

SphereDiff: Tuning-free 360° Static and Dynamic Panorama Generation via Spherical Latent Representation

Minho Park*, Taewoong Kang*, Jooyeol Yun, Sungwon Hwang, Jaegul Choo

KAIST AI, South Korea
{m.park, keh0t0, blizzard072, shwang.14, jchoo}@kaist.ac.kr

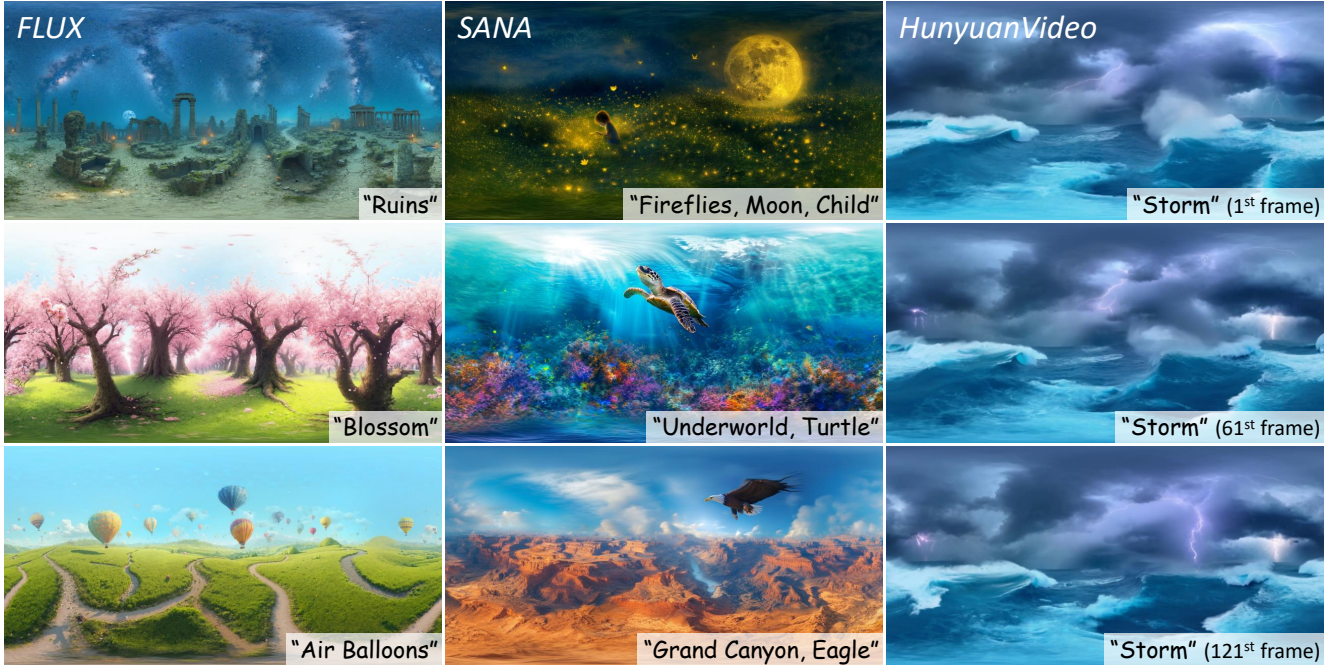


Figure 1: **SphereDiff enables tuning-free 360° panorama generation via spherical latent.** It is compatible with various diffusion backbones, including FLUX (Labs 2024), SANA (Xie et al. 2024), and HunyuanVideo (Kong et al. 2024).

Abstract

The increasing demand for AR/VR applications has highlighted the need for high-quality content, such as 360° live wallpapers. However, generating high-quality 360° panoramic contents remains a challenging task due to the severe distortions introduced by equirectangular projection (ERP). Existing approaches either fine-tune pretrained diffusion models on limited ERP datasets or adopt tuning-free methods that still rely on ERP latent representations, often resulting in distracting distortions near the poles. In this paper, we introduce *SphereDiff*, a novel approach for synthesizing 360° static and live wallpaper with state-of-the-art diffusion models without additional tuning. We define a spherical latent representation that ensures consistent quality across all perspectives, including near the poles. Then, we extend MultiDiffusion to spherical latent representation and propose a dynamic spherical latent sampling method to enable direct use of pretrained diffusion models. Moreover, we intro-

duce distortion-aware weighted averaging to further improve the generation quality. Our method outperforms existing approaches in generating 360° static and live wallpaper, making it a robust solution for immersive AR/VR applications.¹

1 Introduction

The growing demand for AR/VR applications has significantly increased the need for high-quality immersive content. AR/VR technologies offer highly engaging environments, providing a sense of presence that traditional displays (e.g., phones and laptops) cannot. A key element in delivering such experiences is the 360° × 180° panoramic scene, or 360° panorama, which provides an omnidirectional view of the virtual world. This allows users to explore

¹Links: [Github Code](#) | [Project Page](#)

* indicates equal contributions.

their surroundings from any perspective, setting it apart from standard visual content. However, because capturing 360° panoramas requires specialized cameras, their availability is limited, especially for videos. As a result, available content is dominated by simulation-based graphics, which can be rudimentary, while users increasingly seek realistic experiences. In this paper, we aim to generate realistic 360° static and live wallpapers (Fig. 1), enabling the creation of countless immersive scenes without specialized cameras.

360° panoramas are typically represented using an equirectangular projection (ERP), which maps spherical imagery onto a 2D rectangular plane, *e.g.*, mapping a 3D globe to a 2D world map. Due to the limited representational capacity of a 2D plane, an ERP inevitably introduces severe nonlinear distortions, known as ERP distortion, where high-latitude regions appear disproportionately large. For example, as shown in Fig. 1, the content near the poles appear significantly larger than the others since we visualize the 360° wallpapers in ERP. Due to this ERP distortion, 360° panoramas lie in a significantly different distribution from standard perspective images or videos, making it challenging to leverage standard pretrained image or video diffusion models (Rombach et al. 2022; HaCohen et al. 2024).

To handle this gap, several previous studies have fine-tuned pretrained diffusion models using ERP datasets (LatentLabs360 2023; Wang et al. 2024; Chen, Wang, and Liu 2022; Zhang et al. 2024; Li et al. 2024). However, due to the limited availability of text-ERP pairs, data-driven approaches often fail to generate seamless 360° panoramas, particularly near the poles, as shown in Fig. 2. Notably, this issue is more pronounced for generating dynamic 360° panoramas due to the *severely limited availability of 360° panoramic videos*. Thus, tuning-free approaches offer an alternative approach for generating 360° live wallpapers.

Previous tuning-free approaches are based on MultiDiffusion (Liu et al. 2024; Bar-Tal et al. 2023), which denoises large panoramic latents by dividing them into small overlapping patches and blending them in the overlapping regions. They use an ERP representation for latents, that distributes latents uniformly over the ERP. Despite its simplicity, it causes significant differs the density of latents in spherical representation, resulting severe pole-stretching artifacts, as shown in Fig. 2. In this paper, we present a novel tuning-free framework, *SphereDiff*, that does not rely on the ERP representation, and generates seamless 360° static and live wallpapers with minimal distortion, even near the poles.

To this end, we define a spherical latent representation that *uniformly distributes latents over the sphere*, ensuring consistent generation quality across all view directions. We then extend the tuning-free MultiDiffusion framework (Bar-Tal et al. 2023) to operate within this spherical latent representation. In addition, we propose a dynamic latent sampling algorithm that effectively arranges spherical latents onto a 2D perspective grid. Finally, we introduce a distortion-aware weighted averaging scheme to further reduce minor distortions caused by spherical-to-perspective projection. Extensive experiments demonstrate that SphereDiff outperforms existing methods in generating static and live 360° wallpapers, in terms of visual quality and distortion reduction.

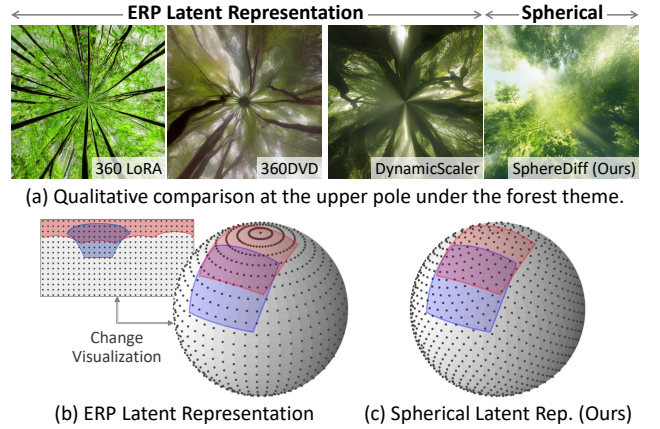


Figure 2: **Motivation.** Both ERP-based finetuning (LatentLabs360 2023; Wang et al. 2024) and tuning-free (Liu et al. 2024) approaches often fail to generate seamless scenes near the poles, as their latents are unevenly distributed over the spherical surface. In contrast, our method produces seamless results by leveraging a spherical latent representation.

In summary, our contributions are threefold:

- We propose *SphereDiff*, a tuning-free framework with a *spherical latent representation* for generating high-quality 360° wallpapers, especially near the poles.
- We extend MultiDiffusion for 360° panoramas with *dynamic latent sampling* for seamless integration with standard diffusion models, allowing tuning-free generation.
- *Distortion-aware weighted averaging* further mitigates the minor distortion from spherical-to-perspective projection, and significantly enhance visual quality.

2 Related Work

Latent Diffusion Models. Recent advancements in diffusion models have enabled the generation of high-quality images (Rombach et al. 2022; Labs 2024; Xie et al. 2024; Chen et al. 2023) and videos (HaCohen et al. 2024; Yang et al. 2024; Kong et al. 2024; Team 2024; Zhang et al. 2025; Liu et al. 2025), achieving impressive visual results across various video generation tasks within the standard perspective of visual content. However, generating content beyond the standard perspective, such as regular or 360° panoramas, remains relatively underexplored. In this paper, we aim to generate 360° panoramas, which differ significantly from standard perspective scenes, by solely leveraging pretrained diffusion models designed for standard perspectives.

360° Panoramic Scene Generation. Most panoramic generation methods rely on equirectangular projection (ERP), which maps spherical coordinates onto a 2D rectangular plane, with latitude and longitude as the vertical and horizontal axes. However, ERP inherently introduces severe nonlinear distortions, particularly near the poles, resulting in a significant gap between ERP and standard perspective views. Although previous studies (Wang et al. 2024; Zhang et al. 2024; Chen, Wang, and Liu 2022; Li et al. 2024) attempt to address this issue by fine-tuning on panoramic ERP

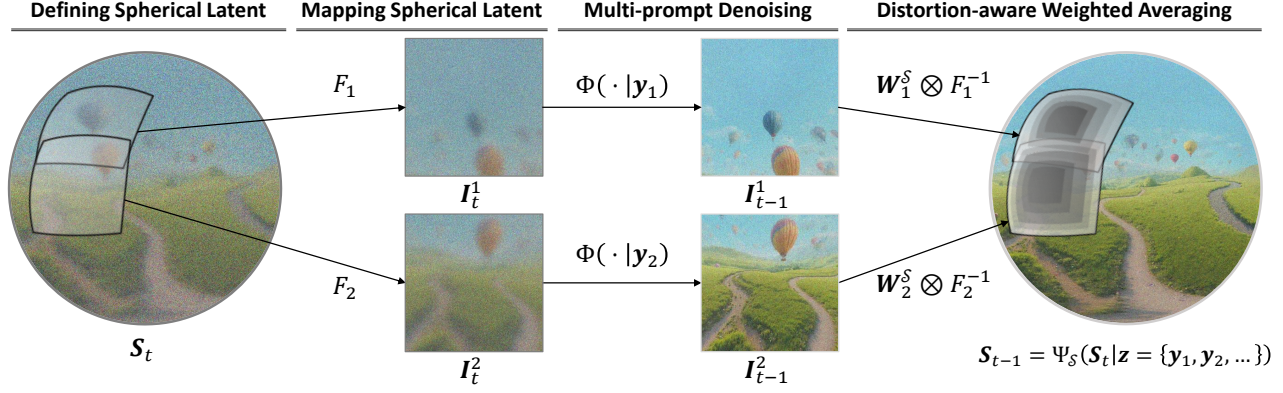


Figure 3: **Overall Pipeline.** We begin by initializing uniformly distributed spherical latents. Next, we map these latents to perspective latents corresponding to multiple view directions. Each view is then denoised using its corresponding prompt. The denoised views are subsequently fused via distortion-aware weighted averaging.

datasets, they often fail to generate seamless panoramas, especially near the poles, or struggle with text controllability due to the domain-specific nature of the datasets (e.g., indoor environments). Recently, CubeDiff (Kalischek et al. 2025) introduces an alternative approach using cube map representations for panoramic image generation. While this method effectively reduces distortions near the poles by training on the large-scale 360° panoramic images, it still struggles with discontinuities at cube-face boundaries. In addition, the data-driven approaches often struggle on generating 360° video, due to the extreme data scarcity of 360° panoramic videos. In contrast, we replace the ERP latent representation with a *spherical latent representation*, providing a natural solution to eliminate distortion across all perspectives, and it does not require any additional tuning.

