# PanFlow: Decoupled Motion Control for Panoramic Video Generation

**Cheng Zhang[1,2], Hanwen Liang[3], Donny Y. Chen[1], Qianyi Wu[1], Konstantinos N. Plataniotis[3], Camilo Cruz Gambardella[2,4], Jianfei Cai[1]**

[1]Department of Data Science and Artificial Intelligence, Monash University, Clayton, Victoria, Australia
[2]Building 4.0 CRC, Caulfield East, Victoria, Australia
[3]The Edward S Rogers Sr. ECE Department, University of Toronto, Toronto, M5S3G8, Canada
[4]Future Building Initiative, Monash University, Caulfield East, Victoria, Australia
{cheng.zhang,qianyi.wu,camilo.cruzgambardella,jianfei.cai}@monash.edu, donny.chen@outlook.sg,
hw.liang@mail.utoronto.ca, kostas@ece.utoronto.ca

## Abstract

Panoramic video generation has attracted growing attention due to its applications in virtual reality and immersive media. However, existing methods lack explicit motion control and struggle to generate scenes with large and complex motions. We propose *PanFlow*, a novel approach that exploits the spherical nature of panoramas to decouple the highly dynamic camera rotation from the input optical flow condition, enabling more precise control over large and dynamic motions. We further introduce a spherical noise warping strategy to promote loop consistency in motion across panorama boundaries. To support effective training, we curate a large-scale, motion-rich panoramic video dataset with frame-level pose and flow annotations. We also showcase the effectiveness of our method in various applications, including motion transfer and video editing. Extensive experiments demonstrate that PanFlow significantly outperforms prior methods in motion fidelity, visual quality, and temporal coherence.

## Introduction

Generating realistic and immersive videos depicting the world around us is among the most compelling challenges in content creation. Panoramic videos, in particular, provide a natural representation by offering continuous 360° views that fully capture the surrounding environment, making them ideal for applications in virtual reality, immersive storytelling, and cinematic content creation. Enabled by advancements in diffusion-based generative models and the availability of panoramic datasets (Wang et al. 2024b), recent works have shown promising results in synthesizing photorealistic panoramic scenes from simple text or image prompts (Liu et al. 2025; Zhou et al. 2025).

Despite these advances, precise motion control in panoramic video generation remains an open challenge (Wang et al. 2024b; Wu et al. 2024b). Unlike traditional perspective videos, panoramic videos require loop consistency across boundaries, both in appearance and motion. Motion in panoramic videos involves complex interactions between two distinct yet intertwined components: *object motion* and *camera motion*. Accurate and coherent control of these motion components is crucial for an immersive viewing experience. Any inaccuracies can introduce noticeable artifacts, such as unnatural object trajectories, discontinuous motion across panorama seams, severely degrading the realism and practicality of generated content.

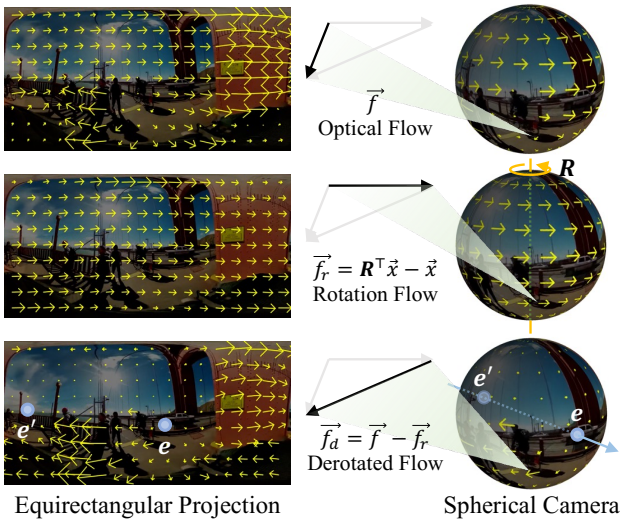


Figure 1: **Spherical Camera Optical Flow.** The optical flow from a panoramic video (left) can be interpreted as a spherical camera optical flow (right). For complex motion $\vec{f}$, the camera rotation yields an analytic rotation flow $\vec{f}_r$ on the sphere. By decomposing $\vec{f}$ into $\vec{f}_r$ and its residual, we obtain a derotated flow $\vec{f}_d$ that more clearly captures camera translation and object dynamics.

In this paper, we address this critical challenge by proposing PanFlow, a novel motion-controllable panoramic video generation framework conditioned on optical flow. We exploit the spherical nature of panoramic imagery and analytically decompose the input optical flow into two fundamental components: a deterministic *rotation flow* induced by camera rotation on the sphere, and a residual *derotated flow* that captures camera translations and object dynamics (Fig. 1). This decomposition is not only theoretically grounded, as camera-induced rotation flow is independent of scene content in panoramic view (Gluckman and Nayar 1998), but

also practically beneficial. It allows us to simplify the motion modeling by conditioning the diffusion model solely on the derotated flow, followed by an inverse derotation to recover the full motion. This decoupled motion control mechanism enhances synthesis stability, improves motion fidelity, and facilitates more coherent and realistic generation of dynamic panoramic scenes. To effectively condition the model on flow, we further introduce a *spherical noise warping* strategy. Instead of conditioning the model on optical flow embeddings, we warp the initial Gaussian noise across frames using optical flow fields. Importantly, in our design, noise pixels can be propagated across panorama boundaries to ensure seamless motion continuity under spherical geometry. Finally, to support high-quality training, we meticulously curate a large-scale dataset featuring dynamic content, diverse motion, and clean equirectangular projections, with frame-level pose and flow annotations. Leveraging proposed techniques and the curated dataset, PanFlow achieves state-of-the-art performance in panoramic video generation, exhibiting significantly improved motion fidelity, perceptual realism, and temporal coherence over existing approaches.

