Fig. 37. 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.