360° Live Wallpaper Generation. Due to the limited availability of 360° video datasets, recent research on 360° live wallpapers has increasingly favored using perspective-based models without additional training. DynamicScaler (Liu et al. 2024) attempts to mitigate ERP’s inherent distortions through panoramic-projected denoising, leveraging the MultiDiffusion framework (Bar-Tal et al. 2023) with adjusted windows. 4K4DGen (Li et al. 2024) seeks to avoid distortion by utilizing the input ERP images with an image-to-video model, which makes it less suitable for generating novel or highly creative content directly from text descriptions. Unlike these methods, we overcome these limitations by directly using uniformly distributed spherical latents, thereby ensuring efficiency and reduced distortion without requiring additional training.

3 Proposed Method

In this section, we introduce *SphereDiff*, a novel tuning-free framework for generating 360° live wallpapers (Fig. 3). First, we present the spherical latent representation and spherical-to-perspective projection (Section 3.1). Next, we extend the MultiDiffusion framework (Bar-Tal et al. 2023) for 360° panoramas (Section 3.2). We then introduce spherical latent sampling methods, which discrete the continu-

ous coordinates of the spherical latent onto a 2D grid (Section 3.3). Finally, we propose a distortion-aware weighted averaging method to mitigate minor distortions from the spherical-to-perspective projection (Section 3.4), and introduce the multi-prompt inference method (Section 3.5).

3.1 Spherical Latent Representation

Definition. We introduce a spherical representation of latent features for generating 360° live wallpapers. We define a latent feature $\mathbf{f} \in \mathbb{R}^C$ paired with the corresponding spherical coordinate \mathbf{d} on a spherical surface. The set of spherical coordinates can be represented as follows:

$$\mathbb{S}^2 = \{\mathbf{d} = (x, y, z) \mid x, y, z \in \mathbb{R}, \|\mathbf{d}\| = 1\}. \quad (1)$$

Then, we pair each latent feature with its associated position, i.e., $\mathbf{s} = (\mathbf{d}, \mathbf{f})$, referred to as *spherical latent*. For N spherical latents, we define the spherical latents \mathcal{S} as:

$$\mathcal{S} = \{\mathbf{s}_i = (\mathbf{d}_i, \mathbf{f}_i) \mid \mathbf{d}_i \in \mathbb{S}^2, \mathbf{f}_i \in \mathbb{R}^C, \text{ for } i \in [1, N]\}. \quad (2)$$

which is now composed of multiple latents similar to standard 2D or 3D latent features. We refer to the domain of spherical latents as \mathcal{S} , i.e., $\mathcal{S} \in \mathcal{S}$.

Equirectangular Projection (ERP) latents also can be written in our spherical latent representation. However, as shown in Fig. 2 (b), due to the 2D grid constraint of ERP latent, its spherical coordinates are not uniformly distributed on the sphere’s surface. In contrast, we define the spherical latents using the Fibonacci Lattice (Hardin, Michaels, and Saff 2016), which offers the number of spherical latents is nearly equal across all perspectives, as shown in Fig. 2 (c).

Perspective Latent Representation. Since standard diffusion models operate in perspective space, we utilize a *spherical-to-perspective projection* which transforms the spherical coordinate to the perspective coordinate. To achieve this, we first define the domain of perspective coordinates as a discretized 2D plane as

$$\mathbb{P}^2 = \left\{ \mathbf{u} = \left(\frac{2j}{H}, \frac{2k}{W} \right) \mid j \in \left[-\frac{H}{2}, \frac{H}{2} \right], k \in \left[-\frac{W}{2}, \frac{W}{2} \right] \right\}, \quad (3)$$

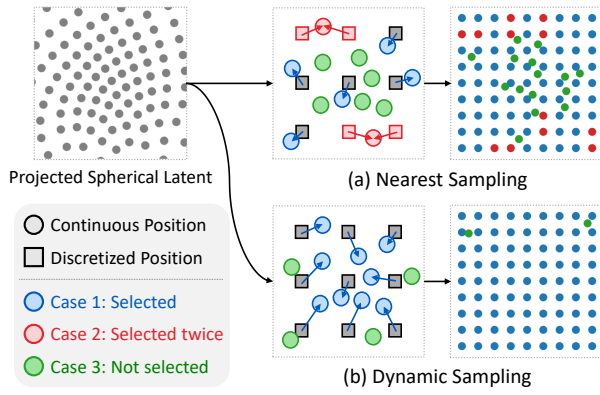


Figure 4: **Comparison of Nearest and Dynamic Sampling.** Nearest sampling often resamples the selected latents or omits central ones, while dynamic sampling selects latents from the center outward, discarding only the outermost ones.

where H, W indicates the height and width of the bounded 2D perspective plane, respectively. We use a view direction $\mathbf{v} \in \mathbb{S}^2$ and a predefined focal length f to define the spherical-to-perspective projection function $\mathbf{u} = \mathcal{T}_{\mathbb{S}^2 \rightarrow \mathbb{P}^2}(\mathbf{d}|\mathbf{v}, f)$. For completeness, the formula of the projection function is provided in Section A.

3.2 MultiDiffusion for Spherical Latent

The MultiDiffusion (Bar-Tal et al. 2023) framework is often utilized for generating arbitrary-shaped images by leveraging pretrained diffusion models (Rombach et al. 2022) trained on standard perspective images. In this section, we introduce an extension of the MultiDiffusion framework to the spherical latent representation.

The goal of this framework is to construct the *Spherical MultiDiffuser* $\Psi_S : \mathcal{S} \times \mathcal{Z} \rightarrow \mathcal{S}$, which takes a noisy spherical latent \mathbf{S}_t and a set of text conditions \mathbf{z} as inputs and produces the denoised spherical latent \mathbf{S}_{t-1} , as illustrated in Fig. 3. Based on the MultiDiffuser, a clean spherical latent \mathbf{S}_0 can be obtained from pure noise \mathbf{S}_T through an iterative denoising process using diffusion models as:

$$\mathbf{S}_T, \mathbf{S}_{T-1}, \dots, \mathbf{S}_0 \quad \text{s.t.} \quad \mathbf{S}_{t-1} = \Psi_S(\mathbf{S}_t|\mathbf{z}). \quad (4)$$

To construct MultiDiffuser Ψ_S , we first leverage a pretrained diffusion model trained on standard perspective latents $\Phi : \mathcal{I} \times \mathcal{Y} \rightarrow \mathcal{I}$, which takes a noisy latent \mathbf{I}_t and a text condition \mathbf{y} as inputs and produces the denoised latent \mathbf{I}_{t-1} . The pretrained diffusion model gradually denoises the pure Gaussian noise $\mathbf{I}_T \sim \mathcal{N}$ into a clean image \mathbf{I}_0 .

$$\mathbf{I}_T, \mathbf{I}_{T-1}, \dots, \mathbf{I}_0 \quad \text{s.t.} \quad \mathbf{I}_{t-1} = \Phi(\mathbf{I}_t|\mathbf{y}) \quad (5)$$

Next, we define a mapping function between the spherical and perspective latent spaces, $F_i : \mathcal{S} \rightarrow \mathcal{I}$, along with a corresponding condition mapping $\lambda_i : \mathcal{Z} \rightarrow \mathcal{Y}$, where $i \in \{1, \dots, n\}$. The mapping functions F_i and λ_i can be formulated in various ways, which will be discussed in Sections 3.3 and 3.5, respectively.

$$\mathbf{I}_t^i = F_i(\mathbf{S}_t), \quad \mathbf{y}_i = \lambda_i(\mathbf{z}) \quad (6)$$

Algorithm 1 Dynamic Latent Sampling

Input : Projected Latent $\mathbf{P} = \mathcal{T}_{\mathcal{S} \rightarrow \mathcal{I}}(\mathbf{S}|\mathbf{v}, f)$
Output: Arranged Perspective Latent \mathbf{I}
 $\mathbf{I}' \leftarrow$ Sort the latents of \mathbf{P} by $\|\mathbf{u}_i\|$.
 $M \leftarrow |\mathbf{P}|$ ▷ Get the number of latents
 $H, W \leftarrow \lfloor \sqrt{M} \rfloor, \lfloor \sqrt{M} \rfloor$ ▷ Dynamic H, W
 $\mathbf{I} \leftarrow \emptyset^{H \times W}$ ▷ Initialize a queue
for $i \in [1, H/2]$ **do**
 $n \leftarrow (2i)^2 - (2i - 2)^2$ ▷ Counts of i -th border
 $\mathbf{l} \leftarrow$ first n latents from the sorted \mathbf{I}' ▷ center-first
 Set i -th border of \mathbf{I} to \mathbf{l}
 Pop first n latents from \mathbf{I}'
end
return \mathbf{I} ▷ Ignore $M - H \times W$ elements of \mathbf{I}

Finally, The denoising step in of MultiDiffuser can be formulated by a closed-form (Bar-Tal et al. 2023).

$$\Psi(\mathbf{S}_t|\mathbf{z}) = \sum_{i=1}^n \mathbf{W}_i^S \otimes F_i^{-1}(\Phi(\mathbf{I}_t^i|\mathbf{y}_i)). \quad (7)$$

where \mathbf{W}_i^S are the per-pixel weights and \otimes is the Hadamard product. In the following sections, we define the latent mapping function F_i (Section 3.3), the per-pixel weights \mathbf{W}_i^S (Section 3.4), and the condition mapping function λ_i (Section 3.5), which together extend MultiDiffusion to the spherical latent representation.

3.3 Mapping Spherical Latent

We define the latent mapping function F that transforms a spherical latent representation into a perspective latent space \mathcal{I} . To define mapping function F , we first apply the transformation $\mathcal{T}_{\mathcal{S} \rightarrow \mathcal{I}}$ based on the view direction $\mathbf{v} \in \mathbb{S}^2$ and focal length f , which projects the coordinates of the spherical latents onto the perspective plane \mathcal{P} . Formally, the spherical-to-perspective latent transformation can be written as

$$\mathcal{T}_{\mathcal{S} \rightarrow \mathcal{P}}(\mathbf{S}|\mathbf{v}, f) = \mathbf{P} = \{\mathbf{p}_i = (\mathbf{u}_i, \mathbf{f}_i) | \mathbf{u}_i \in [-1, 1]^2\}, \quad (8)$$

where $\mathbf{u}_i = \mathcal{T}_{\mathbb{S}^2 \rightarrow \mathbb{P}^2}(\mathbf{d}_i|\mathbf{v}, f)$. Note that, \mathbf{P} does not contain N elements since the perspective are cropped by $[-1, 1]$. We next introduce simple yet effective latent sampling methods for discretizing perspective coordinates.

Nearest Point Sampling. A straightforward approach to discrete continuous coordinates is nearest-neighbor sampling, where the nearest projected spherical latent is selected for each pixel position. Specifically, the latent closest to the center of a $H \times W$ grid is retrieved and used as input for denoising, as illustrated in Fig. 4 (a). Despite its simplicity, this method introduces two critical issues. First, the same latent may be selected multiple times, altering the latent distribution, which often degrades generation performance (Chang et al. 2024). Second, some spherical latents may not be chosen even if they fall within the field of view of the current camera view direction. This phenomenon, referred to as the *undersampling problem*, has particularly detrimental consequences for generating a seamless panorama.

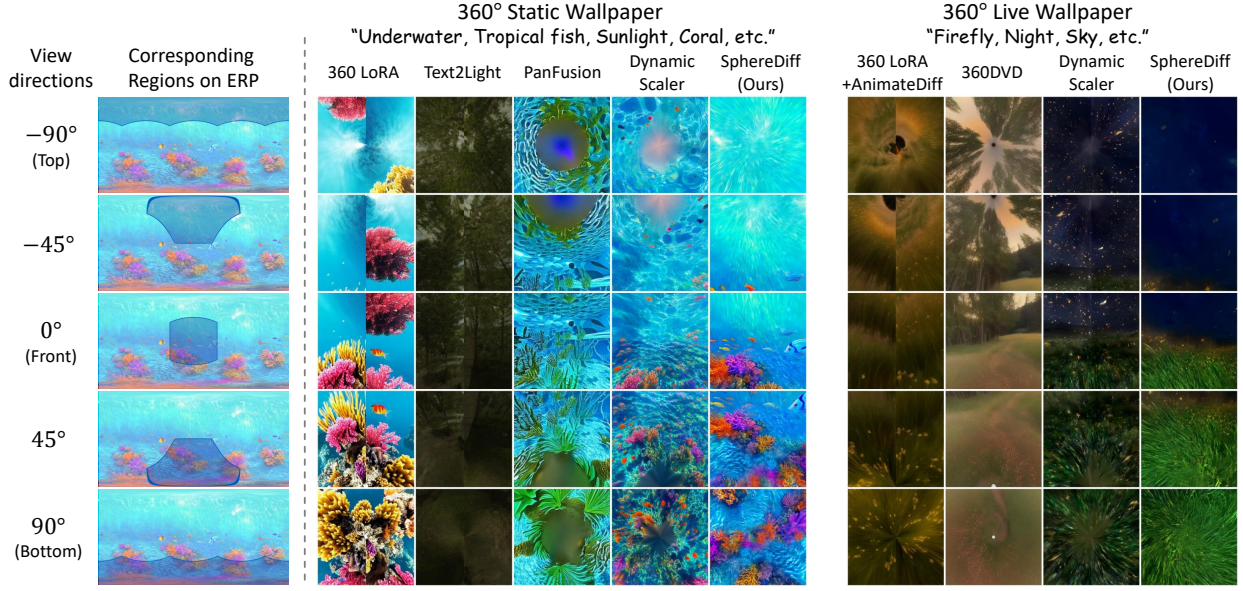


Figure 5: **Qualitative comparison.** Each sample shows perspective views from top to bottom, highlighting end-to-end continuity and distortion. Other methods exhibit artifacts such as seams, pole distortions, blurriness, or spots, while ours produces seamless, high-quality panoramas without these issues. The entire ERPs are available in Section D.

