PanoDiT: Panoramic Videos Generation with Diffusion Transformer

Muyang Zhang[1,2*], Yuzhi Chen[1,2*], Rongtao Xu[2,1], Changwei Wang[3], JinMing Yang[2,1], Weiliang Meng[2,1†], Jianwei Guo[4,2†], Huihuang Zhao[5], Xiaopeng Zhang[2,1]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Qilu University of Technology (Shandong Academy of Sciences), Shandong, China
[4]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[5]College of Computer Science and Technology, Hengyang Normal University, Hunan, China
muyang.zhang@mais.ia.ac.cn, chenyuzhi2022@gmail.com, xurongtao2019@ia.ac.cn, changweiwang@sdas.org, yangjinming2023@ia.ac.cn, weiliang.meng@ia.ac.cn, jianwei.guo@bnu.edu.cn, happyday.huihuang@gmail.com,xiaopeng.zhang@ia.ac.cn

Abstract

As immersive experiences become increasingly popular, panoramic video has garnered significant attention in both research and applications. The high cost associated with capturing panoramic video underscores the need for efficient prompt-based generation methods. Although recent text-to-video (T2V) diffusion techniques have shown potential in standard video generation, they face challenges when applied to panoramic videos due to substantial differences in content and motion patterns. In this paper, we propose PanoDiT, a framework that utilizes the Diffusion Transformer (DiT) architecture to generate panoramic videos from text descriptions. Unlike traditional methods that rely on UNet-based denoising, our method leverages a transformer architecture for denoising, incorporating both temporal and global attention mechanisms. This ensures coherent frame generation and smooth motion transitions, offering distinct advantages in long-horizon generation tasks. To further enhance motion and consistency in the generated videos, we introduce DTM-LoRA and two panoramic-specific losses. Compared to previous methods, our PanoDiT achieves state-of-the-art performance across various evaluation metrics and user study, with code is available in the supplementary material.

Introduction

With the rapid advancement of VR/AR technology, 360-degree panoramic videos have become increasingly prevalent, which offers an immersive experience from every angle, making them valuable across various domains, including entertainment, education, and communication. Despite the emergence of many leading technologies in the field of computer vision (Xu et al. 2023b,a, 2022; Wang et al. 2023a,b; Wu et al. 2023). Traditionally, capturing these videos requires the use of high-resolution fisheye camera arrays to cover a wide field of view, a process that is both time-consuming and resource-intensive. There is an urgent need for more efficient panoramic video generation methods to reduce costs, yet research in this area remains limited.

Figure 1: Panoramic videos generated by **PanoDiT**. For example, the first video is generated with the prompt: "a panoramic view of ice and icebergs floating on water."

With the advent of emerging diffusion strategies (Ho, Jain, and Abbeel 2020), text-to-image (T2I) diffusion models (Rombach et al. 2022; Mou et al. 2024; Ramesh et al. 2022; Nichol et al. 2022; Peebles and Xie 2023) have shown remarkable capabilities in generating images from a wide range of user prompts, including text and images. These advancements in image generation have paved the way for text-to-video (T2V) generation. Recently, several T2V diffusion models (Guo et al. 2023; Deng et al. 2023; An et al. 2023; Singer et al. 2022; Wang et al. 2023d; Xing et al. 2024; Zhou et al. 2022; Yang et al. 2024) adopt a spatio-temporal separable architecture, which leverages pre-trained 2D diffusion models for spatial operations, thereby reducing the complexity of constructing a spatio-temporal model from the ground up. Among these, AnimateDiff (Guo et al. 2023) has made significant contributions by introducing a U-Net-based Motion module and a lightweight LoRA (Hu et al. 2021), enabling user-customized motion patterns tailored to specific tasks, which enhances the controllability

*These authors contributed equally.
†Weiliang Meng and Jianwei Guo are corresponding authors.

of T2V models and achieves notable consistency in both temporal and content aspects. Additionally, 360DVD (Wang et al. 2024) explored the application of similar architectures for 360-degree panoramic videos. However, it still suffers from significant shortcomings in high-frequency detail consistency and long-range generation. Traditional video generation approaches have predominantly relied on U-Net-based denoising schemes. However, due to the extended temporal sequences inherent in video data, the U-Net architecture often struggles with maintaining consistency over large temporal spans, making it difficult to preserve coherence in both time and content. In contrast, Diffusion Transformers (Peebles and Xie 2023) (DiT) achieve high consistency in long-horizon tasks. Implementing 360-degree panoramic video generation to meet industrial standards using DiT poses several challenges: (i) The DiT architecture, lacking the strong priors available in T2I models, requires a high-quality panorama dataset. (ii) No existing work has introduced a controllable Motion LoRA within the DiT framework to enable the generation of controllable content. (iii) Although the DiT architecture can directly generate equirectangular projection (ERP) videos and subsequently synthesize panoramic videos through spherical projection, significant frequency domain differences between ERP videos and standard formats necessitate targeted optimization.

In this paper, we propose PanoDiT, a DiT-based model for text-to-panoramic video generation. We curated the PHQ360 dataset from WEB360 to improve video quality. To control camera movements, we design DTM-LoRA, which fine-tunes the Motion Module in DiT blocks. To overcome standard diffusion loss limitations in panoramic settings, we added two loss functions that optimize spatial and frequency domain weights, improving video quality. As shown in Figure 1, PanoDiT excels in both visual and motion dynamics, achieving state-of-the-art performance in panoramic video generation. The main contributions are summarized as follows.

- We propose PanoDiT, which leverages the Diffusion Transformer to generate high-quality, temporally consistent panoramic videos from text prompts, capitalizing on the Transformer's strengths in continuous feature extraction.
- We construct a novel Panoramic High-Quality 360 (PHQ360) Dataset based on WEB360, which has been meticulously refined using aesthetic and motion scoring, along with Likert scale-based human evaluation.
- We design a Motion LoRA within a specialized DiT architecture to facilitate controlled video generation with minimal computational cost.
- We introduce two novel loss functions for 360-degree panoramic videos to improve dynamic realism and stability of low-frequency details.

Related Works

Diffusion Models

In recent years, diffusion models have made a significant impact in the field of image generation, achieving faster processing times and improved image quality and resolution (Ho, Jain, and Abbeel 2020; Rombach et al. 2022). However, in the realm of text-to-video generation, diffusion models have encountered various intricate challenges. Despite various recent efforts and optimizations (Blattmann et al. 2023; Guo et al. 2023; Xu et al. 2024; Chen et al. 2023; Yu et al. 2023), these issues remain difficult to fully resolve. A notable work in this domain is SVD (Blattmann et al. 2023), which demonstrates the necessity of carefully curated pre-training datasets for generating high-quality videos. AnimateDiff (Guo et al. 2023) introduces a Motion Module based on the Unet framework, along with a corresponding LoRA, significantly contributing to the low-cost and efficient control of motion patterns in video generation. While the success of text-to-image generation has influenced the field of panoramic generation, directly transferring the rich, high-quality priors from text-to-image models remains a significant challenge.

Panorama Generation

With the rise of VR/AR technologies, panoramic generation has become increasingly important, encompassing tasks like panoramic outpainting and text-conditioned generation. For example, PanDiff (Wang et al. 2023c) uses a two-stage angle prediction module and a latent diffusion-based model to generate 360-degree panoramas from unregistered NFoV images and text prompts. Similarly, Spherical-VQGAN (Ai et al. 2024) employs a VQGAN-based architecture to create a sphere-specific codebook from spherical harmonic values, improving the representation of spherical data and enhancing both semantic consistency and visual fidelity.

Generative models have advanced significantly in synthesizing immersive content from text, with text-to-panorama generation gaining attention. Methods such as those in (Zhang et al. 2023a; Lee et al. 2023) combine diffusion pathways for high-resolution images but struggle with equirectangular projection for 360-degree panoramas. MVDiffusion (Deng et al. 2023) uses correspondence-aware attention to generate multi-view images but often results in repetitive elements and inconsistencies. In contrast, 360DVD (Wang et al. 2024) enhances panoramic video fidelity and consistency by incorporating the 360Adapter and Latitude-aware Loss into the denoising Unet architecture from animateDiff (Guo et al. 2023). Despite these advancements, issues with fidelity and artifacts in longer video sequences still impact motion consistency and visual quality.

Validation with models like Sora (Liu et al. 2024) has shown that Diffusion Transformer (DiT) models can effectively compute correlations between frames along the temporal axis, providing a natural advantage over the U-Net architecture in generating long and high-quality videos. This finding has prompted us to explore the adaptation of the DiT architecture for panoramic video generation, which, to the best of our knowledge, is the first attempt within the research community.

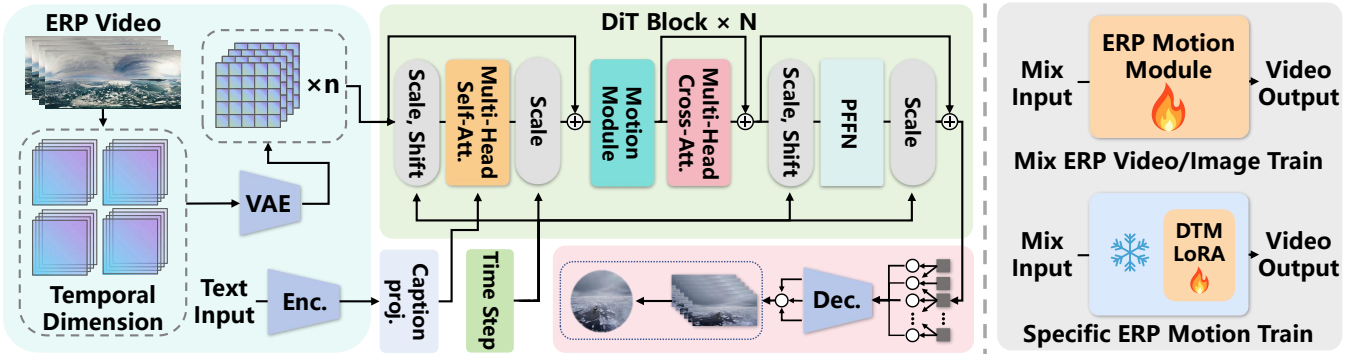

Figure 2: [Left]: Our PanoDiT Framework. Our PanoDiT generates high-fidelity ERP videos from simple text prompts and utilizes spherical projection to produce 360-degree panoramic video outputs. [Right]: Training procedure. The training begins with the high-resolution ERP image motion module, using a combination of images and panoramic videos. Following this, the DTM-LoRA is trained on panoramic videos with varying camera movements.

Method

Preliminaries

Diffusion Model. As the basis of our PanoDiT, Latent Diffusion Models (LDM) (Rombach et al. 2022) consists of three main components: a variational autoencoder (VAE) (Kingma and Welling 2013) with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, a denoising network ϵ_θ, and a condition encoder τ_θ.

In the widely used Stable Diffusion (SD) model, high-resolution images $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ are first projected into a lower-dimensional latent space $\mathbf{Z} = \mathcal{E}(\mathbf{x})$, where $\mathbf{Z} \in \mathbb{R}^{h\times w\times c}$. The down-sampling factor $f = H/h = W/w$ is typically set to 8. These latents are then decoded back into the image space using $\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{Z})$. The training objective for LDM is expressed as:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x,y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(Z_t, t, \tau_\theta(y))\|_2^2\right] \quad (1)$$

where t is uniformly sampled from 1 to T, and $\mathbf{Z}_t$ denotes the noisy latent at time step t. The denoising network ϵ_θ, which follows a UNet architecture (Dhariwal and Nichol 2021), is enhanced with cross-attention mechanisms to incorporate the optional condition encoder $\tau_\theta(\mathbf{y})$. The condition $\mathbf{y}$ can be a text prompt, an image, or any other user-specified input.

Diffusion Transformer. The Diffusion Transformer, introduced by (Peebles and Xie 2023), represents a key advancement by merging diffusion models with transformer architectures (Vaswani et al. 2017). This integration overcomes the limitations of U-Net-based latent diffusion models, improving efficiency, flexibility, and scalability. DiT surpasses U-Net in denoising tasks by capturing long-range dependencies and complex interactions, essential for accurate noise reduction. The self-attention mechanism enables DiT to dynamically focus on different input parts, enhancing context-aware denoising. These improvements result in higher quality video generation, preserving temporal consistency and fine details across frames.

Panoramic videos. Panoramic video typically refers to 360-degree spherical panoramas. The consensus in the research community is that directly generating spherical videos is highly challenging. An alternative approach is to generate equirectangular projection (ERP) videos with a 2:1 aspect ratio, which are then synthesized into 360-degree videos through spherical projection. When projecting panoramic videos onto ERP, meridians are mapped as vertical lines with constant spacing, while parallels are mapped as horizontal lines with constant spacing. The ERP video can be divided into three sections based on latitude: the central region represents low latitudes (equatorial area) with low-motion backgrounds, while the upper and lower regions correspond to high latitudes (polar areas), which contain significant motion and distortions.

Overview

Our PanoDiT framework is designed to generate high-fidelity, 360-degree panoramic videos from text prompts. It leverages the DiT (Peebles and Xie 2023) backbone and is trained on our PHQ360 dataset, as illustrated in Figure 2. To improve upon the traditional DiT denoising architecture, we introduce a novel Temporal VAE structure. This approach segments the input ERP video into time-based batches for encoding, which are then denoised and fused before being fed into the VAE decoder to produce the final output.

To optimize PanoDiT's performance, including the motion module, we employed mixed training with both image and video data, mitigating the risk of introducing artificial artifacts from video-only training. The framework also includes a high-resolution ERP motion module that utilizes a Temporal Transformer. This module projects the input data with positional encoding to capture sequence order, and processes it through self-attention layers to learn dependencies across time steps, thereby providing explicit motion guidance for the DiT model.

Furthermore, PanoDiT integrates DTM-LoRA to fine-tune the ERP motion module, allowing for dynamic control over video motion patterns based on specific ERP input sequences. During the denoising and generation of ERP videos, we apply a post-processing technique known as rotation padding (Wang et al. 2023c, 2024), which en-

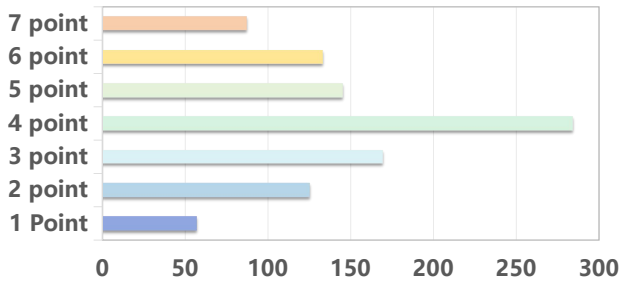

Figure 3: Human Filtration. We extract those Likert scores greater than or equal to 4 to be used in the construction of the PHQ360 dataset.

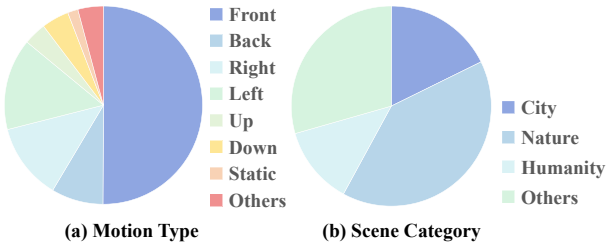

Figure 4: The statistical details of our PHQ360 dataset. (a) shows the distribution of camera motion directions in our PHQ360 dataset. (b) presents the types and distribution of video subjects within the dataset.

sures physical and pixel continuity when stitching meridians into a 360-degree spherical video. PanoDiT offers an efficient framework for generating immersive and controllable panoramic videos from text prompts.

PHQ360 Datasets

A diverse collection of text-video paired datasets is essential for training open-domain text-to-video generation models. However, the lack of high-quality textual annotations in existing panoramic video datasets limits their effectiveness in generating high-quality videos.

In previous work introducing text-to-panoramic video datasets, datasets like ODV360 (Cao et al. 2023) and WEB360 (Wang et al. 2024) were developed. WEB360 that uses annotations based on BLIP (Li et al. 2022) is built upon the ODV360 dataset and contains 2,114 pairs of text and video from open-domain content in high-definition ERP format. However, the text fusion method in the WEB360 dataset only considers multi-view fusion of single frames and lacks comprehensiveness and detail for capturing motion changes in panoramas. To address these shortcomings, we utilize the VILA (Lin et al. 2024) model, which integrates all frames over the temporal axis for improved video understanding, combining the strengths of fine-grained and dynamic annotations.

To enhance the quality of panoramic video datasets, we employ four filtering methods: Duration, Aesthetic, Text, and Motion, to eliminate nearly 1,000 low-quality videos from the WEB360 dataset. Despite this, some videos still had unclear motion or layouts. Therefore, we conducted a survey with 30 participants, each rating about 100 videos on a 7-point Likert scale. Each video's score reflects the average of three evaluations. The final dataset is our Panorama High-Quality (PHQ360) which consists of 649 text-video pairs as shown in Fig. 3.

We also analyzed camera displacements and content themes in our PHQ360 dataset, which aids our PanoDiT in adjusting video motion and themes using LoRA control methods. The statistical details are given in Fig. 4, which are expected to benefit future research.

DTM-LoRA

Guiding fixed video motion through text prompts alone is widely recognized as a highly challenging task. In the community around models like Stable Diffusion, users prefer generative methods that provide reliable control. Thus, developing effective and efficient motion control mechanisms for video generation models has become a key focus. Previous works like AnimateDiff (Guo et al. 2023) and SVD (Blattmann et al. 2023) have integrated motion modules into video generation models for controllable motion. However, these modules are parameter-heavy and require extensive data for fine-tuning. Other studies have explored LoRA-based methods for fine-tuning motion modules (Guo et al. 2023). Therefore, we propose DTM-LoRA, which integrates motion LoRA into the DiT framework by fine-tuning the embedded motion module within transformer blocks using low-rank adaptation.

Generating 360-degree panoramic videos from text presents unique challenges compared to traditional text-to-video (T2V) tasks. These challenges arise mainly from the distinct camera movement patterns and the high-frequency variations in high-latitude regions inherent in the equirectangular projection (ERP) format. Typically, panoramic videos are captured in real-time using specialized 360-degree cameras mounted on moving platforms, such as autonomous vehicles or drones, introducing varied motion characteristics. However, by training DTM-LoRA on a minimal set of samples (approximately 20-30 videos) with specific camera motion directions, the motion direction in ERP videos can be effectively controlled through weight scaling.

Panorama Generation Optimization

360-degree panoramic videos present significant challenges for text-to-video (T2V) generation due to their extensive scale and wide range of viewpoints compared to conventional videos. In ERP panoramic videos derived from 360-degree footage, the low-latitude regions remain stable with minimal variation, while the polar regions, despite their smaller area, exhibit greater randomness, complicating the denoising process. In the frequency domain, polar regions typically correspond to high-frequency areas, whereas low-latitude regions correspond to low-frequency areas. These differences can be optimized by applying filters specifically designed to enhance performance in each region. Traditional

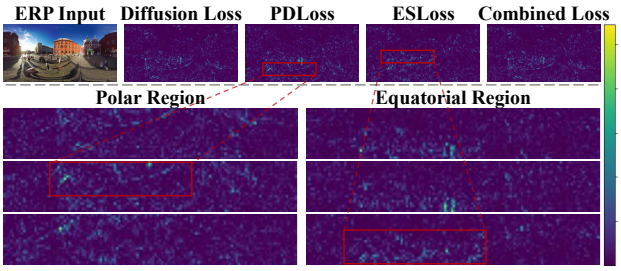

Figure 5: Visualization of our loss design. We introduce a diffusion loss, PD Loss, and ES Loss corresponding to the second frame of the input video. It is noteworthy that PD Loss and ES Loss exhibit more pronounced distribution in the polar and equatorial regions, respectively. Our final loss function is a weighted combination of these three losses.

approaches use diffusion loss to uniformly supervise all outputs but fail to address these challenges. To improve this, we propose two novel loss functions.

Polar Dynamic Loss. To enhance motion consistency in the high-dynamic polar regions of ERP video generation, we introduce the Polar Dynamic Loss (PDLoss). ERP videos captured by panoramic cameras often exhibit varying motion patterns between adjacent frames in the polar regions. To better capture dynamics in these critical areas, we use a polar dynamic-aware weight $\mathbf{w} = \{w_2, w_3, \ldots, w_k\} \in \mathbb{R}^{K-1\times C\times H\times W}$ that emphasizes discrepancies between predicted and actual values in these regions:

$$w_i = \|\Delta\hat{z}_i - \Delta z_i\|^2, \tag{2}$$

where $\Delta\hat{z}_i = \hat{z}_i - \hat{z}_{i-1}$ and $\Delta z_i = z_i - z_{i-1}$ denote the frame-wise temporal differences of the predicted and true values, respectively.

We normalize $\mathbf{w}_i$ within each ERP video segment to prevent issues like gradient explosion or vanishing during the training, thereby improving the model's learning effectiveness and stability. This normalization amplifies significant motion differences, focusing on regions with minimal dynamic activity while excluding the more stable middle and low-latitude areas. To account for the perspective shifts caused by camera displacement in ERP videos, which introduce a degree of causality, we penalize the subsequent frame of each video pair using a new polar dynamic loss:

$$\mathcal{L}_{\text{polar}} = \frac{1}{N}\sum_{i=2}^{K}\frac{\|w\odot(\hat{z}_i - z_i)\|^2}{1+\sqrt{w_i}}, \tag{3}$$

The term w denotes the dynamic weights assigned to each feature location.

Equatorial Stability Loss. Another challenge in ERP video generation is maintaining high consistency in both semantic and temporal aspects of object structures within low-latitude regions. As 360-degree videos are largely composed of low-latitude areas from a human perspective, consistency in these regions directly impacts the video's fidelity. Previous research has highlighted the trade-off between perceptual quality and motion intensity in general video generation (Bar-Tal et al. 2024; Zhang et al. 2023b; Kondratyuk et al. 2023; Bar-Tal et al. 2024). Specifically, in panoramic video generation, it has been observed that high-resolution 512×1024 ERP videos often suffer from significant degradation in structural details, particularly affecting smooth objects or sharp edges. This phenomenon occurs because the edges and textures in the low-latitude regions of ERP frames are concentrated in the high-frequency components of the frequency domain. To address this issue, we identify these components in the frequency domain, as follows:

$$\mathcal{F}(z_i) = \mathcal{F}^{-1}\left(\mathcal{H}\cdot\mathcal{F}(z_i)\right), \tag{4}$$

where $\mathcal{F}(z_i)$ and $\mathcal{F}^{-1}(\cdot)$ denote the frequency domain transformation and its inverse, respectively, and $\mathcal{H}$ represents a high-pass filter defined as:

$$\mathcal{H}_{h,w} = \begin{cases} 1, & \text{if } \sqrt{\left(\frac{2h}{H}-1\right)^2 + \left(\frac{2w}{W}-1\right)^2} \geq d_s, \\ 0, & \text{otherwise.} \end{cases}$$

Using formula 4, we can extract latent features that meet frequency domain requirements. Additionally, utilizing these extracted high-frequency features, we have developed a novel equatorial stability loss, defined as follows:

$$\mathcal{L}_{\text{equatorial}} = \frac{1}{N}\sum_{i=1}^{K}\|\mathcal{F}(\hat{z}_i) - \mathcal{F}(z_i)\|^2, \tag{5}$$

where $\mathcal{F}(\cdot)$ represents the high-frequency component extracted using the high-pass filter $\mathcal{H}$.

The loss function minimizes the difference between the predicted and true high-frequency features in the ERP video, thereby ensuring the stability of generated object edges and textures.

Our final training objective is a weighted sum of Equation 1, Equation 3, and Equation 5, with λ_1 and λ_2 serving as trade-off weights to balance the optimization.

$$\mathcal{L}_{\text{panodit}} = \mathcal{L}_{\text{diffusion}} + \lambda_1\mathcal{L}_{\text{polar}} + \lambda_2\mathcal{L}_{\text{equatorial}}. \tag{6}$$

Rotation Padding. Inspired by previous work (Wang et al. 2024, 2023c), we can maintain semantic and pixel consistency at the ends of ERP videos using a simple rotation padding method, which involves rotating the flattened last dimension of the tensor to the opposite end during each denoising step.

Experiments

Settings

The training configuration featured a resolution of 512×1024, a frame length of 144, a batch size of 2, a learning rate of 5×10^{-6}, and a total of 100,000 training steps. We trained PanoDiT at three different scales: Small (S), Base (B), and Large (L) using our PHQ360 dataset. In the subsequent sections, we focus on presenting the inference results of PanoDiT-B, as it demonstrates optimal performance at a resolution of 512×1024.

Figure 6: Main Results. Our PanoDiT produces panoramic images with higher fidelity and stable high-frequency detail than previous leading work.

Methods	WEB360						Our PHQ360					
	FID(↓)	IS(↑)	FVD(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)	FID(↓)	IS(↑)	FVD(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
AnimateDiff	273.14	5.19	2586.51	6.11	0.21	0.92	207.77	6.53	2319.45	6.34	0.24	0.89
360DVD	131.16	8.71	1895.29	8.13	0.34	0.81	155.21	9.65	1305.41	8.47	0.41	0.73
SVD	102.45	9.32	1290.17	9.11	0.44	0.72	98.67	10.24	1032.56	9.37	0.48	0.69
PanoDiT(ours)	**81.42**	**10.62**	**761.72**	**10.03**	**0.52**	**0.64**	**75.73**	**12.31**	**584.36**	**10.38**	**0.59**	**0.61**

Table 1: Quantitative results conducted on the WEB360 and our PHQ360 datasets for the evaluation of single-frame and video generation. The evaluation metrics for the baseline models are derived from the official implementations, and the re-trained results strictly adhered to the original training settings.

Comparison

We conducted comparative experiments using AnimateDiff, 360DVD, and SVD, all of which were trained under identical conditions on WEB360 and our PHQ360 dataset to ensure a fair comparison.

Qualitative Results. As illustrated in Figure 6, PanoDiT results exhibit pronounced panoramic dynamics. In the rock region on the left and the town region on the right, the ERP video demonstrates the characteristic distortions of polar areas while maintaining excellent semantic consistency. Furthermore, unlike the previous models, which are constrained to generating 16-frame videos, PanoDiT is capable of producing 144 frames at a rate of 24 frames per second. The qualitative results demonstrate that our PanoDiT model significantly surpasses previous models in terms of visual fidelity and dynamic performance.

Quantitative Results. The quantitative results are given in Table 1, We report not only standard metrics for video evaluation, such as Fréchet Video Distance (Unterthiner et al. 2018) (FVD), but also Fréchet Inception Distance (Heusel et al. 2017) (FID) and Inception Score (IS) for individual frames of ERP videos. We improved image quality by at least 2 percentage points and nearly halved the previous values for video metrics.

Methods	FID(↓)	IS(↑)	FVD(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
w/o ESLoss	80.61	10.82	641.2	10.25	0.58	0.62
w/ ESLoss	**75.51**	**12.44**	**517.9**	**11.07**	**0.62**	**0.59**

Table 2: Quantitative Ablation of the ESLoss. ESLoss demonstrates its effectiveness in enhancing panoramic video quality on qualitative metrics.

Ablation Study

DTM-LoRA. To validate the effectiveness of DTM-LoRA in the diffusion transformer architecture, we conduct ablation experiments across three PanoDiT model scales. Optical flow and motion magnitude between consecutive frames are calculated using the Farneback algorithm (Farnebäck 2003), forming the motion score. The final score is the average motion magnitude across all frames (averaged over three inferences). The line graphs in Figure 7 illustrate the relationship between the DTM-LoRA-applied weights and the motion score, which demonstrate DTM-LoRA's significant capability in controlling the generation of high-motion-coefficient videos across the three PanoDiT model scales.

PDLoss. To evaluate the effectiveness of our PDLoss in

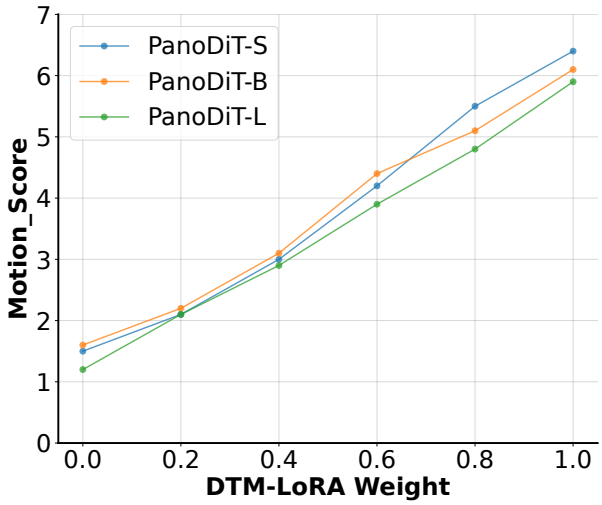

Figure 7: Quantitative Ablation of DTM-LoRA. We report the motion scores for DTM-LoRA inference across all three PanoDiT scales. Higher LoRA weights yield panoramic videos with increased motion scores a nd optical flow.

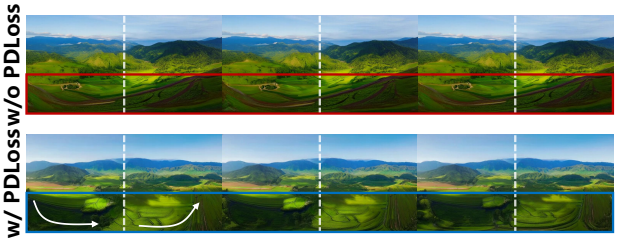

Figure 8: Qualitative Ablation of the PDLoss. The white dashed line indicates the motion reference, while the white arrow denotes the motion direction aligned with panoramic logic.

improving the rationality of motion in high-frequency regions, we conduct qualitative ablation experiments shown in Figure 8 , where ERP video frames in the high-latitude (Antarctic) regions displayed less coherent motion directions before applying PDLoss. In contrast, frames with PDLoss applied exhibit motion directions that align more consistently with the camera's movement.

ESLoss. To validate the effectiveness of our ESLoss in preserving semantic and temporal consistency in low-latitude areas, we conduct a quantitative ablation experiment shown in Table 2 and Figure 9, where the videos employing ESLoss outperform the baseline in all metrics.

User Study

The research community consensus is that evaluating video quality using metrics such as FVD provides a limited perspective. To thoroughly assess the quality of 360-degree videos generated by PanoDiT, we conducted a user study with 20 participants. Inspired by a prior study (Ai et al. 2024), participants performed both objective and subjective evaluations using a Meta Quest 2 VR headset and Oculus

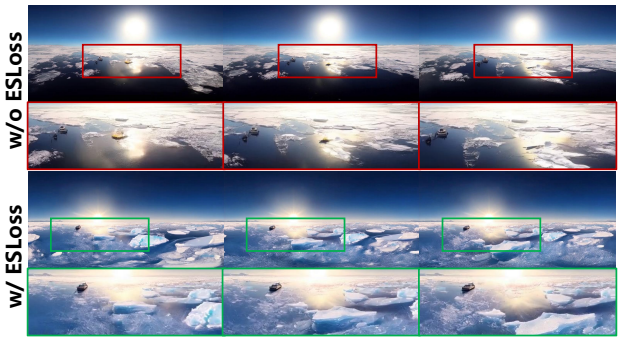

Figure 9: Qualitative Ablation of the ESLoss. Note that the ships in the frames marked with the red boxes exhibit distortions which did not use ESLoss, whereas those marked with the green boxes maintain good semantic and temporal consistency.

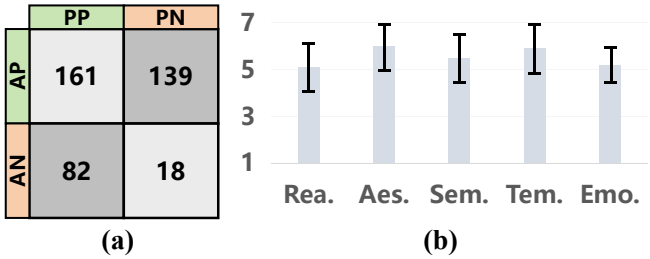

Figure 10: User Study Statistics. (a) Confusion matrix from the user study, where fewer Predicted Negatives indicate higher fidelity. (b) User study results based on subjective evaluation metrics.

CenarioVR software.

Find Fake. Each participant completed 10 test sets, each containing 3 real samples and 1 fake sample. They were asked to identify the sample they believed to be fake. The fake samples, generated by our PanoDiT-B, were randomly selected, while the real videos came from a private dataset. Based on the results, we computed the confusion matrix, as illustrated in Figure 10 (a).

Viewing Rating. Participants evaluated the generated videos on realism, aesthetics, motion, semantics, and temporal consistency during a virtual tour, with results shown in Figure 10 (b).

Conclusion

We propose PanoDiT, a panoramic video generation model with high-fidelity output. PanoDiT employs the exceptional scaling properties of transformer models, leading to significant advancements, particularly in long-horizon panoramic video generation. The DTM-LoRA component in PanoDiT excels in panoramic motion control, with customized loss functions for panoramic videos validated through extensive ablation experiments. Experiments validate that our PanoDiT outperforms previous U-Net-based models in both visual quality and quantitative metrics, achieving SOTA performance in panoramic video generation.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. U21A20515, 62376271, U22B2034, 62262043, 12442056, 62365014, and 62172416), and in part by Beijing Natural Science Foundation (Nos. L231013, L241056). We also thank Miao Yang from Beijing Forestry University for generating visual effects using VR equipment.

References

Ai, H.; Cao, Z.; Lu, H.; Chen, C.; Ma, J.; Zhou, P.; Kim, T.-K.; Hui, P.; and Wang, L. 2024. Dream360: Diverse and Immersive Outdoor Virtual Scene Creation via Transformer-Based 360° Image Outpainting. *IEEE Transactions on Visualization and Computer Graphics*.

An, J.; Zhang, S.; Yang, H.; Gupta, S.; Huang, J.-B.; Luo, J.; and Yin, X. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Li, Y.; Michaeli, T.; et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Cao, M.; