Imagine360: Immersive 360 Video Generation from Perspective Anchor

Jing Tan[1*] **Shuai Yang**[2,5*] **Tong Wu**[3†] **Jingwen He**[1] **Yuwei Guo**[1]
Ziwei Liu[4] **Dahua Lin**[1,5†]

[1]The Chinese University of Hong Kong [2]Shanghai Jiao Tong University
[3]Stanford University [4]S-Lab, Nanyang Technological University
[5]Shanghai Artificial Intelligence Laboratory

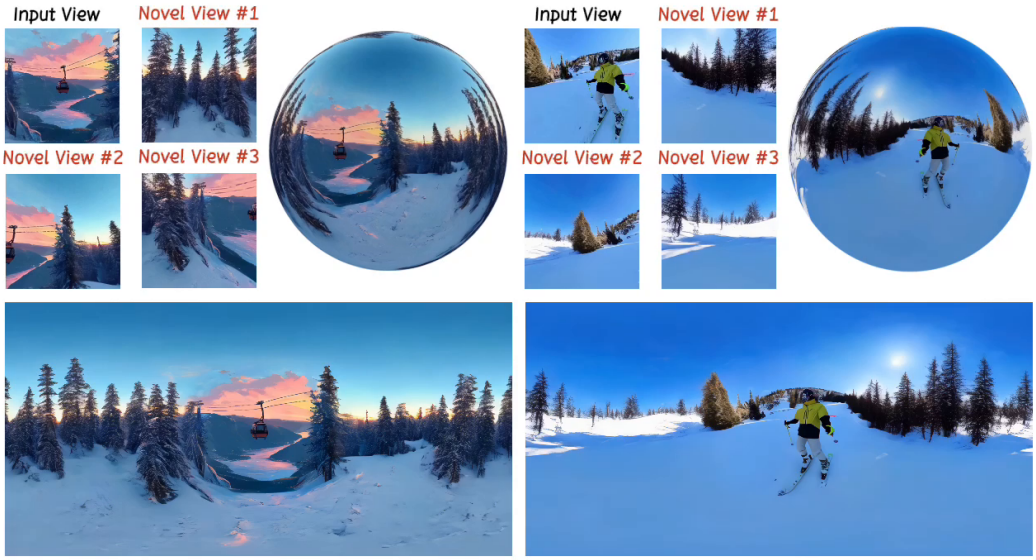

Figure 1: **Overview of Imagine360.** Imagine360 lifts standard perspective video into 360° video, enabling dynamic scene experience from full 360 degrees. Our approach achieves high-quality texture and plausible **spherical** motion patterns for both generated video (left) and in-the-wild video (right). ***Best viewed with Acrobat Reader for the animated 360 videos.*** More examples please visit our webpage.

Abstract

360° videos offer a hyper-immersive experience that allows the viewers to explore a dynamic scene from full 360 degrees. To achieve more accessible and personalized content creation in 360° video format, we seek to lift standard perspective videos into 360° equirectangular videos. To this end, we introduce **Imagine360**, the first perspective-to-360° video generation framework that creates high-quality 360° videos with rich and diverse motion patterns from video anchors. Imagine360 learns fine-grained spherical visual and motion patterns from limited 360° video data with several key designs. **1)** Firstly we adopt the dual-branch design, including a perspective and a panorama video denoising branch to provide local and global constraints for 360° video generation, with motion module and spatial LoRA layers fine-tuned on 360° videos. **2)** Additionally, an antipodal mask is devised to capture long-range motion dependencies, enhancing the reversed camera motion between antipodal pixels across hemispheres. **3)** To handle diverse perspective video inputs, we propose rotation-aware designs that adapt to varying video masking due to

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

changing camera poses across frames. **4)** Lastly, we introduce a new 360 video dataset featuring 10K high-quality, trimmed 360 video clips with structured motion to facilitate training. Extensive experiments show Imagine360 achieves superior graphics quality and motion coherence with our curated dataset among state-of-the-art 360° video generation methods with both real and generated videos. We believe Imagine360 holds promise for advancing personalized, immersive 360° video creation.

1 Introduction

Imagine embarking on a journey through the heart of a bustling city, a serene beach, or a cherished place of your own. It would be wonderful to record this 360-degree dynamic experience for future viewing. 360° video offers an interactive, immersive format that captures a living, breathing world as if the viewer were part of the experience. With the rapid development of VR devices, the demand for 360° videos is increasing, leading to growing focus on 360° video generation research.

Recent advancements in 360° video generation focused on text-guided [34] and image-guided [18] models. While these methods produce plausible 360° videos, they require panoramic optical flow [34] or high-quality panoramic images [18] as guidance, which are hard to obtain for users. In contrast, traditional perspective videos are more accessible, as they can be easily captured using smartphone cameras or generated by advanced video synthesis models. To enable more user-friendly and personalized 360° video creation, we propose a new task: perspective-to-360° video generation, which transforms standard video inputs into 360° equirectangular videos. Specifically, we take a perspective video with narrow FOV as the anchor video, project it to a $360^\circ \times 180^\circ$ FOV video canvas, and synthesize the surrounding pixels.

One relevant task is video outpainting [2; 30; 8], which aims to fill in missing regions outside the edges of video frames in a larger canvas, typically in the perspective domain with fixed video masking across frames. Simply applying standard video outpainting methods does not achieve satisfactory results, as our perspective-to-360° video generation exhibits more challenges. First, due to the large domain gap between perspective and 360 videos, learning the spherical visual and motion patterns requires sophisticated design when trained on limited 360 video data. Second, as videos exhibit different camera poses across frames, after mapping to the 360 canvas, the video mask would change significantly in shape, size, and location, requiring rotation-aware designs for robust generation.