Undersampling Problem. The undersampling of spherical latents disrupts information flow across neighboring windows. As illustrated in Fig. 4, the green points lack information from the current field of view (FoV) since they are not denoised in this step. If the next window’s FoV captures a green point, not the blue point, it receives no information from the current window, *causing discontinuities even when there is a large overlap*.

Dynamic Latent Sampling. To address the undersampling problem, we aim to ensure that all points within the FoV are selected, especially those near the center. We propose a dynamic latent sampling strategy, which comprises three components: (1) a queue, (2) a dynamic number of latents, and (3) center-first selection. First, we avoid selecting the same spherical latent more than once by using a queue: once a latent is selected, it is immediately removed from the queue. Then, we dynamically adjust the number of latents so that H and W are not fixed, thereby reducing the number of latents within the FoV that remain unselected. Lastly, we prioritize selecting center-positioned latents first and then ignore the remaining points at the outermost region. The entire algorithm is demonstrated in Algorithm 1 and Fig. 4 (b).

3.4 Distortion-Aware Weighted Averaging

While the spherical-to-perspective distortion is relatively smaller than the ERP-to-perspective distortion, it can still cause latent position misalignment for other viewpoints.

To address this, we proposed distortion-aware weighted averaging within the MultiDiffusion framework (Bar-Tal et al. 2023). Specifically, we adjust the per-pixel weight W_i^S to account for the spherical-to-perspective distortion. Since distortion increases with distance from the center of the per-

spective image, we introduce a simple yet effective exponential weighting function in the image space \mathcal{I} .

$$W_i^{\mathcal{I}} = [W_i^{jk}]_{j \in [1, H], k \in [1, W]} \in \mathbb{R}^{H \times W}, \quad (9)$$

$$W_i^{jk} = \exp(-\|u_{jk}\|/\tau), \quad (10)$$

where $\|u_{jk}\|$ is the distance from the center of the perspective image, and τ is a scaling factor controlling how quickly the weight decays toward the edges. Then, the weighting function can be represented as $W_i^S = F_i^{-1}(W_i^{\mathcal{I}})$.

3.5 Multi-prompt Inference

We generate 360° wallpapers using multiple prompts that correspond to specific regions on the spherical surface. The use of region-specific prompts helps reduce semantic inconsistencies, such as generating ground in sky regions. To this end, we define a simple condition mapping function λ_i that selects a region-specific prompt from the prompt set \mathbf{z} , according to the elevation of the view direction, $\phi_i = \text{elevation}(\mathbf{d}_i)$. Specifically, λ_i selects the prompt whose elevation is closest to ϕ_i among the discrete elevations $\{-90^\circ, -10^\circ, 0^\circ, +10^\circ, +90^\circ\}$, with the corresponding prompts given by $\mathbf{z} = \{\mathbf{y}_{\text{top}}, \mathbf{y}_{\text{upper}}, \mathbf{y}_{\text{middle}}, \mathbf{y}_{\text{lower}}, \mathbf{y}_{\text{bottom}}\}$. We allocate denser text prompts near the horizon to capture richer visual complexity in that region.

For enhanced scene complexity, we employ an additional foreground prompt $\mathbf{y}_{\text{foreground}}$, conditioned on the corresponding view direction, to generate a foreground object at a specific location (e.g., $\theta_i = \text{azimuth}(\mathbf{d}_i) = 30^\circ, \phi_i = -30^\circ$). The second column of Fig. 1 shows an example generated in this manner. Visual illustrations of the multi-prompt inference are provided in Section B.

Wallpaper Type	Method	Panoramic Criteria		Image Criteria	Text Adherence	Video Criteria	
		Distortion \uparrow	End Continuity \uparrow	Image Quality \uparrow	Text Alignment \uparrow	Motion Smoothness \uparrow	Temporal Flickering \uparrow
360° Static	360 LoRA	21.43	23.81	21.43	20.24	-	-
	Text2Light	5.95	4.76	8.33	5.95	-	-
	PanFusion	14.29	10.71	10.71	16.67	-	-
	DynamicScaler	20.24	25.00	25.00	27.38	-	-
	SphereDiff (Ours)	38.10	35.71	34.52	29.76	-	-
360° Live	360 LoRA + AnimateDiff	25.00	27.38	27.38	27.38	27.38	23.81
	360DVD	11.90	13.10	16.67	20.24	8.33	14.29
	DynamicScaler	30.95	26.19	28.57	22.62	<u>28.57</u>	29.76
	SphereDiff (Ours)	32.14	33.33	<u>27.38</u>	29.76	35.71	32.14

Table 1: **User study results.** The 360° static and live wallpapers generated by SphereDiff have achieved state-of-the-art performance in user preference across most metrics, particularly in panoramic criteria such as distortion and end continuity.

4 Experiments

4.1 Experimental Setup

Implementation Details. For all experiments and comparisons, we adopt SANA (Xie et al. 2024) and LTX-Video (HaCohen et al. 2024) as the base T2I and T2V models, respectively. We use 2,600 spherical latent points, and we conduct the MultiDiffusion framework with 89 view directions with an 80° FoV, where adjacent views overlap by 60%. The additional details are available in Section E.

Evaluation Criteria. We evaluate our 360° static and live wallpapers using four criteria: panoramic, image, video, and text-level aspects. Image quality, text adherence, and temporal smoothness are standard metrics in image and video generation, and we adopt them to assess the realism and consistency of our results. For panoramic evaluation, we consider distortion and end continuity. Distortion measures the geometric deformation introduced when the equirect-angular projections (ERPs) are converted into perspective, which tends to increase if models fail to account for ERP-specific constraints. End continuity (also known as loop-consistency) evaluates the seamless alignment of the left and right borders in ERPs, indicating whether the scene wraps smoothly into a loop.

Evaluation Process. We use 20 predefined text prompt sets designed for immersive outdoor scenes. We assess the criteria through a user study, following prior studies (Wang et al. 2024; Liu et al. 2024), where participants select the sample among the baselines that best fits the given criteria. The assessments are conducted on perspective images captured from 14 predefined view directions with a 90° field of view. Additionally, we conduct automatic evaluations for text and video criteria by leveraging VBench (Huang et al. 2024), which has been widely validated. In contrast, image and panoramic criteria cannot be accurately measured by classical methods such as FID (Heusel et al. 2017) due to the absence of a source domain. Thus, we instead employ vision-language models (Hurst et al. 2024) within the LLM-as-a-judge framework (Zheng et al. 2023). The reliability of the VLM-based evaluation, supported by comprehensive experiments, is provided in Section F.

Baselines. For 360° static wallpaper generation, we compare with open-sourced baselines including 360 LoRA (La-

tentLabs360 2023), Text2Light (Chen, Wang, and Liu 2022), Panfusion (Zhang et al. 2024), and DynamicScaler (Liu et al. 2024). Although DynamicScaler (Liu et al. 2024) was originally designed for panoramic video generation, we extend it for image generation and use it as a baseline. For 360° live wallpaper generation, we use 360DVD (Wang et al. 2024) and DynamicScaler as primary baselines. DynamicScaler is reimplemented with SANA and LTX-Video to enable a fair comparison as a tuning-free method. In addition, we combine 360 LoRA with AnimateDiff (Guo et al. 2023) as an extra baseline. The multi-prompt methods, DynamicScaler and ours, share identical text prompt sets ($z = \{y_{top}, y_{upper}, y_{middle}, y_{lower}, y_{bottom}\}$) for generation, whereas the other single-prompt methods rely on the main reference prompt (y_{middle}).

4.2 Results

Qualitative Comparison. As shown in Fig. 5, 360° static and live wallpaper generation baselines appear noticeable artifacts near the poles, such as distortion, blurriness, and speckling due to the limitation of the ERP latents as we discussed. These artifacts significantly undermine the immersive experience of 360° wallpapers. In contrast, our method ensures consistent quality at all viewing angles and eliminating distortions and discontinuities, since we utilize uniformly distributed spherical latents. Additional qualitative results, including samples generated by stronger diffusion backbones, namely FLUX (Labs 2024) and Hunyuan-Video (Kong et al. 2024), are provided in Section D.

User Study. To evaluate the effectiveness of our panorama generation method, we conduct a user study. A total of 21 participants compared 20 pairs of samples, both images and videos, based on six aspects: visual quality, text alignment, distortion, end continuity, motion smoothness, and temporal flickering. As shown in Tab. 1, our method consistently outperforms all baselines on most metrics for both static and live 360° wallpaper generation. While DynamicScaler (Liu et al. 2024) received slightly higher scores in image quality, our method achieves significantly better performance in panoramic metrics. Overall, the user study strongly supports the effectiveness of the proposed method, as it demonstrates the best results in panoramic criteria, text adherence and video criteria, even without any additional finetuning.

Wallpaper Type	Method	Panoramic Criteria		Image Criteria		Text Adherence		Video Criteria	
		Distortion \uparrow	End Continuity \uparrow	Image Quality \uparrow	Aesthetic Appearance \uparrow	Scene \uparrow	CLIP-Score \uparrow	Motion Smoothness \uparrow	Temporal Flickering \uparrow
360° Static	360 LoRA	2.027	3.423	2.965	3.492	<u>0.2875</u>	26.40	-	-
	Text2Light	2.381	3.454	2.415	2.777	0.0250	20.46	-	-
	PanFusion	1.965	3.696	2.819	3.450	0.2125	25.70	-	-
	DynamicScaler	<u>2.854</u>	<u>3.985</u>	4.496	<u>4.577</u>	0.2750	<u>26.63</u>	-	-
	SphereDiff (Ours)	3.238	4.892	4.496	4.685	0.5875	28.65	-	-
360° Live	360 LoRA + AnimateDiff	1.939	<u>3.482</u>	3.179	3.571	0.2914	26.34	0.9908	0.9847
	360DVD	<u>2.086</u>	3.246	2.929	3.396	0.2570	25.54	0.9857	0.9798
	DynamicScaler	1.971	2.971	2.711	3.236	<u>0.4836</u>	<u>26.89</u>	<u>0.9943</u>	<u>0.9918</u>
	SphereDiff (Ours)	2.579	4.496	<u>3.050</u>	3.593	0.5703	27.52	0.9956	0.9941

Table 2: **Automated Quantitative Evaluation.** We conduct automatic evaluations of 360° static and live wallpaper synthesis across four criteria. SphereDiff consistently outperforms existing methods except image quality, where it ranks second.

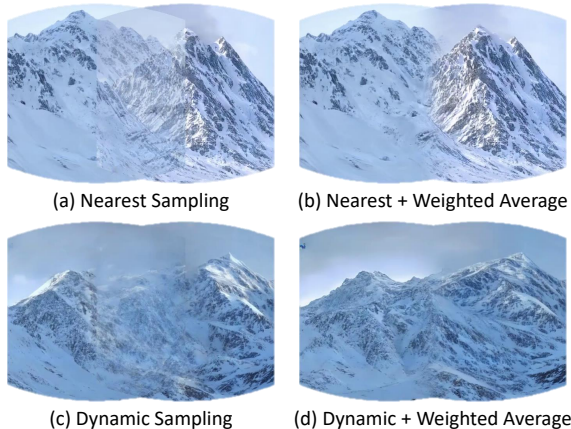


Figure 6: **Visual Ablation Study.** Nearest sampling causes view inconsistencies and overlap artifacts due to undersampling, while dynamic sampling enables better information sharing for more integrated outputs. In addition, weighted averaging significantly improves seamlessness.

Automatic Quantitative Results. As presented in Tab. 2, our method outperforms all baselines across most evaluation metrics in both static and live 360° wallpaper generation, consistent with the results of the user study. Notably, it achieves significantly better scores in distortion and end continuity, demonstrating its effectiveness in producing high-quality panoramic content. While the image quality of our video generation is lower than that of 360 LoRA + AnimateDiff (LatentLabs360 2023; Guo et al. 2023), it remains comparable overall. The overall score of 360° live wallpaper generation is lower than the 360° static wallpaper generation, which may be due to the performance of the underlying diffusion model. Nevertheless, SphereDiff shows the state-of-the-art performance in the most metrics and its performance could be further improved by leveraging a more advanced denoising model.