# Related Work

## Video Diffusion Models

Diffusion-based generative models have significantly improved the realism and diversity of synthesized videos. Early efforts extended image diffusion models (Rombach et al. 2022; Guo et al. 2024b) with temporal modules, while recent approaches (Yang et al. 2025; Wan et al. 2025) embed videos with 3D-VAE and leverage diffusion transformers for denoising. For controllability, most models incorporate conditioning signals such as texts or reference images (Ho et al. 2022; Yang et al. 2025). Beyond semantic controls, some works employ structural guidance like depth maps or edge cues (Wang et al. 2023; Guo et al. 2024a). Our method focus on motion control by leveraging synthetic optical flow as an explicit, scene-agnostic motion condition, enabling precise and temporally coherent control over dynamic content.

## Panoramic Image and Video Generation

More efforts have adapted diffusion models to panoramic domains for immersive scene synthesis. PanoGen (Li and Bansal 2023) employed latent diffusion to synthesize indoor panoramas from text, while StitchDiffusion (Wang et al. 2024a) introduced structural priors to enforce boundary alignment. PanFusion (Zhang et al. 2024) proposed a dual-branch design to handle global context and local detail. In video domains, DynamicScaler (Liu et al. 2025) designed training-free diffusion for panoramic and arbitrary aspect-ratio video generation, while HoloTime (Zhou et al. 2025) employed 4DGS (Kerbl et al. 2023) to generate explorable 4D scenes. PanoDiT (Zhang et al. 2025) and PanoWan (Xia et al. 2025) finetuned diffusion transformers for panoramic video generation. However, these works lack explicit motion control and often entangle camera and object motions with noticeable artifacts. The closest work to ours is 360DVD (Wang et al. 2024b), which extended perspective video diffusion models using 360-Adapter, allowing optional motion conditions. However, it exhibits limited motion control compared to our method, which enables fine-grained, user-customizable motion generation with superior visual quality and motion fidelity.

## Motion-conditioned Video Generation

Precise motion control is essential for interactive and editable video synthesis. Prior works explored object-level control, scene-level control, and motion transfer. Object-level control focuses on directing specific elements using inputs like masks or trajectories. SG-I2V (Namekata et al. 2025) and DragAnything (Wu et al. 2024b) manipulate localized motion using learned trajectories or entity representations. Scene-level control involves guiding global dynamics via camera trajectories, either through explicit pose supervision (He et al. 2024; Wang et al. 2024c) or coarse directional categories, e.g., panning or zooming (Guo et al. 2024b). Motion transfer aims to adapt temporal dynamics from reference videos to new content. MotionClone (Ling et al. 2025) introduced attention-based motion encoding to improve transferability. Go-with-the-Flow (Burgert et al. 2025) leveraged warped noise for detailed motion control. While effective, these methods are limited to perspective videos and falls short to generalize to the spherical topology of panoramic content, which requires structural consistency across seams and poles. We propose the first framework for motion-controlled panorama-to-video generation, introducing tailored spherical noise warping and decoupled motion, enabling loop consistency and better motion fidelity.

# Method

PanFlow is a motion-controlled video generation framework that synthesizes dynamic panoramic scenes from an input panorama. The core of PanFlow is a fine-tuned video diffusion model that takes in spherical warped noise encoded for motion conditions. To enhance fidelity in motion control, we introduce a motion decoupling mechanism to separate camera rotation from other motion components. Additionally, we construct a specialized dataset for training PanFlow, composed of highly dynamic panoramic videos and diverse motion patterns. The overall framework is shown in Fig. 2.

## Loop Consistent Motion-Controllable Panoramic Video Diffusion

Unlike conventional video diffusion models that generate videos in a perspective view, PanFlow targets the generation of 360° panoramic videos. Direct application of perspective video diffusion techniques to panoramas often results in loop inconsistencies, particularly related to motion, semantics, and low-level pixels. Specifically, moving objects exiting boundaries of the panorama must seamlessly re-enter from the opposite sides (*motion consistency*), latent representations must maintain semantic coherence across panorama boundaries (*semantic consistency*), and decoded videos in pixel-space must form a seamless loop at the boundaries (*pixel consistency*). To ensure these consistencies, we introduce separate techniques to address each aspect.

