SPHEROPE: ZERO-SHOT OPTIMIZATION-FREE 360° PANORAMA GENERATION WITH SPHERICAL ROPE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a zero-shot, training-free and optimization-free framework for generating 360° panoramic images and videos by directly injecting spherical priors into pre-trained diffusion transformers. Existing methods either rely on costly fine-tuning on scarce panoramic data that limits generalization, or leverage multi-step optimization that incurs prohibitive inference latency. We observe that contemporary generative models natively exhibit some panoramic priors from large-scale training. However, these emergent capabilities are insufficient, as the models fundamentally fail to satisfy the rigorous topological constraints imposed by equirectangular projection (ERP). We introduce a zero-shot and optimization-free approach that resolves these constraints at inference time. **Spherical RoPE** replaces standard rotary position embeddings: low-frequency channels are re-parameterized as 3D Cartesian coordinates to natively encode the spherical manifold, while high-frequency channels are harmonically quantized to enforce exact 2π periodicity. Coupled with complementary Semantic Distortion classifier-free guidance (CFG) that explicitly steers geometry, we avoid retraining and inherit the full creative breadth of state-of-the-art models. Our approach generalizes across diverse backbones and 360°generation modalities. We demonstrate this across text-to-panorama using Flux.1, Flux.2, and LTX-Video backbones, achieving competitive performance against baselines, all while remaining training-free.

1 INTRODUCTION

The rapid evolution of image and video diffusion models (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Ho et al., 2022) has fundamentally shifted the landscape of content creation, enabling the synthesis of high-fidelity visual assets from natural language descriptions. While these models have achieved remarkable success in generating planar, perspective-frame images and videos, there is an increasing demand for omnidirectional generative frameworks capable of producing immersive 360° panoramic environments. Such content is critical for applications in virtual reality (VR) (Brivio et al., 2021) and robotics simulation, where a narrow field-of-view is insufficient to capture the global context of a scene (Wang et al., 2025b). In this work we address zero-shot and optimization-free text-to-panorama generation across different modalities, including static images, dynamic videos, and integrated audio-video environments.

360° panoramas are typically represented using an equirectangular projection (ERP), which maps spherical imagery onto a 2D rectangular plane. Diffusion transformers (DiTs) (Peebles and Xie, 2023), as employed in recent models such as Flux (Labs et al., 2025) and LTX-Video (HaCohen et al., 2026), rely on rotary position embeddings (RoPE) (Su et al., 2024) and planar attention. These mechanisms inherently assume a Euclidean grid, leading to visible seams at the boundaries and geometric incoherence at the poles when associated with ERP-based generation. Existing research attempting to bridge this gap generally follows two paradigms. *Training-based methods* fine-tune pre-trained models on specialized panoramic datasets, either through full fine-tuning (Feng et al., 2023), parameter-efficient adaptation via LoRA (Hu et al., 2022; Wang et al., 2024a; Zhang et al., 2024; Ye et al., 2024), or specialized architectures such as spherical manifold convolutions (Sun et al., 2025). However, these approaches are constrained by the scarcity of high-quality 360° data. They further suffer from the immense computational cost of retraining large-scale models, degraded generalization in out-of-distribution (OOD) scenarios (Wang et al., 2025b), and the need to re-train

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



Figure 1: **Zero-shot 360°Generation.** Our inference-only approach enables seamless text-to-panorama synthesis across image (FLUX.1, FLUX.2) and audio-video (LTX 2.3) backbones. Each section displays the generated ERP panorama (top) and perspective crops re-projected from the highlighted regions (bottom), demonstrating strict geometric consistency.

whenever the backbone is updated or extended to new modalities. *Optimization-based methods*, such as PanoFree (Liu et al., 2024) and SphereDiff (Park et al., 2026), avoid retraining but rely on iterative warping-inpainting pipelines or patch-based MultiDiffusion (Bar-Tal et al., 2023) in spherical latent spaces, introducing substantial inference-time overhead that makes them impractical for real-time or video applications.

We propose **SpheRoPE**, a zero-shot, training-free and optimization-free framework that generates high-fidelity ERP panoramas by realigning the internal inductive biases of pre-trained DiTs with spherical geometry. Our approach builds upon the insight that pre-trained foundational models already possess rich priors for panoramic environments. When conditioned on panoramic prompts, these models natively generate equirectangular characteristics, yet their standard architectures fail to enforce the rigorous topological constraints imposed by equirectangular projection (ERP). To bridge this gap, we introduce **Spherical RoPE**, a frequency-aware reformulation of the standard transformer RoPE (Su et al., 2024). Under this formulation, high-frequency channels are harmonically quantized to guarantee 2π horizontal periodicity, whereas low-frequency channels utilize 3D Cartesian coordinates on the unit sphere to enforce both polar convergence and periodicity. In addition, we leverage Semantic Distortion CFG that extends standard CFG (Ho and Salimans, 2021) to a three-way scheme using an anchored geometric prompt, steering the denoising process toward valid ERP projections without sacrificing semantic detail. Figure 1 shows results of our method.

A defining advantage of our framework is its architectural generality. By isolating modifications to the positional encoding and guidance logic, our approach generalizes across modern diffusion

transformers without requiring task-specific adaptation. We demonstrate this versatility by leveraging Flux (Labs et al., 2025) for high-fidelity static environments and LTX 2.3 (HaCohen et al., 2026) for 360° video generation. Furthermore, because we do not alter the underlying model weights, the framework strictly preserves the model's foundational capabilities. It seamlessly inherits built-in conditioning pipelines, enabling advanced functionalities such as zero-shot image-to-panorama translation, or creating both video and audio for a scene. Moreover, operating as zero-shot, our method yields highly robust out-of-distribution performance across diverse styles and visual domains.

We evaluate our method on both image and video generation tasks. For static 360° panorama synthesis, we benchmark on ODI-SR (Deng et al., 2021) against training and optimization-based methods. For video generation, we assess performance on two prompt sets using VBench (Huang et al., 2024). Additionally, we conduct an LLM-based perceptual evaluation and a user study, showing a clear preference for the results of our method. Our zero-shot approach achieves state-of-the-art results on several metrics and competitive performance on others, all without any training, optimization, or model-specific adaptation. Finally, we validate our design choices through extensive ablation studies.

Our contributions are summarized as follows:

- We present the first zero-shot, training and optimization-free framework for seamless 360° image and video generation. Our approach modifies only the inference-time spatial priors of pre-trained DiTs, making it plug-and-play for diverse backbones.
- We introduce **Spherical RoPE** to geometrically induce the topological invariants of the sphere (S^2) via spectral decomposition, alongside **Semantic Distortion CFG** to amplify the native panoramic priors of the model during inference.
- We demonstrate the efficacy of our zero-shot approach across image and video benchmarks, achieving highly competitive performance against fine-tuned baselines while consistently outperforming them in panoramic coherence and human preference.

2 Related Work

Text-driven 360 panorama generation. Following early VQGAN-CLIP models (Chen et al., 2022), latent diffusion (Rombach et al., 2022) for panorama generation split into training and optimization paradigms. Training-based methods fine-tune on panoramic datasets via DreamBooth (Ruiz et al., 2023; Feng et al., 2023) or LoRA (Ni et al., 2025; Wang et al., 2025a). To improve spatial consistency, architectures have been extended with MultiDiffusion stitching (Bar-Tal et al., 2023; Wang et al., 2024a), dual-branch networks (Zhang et al., 2024), specialized attention (epipolar (Ye et al., 2024), cubemap (Kalischek et al., 2025)), spherical convolutions (Sun et al., 2025), distortion decoupling (Zheng et al., 2025), and hybrid DiTs (Feng et al., 2025). However, these remain bottlenecked by data scarcity, retraining costs, and limited out-of-distribution generalization. Optimization-based methods avoid fine-tuning. Approaches include factor-graph inference over patches (DiffCollage (Zhang et al., 2023)), iterative warping (PanoFree (Liu et al., 2024)), and distortion-aware MultiDiffusion in spherical latent space (SphereDiff (Park et al., 2026)). While 360PanT (Wang and Xue, 2025) tackles tiling for panorama translation, these training-free methods generally incur high latency and risk global structural inconsistencies due to their patch-based or iterative nature. StochSyn (Yeo et al., 2025) iteratively minimizes the disagreement between many perspective denoising processes projected onto a shared canvas, a formulation its authors derive as an equivalent of score distillation. Each step therefore invokes the backbone once per view rather than once per sample. The extension to 360° video is rapidly emerging. Training-based models introduce adapters (360DVD (Wang et al., 2024b)), DiT backbones (PanoDiT (Zhang et al., 2025)), latitude-aware sampling (PanoWan (Xia et al., 2025)), dual-branch lifting (Imagine360 (Tan et al., 2025)), and geometry-aware conditioning (Argus (Luo et al., 2025)). Training-free video methods include offset-shifting (DynamicScaler (Liu et al., 2025)) and temporal extensions of spherical latents (SphereDiff (Park et al., 2026)). To the best of our knowledge, we are the first to achieve panorama synthesis that is both training-free and optimization-free. Our approach relies solely on inference-time modifications, bypassing resource-intensive training and slow multi-step optimization.

Positional encoding adaptation in transformers. Our training-free approach modifies rotary position embedding (RoPE) (Su et al., 2024) at inference time to encode spherical geometry. Several prior works have explored modifications to RoPE structure. This includes frequency rescaling for

NLP contexts (Chen et al., 2023; Peng, 2023; Peng et al., 2024), adaptations for DiT spatial/temporal axes (Lu et al., 2024; Wang et al., 2024c; Zhao et al., 2025), and dynamic adjustments across denoising timesteps (Zhuo et al., 2024; Issachar et al., 2025). For non-Euclidean spaces, RoPE has been adapted for geolocation tokens in NLP (Unlu, 2023), and classification (van de Geijn et al., 2025). Recently, IaaW (Gui et al., 2025) introduced a spherical RoPE formulation for image-to-panoramic-video generation by multiplicatively entangling latitude, longitude, and time across all channels. This disrupts the pretrained backbone's factorized rotary structure, preventing zero-shot generation and requiring fine-tuning. In contrast, **SpheRoPE** encodes positions on the 2D ERP plane while respecting spherical topology. By modifying only the horizontal RoPE axis and restricting spherical encoding to low-frequency channels via spectral partitioning, we preserve the pretrained positional structure to enable true zero-shot generation.

3 METHOD

We present a zero-shot optimization-free framework for generating ERP panoramas from pre-trained diffusion transformers, requiring no fine-tuning or architectural changes. We begin with our problem formulation and setting in Section 3.1. We then introduce SpheRoPE, our spherical rotary position embedding that encodes horizontal periodicity and polar convergence directly into the attention mechanism in Section 3.2. Finally, we describe Semantic Distortion CFG, a dual-guidance scheme that steers generation toward geometrically valid panoramic structure in Section 3.3. In the supplementary, Section A, we provide detailed formulation of CFG and RoPE.

3.1 PROBLEM FORMULATION

To leverage 2D priors for 360° generation, spherical content is typically projected via ERP (Wang et al., 2025b). Although standard for panoramic formats, ERP requires mapping a sphere to a 2D plane, introducing rigorous geometric demands. Standard diffusion models struggle with two core topological invariants imposed by ERP: *horizontal periodicity* (perfect continuity between the left and right edges) and *polar convergence* (a coordinate singularity where all columns meet at the poles).

Formally, let $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ denote an ERP panoramic image (or video frame) with height H, width $W = 2H$, and C channels. The ERP maps spherical coordinates (θ, ϕ), latitude $\theta \in [-\pi/2, \pi/2]$ and longitude $\phi \in [-\pi, \pi)$, to pixel coordinates (r, c) via

$$r = \frac{\theta + \pi/2}{\pi} \cdot (H - 1), \quad c = \frac{\phi + \pi}{2\pi} \cdot W \tag{1}$$

This mapping introduces two geometric constraints that any valid ERP must satisfy:

- **C1. Horizontal periodicity**: $\mathbf{x}[r, c] = \mathbf{x}[r, c \bmod W]$ for all rows r, since longitudes $-\pi$ and π correspond to the same meridian.
- **C2. Polar convergence**: $\mathbf{x}[0, c_1] = \mathbf{x}[0, c_2]$ and $\mathbf{x}[H-1, c_1] = \mathbf{x}[H-1, c_2]$ for all columns c_1, c_2, since all longitudes converge to a single point at each pole.

Given a text prompt $\mathbf{p}$ and a pre-trained diffusion model, our goal is to generate $\mathbf{x}$ that is both semantically faithful to $\mathbf{p}$ and satisfies constraints **C1** and **C2**, without modifying any model parameters.