4.3 Ablation Study

We conducted ablation studies to evaluate the impact of each component, namely dynamic latent sampling and distortion-aware weighted averaging. For clarity, we performed denoising on only two views, which simplifies visualization

while preserving the validity of the comparison, and we visualized the results in ERP format. As shown in Fig. 6, nearest sampling fails to facilitate information exchange between different views, resulting in noticeable artifacts and unnatural transitions in overlapping regions. In contrast, our dynamic latent sampling improves information exchange across views, producing more seamless and coherent images. Furthermore, incorporating our distortion-aware weighted averaging technique yields significantly clearer and more consistent outputs for both sampling methods. These results demonstrate that each component plays a critical role in our framework by integrating multiple perspectives and ensuring high-quality 360° wallpaper generation. For further evidence, a comprehensive quantitative comparison of the ablation study results is provided in Section C.

5 Conclusion and Discussion

We introduce *SphereDiff*, a tuning-free framework for 360° live wallpaper generation that effectively leverages state-of-the-art image and video diffusion models. The proposed spherical latent representation inherently supports consistent generation quality across all view directions, including near the poles. Our extended MultiDiffusion framework for 360° panoramas with dynamic latent sampling facilitates a tuning-free approach. Lastly, distortion-aware weighted averaging significantly enhances the quality of panoramic content. In summary, we achieve state-of-the-art performance in generating 360° live and static wallpapers, as demonstrated through comprehensive experiments.

Limitations. Our approach achieves strong results in producing high-quality static and live 360° wallpapers for a broad range of outdoor. However, it cannot yet generate highly complex scenes, such as indoor environments, without additional training. Recent panoramic image generation approaches (Kalischek et al. 2025; Çapuk et al. 2025) handle more complex scenes by relying on large-scale panoramic image datasets, but they cannot be readily adapted to panoramic videos due to the limited availability of panoramic video datasets. In contrast, our approach achieves state-of-the-art performance in generating 360° live wallpapers without additional tuning, highlighting its potential as a strong foundation for future work on more complex 360° panoramic video generation.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621) This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government [25ZB1200, Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems]

References

- Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, 1737–1752. 2, 3, 4, 5, 16, 18
- Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR*. 19
- Çapuk, H.; Bond, A.; Kızıl, M. B.; Göçen, E.; Erdem, E.; and Erdem, A. 2025. TanDiT: Tangent-Plane Diffusion Transformer for High-Quality 360 {deg} Panorama Generation. *arXiv preprint arXiv:2506.21681*. 7, 12
- Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*. 15
- Chang, P.; Tang, J.; Gross, M.; and Azevedo, V. C. 2024. How I Warped Your Noise: a Temporally-Correlated Noise Prior for Diffusion Models. In *The Twelfth International Conference on Learning Representations*. 4, 13
- Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*. 2
- Chen, Z.; Wang, G.; and Liu, Z. 2022. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16. 2, 6, 13, 14, 15
- Daras, G.; Nie, W.; Kreis, K.; Dimakis, A.; Mardani, M.; Kovachki, N.; and Vahdat, A. 2024. Warped diffusion: Solving video inverse problems with image diffusion models. *Advances in Neural Information Processing Systems*, 37: 101116–101143. 13
- Feng, M.; Liu, J.; Cui, M.; and Xie, X. 2023. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*. 15
- Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*. 6, 7, 14
- HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*. 2, 6, 15, 16, 18, 19
- Hardin, D. P.; Michaels, T.; and Saff, E. B. 2016. A Comparison of Popular Point Configurations on \mathbb{S}^2 . *arXiv preprint arXiv:1607.04590*. 3
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637. 6, 19
- Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818. 6
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. 6, 17, 18, 19
- Kalischek, N.; Oechsle, M.; Manhardt, F.; Henzler, P.; Schindler, K.; and Tombari, F. 2025. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*. 3, 7, 15
- Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*. 1, 2, 6, 14, 15, 16, 18
- Labs, B. F. 2024. FLUX. <https://github.com/black-forest-labs/flux>. 1, 2, 6, 14, 15, 16, 18
- LatentLabs360. 2023. LatentLabs360. <https://civitai.com/models/10753/latentlabs360>. 2, 6, 7, 14, 15, 16
- Li, R.; Pan, P.; Yang, B.; Xu, D.; Zhou, S.; Zhang, X.; Li, Z.; Kadambi, A.; Wang, Z.; Tu, Z.; et al. 2024. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*. 2, 3, 16
- Liao, K.; Xu, X.; Lin, C.; Ren, W.; Wei, Y.; and Zhao, Y. 2023. Cylind-painting: Seamless 360 panoramic image out-painting and beyond. *IEEE Transactions on Image Processing*, 33: 382–394. 15
- Liu, D.; Li, S.; Liu, Y.; Li, Z.; Wang, K.; Li, X.; Qin, Q.; Liu, Y.; Xin, Y.; Li, Z.; et al. 2025. Lumina-Video: Efficient and Flexible Video Generation with Multi-scale Next-DiT. *arXiv preprint arXiv:2502.06782*. 2
- Liu, J.; Lin, S.; Li, Y.; and Yang, M.-H. 2024. DynamicScaler: Seamless and Scalable Video Generation for Panoramic Scenes. *arXiv preprint arXiv:2412.11100*. 2, 3, 6, 12, 14, 15, 16
- Lu, T.; Shu, T.; Xiao, J.; Ye, L.; Wang, J.; Peng, C.; Wei, C.; Khashabi, D.; Chellappa, R.; Yuille, A.; et al. 2024. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*. 16
- Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748*. 13

- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695. 2, 4
- Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389. 15
- Team, G. 2024. Mochi 1. <https://github.com/genmoai/models>. 2
- Wang, Q.; Li, W.; Mou, C.; Cheng, X.; and Zhang, J. 2024. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6913–6923. 2, 6, 16
- Wang, Y.; He, X.; Wang, K.; Ma, L.; Yang, J.; Wang, S.; Du, S. S.; and Shen, Y. 2025. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13629–13638. 19
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837. 17
- Wu, T.; Li, X.; Qi, Z.; Hu, D.; Wang, X.; Shan, Y.; and Li, X. 2024. Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6126–6134. 15
- Wu, T.; Zheng, C.; and Cham, T.-J. 2024. PanoDiffusion: 360-degree Panorama Outpainting via Diffusion. In *The Twelfth International Conference on Learning Representations*. 15
- Xie, E.; Chen, J.; Chen, J.; Cai, H.; Tang, H.; Lin, Y.; Zhang, Z.; Li, M.; Zhu, L.; Lu, Y.; et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*. 1, 2, 6, 15, 16, 18, 19
- Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*. 2
- You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Dong, C.; and Xue, T. 2024a. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*. 19
- You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; and Dong, C. 2024b. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, 259–276. Springer. 19
- Zhang, C.; Wu, Q.; Gambardella, C. C.; Huang, X.; Phung, D.; Ouyang, W.; and Cai, J. 2024. Taming Stable Diffusion for Text to 360 Panorama Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6347–6357. 2, 6, 14, 15
- Zhang, S.; Li, W.; Chen, S.; Ge, C.; Sun, P.; Zhang, Y.; Jiang, Y.; Yuan, Z.; Peng, B.; and Luo, P. 2025. FlashVideo: Flowing Fidelity to Detail for Efficient High-Resolution Video Generation. *arXiv preprint arXiv:2502.05179*. 2
- Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; and Zhou, Z. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, 519–535. Springer. 15
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623. 6, 19
- Zhou, F.; Gu, T.; Huang, Z.; and Qiu, G. 2024. Vision language modeling of content, distortion and appearance for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*. 19

SphereDiff: Tuning-free 360° Static and Dynamic Panorama Generation via Spherical Latent Representation

Supplementary Material

In the appendix, we include method descriptions, comparisons, ablations, and evaluation details. First, we present method details with mathematical expressions and visual illustrations for completeness (Sections A and B). Next, we present in-depth ablation studies and detailed comparisons with baselines, including additional qualitative and quantitative results (Sections C and D). Finally, we provide implementation and evaluation details, along with the text prompts used to generate the results (Sections E, F and G). The table of contents for our appendix is provided below.

Table of Content

A. Spherical-to-Perspective Projection

- A.1. Spherical-to-Perspective Mapping Function
- A.2. Spherical-to-Perspective Distortion
- A.3. Coordinate Systems

B. Details of Multi-Prompt Inference

- B.1. Visual Illustration of Multi-prompt Inference
- B.2. Foreground Generation

C. In-depth Ablations

- C.1. Ablation Study with Automated Quantitative Metrics
- C.2. Alternatives for Mapping Spherical Latent

D. Detailed Comparison with Baselines

- D.1. Qualitative Comparison in ERP format
- D.2. Detailed Comparison with the Other Panorama Generation Approaches

E. Implementation Details

F. Evaluation Details

- F.1. User Study Details
- F.2. VLM-based Visual Score

G. Used Text Prompts

A Spherical-to-Perspective Projection

In this section, we describe the formulation of the spherical-to-projection mapping functions (Section A.1) and provide further explanation of the spherical-to-perspective distortion (Section A.2) for completeness.

A.1 Spherical-to-Perspective Mapping Function

For perspective latent representation, we define a virtual camera centered at the origin. The points in world coordinates, denoted as $d = (x, y, z)^\top \in \mathbb{S}^2$, are projected onto image space using the projection matrix $P = K[R|t]$. Here, K is the intrinsic camera matrix derived from a predefined focal length f , R represents the viewing direction, and t is

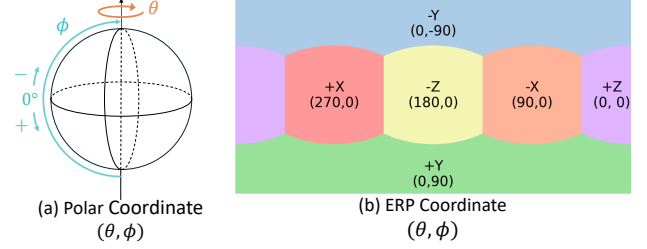


Figure 7: **Our Coordinate System.** (a) The spherical coordinate system with azimuth θ and elevation ϕ . (b) The corresponding layout in the equirectangular projection (ERP) format, showing the mapping of Cartesian axes to spherical coordinates (θ, ϕ) .

set to zero. The spherical-to-perspective projection function $\mathcal{T}_{\mathbb{S}^2 \rightarrow \mathbb{P}^2}$ can be formulated as:

$$\tilde{u} = K[R|t]\tilde{d}, \quad (11)$$

where $\tilde{d} = (x, y, z, 1)^\top$ is the homogeneous coordinate representation of the 3D point d , and $\tilde{u} = (u', v', w')^\top$ represents the projected homogeneous coordinates in image space. The final 2D perspective coordinates $u = (u, v)^\top \in \mathbb{P}^2$ are obtained via perspective division:

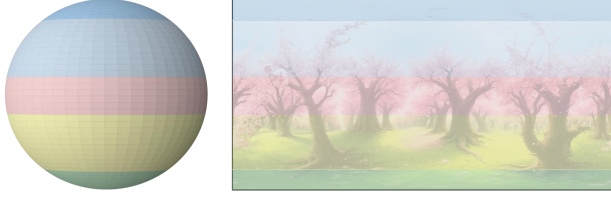
$$u = \left(\frac{u'}{w'}, \frac{v'}{w'} \right). \quad (12)$$

To ensure proper visibility, points located behind the view direction are masked out using their inner product values, retaining only the points in the frontal view.

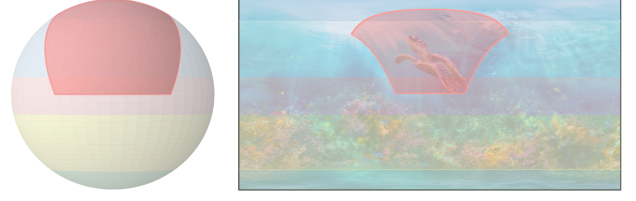
A.2 Spherical-to-Perspective Distortion

The core methodology of our model involves approximating the spherical surface with a local tangent plane for computational purposes. However, this approach inherently introduces a form of geometric distortion. The tangent plane approximation is highly accurate at its central point, which corresponds to the view direction, but the deviation between the true spherical surface and the approximated plane systematically increases with the radial distance from this center. Consequently, a greater degree of error accumulates toward the periphery of the image.

Geometric Formulation. To formalize this, let us consider a sphere of radius R centered at the origin $(0, 0, 0)$. We define a tangent plane to be in contact with the sphere at its north pole, $P_N = (0, 0, R)$. The equation of this plane is therefore $z = R$. A point P_s on the surface of the sphere can be described by its angle θ from the positive z-axis. For simplicity, we can analyze the geometry in a 2D cross-section. In this view, the coordinates of the point P_s are $(R \sin \theta, R \cos \theta)$. The projection ray originates from



(a) Visualizing Multi-prompt Inference



(b) Visualizing Multi-prompt Inference w/ foreground object

Figure 8: Visual Illustration of Multi-Prompt Inference. We illustrate the multi-prompt inference process using visual aids. (a) Three different prompts are assigned based on the elevation angle (ϕ), as indicated by the color-coded regions in both the spherical projection and ERP, for generating 360° wallpapers. (b) A foreground object is generated using an additional foreground prompt ($y_{\text{foreground}}$) conditioned on the corresponding view direction ($\theta = 180^\circ$, $\phi = -40^\circ$).

the center of the sphere $(0, 0, 0)$ and passes through P_s to intersect the tangent plane ($z = R$) at a projected point, P_p . Let the distance of this projected point from the center of the plane be d_p . The coordinates of P_p are thus (d_p, R) in the 2D cross-section. By the property of similar triangles formed by the origin, the z -axis, and the projection ray, the ratio of the horizontal to the vertical component is constant:

$$\frac{\text{horizontal}}{\text{vertical}} = \frac{R \sin \theta}{R \cos \theta} = \frac{d_p}{R}$$

This simplifies to $\tan \theta = d_p/R$. Solving for the projected distance d_p , we get:

$$d_p = R \tan \theta$$

Analysis of Distortion. The distortion arises from the non-linear relationship between the true distance on the sphere’s surface and the projected distance on the plane.