To address these challenges, we introduce Imagine360, the first framework to generate high-quality 360° videos from standard perspective video inputs. Inspired by the dual-domain concept [42] in text-to-image generation, our model employs **dual-branch video outpainting** Unets in the perspective and panoramic domains. These two branches jointly denoise the partially masked 360 spacetime canvas in the global scope and in each perspective window. A cross-domain attention module establishes interactions between dual-branch latents that map to the same position in the 360 sphere, enabling high-quality and plausible spherical video patterns. While positionally aligned latents offer local consistency, they fall short in accounting for a unique characteristic of panoramic videos, where each pixel undergoes reverse camera translation of its antipodal counterpart. Hence, we improve the cross-domain attention with an **antipodal mask**, to extend each pixel's receptive field from its local neighborhood to its antipodal region across the 360 sphere, making it easier to learn long-range panoramic motion dependencies. Another key challenge lies in the varying input camera poses. Handling general videos as anchors requires the framework to accommodate changing camera poses across frames. In practice, we incorporate **rotation-aware designs**, including rotation-aware data sampling in training and a camera pose estimation module in inference. Through explicit tracking of the rotating input video in the 360 spacetime canvas, our model streamlines robust generation from customized video inputs. Effective training of our framework also relies on high-quality 360° video data. However, existing datasets are either small in scale [34; 37; 38] or consist of large-scale, unfiltered web-scraped content that requires extensive cleaning [28]. In this regard, we introduce YouTube360, a ready-to-train 360° video dataset comprising 10K curated clips from YouTube. Our dataset incorporates manual quality control and sophisticated data cleaning to select high-quality training segments with diverse and structured motion.

With these three key designs and our curated dataset, Imagine360 pioneers end-to-end, high-quality 360° video generation from perspective inputs. Extensive experiments show that our model achieves

state-of-the-art performance in both frame quality and motion consistency. As a bonus, our pipeline also demonstrates superior results in panoramic image outpainting. We believe Imagine360 can empower the 360° video generation community to create personalized, hyper-immersive experiences for real-world applications.

2 Related work

2.1 Video outpainting

Video Outpainting aims to fill in the missing regions at the edges of source videos. Compared to advanced image-level outpainting, video-level outpainting remains under-explored due to its challenges in maintaining both spatial and temporal fidelity and consistency. Recent video outpainting methods leverage diffusion to generate high-quality pixels in the missing regions. M3DDM [8] proposes a frame-guided Masked 3D diffusion model and a coarse-to-fine inference strategy to tackle artifact accumulation in long video outpainting. MOTIA [30] employs a per-case optimization strategy to learn the data-specific patterns of source video for better outpainting quality. Follow-Your-Canvas [2] divides the canvas into multiple windows and achieves outpainting of different sizes and resolutions by merging each outpainted window. These methods focus on handling perspective video outpainting with fixed masking in each frame. Despite their appealing outpainting results, it remains difficult for them to handle perspective-360° video generation that requires high-quality outpainting in panoramic distribution and continuously changing video masks from varying camera poses. In contrast, our Imagine360 handles the perspective-to-360° video generation with global and local constraints from dual-branch diffusion and rotation-aware designs to handle changing video masks, producing high-quality 360° video generations from perspective video anchors.

2.2 360 panorama generation

Early methods [31; 4; 1; 5; 15; 21; 35] exploit GAN-based framework for panorama image generation [19; 27]. OmniDreamer [1] proposes transformer-based framework for 360-degree outpainting and devises circular inference to obtain 360° close-loop continuity. Recently, diffusion-based methods [33; 17; 36; 42; 32; 9; 16; 26; 43] have dominated image-level panorama generation. Due to the data scarcity of large panorama image datasets, methods [33; 32] that directly fine-tune LDM to generate the panorama results in low-quality images with simple structures and sparse assets. PanFusion [43] introduces panorama and perspective branches to leverage the synergy from both global and local constraints for text-to-panorama generation.

Despite numerous efforts in image-level panorama generation, there are few works [34; 18; 20] focusing on panorama video generation. 360DVD [34] takes text prompts and additional panorama video optical flow as guidance and learns a 360-Adapter on standard T2V models to generate plausible 360° videos. 4K4DGen [18] animates a static panoramic image at user-selected regions with I2V pre-trained prior in a training-free manner. These fine-tuning-based or training-free approaches struggle to bridge the distribution gap between panoramic and perspective videos, resulting in simple natural perturbations, such as clouds moving and water running. In contrast, our Imagine360 benefits from the dual-branch video denoising structure with antipodal relation modeling and rotation-aware designs, resulting in more dynamic 360° videos with rich and structured motions.

3 Our approach

Given a perspective video, we aim to generate an equirectangular (ERP) 360 video that replicates its visual appearance and motion, and extends to the complete $360^\circ \times 180^\circ$ field of view.

We propose Imagine360, the first perspective-to-360° video generation framework, incorporating three key designs to synthesize high-quality panoramic videos, as illustrated in Fig. 2. First, to learn the spherical visual and motion patterns based on pre-trained perspective generative prior, we employ the dual-branch design, consisting of a panorama branch and a perspective branch, to jointly denoise the 360° spacetime latent. Second, to obtain more fine-grained and plausible panoramic motion, we refine the cross-domain attention to highlight antipodal masking that captures long-range motion dependencies in antipodal directions. Lastly, to handle diverse video inputs with varying camera