3.2 SPHEROPE: SPHERICAL ROPE

The standard linear encoding $\alpha_i(c) = c \cdot \omega_i$ for $\omega_i = \theta_{\text{base}}^{-2i/d}$ in RoPE fundamentally violates both ERP constraints. It assigns different embeddings to columns 0 and W (breaking horizontal periodicity, **C1**), and distinct embeddings to all columns at the poles (breaking polar convergence, **C2**). We propose to replace the width-axis channels of RoPE with a geometry-aware encoding, preserving the height and temporal axes identically to the original model.

A naïve approach would apply a uniform geometric transformation across all channels. However, the wide spectral range of RoPE dictates that different channels govern distinct spatial behaviors within the self-attention mechanism. This spectral diversity creates a natural functional division: high-frequency channels specialize in fine-grained local texture coherence, while low-frequency channels act as a global compass for the spatial layout. To avoid disrupting these specialized roles

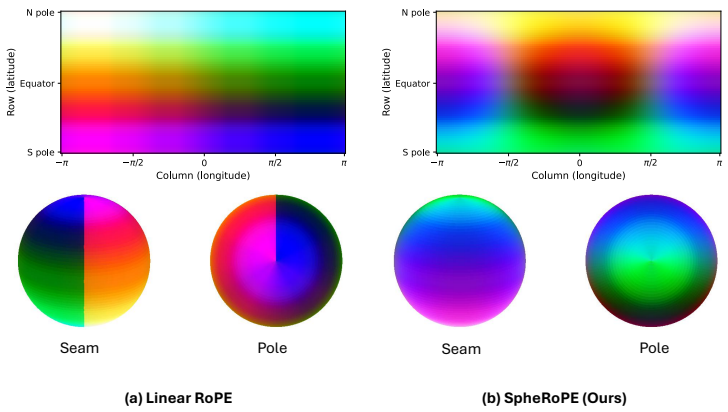

Figure 2: **PCA visualization of RoPE.** (a) Linear RoPE creates a seam at the $\pm\pi$ meridian and disjoint polar embeddings. (b) Our method wraps seamlessly with uniform polar convergence.

with a one-size-fits-all geometric fix, we partition the channels based on their harmonic alignment with the image width W_{tokens}. This allows us to transform the RoPE components to encode spherical geometry while preserving the unique spatial characteristics of each frequency band.

Specifically, for each RoPE channel i, we compute the ratio $k_i = \omega_i/\omega_{\text{fund}}$, where $\omega_{\text{fund}} = 2\pi/W_{\text{tokens}}$ is the fundamental frequency for horizontal wrap-around. A frequency is quantizable if it completes at least one full cycle ($k_i \geq 1$) and is near-integer ($|k_i - \text{round}(k_i)|/k_i \leq \epsilon$). Rather than masking individual channels, we identify the first channel index i_{split} that violates this condition, partitioning the spectrum into two continuous bands. For high-frequency channels ($i < i_{\text{split}}$), we enforce strict cyclic periodicity by using Cyclic Linear Encoding. For the remaining low-frequency channels ($i \geq i_{\text{split}}$), we apply the spherical Cartesian encoding to anchor the global layout. Our solution, illustrated in Figure 2, allows the model to respect spherical topology without sacrificing high-frequency priors.

Spherical Cartesian Encoding (Low-Frequency Channels). Low-frequency channels vary slowly across the image, dictating the global layout of the scene. They cannot be made cyclic through simple quantization: because they accumulate less than one full oscillation across W_{tokens}, snapping them to the nearest harmonic would require a massive spectral shift, shattering the global distance metric of the model. Instead, we abandon the linear parameterization for these channels and replace it with Cartesian coordinates on the unit sphere. For a token at row r and column c, we compute:

$$\begin{aligned} \theta(r) &= \frac{r}{H_{tokens} - 1}\pi - \frac{\pi}{2}, \quad \phi(c) = \frac{2\pi c}{W_{tokens}} - \pi \\ X(r,c) &= (\cos\theta(r)\cos\phi(c) + 1)\, R, \quad Y(r,c) = (\cos\theta(r)\sin\phi(c) + 1)\, R \end{aligned} \tag{2}$$

where R is a scale-dependent radius. To encode the full circular topology and resolve symmetric ambiguity (e.g., distinguishing ϕ from $-\phi$), we replace the single scalar column index c in the encoding function $\alpha_i(c)$ with 2 scalars - Cartesian X and Y coordinates, interleaved into even and odd frequency slots. One coordinate is not enough, since X repeats every value twice, so Y, the position along the perpendicular axis, resolves the ambiguity. We interleave X and Y across adjacent RoPE frequencies so that both are present at every scale.

This encoding mathematically satisfies the topological invariants of the S^2 manifold. For periodicity (**C1**), the trigonometric components trace a closed circle as longitude wraps from $-\pi$ to π, guaranteeing $X(r,0) = X(r,W)$ and $Y(r,0) = Y(r,W)$. For polar convergence (**C2**), as latitude approaches the poles ($\theta \to \pm\pi/2$), X and Y converge to R, independent of the column index.

Cyclic Linear Encoding (High-Frequency Channels). While spherical encoding captures global topology, applying it uniformly to high-frequency channels introduces severe spatial aliasing. Pre-trained models rely on constant phase shifts between adjacent pixels for local texture. Multiplying a non-linear spherical projection by high frequencies catastrophically amplifies phase variance, destroying local coherence and introducing moiré artifacts. To preserve local structure, we maintain a linear Euclidean parameterization for high-frequency channels but enforce strict cyclicity by snapping

each to the nearest integer harmonic:

$$\hat{\omega}_i = \text{round}(k_i) \cdot \omega_{\text{fund}}, \quad \alpha_i(c) = c \cdot \hat{\omega}_i. \tag{3}$$

Because high frequencies destructively interfere over long distances, they do not impact global layout. Harmonically quantizing them ensures phase alignment modulo 2π at the boundaries (**C1**). This flawlessly stitches local textures without disrupting the global spherical structure. While this linear form technically violates polar convergence (**C2**) in the high-frequency subspace, we observe that generation in these areas is dominated by the low frequency correspondences. This formulation allows the model to maintain consistent local texture density across the sphere.

3.3 Semantic Distortion CFG

We observe that pre-trained diffusion models natively exhibit equirectangular projection (ERP) characteristics, such as polar stretching and horizon curvature, when prompted with 360° descriptions. To amplify this inherent prior and complement the hard geometry of Spherical RoPE, we extend standard classifier-free guidance (CFG) to a three-way formulation. We introduce a geometric prompt $\mathbf{p}_{\text{geo}}$ that is anchored to the user prompt $\mathbf{p}$ via concatenation, $\mathbf{p}_{\text{anchor}} = [\mathbf{p}, \mathbf{p}_{\text{geo}}]$. This anchoring isolates the pure effect of ERP geometry without introducing conflicting semantic content. At each denoising step, we compute three noise predictions: $\varepsilon_{\text{cond}}$ (user prompt), $\varepsilon_{\text{uncond}}$ (empty prompt), and ε_{geo} (anchored prompt). The final prediction combines these directions:

$$\hat{\varepsilon} = \varepsilon_{\text{uncond}} + w_{\text{sem}} \cdot (\varepsilon_{\text{cond}} - \varepsilon_{\text{uncond}}) + \gamma \cdot (\varepsilon_{\text{geo}} - \varepsilon_{\text{cond}}) \tag{4}$$

where w_{sem} and γ independently control the semantic and geometric scales. This orthogonal decomposition provides precise control over the trade-off between prompt fidelity and geometric validity, gracefully recovering standard CFG when $\gamma = 0$.

4 Experiments

We evaluate our framework on two 360° generation tasks: text-to-image and text-to-video. We first describe the experimental setup, then present qualitative comparisons highlighting out-of-distribution generalization and image-to-panorama conditioning, followed by quantitative evaluations on both image and video benchmarks. Finally we present an ablation study validating each component of our method. Extended results including a LLM-based perceptual assessment, ablations, limitations, along with our project page which includes interactive panorama viewers are in the supplementary.

4.1 Experimental Setup

Datasets. For image evaluation, following Wang et al. (2025b), we use the ODI-SR dataset (Deng et al., 2021): 1,200 ERP panoramas spanning diverse indoor and outdoor scenes, captioned with Qwen3-VL. Captions are constrained to short descriptions for compatibility with CLIP-based text encoders. None of the evaluated methods were trained on this dataset, ensuring a fair generalization test (Deng et al., 2021). For video evaluation, we use two prompt sets: *SphereDiff-20*, a 20-scene benchmark from Park et al. (2026), and *Stress-20*, a set of 20 prompts generated by an LLM (Claude Opus 4.7) to stress-test panoramic generation with high-motion, dynamic content (Section C.2). Prompts are converted to match each baseline's native conditioning format (Section C.3).

Baselines. For both text-to-image and text-to-video, we benchmark against recent training and optimization-based methods. However, due to its inference latency, we exclude SphereDiff from the ODI-SR benchmark, evaluating it instead via VLM and a user study on a reduced prompt set.

Evaluation metrics. For image generation we compute universal metrics (FID (Heusel et al., 2017), KID (Binkowski et al., 2018), IS (Salimans et al., 2016), CS (Radford et al., 2021)) on perspective crops to evaluate quality and diversity. We assess panoramic fidelity using distortion-aware features (FAED (Oh et al., 2022)) and quantify wrap-boundary artifacts via DS (Christensen et al., 2024). For video generation, we evaluate six VBench (Huang et al., 2024) dimensions: imaging quality (Ke et al., 2021), text-alignment (CLIP Mean (Radford et al., 2021)), temporal stability (temporal flickering, motion smoothness (Li et al., 2023)), and semantic persistence (subject (Caron et al., 2021) and background consistency). Full details are in Section B.2.

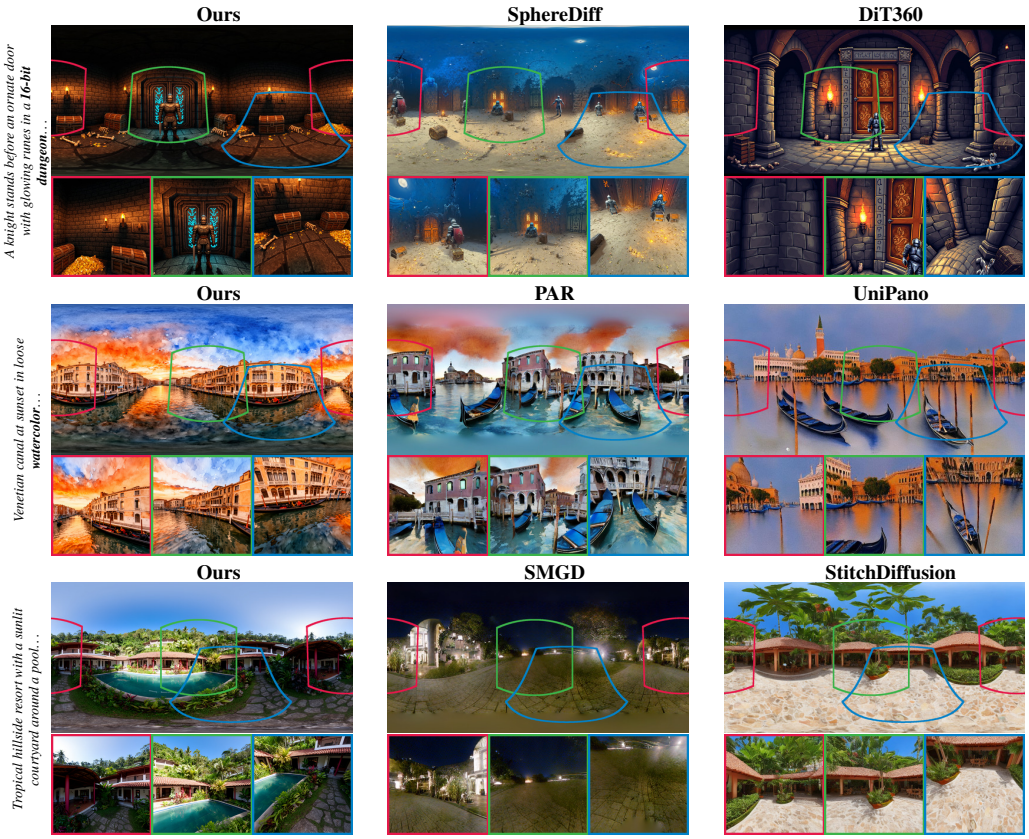

Figure 3: **Text-to-360 Image Qualitative Results.** Each scene shows the ERP panorama (top) and three perspective crops (bottom): seam region (red), horizon (green), and ground (blue). Our method, while being zero-shot, produces more coherent structure and successfully handles OOD prompts.