1. **True Surface Distance:** The actual distance from the pole P_N to the point P_s along the arc of the sphere is the arc length, d_s , given by:

$$d_s = R\theta \quad (\text{where } \theta \text{ is in radians})$$

2. **Projected Distance:** As derived above, the distance from the center of the plane to the projected point P_p is:

$$d_p = R \tan \theta$$

Comparing these two distances reveals the nature of the distortion. For small angles near the center of projection ($\theta \approx 0$), the approximation $\tan \theta \approx \theta$ holds, which means $d_p \approx d_s$. This indicates that there is minimal distortion near the point of tangency (i.e., the view direction). However, as θ increases, the value of $\tan \theta$ grows much more rapidly than θ . This non-linear relationship, $d_p/d_s = \tan \theta/\theta > 1$, causes a radial stretching effect. Consequently, features that are uniformly spaced on the spherical surface become increasingly spread out as they are projected farther from the center of the plane. This effect is precisely the spherical-to-perspective distortion observed the left image of Fig. 4, where the periphery of the image appears magnified relative to the center.

To mitigate this distortion, we employ a weighted average with a standard exponential function. While this weighting scheme is not a perfect theoretical inverse of the distortion, our empirical results indicate that more complex models do not yield a significant improvement in quality. Therefore, we just adopted the simplest exponential form, $W_i^{jk} = \exp(-\|u_{jk}\|/\tau)$.

A.3 Coordinate Systems.

We employ a left-handed Cartesian coordinate system where the $+X$, $+Y$, and $+Z$ axes point right, down, and forward (towards the camera), respectively. For panoramic representation, we use a spherical coordinate system defined by an azimuth angle $\theta \in [0^\circ, 360^\circ)$ and an elevation angle $\phi \in [-90^\circ, 90^\circ]$. As illustrated in Fig. 7, the azimuth θ represents the horizontal rotation, starting from the $+Z$ axis and increasing clockwise. The elevation ϕ represents the vertical angle from the horizontal plane; a value of $\phi = 0^\circ$ corresponds to the horizon, while negative and positive values correspond to the upper (upward-looking) and lower (downward-looking) hemispheres, respectively.

B Details of Multi-Prompt Inference

We further describe the multi-prompt inference method in Section 3.5, accompanied by a visual illustration (Section B.1), and provide additional details on foreground object generation (Section B.2).

B.1 Visual Illustration of Multi-prompt Inference

Unlike standard images and videos, 360° panoramic content exhibits substantial variation depending on spatial location. For instance, as shown in Fig. 8 (a), a 360° panorama exhibits significant variation depending on the vertical position in the ERP representation—lower regions often depict the ground, middle regions show elements like blossoms, and upper regions correspond to the sky. To address this, we provide three distinct prompts corresponding to different elevation levels of the spherical latents. A visual illustration of the condition mapping function is shown in both spherical and ERP formats in Fig. 8, where each color represents a distinct prompt. Examples of the prompts are listed in Section G and visualized in Fig. 17. For a fair comparison, we

Method	Runtime (s)	Panoramic Criteria		Image Criteria		Text Adherence	
		Distortion \uparrow	End Continuity \uparrow	Image Quality \uparrow	Aesthetic Appearance \uparrow	Scene \uparrow	CLIP-Score \uparrow
Nearest sampling	161	2.039	3.400	3.014	4.057	0.3250	27.19
+ Weighted Avg.	162	2.829	4.625	4.139	4.782	0.5125	28.26
Dynamic sampling	184	2.421	4.086	3.454	4.111	0.1375	25.92
+ Weighted Avg.	185	3.238	4.892	4.496	4.685	0.5875	28.65

Table 3: Automated quantitative ablation study in generating 360° static wallpaper generation (SANA, A100-40GB) including runtime. Dynamic latent sampling improves distortion and end-continuity, while the distortion-aware weighted averaging significantly improves the image quality and text adherence.

apply the same multi-prompt inference strategy and prompt settings to the applicable baseline method (Liu et al. 2024).

B.2 Foreground Generation

To synthesize more complex scenes, our method incorporates additional view directions specifically designated for foreground subjects. While the core generation pipeline remains unchanged, it is augmented with one or two foreground-specific views, as illustrated in Fig. 8(b). For example, a turtle is added as a foreground object in the red-highlighted region.

To ensure that the foreground prompt dominates within the specified region and achieves a clearer separation between the subject and the background, we replace the exponential weighting function with a beta kernel, applied only to the foreground prompt. The weight is defined as:

$$W_i^{jk} = \mathcal{B}(\|\mathbf{u}_{jk}^*\|; b) \quad (13)$$

$$\mathcal{B}(x; b) = (1 - x)^{4e^b}, \quad \text{where } x \in [0, 1], b \in \mathbb{R}. \quad (14)$$

Here, \mathbf{u}_{jk}^* is a normalized distance vector, where $x = \min(\|\mathbf{u}_{jk}\|/d_{\max}, 1)$ and d_{\max} denotes the maximum distance. The shape of the kernel is controlled by parameter b ; we set $b = -3.0$ to emphasize the foreground region. Compared to the exponential kernel, the beta kernel produces a flatter distribution, allowing the model to generate foreground objects that remain dominant throughout the entire foreground region.

Refinement Stage. As suggested in previous studies (Liu et al. 2024; Çapuk et al. 2025), we introduce a refinement stage to enhance the visual fidelity and seamlessness of the final output. After the initial panorama generation, we perform a noise-to-denoise process: noise corresponding to a specific timestep is added to the panorama, which is then refined by a pretrained diffusion model. This leverages the model’s generative capabilities to improve fine details and overall visual coherence. The refinement is selectively applied only to panoramas that include foreground elements (*i.e.*, only for SANA in Fig. 1) due to the increased computational cost, but it is particularly effective at reducing seams around the foreground-background boundary.

C In-depth Ablations

In the following sections, we present the quantitative results of the ablation study in Section 4.3, which were pre-

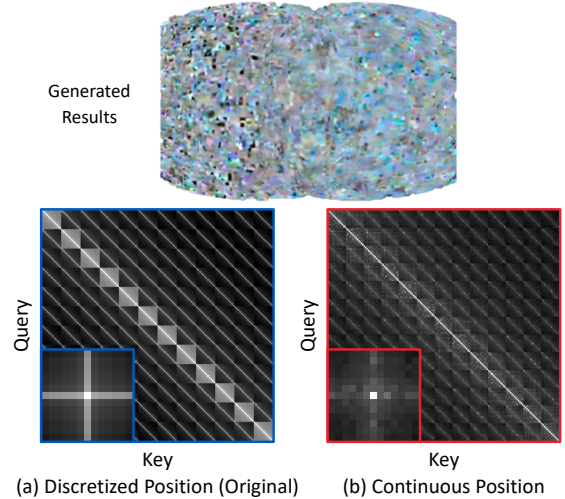


Figure 9: **Similarities Between Positional Embeddings.** The small squares represent the similarity when the central pixel is used as the query. Discretized position originally provides high similarity within the same row or column. In contrast, even slight variations in continuous position result in a significant drop in similarity. As a result, it fails to generate reasonable results due to the significant distribution shift.

viously shown only visually (Section C.1). We then discuss our exploration of reasonable alternatives for mapping spherical latents and the rationale behind our final choice (Section C.2).

C.1 Ablation Study with Automated Quantitative Metrics

As shown in Tab. 3, we conducted a quantitative ablation study on 360° static wallpaper generation (SANA) to isolate and measure the impact of our core components: dynamic sampling and weighted average. Our analysis compares the full model against ablated versions, including a baseline that utilizes nearest point sampling (detailed in Section 3.3) without a weighted average. The results indicate that dynamic sampling is crucial for improving panorama quality metrics. In addition, the weighted average is primarily beneficial for image alignment and text adherence criteria in both nearest and dynamic sampling. Ultimately, our full

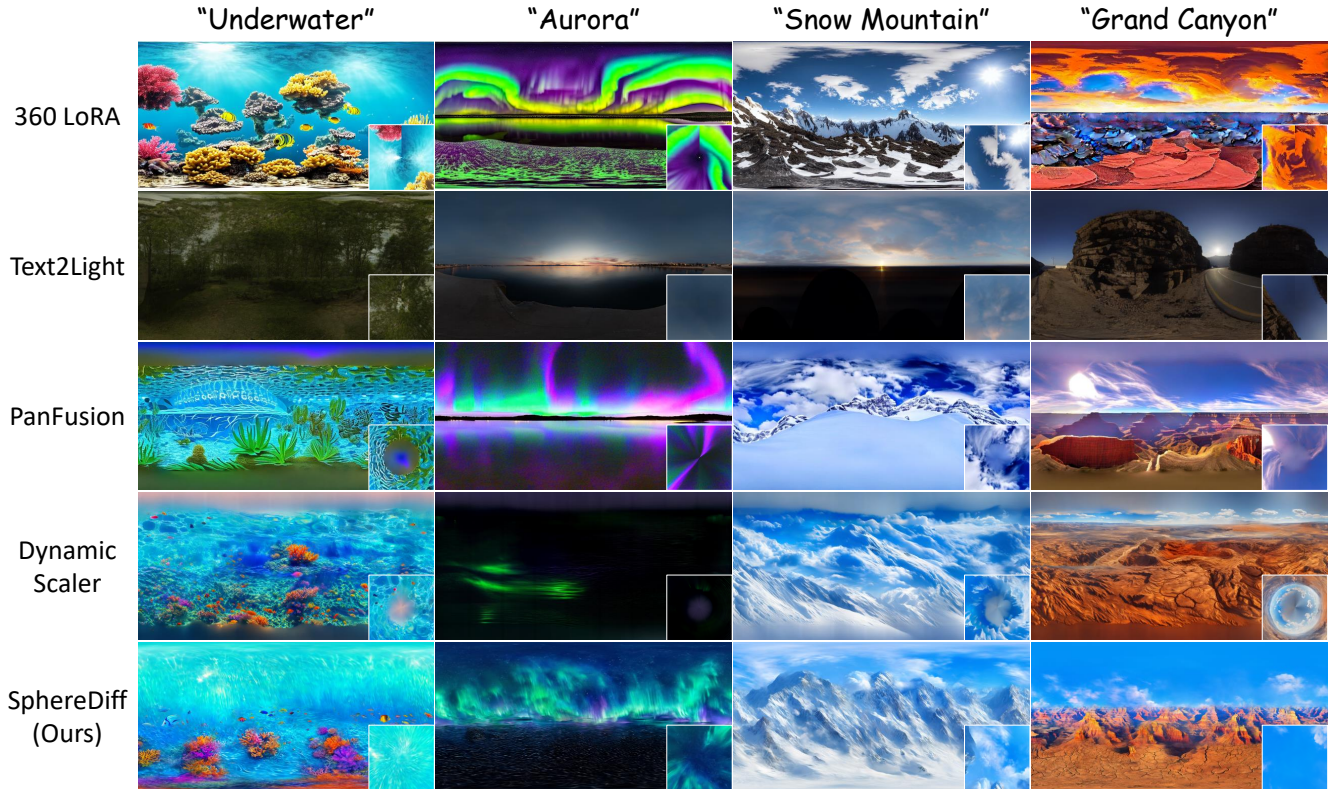


Figure 10: **Qualitative Comparison with Full ERP (360° static wallpapers).** The first column shows the full ERP views corresponding to Fig. 5. The baselines often fail to generate seamless results near the poles, indicating their inability to produce true 360° panoramas. While Text2Light (Chen, Wang, and Liu 2022) generates visually plausible 360° wallpapers, it struggles to adhere to the input text prompts due to the limited diversity of its training dataset. In contrast, our method produces coherent 360° panoramas that better reflect the given text conditions.

model, which combines both techniques, demonstrates the best overall performance on most metrics. On the aesthetic appearance metric, its performance is comparable, confirming the synergistic benefits of our proposed methods without sacrificing visual quality.

C.2 Alternatives for Mapping Spherical Latent

Directly Use Continuous Position. Recent visual generation models, including those based on DiT (Peebles and Xie 2022), support continuous 1D representations through corresponding positional embeddings. A naive approach would be to leverage this property and treat the latent representations as continuous without discretization. However, this approach leads to unstructured outputs due to a distribution shift in positional embeddings. Specifically, in the original positional embedding space, latent similarities are high within the same row or column, ensuring spatial consistency. In contrast, when using continuous positional embeddings, the similarity between two adjacent points is not necessarily high, even if their spatial coordinates are close, as shown in Fig. 9. This discrepancy causes the model to fail in generating structured content. Although DiT can process continuous inputs, discretization remains essential for tuning-free

panoramic visual generation to maintain structured and consistent latent relationships.