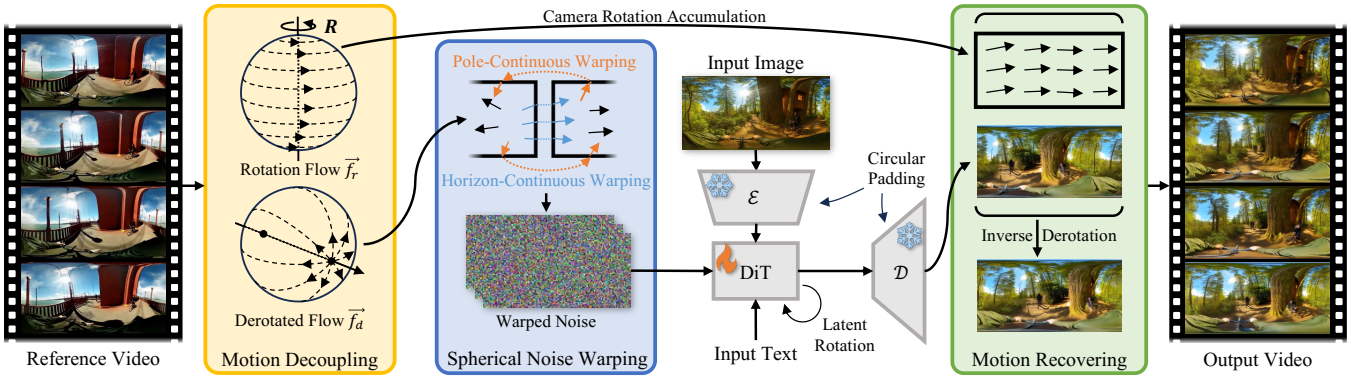


Figure 2: **Our proposed PanFlow pipeline.** Given an input image and text prompt, PanFlow uses a decoupled motion from a video as reference to generate a panoramic video. We first estimate a decoupled optical flow from the reference video, of which the derotated flow is used to generate a latent noise with spherical noise warping. The latent noise then serves as a motion condition for a video diffusion transformer with LoRA fine-tuning to generate derotated videos. Finally, the decoupled rotation is accumulated and applied to the generated video frames to recover the full motion.

**Spherical Noise Warping.** To enable effective motion control in panoramic video generation, we propose a spherical noise warping technique. Our method leverages the fact that Gaussian noise inherently defines image contents in diffusion models (Burgert et al. 2025). We introduce the spherical noise warping where the initial Gaussian noise is warped iteratively to subsequent timestamps using optical flow from a source video. This warped noise maintains distributional similarity across corresponding regions in different frames, achieving both temporal coherence and motion control. Formally, given Gaussian noise in a latent space of size $H \times W$, we denote previous-frame noise as $\boldsymbol{q} \in \mathbb{R}^D$ (where $D = H \times W$) and use noise warping operator (Burgert et al. 2025) to generate next-frame noise $\boldsymbol{q}'$ with forward and backward optical flows $\boldsymbol{f}, \boldsymbol{f}' : D \to \mathbb{N}^2$. To ensure seamless loop motion consistency specifically for panoramic content, noise warping crossing panorama boundaries is performed by defining target pixels $(\hat{i}, \hat{j})$ as:

$$\hat{i} = \begin{cases} -i, & i < 0, \\ i, & 0 \le i < H, \\ 2(H-1) - i, & i \ge H, \end{cases} \qquad (1)$$

$$\hat{j} = \begin{cases} \left(j + \frac{W}{2}\right) \bmod W, & i < 0 \text{ or } i \ge H, \\ j \bmod W, & 0 \le i < H, \end{cases}$$

where $(i, j)$ are the pixel coordinates mapped by the optical flow. This operation passes the content across the left-right boundaries and poles based on the connectivity of the equirectangular projection. We employ this specially warped noise to fine-tune a diffusion transformer equipped with a LoRA adapter (Hu et al. 2022), as illustrated in Fig. 2.

**Latent Rotation on 3D Latent Maps.** We build upon the latest video DiT model (Yang et al. 2025) that leverages transformers for denoising. Previous works (Zhang et al. 2024; Wu, Zheng, and Cham 2023) have shown that the pre-trained priors in the diffusion model conflict with the periodic structure of the panorama, tending to break the loop consistency in the latent space. To mitigate this problem, we apply latent rotation before tokenization on the 3D latent map. Specifically, at each denoising step, we rotate the latent map along the longitude axis with a fixed angle of $\theta$ by rolling for $\theta/360 \times W$ latent pixels along the $W$ dimension. The rolling distance is accumulated during the denoising process to rotate the denoised result back to the original orientation. This rotation helps DiT to blend the seam after it is shifted into the frame, instead of concentrating discontinuities at the boundaries.

**Circular Padding.** Inspired by (Zhuang et al. 2022), we integrate circular padding into both encoder $\mathcal{E}$ and decoder $\mathcal{D}$ of the 3D-VAE to maintain pixel-level loop consistency in both latent maps and decoded videos. Specifically, we pad the left-right boundaries with latents or pixels from opposite sides, enabling convolution kernels to operate seamlessly across panorama edges. After encoding or decoding, outputs are cropped back to their original dimensions. The padding width is set to 8 pixels for the encoder and 1 latent pixel for the decoder, sufficient to cover the receptive field.