4.2 Qualitative Results

Text-to-360 Image. Figure 3 qualitatively compares our text-to-image method against baselines. Our approach consistently generates seamless, globally coherent panoramas, excelling on out-of-distribution stylized prompts that exceed typical training data. In contrast, training-based methods exhibit architectural bias, ignoring stylistic instructions. Although the optimization-based SphereDiff avoids these overfitting artifacts, its patch-based synthesis lacks global coherence.

Plug-and-Play Conditioning. A primary advantage of our zero-shot approach is preserving the full versatility of the foundation model. By modifying only the inference process, we enable seamless 360° generation using built-in conditioning pipelines. Unlike training-based methods that require retraining adapters for new modalities, our framework natively inherits all supported functionalities. Figure 4 demonstrates image-conditioned generation, where our approach maintains high-fidelity identity preservation, accurately capturing fine-grained reference details like the corgi's glasses or the astronaut's rubber duck. Beyond images, applying our pipeline to LTX 2.3 unlocks the first training-free 360° text-to-video-audio generation, yielding panoramic video with synchronized audio. In the supplementary, we demonstrate that SpheRoPE also generalizes to Qwen-Image and Wan2.1. Moreover, while our method is training-free, it seamlessly composes with domain-specific LoRA fine-tuning on panoramic datasets, yielding quantitative gains.

4.3 Quantitative Results

Text-to-360 image generation. Table 1 shows our zero-shot approach matches or outperforms trained baselines on ODI-SR. At the panorama level, we achieve the best FAED score without task-specific training, demonstrating that our outputs most accurately capture the global distribution and

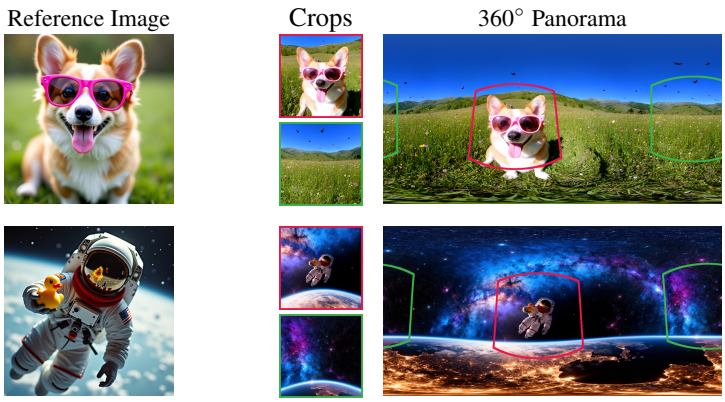

Figure 4: **Image-conditioned 360 generation.** Our zero-shot approach natively inherits built-in model functionalities, such as image-conditioned generation. Using reference images, our framework generates seamless 360° panoramas that preserve identity and style across the entire sphere.

Table 1: **Quantitative comparison**. We compare our zero-shot results with training-based methods on the ODISR dataset. We show both panorama-level metrics and multi-view metrics (computed from 8 equi-spaced horizontal perspective crops). Best in **bold**, second best underlined.

Method	Type	Panorama-Level Metrics		Multi-View Metrics			
		FAED ↓	DS ↓	FID ↓	KID$_{\times 10^2}$ ↓	IS ↑	CS ↑
Text2Light (Chen et al., 2022)	Full Fine-Tune	77.94	3.82	81.01	4.55	6.25	**18.60**
Diffusion360 (Feng et al., 2023)	Full Fine-Tune	71.02	0.87	53.17	2.82	8.92	18.46
StitchDiffusion (Wang et al., 2024a)	LoRA	63.10	1.01	43.73	1.84	10.88	18.50
PanFusion (Zhang et al., 2024)	LoRA	39.29	1.08	40.37	2.12	9.66	17.14
UniPano (Ni et al., 2025)	LoRA	46.19	1.11	34.76	1.78	9.73	16.81
StochSync (Yeo et al., 2025)	Optimization	326.04	**0.67**	67.13	4.72	8.97	18.07
SMGD (Sun et al., 2025)	Full Fine-Tune	33.55	0.79	49.02	2.01	5.77	18.52
LayerPano3D (Yang et al., 2025)	LoRA	55.71	0.73	31.58	1.22	10.48	18.44
DiT360 (Feng et al., 2025)	LoRA	43.42	2.59	**23.28**	0.92	**12.90**	18.52
PAR (Wang et al., 2025a)	Full Fine-Tune	34.79	0.75	25.80	**0.85**	10.47	18.51
Ours (FLUX.1)	**Zero-Shot**	29.94	0.89	45.22	2.63	10.32	18.52
Ours (FLUX.2)	**Zero-Shot**	**25.40**	0.94	25.33	0.90	12.75	18.43

structural integrity of real 360° scenes. This global coherence is further reinforced by a competitive Discontinuity Score (DS), as our method effectively avoid boundary seams. On crop-level metrics, our individual perspective views retain high local realism: we match PAR on KID and IS, and perform competitively alongside DiT360. Ultimately, SpheRoPE delivers strong performance entirely zero-shot, bypassing the massive computational bottlenecks of dataset collection and model fine-tuning.

Text-to-360 video generation. Table 2 reports our results. On SphereDiff-20, our zero-shot approach leads all metrics (temporal stability, imaging quality, and CLIP Mean). On the high-motion Stress-20 set, our method continues to dominate all temporal metrics and imaging quality. While DynamicScaler

Table 2: **Text-to-360 Video Comparison.** We compare our zero-shot results with training and optimization-based methods. Values ×100 except CLIP Mean. Best in **bold**, second-best underlined.

Prompt Set	Method	Type	Sec./ frame ↓	Img. Qual. ↑	CLIP Mean ↑	Temp. Flicker ↑	Motion Smooth. ↑	Subj. Cons. ↑	BG Cons. ↑
SphereDiff-20	360DVD (Wang et al., 2024b)	LoRA	**0.95**	47.99	26.02	99.18	98.98	95.11	96.13
	PanoWan (Xia et al., 2025)	LoRA	1.59	39.10	27.32	99.13	99.43	97.40	96.59
	DynamicScaler (Liu et al., 2025)	Optimization	51.56	53.21	27.38	97.13	98.11	93.79	94.80
	SphereDiff (Park et al., 2026)	Optimization	6.08	40.08	27.03	99.35	99.51	95.90	96.33
	Ours (LTX 2.3)	**Zero-Shot**	1.11	**53.64**	**27.95**	**99.58**	**99.62**	**98.22**	**97.71**
Stress-20	360DVD (Wang et al., 2024b)	LoRA	**0.95**	50.19	28.62	98.66	98.42	93.30	94.82
	PanoWan (Xia et al., 2025)	LoRA	1.59	39.49	28.42	97.11	98.58	91.64	93.66
	DynamicScaler (Liu et al., 2025)	Optimization	51.56	50.65	**32.39**	94.45	96.17	91.58	93.25
	SphereDiff (Park et al., 2026)	Optimization	6.08	33.82	28.20	96.52	97.72	88.55	93.34
	Ours (LTX 2.3)	**Zero-Shot**	1.11	**51.76**	28.91	**98.91**	**99.46**	**96.51**	**96.02**

Table 3: **Human preference study.** We collect 320 blind pairwise judgments from 18 annotators using interactive image panorama viewers. For each pair, raters make a three-way choice (A, B, or tie) on overall preference and text alignment. Our method (using Flux.2) is consistently preferred across all baselines and both criteria.

	Overall Preference			Text Alignment		
Baseline	**Ours**	Tie	Base	**Ours**	Tie	Base
PanFusion (Zhang et al., 2024)	**88.5%**	3.3%	8.2%	**83.6%**	6.6%	9.8%
UniPano (Ni et al., 2025)	**82.7%**	3.8%	13.5%	**65.4%**	17.3%	17.3%
SMGD (Sun et al., 2025)	**95.2%**	2.4%	2.4%	**95.2%**	2.4%	2.4%
SphereDiff (Park et al., 2026)	**93.2%**	3.4%	3.4%	**86.4%**	10.2%	3.4%
DiT360 (Feng et al., 2025)	**56.5%**	10.9%	32.6%	**50.0%**	41.3%	8.7%
PAR (Wang et al., 2025a)	**71.7%**	15.0%	13.3%	**73.3%**	21.7%	5.0%

Table 4: **Components Contribution.** We ablate key components of our method: spherical RoPE and Semantic Distortion CFG. Best in **bold**, second best underlined.

	Panorama-Level Metrics		**Multi-View Metrics**			
Configuration	FAED ↓	DS ↓	FID ↓	$KID_{\times 10^2}$ ↓	IS ↑	CS ↑
Vanilla FLUX.2	33.38	1.37	26.05	1.06	**14.24**	18.42
(+) SpheRoPE	40.90	**0.92**	28.62	1.53	13.77	18.25
(+) SD-CFG	27.41	1.22	28.86	1.16	12.98	18.42
(+) SpheRoPE (+) SD-CFG (Ours)	**25.40**	0.94	**25.33**	**0.90**	12.75	**18.43**

achieves a higher CLIP Mean via multi-pass stitching, it suffers the worst temporal coherence and an order of magnitude slower inference time. Meanwhile, SphereDiff's patch-based synthesis struggles with dynamic scenes, yielding the lowest imaging quality and subject consistency.

User Preference Study. Table 3 summarizes our blind pairwise user study (full protocol in Section C.4) among the six most competitive text-to-image panorama generation. Users consistently prefer our panoramas over all baselines for both overall quality and text alignment. We observe the widest preference margins against other methods. Even against DiT360, the most competitive baseline, our method secures a clear preference advantage in overall quality.

4.4 Ablation Study

We evaluate the quantitative impact of each core component below. Supplementary Section B.5 provides extensive additional qualitative and quantitative ablations.

Component Contributions. Table 4 ablates our framework components on FLUX.2. To ensure a fair comparison, the vanilla baseline is explicitly guided with panoramic text prompts. While scoring reasonably well on unconstrained multi-view metrics, it lacks global structure, evidenced by boundary seams (high Discontinuity Score). When applied independently, Spherical RoPE successfully enforces topological invariants, resolving hard discontinuities and improving DS. However, this geometric constraint alone degrades global image fidelity (FAED increases) as the model struggles with the enforced spherical layout. Conversely, Semantic Distortion CFG (SD-CFG) applied independently compensates for ERP distortion and improves global fidelity (FAED), but cannot close boundary seams. Combining SpheRoPE's spatial topology with SD-CFG's distortion-aware guidance cleanly resolves this tension, achieving the best overall FAED and multi-view metrics while maintaining strict panoramic coherence. Qualitative examples are provided in the supplementary.

5 Conclusion

We present a zero-shot framework for synthesizing seamless 360° images and videos using pre-trained diffusion models. This entirely training-free method achieves competitive performance against fully fine-tuned baselines. Decoupling geometric grounding from training allows our framework to scale directly with future foundation models, paving the way for world models and VR applications.

AI USE STATEMENT

In this work, we used generative AI tools for conducting experiments: GPT-4o as an LLM-based evaluator for perceptual quality assessment following SphereDiff, Qwen3-VL to generate text captions for the ODI-SR evaluation dataset, Claude Opus 4.7 to generate the 20 prompts of the Stress-20 video benchmark (Appendix C.2), and GPT-4o to convert prompts between the single-paragraph and per-elevation formats required by different baselines (Appendix C.3).

We have not used generative AI tools for research ideation, and peer review and data provenance tracking do not apply to this work. Additionally, we used generative AI tools for writing assistance and text editing. We reviewed all AI-assisted work: we fixed the Stress-20 prompts before evaluating any method and used them verbatim; we manually inspected and verified all AI-generated captions, converted prompts, and evaluations on representative samples; and we thoroughly edited and vetted all drafted text for accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of AI.

REPRODUCIBILITY STATEMENT

SpheRoPE is training-free: it modifies only the width-axis rotary position embedding and the guidance rule at inference time, and is fully specified by Section 3. Appendix B.1 lists all hyperparameters (resolution, 50 denoising steps, harmonic quantization tolerance $\varepsilon = 0.06$, radius R, $\gamma = 6.0$, and w_{sem}), the circular VAE padding, the hardware, and the backbones used (FLUX.1 [dev], FLUX.2 [dev], and LTX 2.3), all of which are publicly available. Appendix C gives the exact geometric prompts $\mathbf{p}_{\text{geo}}$, and Tables 8 to 12 reproduce verbatim every LLM prompt used in our evaluation: the LLM-judge prompt, the full Stress-20 prompt set, and the prompt-conversion templates for baselines. Evaluation uses the public ODI-SR dataset and VBench; metric definitions and the resizing protocol are given in Appendix B.2, and the user-study protocol in Appendix C.4. All baselines were run from their official code and released weights, with any deviations documented in Appendix B.1. Our implementation is publicly available and will be linked in the de-anonymized version of the paper.