Latent Interpolation Methods. To map the spherical latents onto perspective, interpolation is typically used for RGB images. Conventional interpolation methods, such as bilinear interpolation, when applied in latent space do not provide satisfactory results due to the lack of interpolation-equivariant properties in VAEs. Thus, stochastic warping methods (Chang et al. 2024; Daras et al. 2024) have been proposed for warping the latents. In our experience, we observed that they also provide suboptimal results in generating 360° live wallpapers, and thus we adopt sampling methods rather than interpolation methods. Nevertheless, further studies on improving spherical latent sampling could be a promising direction for 360° wallpaper generation.

D Detailed Comparison with Baselines

We provide qualitative comparisons with baselines in ERP format (Section D.1) and present a detailed discussion of related work on 360° panorama generation (Section D.2).

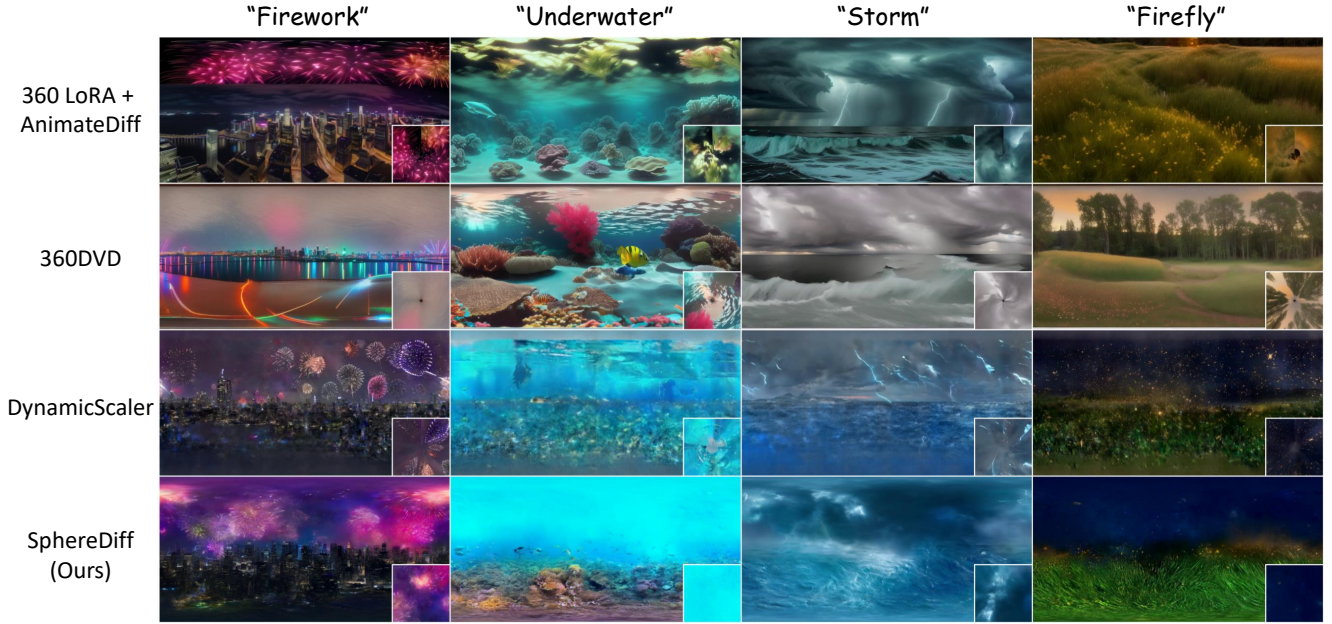


Figure 11: **Qualitative Comparison with Full ERP (360° Live Wallpapers)**. The last column presents the full ERP views corresponding to Fig. 5. Compared to our method, baseline approaches often struggle to maintain spatial continuity near the poles, revealing their limitations in achieving seamless 360° generation in dynamic settings. Please refer to the supplementary material for the full video results.

D.1 Qualitative Comparison in ERP format

Qualitative ERP Comparison of 360° Static Wallpaper Generation Methods. We provide full ERP comparisons for 360° static and live wallpaper generation methods in Fig. 10 and Fig. 11, respectively. In the static wallpaper comparison, although other methods appear to generate reasonable panoramas, they often fail to produce seamless 360° panoramas under the ERP constraints or to faithfully reflect the input text prompts. Specifically, as shown in the bottom-right corner, 360 LoRA (LatentLabs360 2023), PanFusion (Zhang et al. 2024), and DynamicScaler (Liu et al. 2024) often fail to generate seamless upper-view perspectives and exhibit distortion, blurriness, and speckling. While Text2Light (Chen, Wang, and Liu 2022) can generate seamless results even for the upper-view perspective, it struggles to produce content that aligns with the given prompts.

Qualitative ERP Comparison of 360° Live Wallpaper Generation Methods. In the live wallpaper comparison, a similar trend is observed. For example, DynamicScaler (Liu et al. 2024) shows the same limitation, still producing stretched fireworks, as shown in the first column of Fig. 11. We provide video comparisons in the supplementary material. Since 360 LoRA + AnimateDiff (Guo et al. 2023) and 360DVD generate only 16 frames, we interpolate additional frames for visualization, while DynamicScaler and our method generate 121 frames. Although 360 LoRA + AnimateDiff appears to produce reasonable 360° live wallpapers, it struggles to animate the content meaningfully, as shown in our video. For example, fireworks and storms lack dynamic motion; instead, they merely shift from left to right

as the viewpoint changes. 360DVD also fails to generate seamless results near the poles and to faithfully reflect the input text prompts, likely due to the limited diversity in its training 360° video dataset, as evidenced by our extensive quantitative results (Tables 1 and 2 and fig. 13). In contrast, our method faithfully adheres to the input text prompts and strictly respects the ERP constraints, enabled by our use of a spherical latent representation that naturally overcomes the limitations of ERP-based latent spaces.

More Results with Advanced Diffusion Backbones. Given the tuning-free nature and versatility of the proposed approach, we can apply our method to recent advanced diffusion models. First, we apply our method to FLUX (Labs 2024), a powerful text-to-image model, for 360° static wallpaper generation. Our approach is fully compatible with its architecture without requiring any training, enabling the creation of high-quality panoramic images. As illustrated in Fig. 12, this combination yields outputs with superior fidelity while adhering to panoramic constraints. To further demonstrate its versatility, we also extend our method to generate 360° live wallpapers using Hunyuan-Video (Kong et al. 2024), a state-of-the-art open-sourced text-to-video model. This integration is also tuning-free and successfully produces visually compelling videos that preserve panoramic consistency.

Automated Quantitative Evaluation with Standard Deviations. We provide detailed automated quantitative evaluation results, including standard deviations. As shown in Fig. 13, the proposed method consistently demonstrates im-

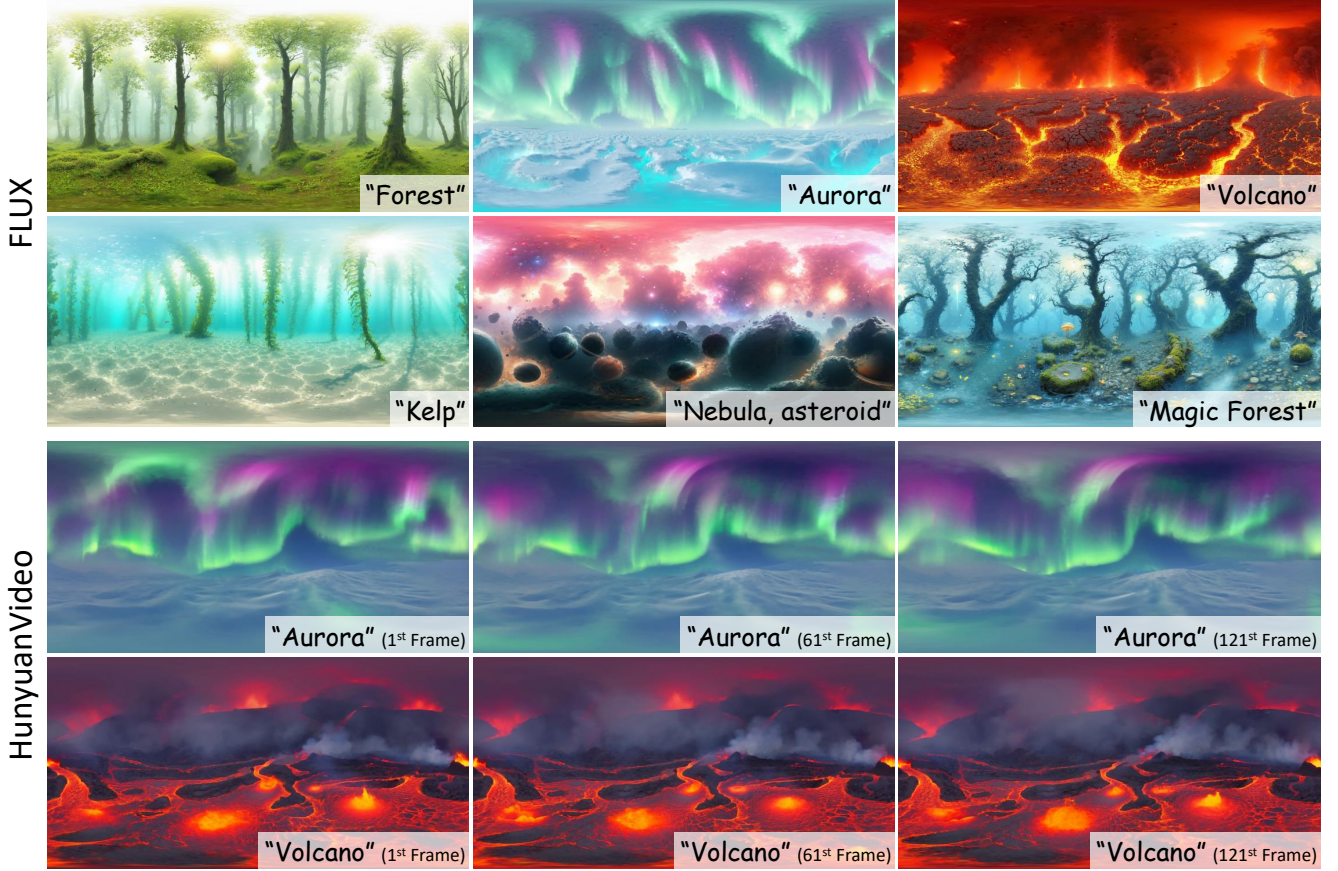


Figure 12: **Qualitative Results with Different Diffusion Backbones.** We extend our method to advanced large-scale diffusion models (Labs 2024; Kong et al. 2024), beyond the efficient backbones used in our main experiments (Xie et al. 2024; HaCohen et al. 2024). These stronger backbones yield significant improvements in output quality, as reflected in the generated results. Notably, FLUX enables the creation of realistic yet fantastical 360° wallpapers (e.g., nebula, asteroid, magic forest), leveraging the creative capacity of text-to-image models to depict scenes beyond the scope of existing ERP datasets. Video results generated with HunyuanVideo are provided in the supplementary materials.

improvements beyond the 1-sigma margin of the second-best method, statistically supporting its superiority. Although baseline methods show comparable performance in image and video criteria, our approach significantly outperforms them in panoramic criteria and text adherence. To ensure a fair comparison with DynamicScaler (Liu et al. 2024), we do not include automated scores for 360° wallpapers generated by advanced diffusion models (Labs 2024; Kong et al. 2024). However, as qualitatively shown in Fig. 12, our method can achieve even better scores when paired with stronger base diffusion models.

D.2 Detailed Comparison with the Other Panorama Generation Approaches

Data-driven 360° Static Wallpaper Generation. Generating images in the Equirectangular Projection (ERP) format is challenging due to its intrinsic properties. As previously discussed in Fig. 2, latents are unevenly distributed over the spherical surface, which leads to distortions and discontinu-

ities when projected to the ERP domain. For instance, latents corresponding to the topmost row in ERP must share the same value, as they represent a single spatial point (the pole). Previous studies have attempted to address these issues through data-driven training with specialized strategies (Zhang et al. 2024; Feng et al. 2023; Wu, Zheng, and Cham 2024; Tang et al. 2023; Chen, Wang, and Liu 2022) or by using alternative representations (Wu et al. 2024; Liao et al. 2023; Kalischek et al. 2025).

However, high-quality 360° panoramic images remain scarce, and most existing datasets focus on indoor scenes, such as Matterport3D (Chang et al. 2017), Structured3D (Zheng et al. 2020), and the dataset used in MVDiffusion (Tang et al. 2023), each containing approximately 10K images. Consequently, existing methods either fail to properly adhere to the 360° constraints (LatentLabs360 2023), or struggle to adhere to the input text prompts (Chen, Wang, and Liu 2022) as shown in Fig. 10. This issue becomes particularly severe for applications like live wallpaper generation, which suffer from an even greater lack of data,

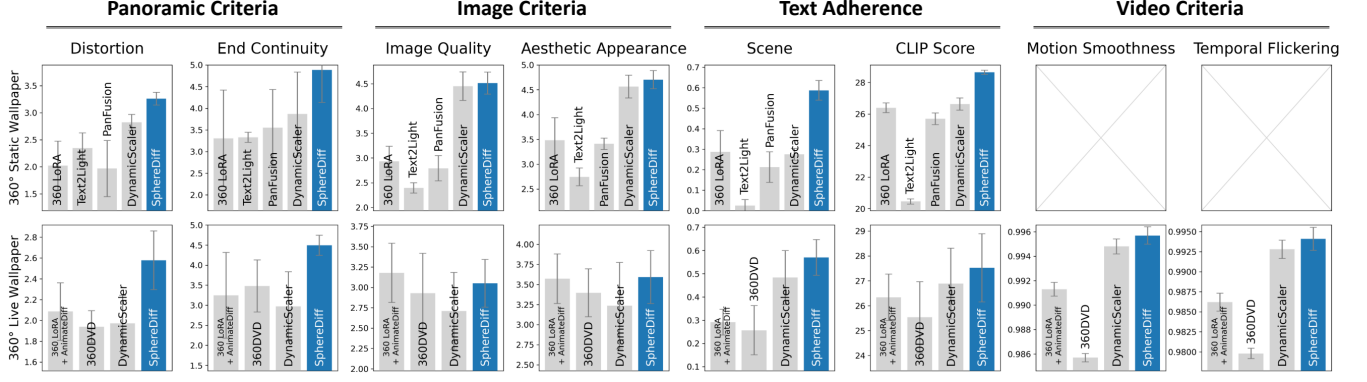


Figure 13: **Automated Quantitative Evaluation with Standard Deviations.** We report the performance of each method across all evaluation criteria, including image, video, panoramic, and text-level aspects. The gray line indicates the 1-sigma error bar. The proposed method consistently surpasses the second-best method in most metrics, especially in panoramic criteria.

as summarized in Tab. 4.

Data-driven 360° Live Wallpaper Generation. The scarcity of data is even more pronounced for 360° live wallpaper content than for static wallpapers. For instance, 360DVD (Wang et al. 2024) extends pretrained text-to-video models using an adapter, but it relies on a small dataset of merely 2,000 video clips, which is insufficient to achieve consistently high-quality results. To overcome the data bottleneck, GenEx (Lu et al. 2024) captures panoramic videos from simulations to build a custom dataset. While this is an effective strategy for data augmentation, it introduces a critical constraint: the model’s output is confined to the distribution of the synthetic simulation, limiting its ability to generate photorealistic content and instead producing results that appear artificial. Our method circumvents these data dependency issues as it can be directly applied to contemporary video diffusion models. This allows for the generation of diverse 360° dynamic content without being constrained by the limitations of a domain-specific, often small or synthetic, training dataset.

Tuning-free 360° Panorama Generation. To address the scarcity of 360° video datasets, recent studies (Li et al. 2024; Liu et al. 2024) have attempted to generate panoramic videos using tuning-free methods. 4K4DGen (Li et al. 2024), for example, adopts an image-to-video pipeline by animating images generated from 360 LoRA (LatentLabs360 2023). However, this approach not only inherits the intrinsic limitations of its static image backbone but also restricts creative flexibility, as it requires a pre-existing source image for generation, as discussed in the related work section. In contrast, direct text-to-video approaches offer greater flexibility but present different technical challenges. DynamicScaler (Liu et al. 2024), for instance, introduces an Offset Shifting Denoiser (OSD) to improve end-to-end continuity in the ERP format. Nevertheless, because it still relies on ERP, the method produces noticeable blurry artifacts in the polar regions, as illustrated in Figures 10 and 11.

In contrast, our method addresses these issues by adopting a spherical representation. This design not only ensures a

Method	Latent Space	Tuning-Free	Wallpaper type
360 LoRA	ERP	✗	static
Text2Light (TOG’22)	ERP	✗	static
PanFusion (CVPR’24)	ERP	✗	static
Cubediff (ICLR’25)	Cube Map	✗	static
360 LoRA + AnimateDiff	ERP	✗	live
360DVD (CVPR’24)	ERP	✗	live
DynamicScaler (CVPR’25)	ERP	✓	static and live
SphereDiff (Ours)	Spherical	✓	static and live

Table 4: **Comparison of 360° static/live wallpaper generation approaches.** Most existing 360° wallpaper generation approaches perform denoising within the equirectangular latent space, whereas our method leverages spherical latents. Among these methods, only DynamicScaler (Liu et al. 2024) and ours support both static and live wallpaper generation due to their tuning-free design.

uniform latent distribution that eliminates polar artifacts, but also facilitates the direct application of text-to-video models. As a result, our approach enables high-quality, seamless, and tuning-free generation of both images and videos.

E Implementation Details

Hyperparameters. In all MultiDiffusion setups (Bar-Tal et al. 2023), the base resolution and temporal length are configured to optimize the performance of each base model. For instance, we use a resolution of 1024×1024 for SANA (Xie et al. 2024), FLUX (Labs 2024), and HunyuanVideo (Kong et al. 2024), and 512×512 for LTX-Video (HaCohen et al. 2024). The temporal length is fixed at 121 frames for both LTX-Video and HunyuanVideo. For the spherical latent representation, we use 2,600 latent points on the sphere and adopt an 80° field of view for each perspective, with 60% overlap between neighboring views across all experiments. The view directions are defined as follows:

- $\phi = \pm 90^\circ: \theta = \frac{360^\circ}{4} \cdot k, ; k \in 0, 1, 2, 3$
- $\phi = \pm 77.5^\circ: \theta = \frac{360^\circ}{8} \cdot k, ; k \in 0, \dots, 7$

Evaluation Prompt

You are an evaluator assessing an image generation model based on a single image at a time. Your evaluation is based on the following four criteria:

1. **Image Quality:** Assess the overall quality of the image.
 2. **Aesthetic Appeal:** Evaluate how visually pleasing the image is.
 3. **Distortion Level:** Determine whether the image appears distorted. If it does not resemble a photo taken with a normal camera, it will receive a lower score.
 4. **Connectivity:** Check if the middle of the image appears disconnected. If there is a noticeable break, the score will be lower.
- Each criterion is rated on a five-point scale: Excellent (5), Good (4), Fair (3), Poor (2), and Awful (1).

You will receive one image at a time. For each criterion, provide a concise reason for the score before listing the rating.

Format your response as follows:

- **Image Quality:** (Brief reason) → Score
 - **Aesthetic Appeal:** (Brief reason) → Score
 - **Distortion Level:** (Brief reason) → Score
 - **Connectivity:** (Brief reason) → Score
- {image}

Please evaluate the image with the given criteria.

Table 5: **Evaluation Prompt for VLM.** The evaluation prompt used to assess the quality of 360° live and static wallpaper generation based on four image and panoramic criteria. We instruct the VLM (Hurst et al. 2024) to provide a score along with a brief reason to encourage chain-of-thought (Wei et al. 2022).

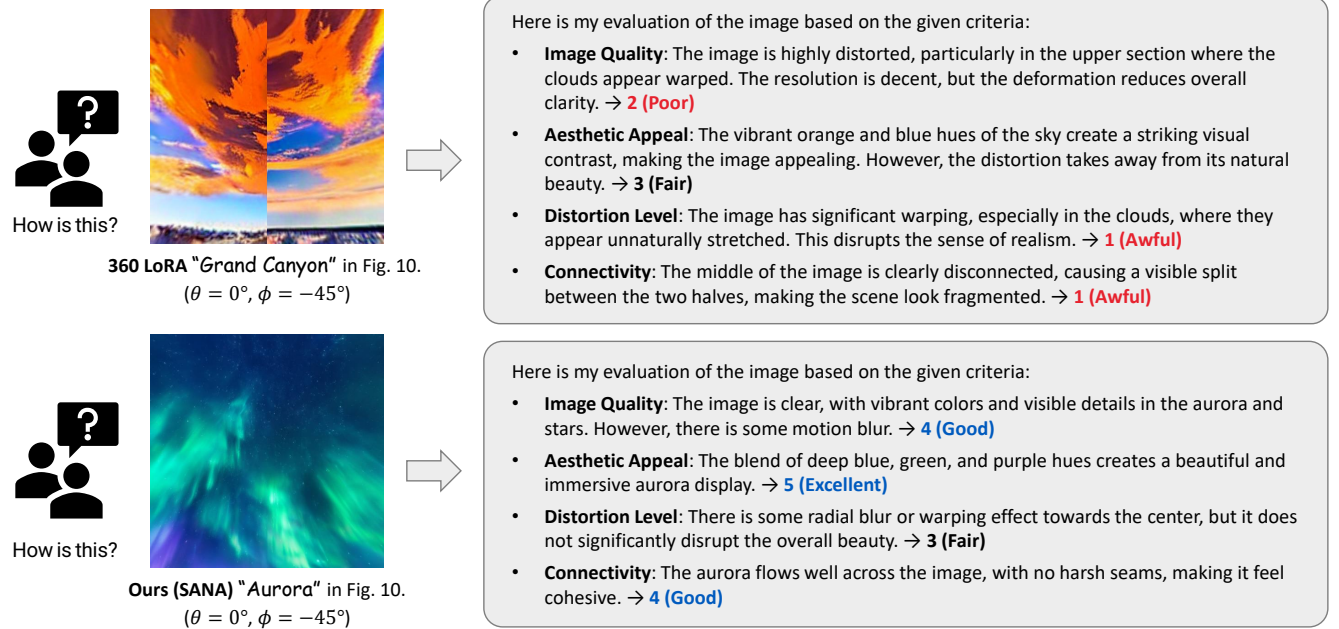


Figure 14: **Example of VLM Responses.** The VLM (Hurst et al. 2024) provides a concise justification along with a corresponding score. Interestingly, we observe that the VLM demonstrates awareness of panoramic-specific artifacts such as distortion and end continuity, despite these aspects being relatively underexplored in prior VLM-based evaluation research.

- $\phi = \pm 45^\circ$: $\theta = \frac{360^\circ}{11} \cdot k$, $k \in 0, \dots, 10$
- $\phi = \pm 22.5^\circ$: $\theta = \frac{360^\circ}{14} \cdot k$, $k \in 0, \dots, 13$
- $\phi = 0^\circ$: $\theta = \frac{360^\circ}{15} \cdot k$, $k \in 0, \dots, 14$

This results in a total of 89 view directions. Although we reduce the number of directions to speed up inference, this reduction can slightly compromise seamlessness due to insufficient overlap. Nonetheless, reducing the number of views while preserving quality remains a promising direc-

tion for future research on efficient panoramic generation.

Tiled VAE decoding. To visualize the results, we decode the denoised latent representation S_0 for each view direction using a VAE decoder and stitch them together to construct an ERP image. During this decoding process, we apply distortion-aware weighted averaging techniques to ensure seamless integration.

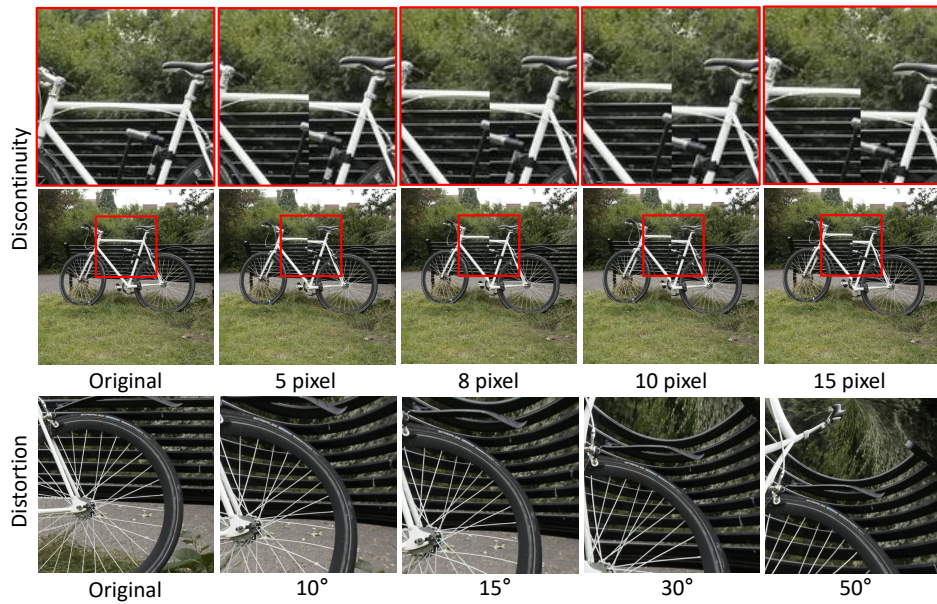


Figure 15: **Exemplars of Synthetic Distortion and Discontinuity.** To evaluate how well the VLM (Hurst et al. 2024) can recognize panoramic artifacts, we construct synthetic examples with controlled levels of distortion and discontinuity. We vary the severity of both artifact types, as shown in the figure.

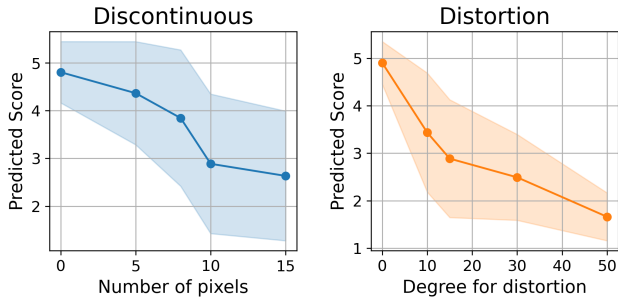


Figure 16: **Predicted Scores by GPT-4o (Hurst et al. 2024) for Distortion and Continuity.** The scores are reported on a 1–5 scale, and the shaded area in the plot represents the 1-sigma range. GPT-4o demonstrates strong sensitivity to the severity of both artifact types, providing scores that closely align with the levels of distortion and discontinuity, thereby indicating the reliability of its score predictions.