## Decoupled Motion Control

Although the proposed techniques above establish a strong baseline for motion control in panoramic video generation, they still face challenges in highly dynamic scenarios. We observe that the motion condition often comprises a complex combination of camera rotation, translation, and object motions. Considering the spherical nature inherent to panoramic videos, we introduce a decoupled motion control to isolate camera rotation from other motion components. This decoupling reduces the modelling burden over the video diffusion model, thereby enhancing its ability to effectively follow complex motion conditions.

**Spherical Camera Optical Flow.** A 360° panorama corresponds to an equirectangular projection (Fig. 1 left) from a spherical camera (Fig. 1 right). Camera motions can thus be represented using spherical optical flow (Gluckman and

Nayar 1998). Specifically, for a motion described by optical flow $\vec{f}$, one independent component is its camera rotation. When the camera rotation is defined as rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$, its optical flow vector $\vec{f} \in \mathbb{R}^3$ at pixel position $\vec{x} = (x, y, z) \in \mathbb{R}^3$ on the unit sphere can be calculated as:

$$\vec{f_r} = \boldsymbol{R}^\top \vec{x} - \vec{x}. \tag{2}$$

This rotation flow follows a circular pattern around the rotation axis, and the magnitude is proportional to the distance to the rotation axis, as shown in the second row of Fig. 1. In contrast, in a static scene, when the camera moves with only translation, the optical flow vectors diverge from one epipole $\boldsymbol{e}$ and converge to opposite epipole $\boldsymbol{e}'$, where $\boldsymbol{e} - \boldsymbol{e}'$ is the direction of the camera translation and the magnitude of the flow vector depends on the scene depth.

**Decoupling Camera Rotation from Motion Condition.** We note that the rotation flow is invariant to the scene content, which can be easily decoupled from the input optical flow with a derotation operation:

$$\vec{f_d} = \vec{f} - \vec{f_r}. \tag{3}$$

The decoupling operation is simple yet effective, as it allows the video diffusion model to focus on the derotated flow, which only contains translation and object motion, while the camera rotation is handled separately. In practice, we directly derotate the input video and then estimate the derotated optical flow with a panoramic flow estimator (Shi et al. 2023) for more stable optical flow conditions, when input video is available in most applications such as motion transfer and object editing.

**Recovering Camera Rotation for Generated Videos.** After the derotated video is generated with the derotated optical flow as condition, we recover the decoupled camera rotation by applying the inverse derotation operation on each frame $\boldsymbol{I}_t$:

$$\hat{\boldsymbol{I}}_t = \text{RotateERP}(\boldsymbol{I}_t, \boldsymbol{R}_t) \tag{4}$$

where $\boldsymbol{R}_t$ is the rotation matrix relative to the first frame, and $\text{RotateERP}(\cdot)$ is the operation that rotates ERP panoramas with bilinear interpolation so that $\hat{\boldsymbol{I}}_t(\vec{x}) = \text{Bilinear}(\boldsymbol{I}_t, \boldsymbol{R}_t^\top \vec{x})$ for each pixel $\vec{x}$ on the unit sphere of $\boldsymbol{I}_t$. $\boldsymbol{R}_t$ can be extracted from the input video with SLAM (Sumikura, Shibuya, and Sakurada 2019) or from the input optical flow by accumulating estimated rotation differences (Kim et al. 2021). We note that this operation is achievable thanks to the omnidirectional nature of the panorama, which allows us to rotate the panorama arbitrarily with minimal information loss.

## Motion-Rich Dataset Construction

Training an effective motion-controllable panoramic diffusion model requires a highly dynamic motion-rich training dataset, including diverse camera motions and object motions. However, existing datasets such as WEB360 (Wang et al. 2024b) exhibit limited motion diversity and magnitude, often featuring static cameras and minimal object movement. To address this, we construct a new dataset by curating a collection of videos from a large corpus dataset, 360-1M (Wallingford et al. 2024). We design a specialized preprocessing pipeline tailored for panoramic videos to filter, segment, and annotate clips for training PanFlow, including:

- **Format Check.** 360-1M contains a portion of videos that are not in the equirectangular projection format. We filter out these videos with handcrafted heuristics, such as computing SSIM (Wang et al. 2004) between the left and right halves to exclude stereo (3D) formats, and applying spherical masks to identify and remove fisheye footage.
- **Transition Detection.** We apply the `scenedetect` (Inc. and contributors 2024) library to identify scene transitions, including hard cuts and black fades.
- **SLAM for Clip and Pose Extraction.** We use a panoramic visual SLAM (Sumikura, Shibuya, and Sakurada 2019) to extract the camera poses. As some fade cuts can not be easily detected by the transition detection, we exploit the "tracking lost" signal from the SLAM to further segment the scene clips.
- **Watermark Filtering.** We apply a deep learning model (LAION-AI 2022) to detect watermarks and captions in the videos, and filter out videos with high watermark scores.
- **Motion Filtering.** To enhance the dynamics of dataset, we compute the optical flow (Bradski 2000), and discard clips with low flow magnitudes and filter out those with minimal camera motion based on estimated poses.

The resulting dataset offers a rich and diverse set of 150k, 3-10s dynamic panoramic clips, with 53% natural, 22% urban, 9% indoor, and 16% others. We detail the dataset construction process in the supplementary material.