REFERENCES

Anthropic. Introducing claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-06.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 1737–1752. PMLR, 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Eleonora Brivio, Silvia Serino, Erica Negro Cousa, Andrea Zini, Giuseppe Riva, and Gianluca De Leo. Virtual reality and 360° panorama technology: a media comparison to study changes in sense of presence, anxiety, and positive emotions. *Virtual Real.*, 25(2):303–311, 2021. doi: 10.1007/S10055-020-00453-7. URL https://doi.org/10.1007/s10055-020-00453-7.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://doi.org/10.1109/ICCV48922.2021.00951.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595, 2023. doi: 10.48550/ARXIV.2306.15595. URL https://doi.org/10.48550/arXiv.2306.15595.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

Anders Christensen, Nooshin Mojab, Khushman Patel, Karan Ahuja, Zeynep Akata, Ole Winther, Mar González-Franco, and Andrea Colaco. Geometry fidelity for spherical images. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, Lecture Notes in Computer Science, pages 276–292. Springer, 2024. doi: 10.1007/978-3-031-72989-8_16. URL https://doi.org/10.1007/978-3-031-72989-8_16.

Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9189–9198. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00907. URL https://openaccess.thecvf.com/content/CVPR2021/html/Deng_LAU-Net_Latitude_Adaptive_Upscaling_Network_for_Omnidirectional_Image_Super-Resolution_CVPR_2021_paper.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. *CoRR*, abs/2510.11712, 2025. doi: 10.48550/ARXIV.2510.11712. URL https://doi.org/10.48550/arXiv.2510.11712.

Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *CoRR*, abs/2311.13141, 2023. doi: 10.48550/ARXIV.2311.13141. URL https://doi.org/10.48550/arXiv.2311.13141.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. Image as a world: Generating interactive world from single image via panoramic video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PA47sKU8CU.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: efficient joint audio-visual foundation model. *CoRR*, abs/2601.03233, 2026. doi: 10.48550/ARXIV.2601.03233. URL https://doi.org/10.48550/arXiv.2601.03233.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Noam Issachar, Guy Yariv, Sagie Benaim, Yossi Adi, Dani Lischinski, and Raanan Fattal. Dype: Dynamic position extrapolation for ultra high resolution diffusion. *CoRR*, abs/2510.20766, 2025. doi: 10.48550/ARXIV.2510.20766. URL https://doi.org/10.48550/arXiv.2510.20766.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M2SsqpxGtc.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5128–5137. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00510. URL https://doi.org/10.1109/ICCV48922.2021.00510.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

Aoming Liu, Zhong Li, Zhang Chen, Nannan Li, Yi Xu, and Bryan A. Plummer. Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVII*, Lecture Notes in Computer Science, pages 146–164. Springer, 2024. doi: 10.1007/978-3-031-73383-3_9. URL https://doi.org/10.1007/978-3-031-73383-3_9.

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, and Ming-Hsuan Yang. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 6144–6153. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00576. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_DynamicScaler_Seamless_and_Scalable_Video_Generation_for_Panoramic_Scenes_CVPR_2025_paper.html.

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 33160–33176. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/lu24k.html.

Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360° panoramic videos from perspective videos. *CoRR*, abs/2504.07940, 2025. doi: 10.48550/ARXIV.2504.07940. URL https://doi.org/10.48550/arXiv.2504.07940.

Jinhong Ni, Chang-Bin Zhang, Qiang Zhang, and Jing Zhang. What makes for text to 360-degree panorama generation with stable diffusion? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16555–16564, 2025.

Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. BIPS: bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI*, Lecture Notes in Computer Science, pages 352–371. Springer, 2022. doi: 10.1007/978-3-031-19787-1_20. URL https://doi.org/10.1007/978-3-031-19787-1_20.

Minho Park, Taewoong Kang, Jooyeol Yun, Sungwon Hwang, and Jaegul Choo. Spherediff: Tuning-free 360° static and dynamic panorama generation via spherical latent representation. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 8305–8313. AAAI Press, 2026. doi: 10.1609/AAAI.V40I10.37779. URL https://doi.org/10.1609/aaai.v40i10.37779.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Bowen Peng. NTK-aware scaled RoPE. 2023. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/ARXIV.2204.06125. URL https://doi.org/10.48550/arXiv.2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02155. URL https://doi.org/10.1109/CVPR52729.2023.02155.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 5824–5834. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00547. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_Spherical_Manifold_Guided_Diffusion_Model_for_Panoramic_Image_Generation_CVPR_2025_paper.html.

Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BcYfsMMpV1.

Eren Unlu. Spherical position encoding for transformers. *CoRR*, abs/2310.04454, 2023. doi: 10.48550/ARXIV.2310.04454. URL https://doi.org/10.48550/arXiv.2310.04454.

Chase van de Geijn, Timo Lüddecke, Polina Turishcheva, and Alexander S. Ecker. A circular argument : Does rope need to be equivariant for vision? *CoRR*, abs/2511.08368, 2025. doi: 10.48550/ARXIV.2511.08368. URL https://doi.org/10.48550/arXiv.2511.08368.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Chaoyang Wang, Xiangtai Li, Lu Qi, Xiaofan Lin, Jinbin Bai, Qianyu Zhou, and Yunhai Tong. Conditional panoramic image generation via masked autoregressive modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=2YxtR50mho.

Hai Wang and Jing-Hao Xue. 360pant: Training-free text-driven 360-degree panorama-to-panorama translation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pages 212–221. IEEE, 2025. doi: 10.1109/WACV61041.2025.00031. URL https://doi.org/10.1109/WACV61041.2025.00031.

Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 4921–4931. IEEE, 2024a. doi: 10.1109/WACV57701.2024.00486. URL https://doi.org/10.1109/WACV57701.2024.00486.

Hai Wang, Xiaoyu Xiang, Weihao Xia, and Jing-Hao Xue. A survey on text-driven 360-degree panorama generation. *CoRR*, abs/2502.14799, 2025b. doi: 10.48550/ARXIV.2502.14799. URL https://doi.org/10.48550/arXiv.2502.14799.

Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6913–6923. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.00660. URL https://doi.org/10.1109/CVPR52733.2024.00660.

Zidong Wang, Zeyu Lu, Di Huang, Cai Zhou, Wanli Ouyang, and Lei Bai. Fitv2: Scalable and improved flexible vision transformer for diffusion model. *CoRR*, abs/2410.13925, 2024c. doi: 10.48550/ARXIV.2410.13925. URL https://doi.org/10.48550/arXiv.2410.13925.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J Fleet, Marcus A Brubaker, and Saurabh Saxena. 360anything: Geometry-free lifting of images and videos to 360 {\deg}. *arXiv preprint arXiv:2601.16192*, 2026.

Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. Panowan: Lifting diffusion video generation models to 360 with latitude/longitude-aware mechanisms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In Ginger Alford, Hao (Richard) Zhang, and Adriana Schulz, editors, *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers 2025, Vancouver, BC, Canada, August 10-14, 2025*, pages 109:1–109:10. ACM, 2025. doi: 10.1145/3721238.3730643. URL https://doi.org/10.1145/3721238.3730643.

Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIOvR40hSo.

Kyeongmin Yeo, Jaihoon Kim, and Minhyuk Sung. Stochsync: Stochastic diffusion synchronization for image generation in arbitrary spaces. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=XPNprvlxuQ.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360° panorama image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6347–6357. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00607. URL https://doi.org/10.1109/CVPR52733.2024.00607.

Muyang Zhang, Yuzhi Chen, Rongtao Xu, Changwei Wang, Jinming Yang, Weiliang Meng, Jianwei Guo, Huihuang Zhao, and Xiaopeng Zhang. Panodit: Panoramic videos generation with diffusion transformer. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 10040–10048. AAAI Press, 2025. doi: 10.1609/AAAI.V39I10.33089. URL https://doi.org/10.1609/aaai.v39i10.33089.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10188–10198. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00982. URL https://doi.org/10.1109/CVPR52729.2023.00982.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhao25m.html.

Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 21610–21619. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02013. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zheng_Panorama_Generation_From_NFoV_Image_Done_Right_CVPR_2025_paper.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Lirui Zhao, Si Liu, Xiangyu Yue, Wanli Ouyang, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-next : Making lumina-t2x stronger and faster with next-dit. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ed2dad593d87ca474a636cba610a29d3-Abstract-Conference.html.

A PRELIMINARIES

Diffusion Models and CFG. Diffusion models Rombach et al. (2022) learn to reverse a gradual noising process: given a clean sample $\mathbf{x}_0$, Gaussian noise is progressively added over T timesteps to produce $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and a neural network ε_θ is trained to predict the noise at each step. Generation proceeds by iteratively denoising from pure noise back to a clean sample. To condition generation on a text prompt $\mathbf{p}$, classifier-free guidance (CFG) Ho and Salimans (2021) interpolates between an unconditional prediction $\varepsilon_{\text{uncond}}$ and a conditional prediction $\varepsilon_{\text{cond}}$

$$\hat{\varepsilon} = \varepsilon_{\text{uncond}} + w_{sem} \cdot (\varepsilon_{\text{cond}} - \varepsilon_{\text{uncond}}), \tag{5}$$

where $w_{sem} > 1$ amplifies the influence of the text condition. This formulation operates along a single guidance direction, from unconditional to semantically conditioned, and encodes no geometric priors about the output domain. The same principle applies to flow-matching models that predict velocity rather than noise, where guidance is performed in the velocity or $\mathbf{x}_0$-prediction space.

Rotary Position Embedding. Modern diffusion transformers Peebles and Xie (2023) encode spatial structure through rotary position embedding (RoPE) Su et al. (2024). The transformer block is permutation-equivariant by design. A positional encoding mechanism is therefore necessary to model the strong spatial dependencies present in natural images. While early approaches used fixed sinusoidal Vaswani et al. (2017) or learned absolute embeddings Devlin et al. (2019), RoPE has emerged as a more effective alternative that encodes *relative* positions directly into the query-key interaction in the attention mechanism. Concretely, RoPE represents a position coordinate as a set of rotations at different frequencies. For a token at spatial position (r, c) in a 2D image, RoPE is applied axially: one subset of the hidden dimension is rotated according to the row coordinate r, and the other according to the column coordinate c, enabling the model to encode relative offsets along each spatial axis independently. Focusing on the width (column) axis, the rotation angle for frequency channel i is

$$\alpha_i(c) = c \cdot \omega_i, \quad \text{for} \quad \omega_i = \theta_{\text{base}}^{-2i/D_w} \tag{6}$$

where D_w is the number of width frequency channels, and θ_{base} determines the geometric series of frequencies. Each frequency channel i produces a 2D rotation matrix $\mathbf{R}(\alpha_i)$ applied to the corresponding pair of dimensions in the query and key vectors

$$\mathbf{R}(\alpha_i) = \begin{pmatrix} \cos\alpha_i & -\sin\alpha_i \\ \sin\alpha_i & \cos\alpha_i \end{pmatrix}. \tag{7}$$

The full RoPE transformation applies these block-diagonal rotations to the query $\mathbf{q}$ and key $\mathbf{k}$ before computing the attention inner product.

B ADDITIONAL EXPERIMENTS

B.1 IMPLEMENTATION DETAILS

All experiments are conducted on NVIDIA H100 GPUs. For image generation, we produce ERP panoramas at a resolution of 1024×2048 using 50 denoising steps. We use a harmonic quantization tolerance of $\varepsilon = 0.06$. For the spherical Cartesian encoding, the radius R is set as $R = W_{\text{tokens}}/2$, and $R = W_{\text{span}}/2$ for LTX 2.3, where W_{span} is the coordinate-normalized width extent. The RoPE dimensionality follows native structure of each backbone. In all cases, only the width axis is modified; the temporal and height axes remain identical to the stock model. The Semantic Distortion CFG scale is set to $\gamma = 6.0$ across all experiments. The semantic guidance scale w_{sem} follows the default CFG value of each backbone.

All assets are used under their original licenses: FLUX.1/2 [dev] under Black Forest Labs' Non-Commercial License, LTX-Video 2.3 under Lightricks' Community License, ODI-SR and VBench under Apache 2.0. Use is strictly for non-commercial academic research.

All baselines are evaluated using their official codebases and released pretrained weights. For StitchDiffusion, we use the diffusers-based reimplementation by Wang et al. (2024a) with the official LoRA weights, as the original codebase does not provide a runnable inference script. Since the original Stable Diffusion 2.1-base weights were removed from HuggingFace by Stability AI, we use a community mirror (`Manojb/stable-diffusion-2-1-base`) for methods that depend

on it. All other methods use their default inference hyperparameters as specified in their respective repositories. Generated panoramas are resized to 512×1024 before evaluation when the native resolution differs.