Runtime Analysis. Inference time per sample is approximately 3 minutes for image generation with SANA (Xie et al. 2024) and 20 minutes for video generation with LTX-Video (HaCohen et al. 2024), measured on an NVIDIA A100-40GB GPU. Since our method does not require additional memory beyond the base requirements of the underlying models (SANA and LTX-Video), we verified that it can also run on GPUs with 24GB of VRAM. Runtime analysis for each ablation is summarized in Tab. 3.

When using more advanced diffusion backbones, computational demands increase significantly. Specifically, FLUX (Labs 2024) requires 40GB of VRAM and takes ap-

proximately 12 minutes per sample on a single H200 GPU, while HunyuanVideo (Kong et al. 2024) consumes about 50GB of VRAM and requires around 3 hours for video generation. Although our method incurs non-trivial inference time, it achieves state-of-the-art performance in both image and video quality. Nevertheless, the relatively long runtime remains a limitation inherited from the MultiDiffusion (Bar-Tal et al. 2023) framework. Reducing inference time, for example by minimizing the overlapping region, presents a promising direction for future work.

F Evaluation Details

In this section, we provide additional evaluation details that could not be included in the main paper due to space constraints. Specifically, we present the details of the user study (Section F.1) and provide the justification for the VLM-based visual assessment (Section F.2).

F.1 User Study Details

We divided the 20 pairs of images and videos into five sets, each containing four pairs, allowing users to select one set for evaluation based on their convenience. To accurately assess distortion and end-continuity, images and videos were presented from a viewpoint with a fixed single azimuth angle ($\theta = 0^\circ$) for evaluating end-continuity while varying the elevation angle (ϕ) across five positions: $-90^\circ, -45^\circ, 0^\circ, 45^\circ, 90^\circ$, as illustrated in Fig. 5. The presentation order was randomized to minimize bias. For image evaluation, participants selected the most suitable model from five available options, while for video evaluation, they chose from four models, selecting the one they found most appropriate for each criterion.

F.2 VLM-based Visual Score

Preprocess for Automated Evaluation. To automatically evaluate the 360° static and live wallpapers, we project the ERP into perspective views to fully leverage pretrained models trained on standard perspective images. To this end, all metrics are computed on perspective images or videos rendered with a 90° field of view (FoV) at 14 predefined view directions. The selected views include four azimuth angles (0°, 90°, 180°, 270°) at each of three elevation angles (45°, 0°, and −45°), along with one view each at the top and bottom poles (90° and −90°).

Justification of VLM-based Evaluation. Since tuning-free 360° wallpaper generation methods do not rely on the source domain, and it cannot easily utilize FID-based metrics (Heusel et al. 2017) for measuring image and panoramic criteria. Specifically, several approaches have attempted to evaluate the quality of the generated images or videos with vision-language models (Wang et al. 2025; You et al. 2024b,a; Zhou et al. 2024), in terms of blur, noise, exposure, or the overall descriptive aesthetic assessment. Thus, we utilize GPT-4o (Hurst et al. 2024) for automatic evaluation for the image criteria, within the LLM-as-a-judge framework (Zheng et al. 2023). Tab. 5 details the prompt that we used for evaluate, and Fig. 14 illustrates the response examples of the VLM.

Automated Assessment for Panoramic Criteria. In a manner analogous to the LLM-as-a-judge (Zheng et al. 2023) paradigm, we also utilize GPT-4o (Hurst et al. 2024) to quantify levels of distortion and discontinuity in images. To validate the reliability of this automated evaluation, we tested GPT-4o on a curated set of 72 images exhibiting these artifacts, which were generated by randomly cropping 8 samples from 9 distinct scenes within the Mip-NeRF 360 (Barron et al. 2022) dataset. To evaluate the reliability of panoramic quality metrics, we introduced two types of synthetic image degradation—distortion and discontinuity—at

five different intensity levels, including a no-degradation case, as visually illustrated in Fig. 15.

Specifically, discontinuity is introduced by shifting the left half of the image vertically, breaking the seamless alignment. We applied four levels of displacement: 5, 8, 10, and 15 pixels, which correspond to about 1.2%, 1.9%, 2.4%, and 3.7% of the total image height (411 pixels), respectively. Distortion is simulated by projecting ERP to perspective views, which introduces curvature and warping artifacts. This reflects the common distortion pattern that arises when 360° panoramas are generated as if they were regular flat images, ignoring the spherical geometry. The intensity levels corresponded to elevation shifts in the view direction of 10°, 15°, 30°, and 50°, respectively.

As shown in Fig. 16, our automated assessment framework demonstrates its capability to measure key panoramic criteria. We observe a clear and systematic trend where GPT-4o (Hurst et al. 2024) continuously assigns lower scores as the intensity of distortion and discontinuity increases. This consistent, inverse relationship validates our LLM-based approach as a reliable method for quantifying such panoramic artifacts.

G Used Text Prompts

Our prompts cover 20 different scene concepts, primarily focusing on natural landscapes. To generate 360° live wallpapers, we use GPT-4o (Hurst et al. 2024) to enrich the prompts following the prompt engineering guidelines provided on its official pages, as LTX-Video (HaCohen et al. 2024) struggles with very short prompts. In contrast, we observed that SANA (Xie et al. 2024) is relatively insensitive to prompt variations, so we use the same set of prompts for generating 360° static wallpapers. While the full prompt set will be publicly released, we provide several representative examples in Fig. 17, which were used to generate the figures in the main paper.

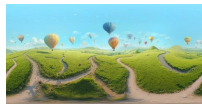


Figure 1.3.
"Air Balloons"
360° Static Wallpaper
generated by FLUX

Upper: clear **blue sky** with clouds. 4k, high-resolution.

Middle: an open countryside landscape with lush **green fields** and a winding gravel path cutting through the terrain. The texture of the small stones on the road is clearly visible, contrasting with the smooth, vibrant grass. The soft shadows of **hot air balloons** above can be faintly seen on the ground, adding depth and movement to the peaceful setting.

Lower: A **high-angle view** of an open countryside landscape with lush **green fields** and a winding gravel path cutting through the terrain. The texture of the small stones on the **road** is clearly visible, contrasting with the smooth, vibrant grass ...



Figure 1.
"Ruins"
360° Static Wallpaper
generated by FLUX

Upper: An **upward view** of the night sky filled with countless **stars and the Milky Way** stretching across, creating a breathtaking cosmic scene. ...

Middle: A grand view of **ancient ruins** under a vast, starry night sky. The weathered stone columns and structures stand as silent witnesses to history, illuminated by the soft glow of moonlight and distant celestial bodies.

Lower: A directly **downward view** of the **ancient ruins**, showing only the moss-covered stone foundations and weathered ground. **Cracked stone pathways** and scattered remnants of fallen pillars blend into the rugged terrain, ...



Figure 1.
"Blossom"
360° Static Wallpaper
generated by FLUX

Upper: A gently **upward-tilted view** captures a dense canopy of **cherry blossoms** in full bloom, with countless delicate pink petals stretching overhead. The intertwining branches frame the sky at an angle, revealing glimpses of bright blue through the soft pink expanse. ...

Middle: A **low-angle view** showcases towering cherry blossom trees in full bloom, their thick, textured **trunks** rising from the ground as branches heavy with pink flowers stretch skyward. Delicate petals drift down in the gentle breeze, covering the soft, lush grass in a gentle pink layer. ...

Lower: From a dramatic **high-angle perspective** above the lush **grass and a root of blossom tree**, a gentle breeze sweeps across the soft, green expanse, causing fallen cherry blossom petals to shift slightly. Near the base of a cherry tree, its roots emerge slightly from the earth, ...

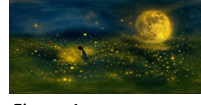


Figure 1.
"Fireflies, Moon, Child"
360° Static Wallpaper
generated by SANA

Foreground 1: A **child** in a grassy meadow at twilight, gently trying to catch glowing **fireflies** in a glass jar, capturing a moment of pure innocence, ...

Foreground 2: In the center of the image a big large **yellow Full moon**, a big **large yellow Full moon** in the dark sky above a field of fireflies, creating a warm and hopeful atmosphere.

Upper: An **upward view** of the vast night sky above an open field, with scattered **fireflies** drifting gently, their faint glows blending with distant stars.

Middle: A serene nighttime meadow, where countless **fireflies** flicker softly, casting a warm, golden light that dances over the swaying grass and wildflowers.

Lower: A dramatic **top-down view** of a vast open **field grass**, where waves of grass ripple in the night breeze, dotted with countless fireflies drifting just above, ...



Figure 1.
"Grand Canyon, Eagle"
360° Static Wallpaper
generated by SANA

Foreground: Photorealistic, dramatic **low-angle** shot of a massive **eagle** in flight, viewed from 50 degrees below, with the clear and blue sky with soft clouds

Upper: clear and **blue sky** with soft clouds

Middle: A breathtaking **panoramic view of the Grand Canyon**, displaying its immense red and orange rock formations carved by the Colorado River. The deep valleys and layered cliffs stretch endlessly, revealing millions of years of geological history under the golden sunlight.

Lower: A **top-down view** of the arid, reddish-brown ground of the **Grand Canyon**, covered in dry, cracked earth and scattered rock formations, radiating heat under the bright sun.



Figure 1.
"Underworld, Turtle"
360° Static Wallpaper
generated by SANA

Foreground: A dynamic **low-angle view** of a **sea turtle** at a 45-degree angle, looking up from below as it glides through sunlit blue water. at the center of the image.

Upper: An **upward view** from **underwater**, looking towards the surface where **sunlight beams** penetrate the clear ocean. Gentle ripples create a shimmering effect, and the water transitions from deep blue to a lighter, ...

Middle: A stunning **underwater** scene filled with vibrant tropical fish swimming gracefully through the crystal-clear water. Various species, from small neon-colored fish to larger, elegant ones, move in harmony...

Lower: A breathtaking **top-down view** of a colorful tropical coral reef. The ocean floor is covered with vibrant **corals**, ranging from bright orange and pink to deep purple and blue. Small fish dart between the coral formations, ...



Figure 1.
"Storm"
360° Live Wallpaper
generated by HunyuanVideo

Upper: Dark **storm clouds** swirl overhead as multiple lightning bolts strike at different moments, briefly illuminating the chaotic sky. The jagged bolts cut through the darkness, revealing shifting cloud formations in flashes of electric blue and white. The perspective is an upward view, ...

Middle: A **mid-angle view** reveals the vast **ocean** meeting the storm-filled sky, where towering waves rise and fall beneath the relentless tempest. Lightning bolts crack ...

Lower: A **high-angle view** captures the vast, turbulent deep blue ocean as powerful waves crash and churn beneath the storm's force. **White foam swirls** atop the restless sea, contrasting against the dark depths. ...

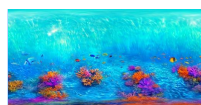


Figure 5.
"Underwater"
360° Static Wallpaper
generated by SANA

Upper: An **upward view** from **underwater**, looking towards the surface where sunlight beams penetrate the clear ocean. Gentle ripples create a shimmering effect, and the water transitions from deep blue to a lighter, ...

Middle: A stunning **underwater** scene filled with vibrant tropical fish swimming gracefully through the crystal-clear water. Various species, from small neon-colored fish. ...

Lower: A breathtaking **top-down view** of a colorful tropical coral reef. The ocean floor is covered with vibrant **corals**, ranging from bright orange and pink to deep purple and blue. Small fish dart between the coral formations, ...

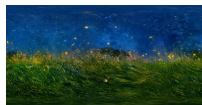


Figure 5.
"Firefly"
360° Live Wallpaper
generated by LTX-Video

Upper: An **upward view** of the vast **night sky** above an open field, with scattered fireflies drifting gently, their faint glows blending with distant stars.

Middle: A serene nighttime meadow, where countless **fireflies** flicker softly, casting a warm, golden light that dances over the swaying grass and wildflowers.

Lower: A dramatic **top-down view** of a vast open field grass, where waves of grass ripple in the night breeze, dotted with countless **fireflies** drifting just above, their soft glow flickering across the landscape.



Figure 2.
"Forest"
360° Static Wallpaper
generated by SANA

Upper: a **low angle view** of **forest**. high-quality, 4k, detailed, realistic.

Middle: a **high angle view** of the Dense, **green forest** interior with sunlight filtering through the canopy, moss-covered ground, towering trees, and a misty, tranquil atmosphere.

Lower: a **top-down view** of **moss-covered ground**. high-quality, 4k, detailed, realistic.

Figure 17: **Prompts Used for Generating the 360° Static and Live Wallpapers in the Main Paper.** We present the text prompts corresponding to the ERP results shown in the main paper. Note that only Fig. 1 (SANA) utilizes an additional foreground prompt; all other examples are generated without it. The complete list of 20 prompts will be released with the code.