# Experiments

## Experimental Setup

**Implementation Details.** We hold out 100 clips from our curated dataset as a test set. To further evaluate the out-of-domain generalizability, we additionally select 100 random clips from the WEB360 dataset (Wang et al. 2024b) for evaluation. Please refer to the supplementary material for more details about the training and inference details.

**Baselines.** To comprehensively evaluate the effectiveness of our method, we compare against a diverse set of strong baselines covering both panoramic and flow-conditioned video generation. For panoramic video generation, we include text-driven model 360DVD (Wang et al. 2024b) and image-conditioned frameworks DynamicScaler (Liu et al. 2025) and HoloTime (Zhou et al. 2025). To assess motion control quality, we further compare against two recent flow-conditioned works, MotionClone (Ling et al. 2025) and GoWithTheFlow (Burgert et al. 2025). These baselines provide a broad and representative benchmark spanning different conditioning modalities and generation paradigms.

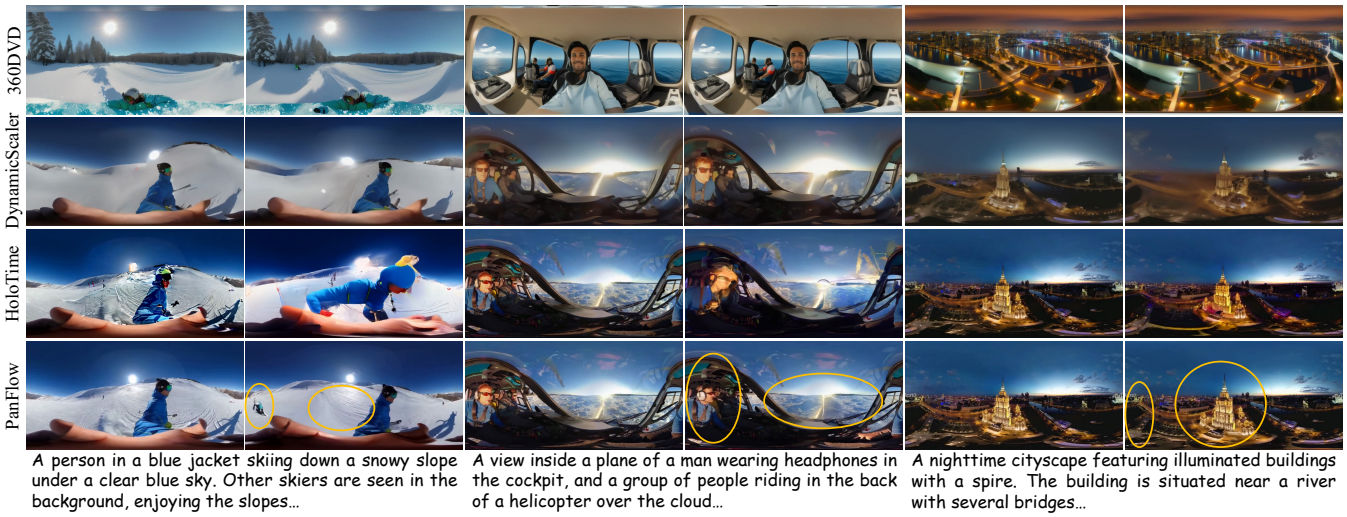


Figure 3: **Comparison with Panoramic Video Generation Methods.** We compare our proposed PanFlow with the baselines conditioned on the same input images and text prompts (360DVD uses text prompts only). Generated frames are shown at the same timestamps. We highlight regions exhibiting more dynamic motion, high-fidelity textures, and consistent geometry.

| | Method | FVD↓ | FID↓ | CLIP↑ | Motion Control | | | | Q-Align | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Flow EPE↓ | LPIPS↓ | PSNR↑ | SSIM↑ | Image Quality↑ | Image Aesthetic↑ | Video Quality↑ |
| 360-1M | 360DVD | 1663.61 | 152.04 | 24.57 | 4.467 | 0.771 | 9.41 | 0.319 | 0.5426 | 0.4231 | 0.6073 |
| | DynamicScaler | 1360.94 | 85.25 | 25.14 | 8.647 | 0.529 | 14.32 | 0.500 | 0.5137 | 0.4142 | 0.5884 |
| | HoloTime | 1013.32 | 59.59 | 24.62 | 6.322 | 0.443 | 14.83 | 0.490 | 0.5218 | 0.4016 | 0.5929 |
| | MotionClone | 1435.66 | 87.60 | 25.71 | 9.613 | 0.577 | 11.91 | 0.412 | 0.5533 | 0.4364 | 0.6147 |
| | GoWithTheFlow | 477.13 | 28.59 | 25.75 | 3.297 | 0.306 | 18.43 | 0.622 | 0.5459 | 0.4069 | 0.5879 |
| | PanFlow | 298.39 | 23.01 | 27.14 | 2.011 | 0.243 | 21.21 | 0.723 | 0.5552 | 0.4397 | 0.6196 |
| WEB360 | 360DVD | 1225.55 | 153.55 | 25.60 | 1.116 | 0.726 | 11.10 | 0.416 | 0.6918 | 0.5980 | 0.7342 |
| | DynamicScaler | 979.30 | 94.84 | 25.30 | 3.132 | 0.464 | 18.20 | 0.634 | 0.5991 | 0.5478 | 0.6880 |
| | HoloTime | 443.56 | 40.85 | 25.31 | 0.899 | 0.250 | 23.97 | 0.771 | 0.6865 | 0.5714 | 0.7407 |
| | MotionClone | 667.19 | 72.53 | 25.71 | 1.864 | 0.405 | 17.03 | 0.634 | 0.6831 | 0.5784 | 0.7466 |
| | GoWithTheFlow | 319.01 | 23.81 | 25.30 | 0.710 | 0.219 | 25.38 | 0.795 | 0.6357 | 0.5187 | 0.6743 |
| | PanFlow | 195.77 | 19.44 | 26.84 | 0.486 | 0.185 | 27.46 | 0.842 | 0.7014 | 0.6165 | 0.7515 |

Table 1: **Comparison with SoTA Methods.** We evaluate the video generation quality with FVD, FID, CLIP scores, and user-centric Q-Align metrics. For motion control, we report Flow EPE (End-Point Error), LPIPS, PSNR, and SSIM scores. PanFlow achieves consistently superior performance across both 360-1M and WEB360 datasets.

**Evaluation Metrics.** Evaluation is conducted using a comprehensive suite of metrics. Visual quality and temporal coherence of the generated videos are evaluated using Fréchet Inception Distance (FID) (Heusel et al. 2017) and Fréchet Video Distance (FVD) (Unterthiner et al. 2019). Text alignment is measured by CLIP (Radford et al. 2021) similarity. Motion controllability and appearance quality is assessed by frame-wise visual similarity measuring PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) between the generated images and ground-truth views. To specifically assess motion control fidelity, we additionally report Flow EPE (Otte and Nagel 1994), which measure the deviation between optical flows estimated from generated and ground-truth videos, isolating motion fidelity from visual quality. For user-centric evaluation, we follow (Liu et al. 2025) and incorporate Q-Align (Wu et al. 2024a), an LLM-based visual evaluator that scores content based on aesthetic appeal and perceptual quality. This metric shows strong alignment with human judgment in visual assessment tasks. When evaluating different loop consistency setups, we also report the end continuity error from (Xia et al. 2025) that measures MSE between the leftmost and rightmost columns of the generated panoramic video.

## Comparison with Previous Methods

**Quantitative Comparison.** Tab. 1 presents the quantitative comparison between our method and the baselines on the 360-1M and WEB360 datasets. PanFlow consistently achieves the best performance across all evaluation metrics on both datasets. In terms of generation quality, PanFlow shows significant improvements in FVD, FID, and CLIP scores, indicating its superior video quality, temporal coherence, and text alignment. On human-aligned Q-Align metrics, PanFlow also outperforms all baselines, demonstrating superior aesthetics and visual quality that align closely with human preferences. In terms of motion control, PanFlow


Figure 4: **Comparison with Motion-controlled Video Generation Methods.** All the methods generate videos conditioned on the same input images and motion flows. PanFlow better follows the flow conditions and aligns more closely with the ground truth. We use rectangles to highlight regions with large motion. (Text prompts omitted without loss of generality.)

| Method | Decoupled Motion | Loop Consistency Setup | End Continuity↓ | FVD↓ | FID↓ | CLIP↑ | Motion Control | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Flow EPE↓ | LPIPS↓ | PSNR↑ | SSIM↑ |
| GWTF | w/o | w/o | 0.1389 | 610.69 | 32.34 | 25.77 | 5.754 | 0.435 | 14.63 | 0.500 |
| | w/ | w/o | 0.0276 | 580.79 | 32.26 | 25.77 | 4.818 | 0.420 | 15.01 | 0.520 |
| PanFlow | w/o | w/o Circular Padding | 0.0512 | 429.29 | 26.01 | 27.07 | 4.254 | 0.370 | 16.86 | 0.573 |
| | | w/o Latent Rotation | 0.0287 | 428.58 | 26.31 | 27.08 | 4.306 | 0.371 | 16.76 | 0.568 |
| | | w/o Spherical Noise Warping | 0.0286 | 441.59 | 26.04 | 27.08 | 4.289 | 0.371 | 16.83 | 0.572 |
| | | Full | 0.0264 | 430.90 | 25.89 | 27.07 | 4.260 | 0.369 | 16.87 | 0.573 |
| | w/ | Full | 0.0257 | 425.78 | 25.21 | 27.21 | 3.626 | 0.357 | 17.22 | 0.593 |

Table 2: **Ablation Study.** We compare ablated variants of PanFlow on 360-1M dataset. We also show that our proposed decoupled motion control module can be integrated into GoWithTheFlow (GWTF) (Burgert et al. 2025) as a plug-and-play module.

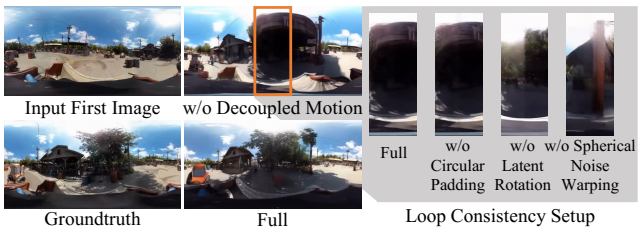


Figure 5: **Ablation Study.** Panoramas are horizontally rotated by 180 degrees to better visualize the seam. We zoom in on the seam to compare different loop consistency setups.

achieves the best performance in Flow EPE, LPIPS, PSNR, and SSIM, demonstrating its effectiveness in generating temporally coherent videos with accurate motion control. We note that our method is, to the best of our knowledge, the first to achieve high-quality panoramic video generation from image and flow motion conditions. Therefore, we focus our comparisons on motion-controllable video generation baselines, i.e. MotionClone and GoWithTheFlow, on these metrics. In addition, as 360DVD generates panoramic video from motion without image condition, we compare it using Flow EPE that solely evaluates motion fidelity, showing that 360DVD falls short significantly. In both tasks, we evaluate directly generalization to WEB360 dataset, where PanFlow still achieves the best performance, highlighting its strong generalization capability to out-of-domain data.

**Qualitative Comparison.** Fig. 3 and Fig. 4 present qualitative comparisons of our method with panoramic video generation and motion-controllable video generation baselines, respectively. In Fig. 3, we show the first and last frames of the generated videos. PanFlow produces high-quality panoramic videos with shaper details and more dynamic motion, showing more preserved textures on the snowy slope, more realistic motion of pilot turning head, and more consistent 3D geometry of the nighttime aerial view. In Fig. 4, we compare PanFlow with two flow-conditioned image-to-video generation methods, MotionClone and GoWithTheFlow. The image condition is shown on the left, and the generated 16th frames are shown on the right. PanFlow demonstrates more accurate motion control, with generated frames more closely matching the target frame. It also produces more coherent panoramic geometry than MotionClone and exhibits fewer boundary artifacts than GoWithTheFlow.

## Ablation Study

We conduct an ablation study on the full generated frames on the 360-1M test set and use a challenging various frame


(a) **Motion Transfer.**


(b) **Video Editing.**

Figure 6: **Applications.** Our method can be applied to motion transfer and video editing tasks. We visualize the direction of motion in the reference video (left) with white arrows. Full prompts and changes are omitted for brevity.

strides (1 to 3) to evaluate the impact on motion fidelity and overall video quality. Tab. 2 and Fig. 5 present the quantitative and qualitative results, respectively.

**Decoupled Motion Control.** Our decoupled motion control module works as a plug-and-play component that can be easily integrated into existing motion-controllable video generation methods. To demonstrate this, we incorporate it into the baseline GoWithTheFlow (GWTF) method, which improves most metrics, particularly those related to motion controllability. We also ablate the module by removing from our full model, observing a significant drop across all metrics. This highlights its crucial role in enhancing motion fidelity. In the very challenging case shown in Fig. 5, where the target frame is temporally distant from the image condition and involves highly dynamic motion, the decoupled motion control improves the semantic consistency, e.g., generating a tree rather than a building in the center, by stabilizing the motion condition. Interestingly, we find that the decoupled motion control also enhances loop-end continuity. This is because it rotates the seam in the generated derotated video away from the boundary, presumably hacking the metric. Therefore, in the following, we evaluate different loop consistency setups based on the ablated version without decoupled motion control.

**Loop Consistency.** We further evaluate the proposed spherical noise warping and other techniques for improving loop consistency. By ablating circular padding, the end continuity error increases dramatically, while other metrics remain similar. Removing latent rotation and spherical noise warping leads to less increase in end continuity error, but also affects Flow EPE and PSNR, indicating that they also contribute to the overall video quality and motion control fidelity. This is especially true for the spherical noise warping, as removing it leads to worse FVD and FID. The qualitative results in Fig. 5 further confirm the results, where the circular padding mostly affects pixel level end continuity, which is hard to see in the generated video, while the removal of latent rotation and spherical noise warping leads to more severe artifacts in the seams.

## Applications Driven by Motion Control

By combining decoupled motion control with spherical noise warping technique, our method can be easily applied to motion transfer and video editing tasks that require more robustness when motion control does not perfectly match the input image. We achieve this by following a random degradation strategy (Burgert et al. 2025), which blends a random amount of noise into the warped spherical noise when training. During inference, we relax the motion condition by blending it with noise of equal strength.

**Motion Transfer.** Fig. 6a presents the motion transfer results, where we transfer the motion from a source video to a drastically different target scene. It is shown that our method can preserve the target scene's geometry and details while accurately following the motion of the source video.

**Video Editing.** Fig. 6b shows examples of video editing. We use ChatGPT 4o's image generation to edit the first frame and the input prompt, and then generate the panorama videos based on the source motion. The results show that our method can adapt existing motion to different geometries while preserving the motion control fidelity.

## Conclusion

In this paper, we presented PanFlow, a novel framework for generating temporally coherent 360° panoramic videos with precise motion control from a single image. We introduced a decoupled spherical motion-flow mechanism that enables more stable and high-fidelity motion control. The spherical noise warping is introduced alongside other loop consistency techniques for seamless motion generation across panorama boundaries. We also curated a large-scale dynamic panorama dataset with frame-level flow and pose annotations to support training and evaluation. Extensive experiments across multiple datasets show that PanFlow significantly outperforms prior methods in motion magnitude and fidelity, temporal coherence, and visual quality.

**Limitations.** While the SLAM-based rotational estimation in our pipeline is robust under large motions, it may degrade when parallax is small—a limitation that could be mitigated by a learned derotation module (Kim et al. 2021), which produces robust rotation outputs, reduces the need for reference videos, and presents a promising direction for future work.

## Acknowledgments

This research is supported by Building 4.0 CRC.

## References

Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Burgert, R.; Xu, Y.; Xian, W.; Pilarski, O.; Clausen, P.; He, M.; Ma, L.; Deng, Y.; Li, L.; Mousavi, M.; et al. 2025. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 13–23.

Gluckman, J.; and Nayar, S. K. 1998. Ego-motion and omnidirectional cameras. In *ICCV*, 999–1005. IEEE.

Guo, Y.; Yang, C.; Rao, A.; Agrawala, M.; Lin, D.; and Dai, B. 2024a. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *ECCV*, 330–348. Springer.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024b. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *ICLR*.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *NeurIPS*, 35: 8633–8646.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Inc., B. A.; and contributors. 2024. PySceneDetect. https://github.com/Breakthrough/PySceneDetect. Accessed: 2025-11-29.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4): 139–1.

Kim, D.; Pathak, S.; Moro, A.; Yamashita, A.; and Asama, H. 2021. Self-supervised optical flow derotation network for rotation estimation of a spherical camera. *AR*, 35(2): 118–128.

LAION-AI. 2022. LAION-5B-WatermarkDetection. https://github.com/LAION-AI/LAION-5B-WatermarkDetection. Accessed: 2025-11-29.

Li, J.; and Bansal, M. 2023. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *NeurIPS*, 36: 21878–21894.

Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2025. Motionclone: Training-free motion cloning for controllable video generation. In *ICLR*.

Liu, J.; Lin, S.; Li, Y.; and Yang, M.-H. 2025. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In *CVPR*, 6144–6153.

Namekata, K.; Bahmani, S.; Wu, Z.; Kant, Y.; Gilitschenski, I.; and Lindell, D. B. 2025. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *ICLR*.

Otte, M.; and Nagel, H.-H. 1994. Optical flow estimation: advances and comparisons. In *ECCV*, 49–60. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Shi, H.; Zhou, Y.; Yang, K.; Yin, X.; Wang, Z.; Ye, Y.; Yin, Z.; Meng, S.; Li, P.; and Wang, K. 2023. PanoFlow: Learning 360 optical flow for surrounding temporal understanding. *T-ITS*, 24(5): 5570–5585.

Sumikura, S.; Shibuya, M.; and Sakurada, K. 2019. OpenVSLAM: A versatile visual SLAM framework. In *ACM MM*, 2292–2295.

Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation.

Wallingford, M.; Bhattad, A.; Kusupati, A.; Ramanujan, V.; Deitke, M.; Kembhavi, A.; Mottaghi, R.; Ma, W.-C.; and Farhadi, A. 2024. From an image to a scene: Learning to imagine the world from a million 360 videos. *NeurIPS*, 37: 17743–17760.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, H.; Xiang, X.; Fan, Y.; and Xue, J.-H. 2024a. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 4933–4943.

Wang, Q.; Li, W.; Mou, C.; Cheng, X.; and Zhang, J. 2024b. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *CVPR*, 6913–6923.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 36: 7594–7611.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024c. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024a. Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In *ICML*, 54015–54029.

Wu, T.; Zheng, C.; and Cham, T.-J. 2023. Ipo-ldm: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. *arXiv preprint arXiv:2307.03177*.

Wu, W.; Li, Z.; Gu, Y.; Zhao, R.; He, Y.; Zhang, D. J.; Shou, M. Z.; Li, Y.; Gao, T.; and Zhang, D. 2024b. Draganything: Motion control for anything using entity representation. In *ECCV*, 331–348. Springer.

Xia, Y.; Weng, S.; Yang, S.; Liu, J.; Zhu, C.; Teng, M.; Jia, Z.; Jiang, H.; and Shi, B. 2025. PanoWan: Lifting Diffusion Video Generation Models to 360 {\deg} with Latitude/Longitude-aware Mechanisms. *arXiv preprint arXiv:2505.22016*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2025. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*.

Zhang, C.; Wu, Q.; Gambardella, C. C.; Huang, X.; Phung, D.; Ouyang, W.; and Cai, J. 2024. Taming stable diffusion for text to 360 panorama image generation. In *CVPR*, 6347–6357.

Zhang, M.; Chen, Y.; Xu, R.; Wang, C.; Yang, J.; Meng, W.; Guo, J.; Zhao, H.; and Zhang, X. 2025. PanoDit: Panoramic videos generation with diffusion transformer. In *AAAI*, volume 39, 10040–10048.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhou, H.; Yu, W.; Guan, J.; Cheng, X.; Tian, Y.; and Yuan, L. 2025. HoloTime: Taming Video Diffusion Models for Panoramic 4D Scene Generation. *arXiv preprint arXiv:2504.21650*.

Zhuang, C.; Lu, Z.; Wang, Y.; Xiao, J.; and Wang, Y. 2022. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *AAAI*, volume 36, 3653–3661.