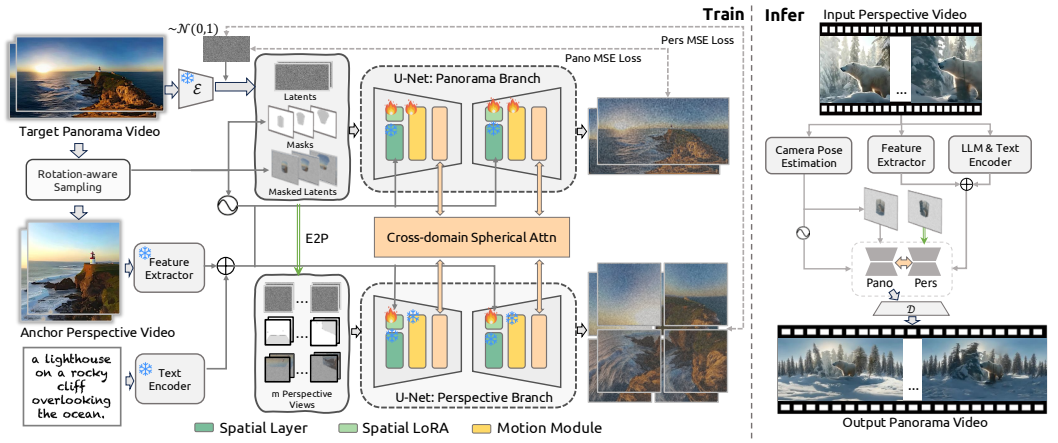

Figure 2: **Pipeline of Imagine360.** Given a perspective anchor video, Imagine360 uses a dual-branch denoising structure across panorama and perspective domains, featuring cross-domain spherical attention with antipodal masking for modeling long-range antipodal motion, and rotation-aware designs to handle varying camera rotations.

poses, we propose rotation-aware training and inference designs to obtain robust generations from different video anchors.

3.1 Preliminary: 360 video format

360° video is a video format that captures a full horizontal 360° field of view (FOV) and vertical 180° FOV. Unlike standard perspective videos, which are limited to a fixed, narrow field of view, 360° video enables viewers to look in any direction at each frame location, creating an immersive experience. For encoding, storage, and playback, 360 video must be projected onto a 2D plane from the 3D sphere. In our work, we use 360 videos encoded by Equirectangular Projection (ERP).

To create perspective video frame from ERP video frame of resolution $H \times W$, we simulate a virtual pinhole camera looking in a specific direction defined by yaw (θ), pitch (ϕ) and roll (ψ) Euler angles and FOV parameter. Each pixel on the image plane is computed with a direction vector $\vec{v}_{ij} = [1, \frac{2j}{W} - 1, -\left(\frac{2i}{H} - 1\right)]^T \cdot \tan\left(\frac{\text{FOV}}{2}\right)$ in 3D space and is rotated according to the desired viewing orientation $\vec{v}' = R_\psi \cdot R_\phi \cdot R_\theta \cdot \vec{v}_{ij}$ given the Euler angles. Then, these vectors are converted into spherical coordinates $\lambda = \arctan 2(\vec{v}'_y, \vec{v}'_x), \phi = \arcsin(\vec{v}'_z)$ and mapped to 2D ERP coordinates $u = \left(\frac{\lambda+\pi}{2\pi}\right) \cdot W, v = \left(1 - \frac{\phi+\frac{\pi}{2}}{\pi}\right) \cdot H$ for sampling. We denote this process as the Equirectangular-to-perspective (E2P) mapping. The inverse mapping, denoted as P2E mapping, maps a perspective video frame back to ERP frame. The mapping begins by computing a 3D directional vector in the spherical coordinates for each pixel in the ERP frame $\lambda = 2\pi \cdot \left(\frac{u}{W}\right) - \pi, \phi = \frac{\pi}{2} - \pi \cdot \left(\frac{v}{H}\right), \mathbf{v} = [\cos(\phi) \cdot \cos(\lambda), \cos(\phi) \cdot \sin(\lambda), \sin(\phi)]^T$, then rotate the vector inversely into the perspective camera's reference frame and projected onto the image plane. Pixels that fall within the camera's FOV sample the perspective pixel value and write back to the ERP frame.

3.2 Video-conditioned 360° video generation

Our model is trained on triplets of 360° videos, text captions, and corresponding 360° video masks. Using E2P (Equirectangular-to-Perspective) mapping, the 360° video mask determines the anchor perspective video extracted from the full 360° video. To generate different training masks, we sample sets of Euler angles $(\theta^{1:T}, \phi^{1:T}, \psi^{1:T})$ over T frames via P2E (Perspective-to-Equirectangular) mapping, where θ denotes yaw, ϕ pitch, and ψ roll. Similar to previous inpainting models [23; 2], our model takes as input a concatenation of the noisy 360° video latent, the video mask, and the masked 360° video latent along the channel dimension. A frozen VAE encodes both the full and masked 360° video inputs into d-dimensional latent channels. Additionally, we project the masked region of the 360° video into the perspective anchor video to provide fine-grained visual guidance. Following [2], the anchor video $I_{\text{anc}}^{1:T}$ is encoded into semantic features using the visual encoder from SAM [14]. These features are then processed by a pre-trained query-based Transformer, which distills high-level

visual and motion cues into compact latent tokens. The visual tokens are concatenated with the text embeddings in the conditioning branch. To guide large-scale spacetime pixel generation, we incorporate IP attention [41], which decouples cross-attention for injecting text prompts and visual tokens in the U-Net layers. This enables our model to effectively propagate visual content and motion patterns across views and into unmasked regions of the spacetime panorama.

3.3 Dual-branch 360° video denoising

Dual-branch design. Extending a perspective video to a 360° canvas requires careful model design due to the large differences between panoramic and perspective distributions. Mainstream methods either train a Latent Diffusion Model (LDM) to directly denoise panorama frames or jointly denoise multiple perspective view projections. The former results in plain visuals and mild motion from limited 360 data, while the latter easily produces local, short-range motions in each view. Inspired by image-level generation [42], we employ a dual-branch video denoising structure. It consists of a global panorama branch and a local perspective branch based on U-Net structure, where each block consists of spatial layers initialized from SD weights and a motion module initialized from [2] weights. The two branches share the same input: the panoramic branch latents of shape $\mathbb{R}^{T\times(2d+1)\times h\times w}$ is projected into $m = 20$ perspective views according to the P2E icosahedron modeling [22; 42], and each view gets latent of shape $\mathbb{R}^{T\times(2d+1)\times h/2\times h/2}$ through projection. In the panoramic branch self-attention, the model gains a holistic modeling of the 360° spacetime canvas for global consistency. In the perspective branch, the latent permutes and flattens the number of views with the batch size for self-attention to preserve the pre-trained generative power in the local window. In each U-Net downsample block, the latents from both branches are aligned via cross-attention to enhance the local-global synergy. The perspective latents $z_P^{1:T} \in \mathbb{R}^{T\times m\times D\times h/2\times h/2}$ are reshaped to $\mathbb{R}^{T\times D\times(m\times h/2\times h/2)}$, then bidirectionally cross-attended with panorama latents $z_E^{1:T} \in \mathbb{R}^{T\times D\times(h\times w)}$, where D is the channel dimension, h, w are the latent spatial dimension. Spherical positional encoding [42] is added to indicate the relative spherical location. A spherical attention mask on the attention map enforces E2P and P2E mapping between the flattened latent sequence from the two branches. As illustrated in Fig. 3, the directly mapped pixels via P2E mapping in the perspective branch are highlighted in orange. Gaussian blur is also applied to the spherical mask to enable neighboring activations in the attention.

Resource-friendly fine-tune strategy. With only thousands of training panorama videos, we seek to fine-tune a limited proportion of parameters to yield good generation results. Thanks to the space-time disentangled design from AnimateDiff [10], we do not need to fine-tune the whole model to learn the distributions of panoramic videos. In practice, we add LoRA on spatial attention layers and IP attention layers for both branches. As for the panoramic motion, we show in supp. that employing motion LoRA layers does not have sufficient influence on the pretrained prior to generate good panorama motion. Therefore, the whole motion module is fine-tuned for the panorama branch to learn the necessary prior for generating spherical motion patterns. Note that, to fully exploit the perspective generative prior, we do not fine-tune the motion module in the perspective branch. Experimental results show that our resource-friendly fine-tuning strategy achieves competitive results for 360° video generation despite the limited data and GPU resources. The training objectives in the two branches are:

$$\mathcal{L}_E = \mathbb{E}_{\mathcal{E}\left(x_E^{1:T}\right),y,\epsilon^{1:T},t}\left[\left\|\epsilon - \epsilon_{E,\Theta}\left(z_{E,t}^{1:T}, t, \tau_\Theta(y), I_{anc}^{1:T}\right)\right\|_2^2\right], \tag{1}$$

$$\mathcal{L}_P = \mathbb{E}_{\mathcal{E}\left(x_P^{1:T}\right),y,\epsilon^{1:T},t}\left[\left\|\epsilon - \epsilon_{P,\Theta}\left(z_{P,t}^{1:T}, t, \tau_\Theta(y), I_{anc}^{1:T}\right)\right\|_2^2\right], \tag{2}$$

where $x_\cdot^{1:T}$ and $\epsilon_{\cdot,\Theta}$ denote the target equirectangular (E) and perspective (P) video and its predicted noise, y is the text prompt. To facilitate the synchronous exchange of information between two branches, we combine the two branches losses as the final training objective $\mathcal{L} = \mathcal{L}_E + \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_P^i$.

3.4 Encouraging fine-grained panoramic patterns

The dual-branch design generates plausible spherical patterns for spatial appearance and motion, however, to achieve fine-grained panoramic videos, i.e., 360° close-loop continuity and reversed antipodal translation, we introduce additional advanced techniques, including circular padding and antipodal mask in cross-domain spherical mask.

Circular padding for close-loop continuity. The 360° closed-loop continuity ensures seamless alignment between the left-most and right-most edges of panoramic videos. To preserve this property and mitigate artifacts from local convolutions, we follow [42; 39] by applying circular padding before each convolutional layer in the U-Net blocks, then unpadding afterwards to restore the original resolution. For improved 360° continuity, we adopt circular padding in the optional video super-resolution [11] to process upsampled latents during decoding.

Antipodal mask for reversed translational motion. One distinctive property of panoramic motion is that antipodal pixels, i.e., those located on opposite sides of the 360° sphere, often exhibit reversed motion under camera translation. Specifically, if the camera moves forward relative to the forward viewing direction, the scene observed from the backward viewing direction should appear to move away from the viewer. However, in both the panoramic and perspective branches, antipodal pixels are typically distant in the token sequence. Since attention modules tend to focus on local neighborhoods, we observe that motion in antipodal directions tends to be less evident compared to motion in nearby pixels of the anchor video. To address this, and to better emphasize antipodal relationships in 360° video generation, we introduce the antipodal mask: a new activation mask for the cross-domain spherical attention. As illustrated in Fig. 3, we identify, for each latent in the perspective domain, its corresponding antipodal latent (shown in blue) in the panorama domain, and vice versa. These antipodal pairs are assigned higher activation values in the attention mask to encourage long-range attention across opposite directions. Additionally, we apply a Gaussian blur to the mask to softly include antipodal neighbors.

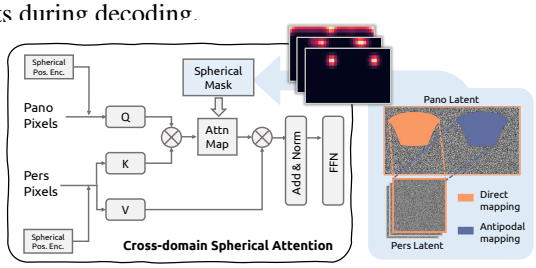

Figure 3: **Cross-domain Spherical Attention** highlights interaction for direct-mapped pixels (spherical mask) and **antipodal-mapped** pixels (antipodal mask) between panorama and perspective domains.

3.5 Rotation-robustness over general video input

Commonly, we assume the video input to be upright with a fixed camera pose, but in-the-wild perspective videos are of various rotation angles ($\theta^{1:T}$, $\psi^{1:T}$ and $\phi^{1:T}$). The change of rotation angle significantly impacts the shape and location of the masking, as shown in Fig. 6. To obtain robust generation for in-the-wild video inputs, we propose rotation-aware data sampling in training and camera pose estimation during inference.

Rotation-aware data sampling. We create flexible masking for each frame with a rotation-aware data sampling strategy. For simplicity, we randomly sample $\phi \in (-20°, 20°)$, $\theta \in (-45°, 45°)$, and $\psi \in (-10°, 10°)$ for each frame, then generate smooth angle sequences that are either monotonic or back-and-forth within the sampling range. Moreover, we add the sinusoidal positional embedding of the mask and the Euler angles to the conditioning tokens, i.e., the visual tokens of the anchor video and the text embedding, in order to indicate the relative position of the mask and the global canvas. For mask positional embedding, we compute based on its maximum inscribed box coordinates. The positional embeddings are concatenated along the channel dimension.

Camera pose estimation for inference. During inference, we employ MonST3R [44] to estimate the camera pose for each frame in the test video. The Euler angles are calculated from the extrinsic matrix at each frame. We do not normalize the Euler angle sequence according to the first frame, as sometimes the first frame does not map to the center region of the 360 canvas. We create the 360 video mask and masked 360 video pixels based on the estimated Euler angles. The masked video pixels are encoded by VAE and concatenated with the 360 video mask and noise as inference input.

3.6 YouTube360 dataset

Fine-tuning for 360° motion generation requires substantial text-video paired data due to the domain gap between panoramic and perspective motions. Existing datasets are insufficient: WEB360 [34] is small and limited to simple landscape videos, while 360-1M [28], although large in scale, contains unfiltered low-quality videos with incomplete views and polar artifacts. Cleaning such data requires large time and computational costs.

Table 1: **Quantitative comparison** on Vbench [12], flow-based Endpoint Error (EPE), OmniFVD [7] and user study results.

Method	Guidance	Vbench				EPE ↓	OmniFVD ↓	User Study		
		IQ ↑	AQ ↑	MS ↑	SC ↑			GQ ↑	SP ↑	TC ↑
Animatediff+LatentLab360	I2V	0.6248	0.5368	0.9866	0.8770	3.5393	285.9	1.2067	1.7588	1.4279
Follow-Your-Canvas	V2V	0.6257	0.5211	0.9799	0.9122	3.1767	315.4	2.7692	2.1298	3.0385
360DVD	T2V	0.5501	0.4359	0.9856	0.9356	3.1904	747.4	1.2067	1.7588	1.4279
Ours	V2V	**0.7372**	**0.5722**	**0.9866**	**0.9649**	**2.5583**	**204.0**	**3.6827**	**3.7260**	**3.3942**

To address this, we curate a high-quality 360 video dataset, YouTube360, from web videos. We collected 360-degree videos from YouTube, covering city tours, wildlife documentaries, and VR game captures. We prioritize long videos with diverse content, as they maintain a consistent format, offer varied scenes, and are easier to discard if polar artifacts are present. All videos are manually reviewed to remove sensitive content and low-quality samples with incomplete views or polar distortions. Then, the videos are converted to equirectangular format, resized to 512×1024, and segmented with TransNet-v2 [25]. For slow-paced clips, we apply 2× speed-up. To filter out static content, we compute optical flow using PanoFlow [24] and discard clips where fewer than 10% of frames exceed an average flow magnitude of 0.1. Captions are generated using VideoLLaMa-2 [6]. The final dataset contains 9,558 five-second clips at 20 fps. Each annotated sample includes a caption, a YouTube video ID, and a time interval. We provide annotations with documentation in the supplementary. Please refer to the supplementary material for additional details.

4 Experiments

4.1 Implementation details

Training settings. The spatial and motion modules are respectively initialized on Stable Diffusion v2.1 and [2]. Training is conducted on 8 NVIDIA A100 GPUs in 50k training steps, with the spatial LoRA rank and α_{LoRA} set to 32 and 1.0. The training resolution $H \times W$ is set to 256×512, the length of frames to 40, the batch size to 1, and the learning rate to 1×10^{-5}. For inference, generating a video of 512×1024 resolution and 32 frames takes approximately 6 minutes on average, using a single NVIDIA A100 GPU with 39 GB VRAM.

Evaluation metrics. We collect a test benchmark by randomly sampling videos from 360-1M [28], RealEstate10K [45], and CogVideoX [40] generations with prompts from GPT-4o [13], including both real videos and generated videos. We also randomly select 15 videos from the benchmark for the user study. Following previous practice, we employ Imaging Quality (IQ), Aesthetic Quality (AQ), Motion Smoothness (MS), and Subject Consistency (SC) metrics from VBench [12] to measure the graphics quality and motion consistency. To measure 360 motion correctness, we report Endpoint Error (EPE) on video optical flow (OF) with groundtruth 360 videos from 360-1M [28]. The EPE metric is adapted from optical flow to evaluate motion accuracy. We compute optical flow for both the generated video V_{gen} and the ground-truth V_{gt} using PanoFlow [24], and define EPE as the pixel-wise Euclidean distance between the two flows: $\text{EPE}(V_{\text{gen}}, V_{\text{gt}}) = |V_{\text{gen}} - V_{\text{gt}}|$. Small EPE indicates the higher similarity of generated video OF with Groundtruth OF. To measure omnidirectional video quality, we report OmniFVD that extends from OmniFID [7]. OmniFVD projects the 360 video into six cubemap views, computes FVD on each view, and reports the average as the final score.

Comparison methods. As the first perspective-to-360° video generation framework, it's infeasible to find a method that has the exact same input condition as ours. We compare with methods that produce 360 videos from various guidance. *360DVD* [34] is the advanced text-guided 360° video generation method that takes text prompts and panorama video optical flow as input for 360° video generation. Here, we feed the same text prompt as ours with GT optical flow into 360DVD for comparison. *Follow-Your-Canvas* [2] is the state-of-the-art video outpainting method that employs tile-based outpainting to handle outpainting of arbitrary size and resolution. Here, we project our perspective anchor video into the panorama canvas and feed the masked canvas into [2] for evaluation. *AnimateDiff+LatentLab360* [10] takes image inputs and animates with the LatentLab360 LoRA. For comparison, we feed the first frame of our generations into AnimateDiff+LatentLab360 and take its animated videos for evaluation.

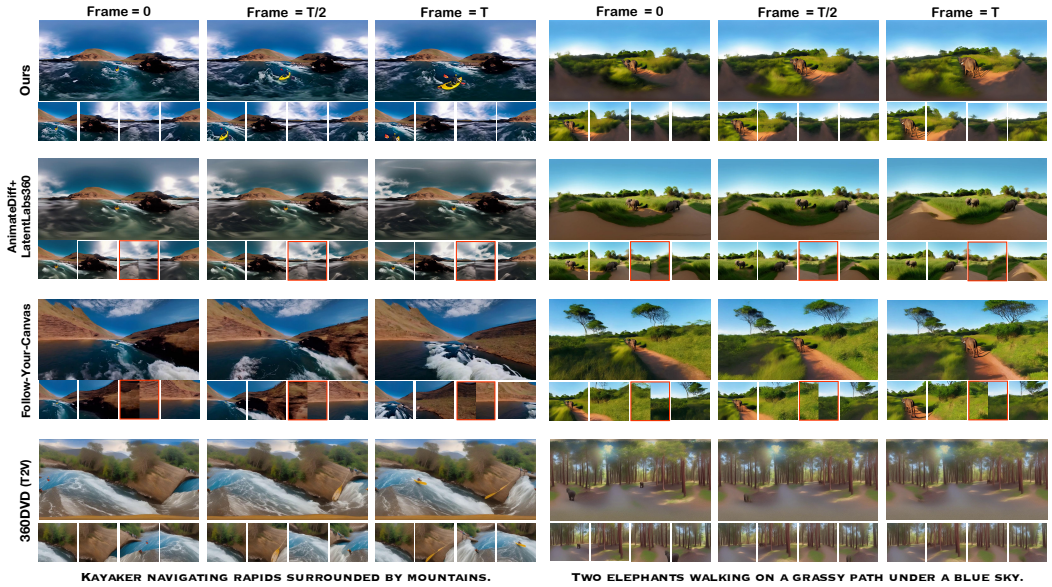

Figure 4: **Qualitative comparisons** on 360° video generations among state-of-the-art methods. Imagine360 generates 360° video generation with superior visual quality and plausible panoramic patterns.

Table 2: **Quantitative ablation studies** on dual-branch designs, antipodal mask, rotation-aware designs under Vbench, EPE, and OmniFVD metrics.

Method	Vbench				EPE ↓	OmniFVD ↓
	IQ ↑	AQ ↑	MS ↑	SC ↑		
Ours(Full model)	**0.7372**	**0.5722**	**0.9866**	**0.9649**	**2.5583**	**204.0**
w/o persbranch	0.7269	0.5464	0.9783	0.9491	3.0264	681.6
w/o panobranch	0.7196	0.5081	0.9760	0.9307	3.5308	866.0
w/o antipodal mask	0.7321	0.5377	0.9737	0.9375	2.7169	257.6
w/o rotation-aware designs	0.7319	0.5518	0.9785	0.9441	2.7276	332.1

4.2 Quantitative comparisons

Tab. 1 provides the quantitative comparison with baseline models. Imagine360 outperforms the existing 360 video generation methods at both ground-level perspective views and overall 360 correctness. 360DVD often produces blurred videos, resulting in lower performance across all image quality metrics. Follow-Your-Canvas performs comparably to AnimateDiff in perspective image quality metrics (IQ, AQ); while it generates more aesthetically pleasing details, it may suffer from distortion in projected perspective views. All methods achieve similar scores in perspective motion quality metrics (MS, SC). For the 360° motion metric EPE, video-conditioned generation methods outperform text- or image-based approaches. 360DVD achieves better EPE scores than AnimateDiff, as it leverages ground-truth optical flow as input.

We also involve a user study to examine the generated 360° videos. We invite 26 users with expertise in video and 3D generation to assess the results across three dimensions: graphics quality, structure plausibility, and temporal coherence in both 360 videos and projected perspective views. Tab. 1 reports the average user ranking of all four methods, and our method achieves the best performance in all three dimensions. Please refer to the supplementary material for more details regarding the setup and metrics. Our method also achieves superior panorama image outpainting performance, and we provide the comparison in the supplementary material.

4.3 Qualitative comparisons

We present the qualitative comparison between Imagine360 and other baseline models in Fig. 4. At each frame, we present 4 projected perspective views ($\phi = 0°$, $\theta = [0°, 90°, 180°, 270°]$) to examine the panoramic structure plausibility. Fig. 4 shows that AnimateDiff+LatentLab360 and Follow-Your-Canvas fail to achieve 360° close-loop continuity (orange box) and they produce mild-

Table 3: **Additional Upgrading Experiments** on backbone and camera pose estimation.

Method	Vbench				EPE ↓	OmniFVD ↓
	IQ ↑	AQ ↑	MS ↑	SC ↑		
Ours	0.7372	0.5722	0.9866	0.9649	**2.5583**	**204.0**
+ TTT3R [3]	0.7406	**0.5731**	0.9831	**0.9659**	2.7276	332.1
+ WAN [29]	**0.7508**	0.4896	**0.9929**	0.9639	2.7385	268.6

Figure 5: **Qualitative ablation on the dual-branch design** shows reduced artifacts and improved 360 patterns with dual-branch video denoising.

Figure 6: **Qualitative ablation on the rotation-aware designs** show plausible, consistent scene geometry with rotation designs.

scale motion, as observed in the change of the canoe location (left-side) and animal size (right-side). 360DVD produces more distorted patterns and blurred visual appearance in both cases. In contrast, our Imagine360 achieves superior visual quality, and plausible, obvious motion in the generated 360° videos. We also provide side-by-side comparison videos in the supplements for clearer visualization.

Table 4: Efficiency analysis with baseline methods under inference time and VRAM.

	AnimateDiff	360DVD	Follow-Your-Canvas	Ours
Inference Time (s)	122.65	93.18	381.89	182.76
max VRAM (GB)	20.83	18.06	27.99	25.59

Table 5: Inference Runtime Breakdown of Imagine360.

	Model Load	IO	Camera Est.	Inference
Runtime (s)	13.29	2.1	47.88	182.76

4.4 Efficiency Analysis

We provide a runtime breakdown table (Tab. 5) and efficiency comparison table (Tab. 4) in terms of runtime and VRAM consumption. We report the average runtime based on tests using 16-frame video clips.

From the efficiency comparison, we observe that compared to another V2V model Follow-Your-Canvas (FYC), our model is faster in inference speed and consumes less VRAM. The main reason is that FYC requires multiple rounds of outpainting to ensure outpaint quality, and they also train and infer in float32. Thanks to our dual-branch structure, Imagine360 achieves good visual quality in a single pass of inference. AnimateDiff and 360DVD have faster inference speeds, which is within expectation, as they are text-guided and image-guided frameworks. In addition, by employing multiple memory optimization techniques (bfloat16,vae slicing,cuda.empty_cache, intermediate tensor cleanup, etc.), our dual-branch structure does not increase VRAM usage by a lot.

In the runtime breakdown table, we observe that the external camera estimation module takes up a considerable amount of time. Although optimizing the inference speed of off-the-shelf camera pose estimator is beyond the scope of this work, we will seek to upgrade this module with a faster and more robust camera estimator.

4.5 Ablative studies

Ablation on dual-branch design. Fig. 5 shows that with limited training data, single panorama branch often proposes non-spherical patterns, and single perspective branch without global guidance produces messy boundary artifacts. Quantitative results in Tab. 2 also support the effectiveness of the dual-branch design. We also study the dual-branch fine-tuning strategy (see supp.), and find that compared to fine-tuning the pano motion module, adding motion LoRA is insufficient to generate good spherical pattern.

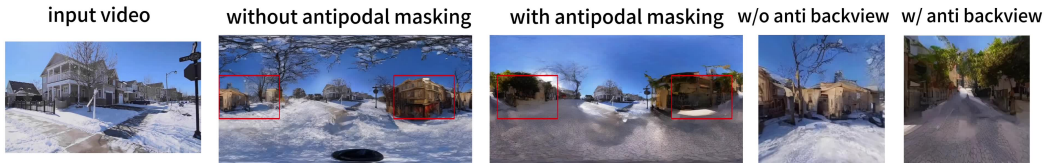

Figure 7: **Qualitative ablation on the antipodal mask** shows improved reversed motion in the antipodal view for forward-moving camera motion. Best viewed with Acrobat Reader for the motion.

Ablation on rotation-aware designs. As shown in Fig. 6, neglecting camera rotations in the input anchor video introduces artifacts in the scene geometry of the resulting 360° video. Without rotation-aware processing, the mountain appears to grow unnaturally taller (see yellow box). In contrast, our rotation-aware design preserves consistent scene geometry. We provide a clearer video comparison in supp. to show the robustness to diverse rotations. Quantitative results in Tab. 2 also validates the benefits of rotation-aware designs.

Ablation on antipodal mask. Fig. 7 illustrates the impact of the antipodal mask on antipodal motion. Thanks to the antipodal mask, as the camera moves forward in the input direction (center view), we can observe clear backward motion in the antipodal views (see red box). Quantitative result in Tab. 2 also proves the effectiveness of adding antipodal activations for general videos. We show the backward view for both cases: the result without antipodal mask exhibits distorted patterns at faraway locations, while with antipodal mask, the camera moves away from the scene in a more plausible way. We further verify that this design does not compromise content quality for *rotation-only* videos. On a filtered subset of 100 such videos from the test set, VBench IQ and AQ scores remain comparable: with the antipodal mask, IQ is **0.6688** and AQ is 0.5553; without it, IQ is 0.6621 and AQ is **0.5649**.

Analysis on camera estimation model. In our default setting, we use MonST3R for camera pose estimation during inference, which provides accurate and stable results in most cases. While precise pose estimation remains challenging in rare corner cases, this is beyond our paper's scope. Nevertheless, as camera pose estimation continues to advance, MonST3R can be seamlessly replaced with more powerful models. For instance, we replace MonST3R with more advanced TTT3R [3] and report the results in Tab. 3. Results with TTT3R improve on vbench IQ, AQ, and SC metrics, indicating more stable pose estimation ability of TTT3R over MonST3R.

Analysis on backbone model. As more advanced video generation backbone emerges, we also upgrade our backbone model from AnimateDiff [10] to Wan [29] with technical modifications from Unet to DiT. We find that this upgrade significantly improves the image quality and temporal consistency, especially solving the temporal flickering issue observed in some of our primary results. We report the results with Wan [29] in Tab. 3, and show qualitative results in the webpage as well as in the supplementary.

5 Conclusion

In this paper, we propose Imagine360, the first framework for video-conditioned 360 video generation with structured 360 motion. Our key contributions are: (1) a dual-branch video denoising network with panorama and perspective branches for global-local constraints; (2) an antipodal mask in cross-domain spherical attention to capture reverse translational camera motion; and (3) rotation-aware designs to handle diverse camera poses. We also introduce *YouTube360*, a high-quality 360 video dataset with diverse, structured foreground motion. Experiments show that Imagine360 achieves superior video quality and panoramic motion realism.

Limitations and future work. During inference, Imagine360 leverages camera pose estimation model to obtain the input camera poses. While these methods [44; 3] generally performs well, it can underestimate rapid movements involving large rotations due to data bias, leading to noticeable distortions (see supp.) that limit the video fidelity in certain scenarios. As research in geometry estimation continues to progress, we plan to keep our model updated with more advanced techniques to alleviate this issue.

Acknowledgments and Disclosure of Funding

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

[1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. *CVPR*, 2022.

[2] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024.

[3] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

[4] 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.

[5] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via GAN inversion. In *CVPR*, pages 11421–11430. IEEE, 2022.

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[7] Anders Christensen, Nooshin Mojab, Khushman Patel, Karan Ahuja, Zeynep Akata, Ole Winther, Mar Gonzalez-Franco, and Andrea Colaco. Geometry fidelity for spherical images. In *European Conference on Computer Vision*, pages 276–292. Springer, 2024.

[8] Fanda Fan, Chaoxu Guo, Litong Gong, Biao Wang, Tiezheng Ge, Yuning Jiang, Chunjie Luo, and Jianfeng Zhan. Hierarchical masked 3d diffusion model for video outpainting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7890–7900, 2023.

[9] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models, 2023.

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

[11] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024.

[12] 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 *CVPR*, pages 21807–21818. IEEE, 2024.

[13] 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.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[15] Dilip Krishnan, Piotr Teterwak, Aaron Sarna, Aaron Maschinot, Ce Liu, David Belanger, and William T. Freeman. Boundless: Generative adversarial networks for image extension. In *ICCV*, pages 10520–10529. IEEE, 2019.

[16] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions, 2023.

[17] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023.

[18] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024.

[19] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14211–14219, 2024.

[20] Jingwei Ma, Erika Lu, Roni Paiss, Shiran Zada, Aleksander Holynski, Tali Dekel, Brian Curless, Michael Rubinstein, and Forrester Cole. Vidpanos: Generative panoramic videos from casual panning videos. In *SIGGRAPH Asia 2024 Conference Papers*, December 2024.

[21] 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 *ECCV (16)*, volume 13676 of *Lecture Notes in Computer Science*, pages 352–371. Springer, 2022.

[22] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3772, 2022.

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

[24] Hao Shi, Yifan Zhou, Kailun Yang, Xiaoting Yin, Ze Wang, Yaozu Ye, Zhe Yin, Shi Meng, Peng Li, and Kaiwei Wang. Panoflow: Learning 360° optical flow for surrounding temporal understanding. *IEEE*

Transactions on Intelligent Transportation Systems, 24(5):5570–5585, 2023.
[25] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.
[26] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. Ldm3d: Latent diffusion model for 3d, 2023.
[27] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023.
[28] Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360° videos. *Advances in Neural Information Processing Systems*, 37:17743–17760, 2024.
[29] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
[30] Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. In *European Conference on Computer Vision*, pages 153–168. Springer, 2025.
[31] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing, 2022.
[32] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, pages 4921–4931. IEEE, 2024.
[33] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *ACM Multimedia*, pages 6811–6821. ACM, 2023.
[34] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6923, 2024.
[35] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Trans. Multim.*, 25:3546–3559, 2023.
[36] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion, 2024.
[37] Yutong Xu, Junhao Du, Jiahe Wang, Yuwei Ning, Sihan Zhou, and Yang Cao. Panonut360: A head and eye tracking dataset for panoramic video. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 319–325, 2024.
[38] Shilin Yan, Xiaohao Xu, Renrui Zhang, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. In *European Conference on Computer Vision*, pages 346–365. Springer, 2024.
[39] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv preprint arXiv:2408.13252*, 2024.
[40] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
[41] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.
[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.
[44] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024.
[45] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper tackles perspective-to-360 video generation via a dual-branch video denoising framework with rotation-aware designs and novel antipodal masking. Evaluation on 360 video quality and projected perspective video quality with a user study demonstrates the effectiveness of the model.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As discussed in supp., MonST3R can underestimate large camera rotations which leads to distorted input video in the canvas and hence distortions in the generated results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We discuss the model design and data collection process in Fig. 2 and Sec. 3.6 and in supplementary. We also provide the data annotations in the supplementary folder.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide YouTube ID, time intervals, and text captions in a csv file in the supplementary material. We also provide the inference code of our proposed model in the supplementary.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain all the training and test details in Sec. 4.1 and in the supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our metrics use standard benchmarks for the tasks we evaluate. These metrics do not include statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss in Sec. 4.1 that our model is trained on 8 80G A100 GPUs for 60 hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in the supplementary.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: As described in the supplementary, the codes and dataset will be made available under CC BY 4.0. We also employ manual efforts to filter out videos with sensitive content.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited all models and datasets we use in the paper. We also describe the licenses in the supplementation.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We describe the data curation and usage in Sec. 3.6 and in supplementary pdf. We also attach a readme file for the annotation file (.csv) in the supplementary.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the details of user study both in the experiment section and in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We follow the standards of the country where user studies were performed and personal data analyzed.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM is used exclusively for writing, editing, and data processing. See the method section and supplementary material.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.