Circular Latent Encoding To address boundary discontinuities introduced by standard zero-padding in convolution-based VAEs, we follow 360Anything Wu et al. (2026). We apply circular padding symmetrically to the VAE decoder, horizontally extending the tensors to simulate periodic boundaries prior to the decoding pass. It is important to distinguish the roles of our two components: SpheRoPE acts inside the transformer, ensuring the generated latent content is consistent across the wrap boundary. However, the VAE decoder is a convolutional network, and without circular padding, it cannot see across the boundary, resulting in a thin decoder-induced seam. By circularly padding the latent prior to the decoding pass, the decoder sees the true circular neighborhood. Because the transformer operates on the unpadded grid, no positional encoding ever sees the padding, cleanly isolating geometric synthesis from pixel-level rendering.

B.2 Evaluation Metrics

We evaluate $360°$ panorama generation using both universal image quality metrics and panorama-specific metrics, as standard measures often fail to account for the global layout and geometric properties unique to equirectangular projection (ERP) Wang et al. (2025b).

Universal metrics. We report Fréchet Inception Distance (FID) Heusel et al. (2017) and Kernel Inception Distance (KID) Binkowski et al. (2018), which measure the distributional distance between generated and real images using Inception-v3 features. We also report Inception Score (IS) Salimans et al. (2016), which evaluates both quality and diversity via the KL divergence between conditional and marginal class distributions, and CLIP Score (CS) Radford et al. (2021), the cosine similarity between CLIP text and image embeddings.

Because these metrics rely on encoders trained on perspective images, they may penalize geometrically correct panoramic features as artifacts Wang et al. (2025b). To mitigate this, we compute all universal metrics on perspective crops extracted via gnomonic projection from the ERP panoramas, ensuring the feature extractors operate on undistorted patches consistent with their training distribution.

We further note that distributional metrics (FID, KID, FAED) inherently favor training-based methods, whose fine-tuning data overlaps with the evaluation distribution. Our zero-shot approach generates from the broader, unconstrained distribution of the foundation model, placing it at a structural disadvantage on these metrics despite producing visually competitive results.

Panorama-specific metrics. FAED Oh et al. (2022) computes Fréchet distances using an autoencoder trained specifically on $360°$ panoramic images, providing a distortion-aware alternative to standard FID that better captures the perceptual and geometric quality of panoramic content. Discontinuity Score (DS) Christensen et al. (2024) quantifies seam artifacts at the horizontal wrap boundary ($\pm\pi$ meridian) using kernel-based edge detection. Lower DS indicates better perceived continuity across the seam - a direct measure of whether horizontal periodicity (**C1**) is satisfied in the generated output.

B.3 Additional Quantitative Results

LLM-based Evaluation. Following the established protocol of SphereDiff Park et al. (2026), we evaluate each panorama with GPT-4o Hurst et al. (2024) as an LLM judge Zheng et al. (2023), scoring 14 perspective views on a 1-5 scale across four criteria: *distortion* (whether the image resembles a natural photograph), *end continuity* (seamlessness at the wrap boundary), *image quality*, and *aesthetic appearance*. The exact prompt used, taken directly from SphereDiff Park et al. (2026), is reproduced in Section C.1. Table 5 reports the results: we achieve the best end continuity, are competitive on distortion, and substantially outperform SphereDiff on image quality and aesthetics, while running over $20\times$ faster.

Table 5: **LLM-based evaluation on** 360° **panorama generation.** Following SphereDiff Park et al. (2026), we use GPT-4o to score 14 perspective views per scene on a 1–5 scale across panoramic and image criteria, using their 20-prompt benchmark. We additionally report inference time per scene (seconds) on an NVIDIA H100. Best in **bold**, second best underlined.

Method	Type	Panoramic Criteria		Image Criteria		Time/scene (s)
		Distortion ↑	End Continuity ↑	Image Quality ↑	Aesthetic Appearance ↑	
360 LoRA	LoRA	2.03	3.42	2.97	3.49	–
Text2Light	Fine-Tune	2.38	3.45	2.42	2.78	36
PanFusion	LoRA	1.97	3.70	2.82	3.45	**28**
DynamicScaler	Optimization	2.85	3.99	**4.50**	**4.58**	–
SphereDiff (Flux1)	Optimization	**3.15**	4.60	3.76	4.24	1274
SpheRoPE (Flux.1)	**Zero-Shot**	3.05	4.56	4.37	4.49	62
SpheRoPE (Flux.2)	**Zero-Shot**	3.13	**4.69**	4.18	4.41	189

†Scores for 360 LoRA, Text2Light, PanFusion, and DynamicScaler are taken from Park et al. (2026). We additionally time all methods on our hardware for fair comparison, except 360 LoRA and DynamicScaler, which are reimplemented by Park et al. (2026) and do not have publicly available inference code. All other scores and timings are ours (NVIDIA H100).

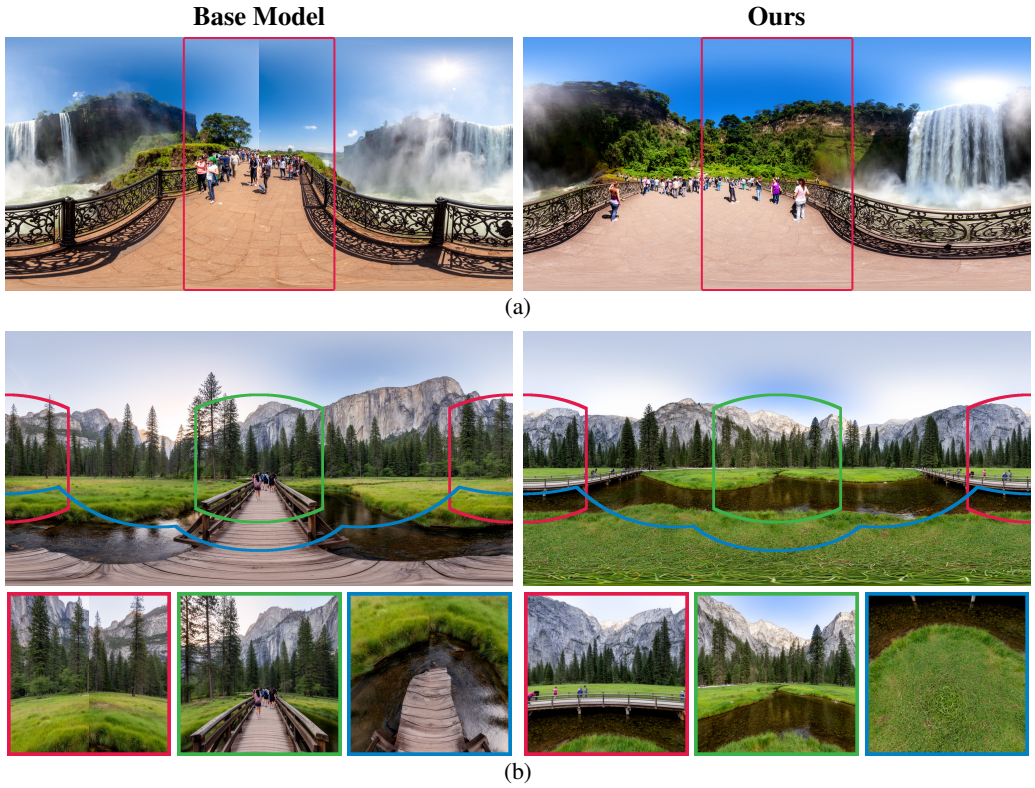

Figure 5: **Analysis of Implicit 360 Capabilities.** While current diffusion models capture some 360 panoramic image characteristics, they fail to satisfy spherical topology. (a) Shifting the panorama horizontally reveals a noticeable vertical seam in the base model, whereas our method maintains seamless periodicity. (b) Perspective crops show that the base model sometimes fails to properly model ERP distortions, highlighting the need for our CFG enhancement.

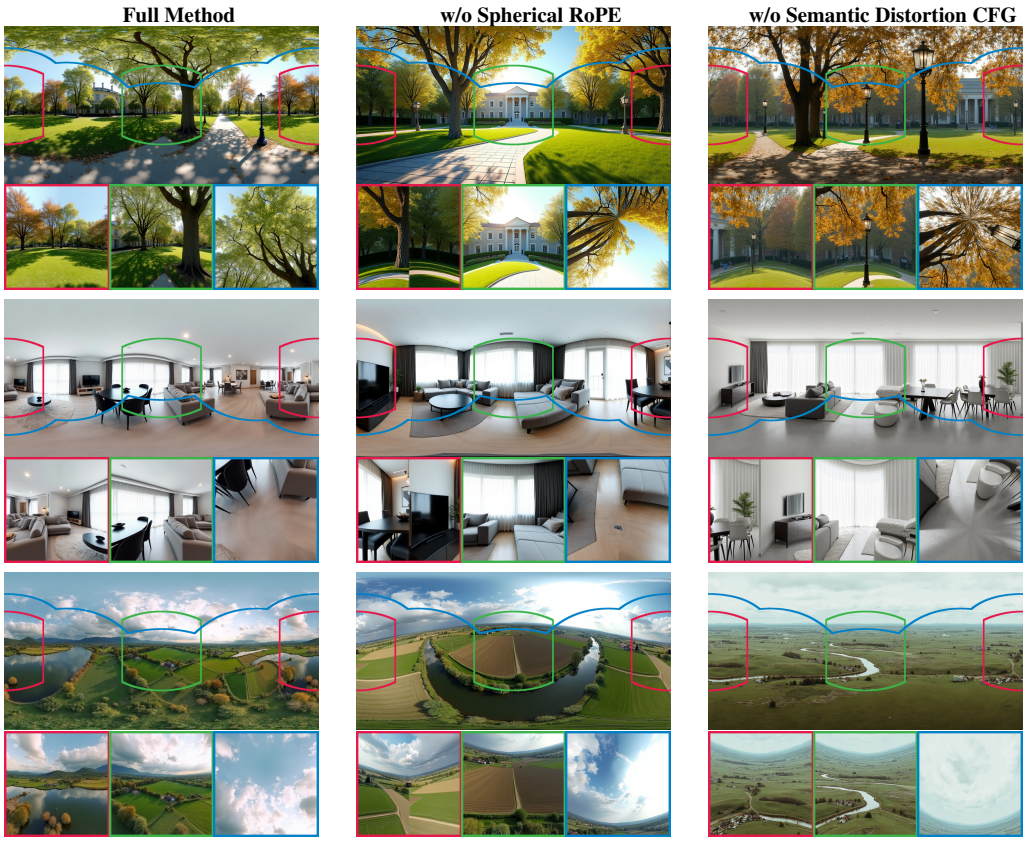

Figure 6: **Component ablation.** Generated using Flux.1. Each column shows the full ERP panorama (top) and three perspective crops (bottom): seam region (red), horizon (green), and zenith/nadir (blue). Without spherical RoPE, the output is a flat perspective image with a hard seam discontinuity and no polar convergence. Without Semantic Distortion CFG, RoPE ensures seamless boundaries but the content remains perspective-like, missing the pole stretching and horizon curvature expected in ERP.

B.4 Additional Qualitative Results

Implicit 360 Capabilities. Figure 5 illustrates the core intuition behind our approach. As shown in the base model output (Flux.2), pre-trained diffusion transformers inherently capture the equirectangular projection (ERP) distribution, naturally synthesizing the stretched poles and wide aspect ratios characteristic of 360° scenes. However, they lack the topological awareness to mathematically close the sphere, leaving disjointed boundaries and polar discontinuities. Our framework provides the critical missing link: Semantic Distortion CFG amplifies the model's latent panoramic priors, while Spherical RoPE explicitly enforces horizontal periodicity. Together, these mechanisms gently guide the base model's native ERP distribution into a seamless panorama.

B.5 More Ablation Studies

Components Ablation. We extend the component ablation from the main paper (Table 4) with qualitative results on FLUX.1 (Figure 6). As shown, without spherical RoPE (middle column), the output lacks spherical characteristics - the seam crop (red) reveals a hard discontinuity where the left and right boundaries depict entirely different content, and the zenith crop (blue) shows artifacts rather than the characteristic polar convergence of a valid ERP. Without Semantic Distortion CFG (right column), the seam is more continuous thanks to RoPE's periodicity encoding, and the zenith exhibits better convergence, but the overall scene lacks the geometric distortion patterns expected in equirectangular imagery, resulting in degraded distributional metrics.

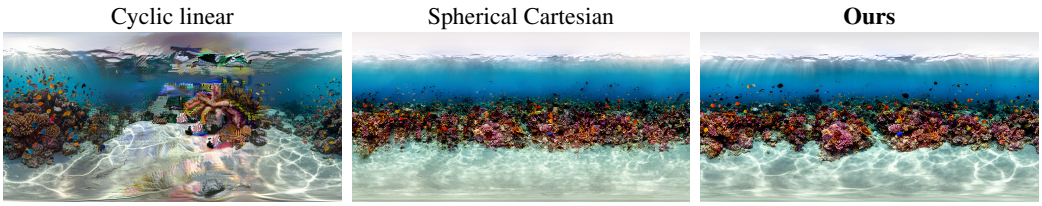

Figure 7: **Spherical RoPE ablation.** (a) Using only Cyclic Linear enforces horizontal periodicity but yields out-of-distribution low-frequency values, resulting in artifacts. (b) Using only Spherical Cartesian satisfies global spherical topology and polar convergence but disrupts local distance metrics, causing blurry or aliased textures. (c) By partitioning the spectrum and using both strategies, we preserve high-frequency texture coherence while anchoring the global spherical manifold.

RoPE Ablation. Figure 7 provides a comprehensive visual ablation that clearly demonstrates the necessity of our dual-path RoPE encoding strategy. When the model relies exclusively on cyclic linear encoding, it successfully achieves seamless horizontal wrapping. However, this comes at the cost of severe artifacts. This failure occurs because the large change in low frequencies RoPE is OOD for the pre-trained model, causing global structural priors to break down. On the other hand, applying only spherical Cartesian encoding successfully enforces the correct global topology and polar convergence. Yet, this approach severely disrupts the relative distance priors that are essential for fine-grained, high-frequency synthesis, ultimately degrading local texture quality and yielding blurry or heavily aliased details. Our proposed spectral partitioning mechanism elegantly resolves this dichotomy by assigning distinct functional roles to each encoding type. By routing the representations through a dual path, the linear encoding is leveraged exclusively to preserve crisp, sharp local textures, while the spherical encoding serves to firmly anchor the coherent global layout of the panorama.

Compatibility with LoRA Fine-Tuning. When a panoramic dataset is available, SpheRoPE composes naturally with fine-tuning. We LoRA-fine-tuned FLUX.2 with the SpheRoPE formulation on Matterport data and evaluated it on ODI-SR. This improved structural metrics significantly: FID dropped from 25.33 to 16.63, KID ($\times 10^2$) from 0.90 to 0.39, and DS from 0.94 to 0.84, demonstrating that our formulation benefits directly from ERP-specific supervision when available.

Generalization to Diverse Base Architectures. To demonstrate backbone-agnostic applicability, we evaluated SpheRoPE on Qwen-Image (text-to-image) and Wan2.1 (text-to-video). SpheRoPE with Qwen-Image achieves superior DS (0.71) compared to our default FLUX.2 backbone. Similarly, applying SpheRoPE to Wan2.1 yields highly competitive imaging quality on both SphereDiff-20 and Stress-20 benchmarks.

Compatibility with High-Resolution RoPE Extrapolation. SpheRoPE is mathematically compatible with NTK-style scaling because the mechanisms act on different factors of the rotary phase (extrapolation rescales frequencies; SpheRoPE remaps positions). Generating at 1024×2048 with NTK-scaled SpheRoPE on FLUX.2 maintained robust seam consistency and fidelity (DS 0.84 vs 0.90, $\text{KID}_{\times 10^2}$ 0.65 vs 0.82 compared to unscaled), proving our Cartesian mapping scales reliably.

Harmonic Quantization Tolerance. SpheRoPE partitions the RoPE channels between the Cyclic Linear and Spherical Cartesian encodings based on whether frequencies ω_i can be harmonically quantized within a tolerance ε. Frequencies are assigned to the Cyclic Linear path as long as they complete at least one full cycle and are near-integer ($|k_i - \text{round}(k_i)|/k_i \leq \varepsilon$). The first channel index i_{split} that violates this condition determines the boundary, with all subsequent lower frequencies ($i \geq i_{\text{split}}$) routed to the Spherical Cartesian encoding. The tolerance ε therefore controls this split: a smaller value triggers the violation earlier, while a larger value extends the cyclic linear path. Crucially, this mechanism has a natural saturation point. Increasing the tolerance beyond a certain threshold (e.g., $\varepsilon \geq 0.10$) yields an identical channel split, as all remaining low frequencies inherently fail the minimum cycle requirement ($k_i < 1.0$) and are strictly routed to the spherical path. Figure 8 shows a sensitivity analysis across $\varepsilon \in \{0.01, 0.03, 0.06, 0.10\}$ on both FLUX.2 and FLUX.1. A very strict tolerance (ε=0.01) triggers an early split, forcing too many channels into the spherical encoding and producing blurrier textures due to the coarser positional resolution of the spherical path.

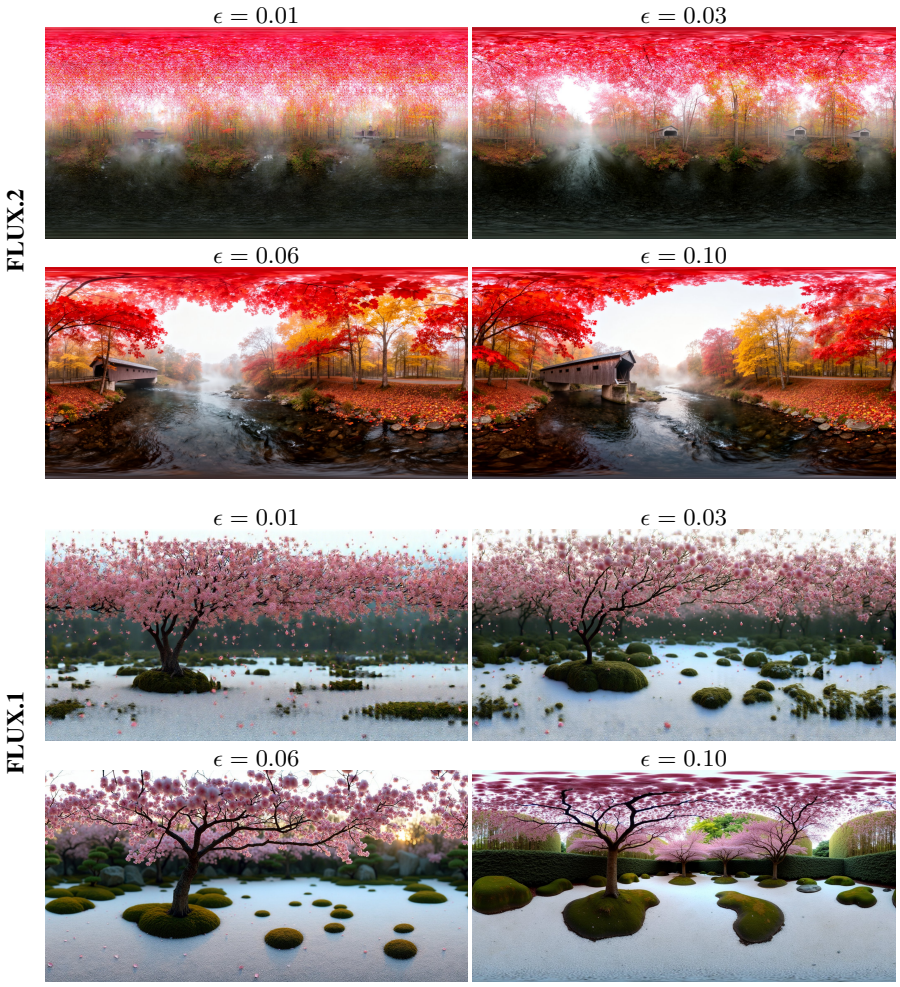

Figure 8: **Harmonic quantization tolerance ϵ.** We vary the tolerance used to partition RoPE frequencies between the cyclic linear path (harmonically quantizable within ϵ) and the spherical Cartesian. Results shown for both FLUX.2 (top) and FLUX.1 (bottom) backbones.

Table 6: **Semantic Distortion Strength.** We sweep the distortion gamma γ controlling how aggressively ERP-aware CFG re-weights patches by their spherical distortion. Best in **bold**, second best underlined.

	Panorama-Level Metrics		Multi-View Metrics			
Configuration	FAED ↓	DS ↓	FID ↓	$\text{KID}_{\times 10^2}$ ↓	IS ↑	CS ↑
$\gamma = 3$	34.48	0.90	29.05	1.15	**13.43**	**18.46**
$\gamma = 4$	34.08	0.88	<u>28.58</u>	1.23	<u>13.07</u>	**18.46**
$\gamma = 8$	<u>32.20</u>	**0.85**	**28.06**	**0.98**	11.95	<u>18.46</u>
$\gamma = 10$	**31.75**	<u>0.85</u>	28.70	<u>1.06</u>	11.59	18.44

A loose tolerance (ε=0.10) delays the split, however the minimum cycle requirement makes sure we avoid OOD artifacts.

Semantic Distortion CFG. We sweep $\gamma \in \{3, 4, 8, 10\}$ to examine how aggressively our ERP-aware CFG branch should re-weight latent patches according to their spherical distortion. Rather than producing a single clear winner, Table 6 reveals a mild but consistent trade-off between panorama-level and multi-view metrics: weaker distortion ($\gamma = 3, 4$) maximises multi-view IS and CS, while

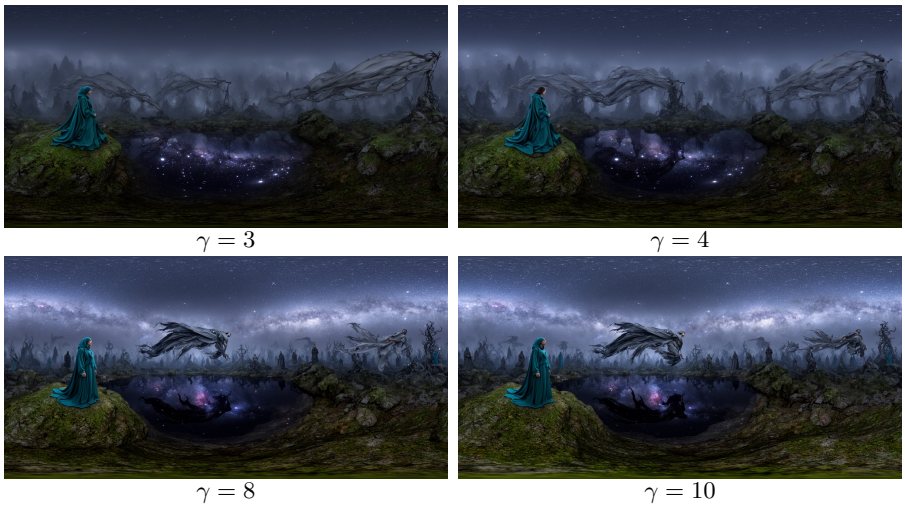

Figure 9: **Visual effect of the distortion strength γ.** Panoramas generated with the same prompt and seed under four different values of the ERP-aware CFG distortion parameter. Visual quality and layout remain stable across the full range: differences are subtle and primarily manifest as slightly sharper, more varied texture at lower γ versus slightly smoother, more geometrically-regularised content at higher γ.

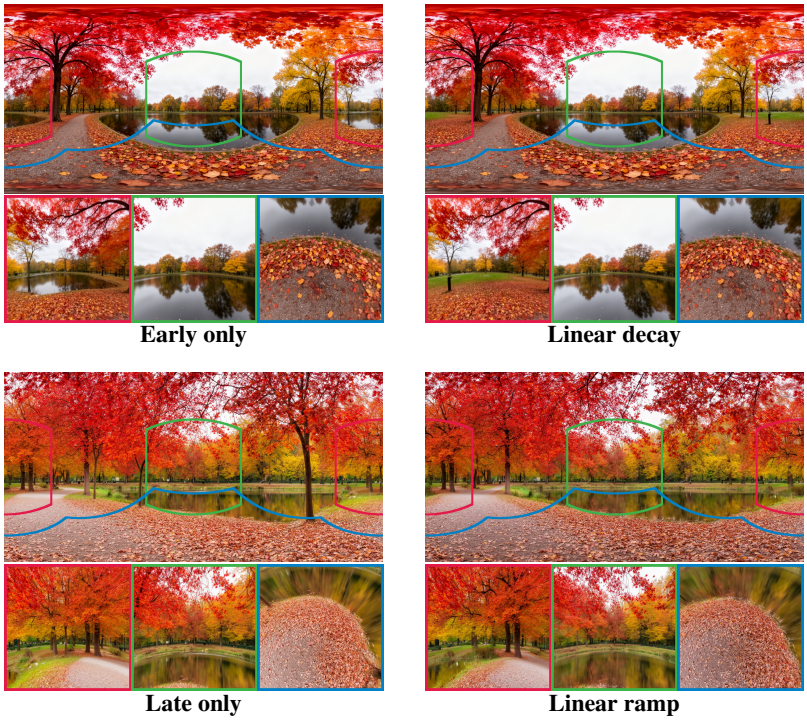

Figure 10: **Semantic Distortion CFG schedule.** We apply the geometric guidance term with different schedules over the N denoising steps. The top row (*early only*, *linear decay*) concentrates guidance in the early steps and produces valid ERP geometry with proper polar convergence. The bottom row (*late only*, *linear ramp*) concentrates guidance in the late steps and yields flat perspective-like panoramas with pole artifacts and incorrect distortion patterns.

stronger distortion ($\gamma = 8, 10$) minimises FAED, DS, and perspective FID, and KID. Critically, *the absolute spread of each metric is narrow* - FAED varies by only 2.7 points, FID by 1.0, and CS is essentially flat (within 0.02). This quantitative stability is mirrored qualitatively in Figure 9:

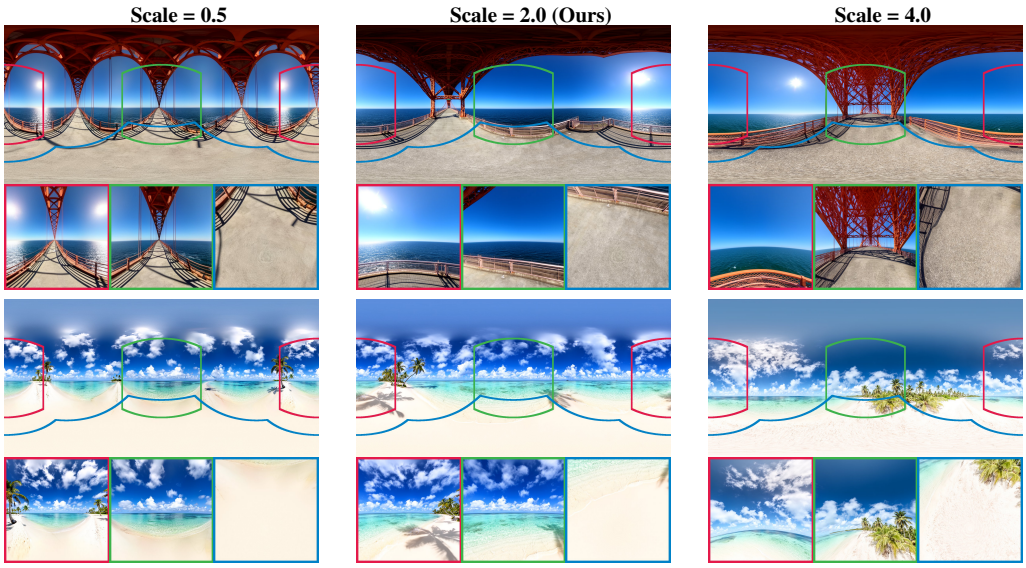

Figure 11: **RoPE scale ablation.** We vary the radius scale s in $R_{\text{width}} = W/s$, which controls the range of positional values fed to RoPE. **Our default** (s=2.0) matches the expected input distribution, preserving both global geometry and local detail.

panoramas generated under all four values of γ preserve the same overall layout and perceptual quality, with differences amounting to slight local texture variations rather than structural changes. Together, these results show that our method is *largely agnostic* to the precise choice of γ within this range: any value between 3 and 10 yields comparable quality. We adopt $\gamma = 6$ as a balanced default that sits inside this stable regime.

We further investigate the optimal scheduling for Semantic Distortion CFG across the denoising process. Figure 10 compares four strategies: *early only*, *late only*, *linear decay*, and *linear ramp*. Results show that panoramic geometry is established during the initial denoising steps. Consequently, early-weighted schedules (*early only*, *linear decay*) successfully enforce valid ERP geometry and yield nearly identical, seamless panoramas. Conversely, late-weighted schedules (*late only*, *linear ramp*) fail to alter the already-crystallized global structure, resulting in the same pole artifacts and flat layouts seen in the unguided baseline. This confirms that Semantic Distortion CFG primarily shapes the early geometric trajectory rather than refining late-stage details. For simplicity, we choose to apply the Semantic Distortion CFG constantly across all denoising steps.

SpheRoPE Radius Scale Ablation. The spherical encoding path in SpheRoPE (Eq. 2) uses a radius $R_{\text{width}} = W/s$, where s is a scale hyperparameter that controls the range of positional values fed to the pre-trained RoPE. Since the model was trained on a specific range of positional values (roughly $[0, W]$), this scale directly affects whether the RoPE frequencies operate in their learned regime. Figure 11 shows the failure modes at both extremes. Our default (s=2.0) matches the expected input distribution of the pre-trained model, producing panoramas with both coherent global structure and fine-grained texture.

Geometric Prompt Ablation. Semantic Distortion CFG is steered by a geometric prompt $\mathbf{p}_{\text{geo}}$ that describes the desired ERP properties. Figure 12 shows the effect of varying this prompt. A minimal prompt ("equirectangular panorama") under-specifies the desired geometry, and the model defaults to flat panoramas with visible pole artifacts. A verbose prompt, enumerating every panoramic property, over-constrains the generation and produces extreme top-down curvature. Our default prompt strikes a balance, providing enough specificity to enforce valid ERP geometry without distorting the natural perspective of the scene.

Circular Latent Encoding. Figure 13 shows the effect of disabling circular VAE padding. Without it, the VAE's zero-padded convolutions introduce a latent-space discontinuity at the horizontal

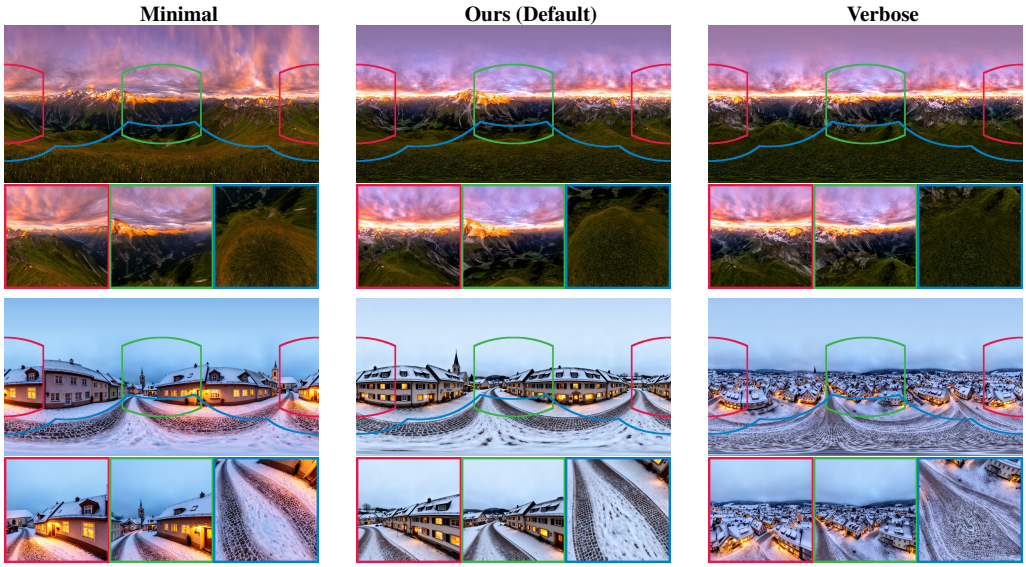

Figure 12: **Geometric prompt ablation.** We vary the geometric prompt $\mathbf{p}_{\text{geo}}$ used in Semantic Distortion CFG. The **minimal** prompt ("equirectangular panorama") under-specifies the desired geometry, producing flat panoramas with pole artifacts. The **verbose** prompt over-specifies, pushing the model toward extreme top-down curvature. **Our default** balances these extremes, yielding valid ERP geometry with natural perspective.



(a) Without circular padding (b) With circular padding

Figure 13: **Circular latent encoding ablation.** Panoramas are shifted by 180° to place the wrap boundary at the center of the frame.(a) Without circular VAE padding: visible seam appears at the boundary. (b) With circular padding: the seam is eliminated.

Table 7: **Semantic CFG Scale.** We vary the guidance scale w_{sem}. Best in **bold**, second best underlined.

Configuration	Panorama-Level Metrics		Multi-View Metrics			
	FAED ↓	DS ↓	FID ↓	$\text{KID}_{\times 10^2}$ ↓	IS ↑	CS ↑
$w_{sem} = 2$	40.60	**0.80**	28.96	1.15	12.60	**18.45**
$w_{sem} = 3$	<u>36.34</u>	<u>0.82</u>	<u>28.53</u>	<u>1.06</u>	12.77	18.41
w_{sem}=4 (Ours)	**25.40**	0.94	**25.33**	**0.90**	12.75	<u>18.43</u>
$w_{sem} = 5$	36.64	0.83	29.67	1.22	**12.88**	18.37

boundary that manifests as a visible vertical seam in the decoded output. Enabling circular padding eliminates this artifact entirely, producing seamless wrap-around at no additional inference cost.

Semantic CFG Scale. We analyze the sensitivity of the semantic guidance scale w_{sem} while maintaining all geometric components (Table 7). Lower values ($w_{sem} = 2$) yield slightly better structural continuity (DS) and marginal gains in text alignment (CS), but significantly underperform in distributional image quality (higher FAED, and FID). Our default ($w_{sem} = 4$) achieves the optimal balance, yielding the highest overall image quality and panoramic fidelity across both panorama-level and multi-view metrics. Higher values ($w_{sem} = 5$) lead to regression in both visual quality and prompt adherence, suggesting that overly aggressive semantic guidance disrupts the balance between content fidelity and geometric validity.

C GEOMETRIC PROMPTS

As described in Sec. 3.3 of the main paper, our Semantic Distortion CFG uses an anchored geometric prompt $\mathbf{p}_{\text{anchor}} = [\mathbf{p};\ \mathbf{p}_{\text{geo}}]$ to steer the denoising process toward valid ERP geometry. Below we list the fixed geometric prompts $\mathbf{p}_{\text{geo}}$ used across all experiments.

Image generation (Flux 1, Flux 2).

> *"Single unified continuous environment, monolithic scene composition, solitary spatial layout, flawlessly stitched 360 panorama, true equirectangular projection, accurate spherical geometry, continuous horizontal wrap, zero parallax error."*

Video generation (LTX 2.3).

> *"True 2:1 equirectangular projection, proper zenith/nadir pole geometry, seamless 360° horizontal wrap. Rigidly locked static tripod camera at a fixed nodal point. All geometry is permanently static: buildings, terrain, walls, floors, ceilings, roads, vegetation trunks, rocks, and all man-made structures maintain absolute pixel-locked positions across every frame. Zero structural deformation, zero background warping, zero surface drift. Only permitted motion: faint atmospheric haze, subtle light caustic shifts, microscopic dust particles. Flawless inter-frame coherence."*

The video prompt is longer than the image prompt because it must additionally enforce temporal stability: the camera must remain static and all scene geometry must be pixel-locked across frames, with only minimal atmospheric motion permitted. This prevents the video model from introducing camera movement or structural warping that would break the ERP constraints.

C.1 LLM-AS-A-JUDGE EVALUATION PROMPT

For the LLM-based evaluation reported in Sec. 4 of the main paper, we use the exact prompt introduced by SphereDiff Park et al. (2026) (Tab. 5 of their paper), which we reproduce verbatim in Tab. 8 for completeness. The prompt elicits a brief chain-of-thought Wei et al. (2022) justification before each score.

Table 8: **Evaluation Prompt for VLM.** The evaluation prompt used to assess panorama quality based on four image and panoramic criteria, taken verbatim from SphereDiff Park et al. (2026). 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).

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.

C.2 STRESS-20 BENCHMARK CONSTRUCTION

Construction protocol. To complement the SphereDiff-20 benchmark Park et al. (2026) with prompts that specifically target temporal coherence under challenging dynamics, we constructed the

Stress-20 prompt set using the following protocol. We provided Claude Opus 4.7 Anthropic (2026) with the 20 SphereDiff-20 scene descriptions as reference examples of typical panoramic video prompts, and instructed it to generate 20 *new*, diverse single-line prompts that stress-test 360° video generation with high-motion content, rapid camera-implied dynamics, multiple interacting subjects, and complex environmental effects (e.g., explosions, weather, crowds). The exact generation prompt was:

> *"Given these 20 panoramic video scene descriptions as reference for style and format, generate 20 new and diverse single-paragraph prompts (each 30–60 words) designed to stress-test 360-degree panoramic video generation. Focus on: rapid full-sphere motion, multiple interacting subjects with independent trajectories, dynamic lighting changes, particle effects (fire, water, debris), and scenes where temporal coherence is difficult to maintain. Each prompt should describe a different environment and action. Do not repeat or paraphrase the reference prompts."*

The 20 returned prompts (shown in Table 9) were used verbatim without any filtering, reordering, or modification. Crucially, the prompts were generated and finalized *before* any method was evaluated on them, no prompts were added, removed, or edited after observing any method's outputs.

Cross-set ranking agreement. To verify that Stress-20 does not selectively favor our method, we compare method rankings across the two prompt sets. The ranking pattern remains stable: our method ranks first on all 6 VBench metrics on SphereDiff-20 and on 5/6 metrics on Stress-20. The only exception is CLIP Mean on Stress-20, where DynamicScaler surpasses our method and we rank second, indicating that the self-designed benchmark does not uniformly inflate our scores.

C.3 Prompt Format Conversion for Video Evaluation

As described in Sec. 4.3, our video evaluation spans two prompt sets - SphereDiff-20 and Stress-20 - and several methods that expect incompatible conditioning formats. Specifically, while our approach, along with training-based approaches expect a single semantic prompt, optimization-based methods (such as SphereDiff Park et al. (2026) and DynamicScaler Liu et al. (2025)) require a per-elevation 5-line format with separate per-view conditioning. To bring them onto a common ground without handicapping any method at generation time, we use two LLM-driven prompt conversions: a $\mathbf{5} \rightarrow \mathbf{1}$ consolidation (for single-prompt methods) and a $\mathbf{1} \rightarrow \mathbf{5}$ expansion (to lift Stress-20 single-paragraph prompts into SphereDiff's canonical per-elevation format). Importantly, our method and all training-based baselines receive only the consolidated single prompt, which necessarily discards some per-elevation detail, while the richer 5-line format is provided only to the optimization-based baselines that require it.

$5 \rightarrow 1$ consolidation. For the $5 \rightarrow 1$ direction, we use GPT-4o Hurst et al. (2024) to consolidate the five elevation lines into a single prompt. The exact consolidation prompt is given in Tab. 10. For 360DVD Wang et al. (2024b), whose CLIP text encoder truncates inputs at 77 tokens, we further shorten the consolidated prompt to approximately 25 words using GPT-4o. The exact prompt is given in Tab. 11.

$1 \rightarrow 5$ expansion. We use GPT-4o Hurst et al. (2024) to expand Stress-20 single-paragraph prompts into the per-elevation 5-line format required by SphereDiff Park et al. (2026) and DynamicScaler Liu et al. (2025). The exact expansion prompt is given in Tab. 12.

C.4 User Study Protocol

We conduct a blind pairwise preference study to evaluate perceptual quality across the full sphere. Each trial presents two anonymized panoramas in interactive 360° viewers with drag-to-pan navigation, allowing raters to freely explore the entire sphere rather than judge a single static viewpoint. The two viewers are synchronized so that both display the same region simultaneously, ensuring direct comparison. Navigation presets allow raters to jump to the seam boundary, poles, and front view, guiding inspection of known failure modes. For each pair, raters make a three-way choice (A, B, or tie) on two questions: (1) "Which panorama do you prefer?" (overall quality) and (2) "Which one has better text alignment?". The left-right presentation order is position-balanced to mitigate side bias.

Table 9: **Stress-20 prompt set.** Twenty prompts designed to stress-test 360° video generation with challenging motion, parallax, occlusions, and lighting transitions.

#	Prompt
1	Rock concert in a packed 360° arena, camera moving forward through the crowd toward the stage, people jumping and pushing past the camera, flashing strobe lights, smoke and confetti, performers moving rapidly across the full field of view with slight handheld roll and tilt.
2	Formula 1 race track from trackside perspective, camera moving laterally along the barrier while cars rush past at high speed in both directions, strong motion blur, rapid depth changes, vehicles crossing the full 360° view.
3	Crowded subway platform, camera walking forward along the platform edge, trains arriving and departing, doors opening, dense crowd crossing left and right across the entire panorama, flickering overhead lights.
4	Busy outdoor marketplace with narrow aisles, camera weaving forward between stalls, people passing very close to the lens, fabrics and flags moving in the wind, sunlight shifting through overhead covers.
5	Dense forest trail, camera moving forward at walking speed, trees passing close on both sides, strong parallax between near trunks and distant background, sunlight flickering through leaves.
6	Underwater coral reef scene, camera drifting forward and slightly tilting, schools of fish swirling in all directions, particles floating, dynamic light caustics moving across the environment.
7	Stormy ocean deck of a ship, camera moving with the motion of the waves, strong vertical and lateral sway, water splashing across the scene, rain and lightning causing rapid lighting changes.
8	Urban city street at night after rain, camera moving forward along the sidewalk, pedestrians passing in both directions, neon reflections shifting on wet ground, cars driving past across the full view.
9	Airport terminal interior, camera moving forward through a crowd, travelers with luggage crossing in front of and behind the camera, escalators moving, dynamic occlusions and reappearances.
10	Large train station hall with repeating arches, camera dollying forward while slightly shifting sideways, people crossing symmetry lines, strong structural geometry with moving crowds.
11	Office interior with glass walls and corridors, camera walking through hallways, reflections on glass surfaces, people entering and exiting rooms, multiple occlusions and reappearances.
12	Library with tall dense bookshelves, camera sliding sideways while moving forward, strong parallax from near shelves and distant aisles, people occasionally crossing between rows.
13	Driving through a forest road from inside a car, camera moving forward with slight natural shake, trees passing quickly on both sides, sunlight flickering through branches.
14	Narrow cave interior transitioning to bright outdoor exit, camera moving forward, exposure adapting from dark to bright, irregular geometry and depth changes, turning slightly to look around.
15	Construction site with heavy machinery, camera moving through the site, cranes rotating, workers walking, dust and debris in the air, multiple moving objects at different depths.
16	Snowy mountain slope with skiers, camera moving downhill, people passing at different speeds and distances, snow particles flying, strong depth variation and motion.
17	Crowded street protest, camera walking through dense moving crowd, flags waving, smoke flares, people crossing the full panorama, chaotic motion and occlusions.
18	Theme park roller coaster POV, camera moving rapidly forward with turns and drops, fast changing depth, structures passing close to the camera, dynamic lighting transitions.
19	Shopping mall interior with multiple floors, camera moving along a corridor while looking around slightly, escalators and people moving at different levels, reflections and glass surfaces.
20	Beach shoreline with waves and people, camera walking along the waterline, waves moving in and out, people running and crossing near and far, strong motion in both foreground and background.

We use 20 text prompts and 6 baselines (DIT360, PanFusion, PAR, SMGD, SphereDiff, and UniPano). Each rater evaluates a subset of pairs with no repeated anchor images to avoid familiarity bias. We collect 320 pairwise judgments from 18 annotators with no filtering applied.

Table 10: **$5 \rightarrow 1$ prompt consolidation.** Used to convert a 5-line per-elevation prompt into a single-paragraph prompt for methods that accept only one prompt (e.g., LTX 2.3).

$5 \rightarrow 1$ Consolidation Prompt
You are a prompt rewriter for a text-to-video model. You will receive a 360° panoramic scene described across five elevation-specific lines: 1. **Top (skyward view)**: what is visible looking straight up. 2. **Above the horizon**: what is visible in the upper hemisphere. 3. **Eye level (horizon)**: the main subject and scene at eye level. 4. **Below eye level**: what is visible in the lower hemisphere. 5. **Bottom (ground-facing view)**: what is visible looking straight down. Your task is to consolidate these five lines into a single coherent paragraph (approximately 40-60 words) that preserves the main subject, the dominant camera motion, and the overall scene atmosphere. Prioritize the eye-level description as the anchor of the consolidated prompt. Drop per-region detail when it does not fit naturally into a single-paragraph description. Do not invent content that is not implied by the five input lines. Do not include section labels, bullet points, or the words "top", "above", "below", or "bottom" in the output. {top} {above} {eye} {below} {bottom} Return only the consolidated paragraph.

Table 11: **Prompt summarization for 360DVD.** Used to condense the consolidated prompt to $\leq$25 words for compatibility with CLIP's 77-token limit.

Summarization Prompt
You are a prompt rewriter for a text-to-video model with a strict token limit. You will receive a scene description for a 360° panoramic video. Your task is to rewrite it as a single sentence of at most 25 words that preserves the main subject, setting, and atmosphere. Do not add any detail not present in the input. Do not use bullet points or labels. {prompt} Return only the shortened sentence.

Table 12: **$1 \rightarrow 5$ prompt expansion.** Used to convert Stress-20 single-paragraph prompts into the per-elevation 5-line format used by SphereDiff Park et al. (2026) and DynamicScaler Liu et al. (2025).

$1 \rightarrow 5$ Expansion Prompt
You are a prompt rewriter for 360° panoramic video generation. You will receive a single-paragraph scene description. Your task is to expand it into five coherent, elevation-specific lines that together describe what a viewer would see across the full sphere at a fixed vantage point. The five lines must be plausibly consistent with each other and with the input paragraph, and must not introduce content that contradicts the input. Produce exactly the following five lines, each 1-2 sentences: 1. **Top (skyward view)**: what is visible looking straight up at the zenith. 2. **Above the horizon**: what fills the upper hemisphere between the horizon and the zenith. 3. **Eye level (horizon)**: the main subject and scene at eye level; preserve the dominant motion. 4. **Below eye level**: what fills the lower hemisphere between the horizon and the nadir. 5. **Bottom (ground-facing view)**: what is visible looking straight down at the nadir. Keep all five lines mutually consistent in setting, lighting, time of day, and style. Do not use the labels in the output; return only the five descriptions, one per line. {paragraph}

D LIMITATIONS AND FUTURE WORK

Failure Cases. We illustrate representative failure modes in Figure 14. For prompts with strong perspective priors (e.g., front-facing compositions), the model may default to generating a conventional perspective image rather than a true equirectangular panorama. Additionally, the model may occasionally repeat structural elements to fill the full 360° field of view.

Limitations. Our method relies on two architectural assumptions. First, the backbone must use Rotary Position Embeddings (RoPE) for spatial encoding, as Spherical RoPE directly modifies the width-axis rotation angles. Architectures using other positional encoding schemes (e.g., learned absolute embeddings or additive sincos) would require a different adaptation strategy. Second, the pre-trained model must have been exposed to panoramic or ERP-like content during training. Our approach amplifies and steers latent panoramic priors that already exist in the model's distribution. But, if the backbone has no such priors, Spherical RoPE and Semantic Distortion CFG alone cannot induce panoramic generation from scratch. Representative residual failure modes are shown in Figure 14. Additionally, our method inherits the resolution and quality ceiling of the backbone model. Furthermore, our extension to video generation is currently restricted by motion constraints. To prevent the model from breaking the enforced ERP topology across frames, we must rely on rigid prompting to ensure a strictly static camera and limit dynamics to minimal atmospheric motion. Consequently, generating 360° videos with complex subject movement or camera trajectories remains an open challenge. Finally, because Semantic Distortion CFG computes three noise predictions per denoising step, it entails a $1.5\times$ increase in Network Function Evaluations (NFE) compared to standard CFG. This introduces a computational trade-off, exchanging inference-time efficiency for strict geometric adherence.

Broader impact. By eliminating the need for model fine-tuning, our approach significantly reduces the computational burden and energy consumption typically associated with adapting large-scale diffusion models. This training-free paradigm not only lowers the carbon footprint of 360° content creation but also democratizes access, enabling independent researchers and smaller studios to generate high-quality immersive environments without requiring massive GPU clusters. However, as with all generative models, there is a potential for misuse in creating fabricated immersive content. We inherit the safety guardrails of the underlying backbone models and do not introduce new capabilities for generating harmful or deceptive content beyond what the base models already permit.

Future work. Our current audio-video generation produces mono audio natively output by LTX 2.3, with no spatial awareness. A natural extension is omnidirectional audio-video generation, where the audio is spatialized to match the 360° visual content, enabling fully immersive audiovisual experiences. More broadly, the principle of reshaping RoPE to encode non-Euclidean geometry is not limited to the sphere - it could be extended to other domains such as cylindrical projections, hyperbolic spaces, or arbitrary manifolds, opening new directions for geometry-aware generation beyond panoramas.



(a) "... claymation-style garden party... soft diffused studio lighting and a shallow depth of field."



(b) "A bedroom in a hotel room."

Figure 14: **Failure cases.** (a) Prompts with strong perspective priors (e.g., studio lighting, shallow depth of field) can produce perspective-like outputs rather than true panoramas. (b) The model may repeat structural elements to fill the full sphere.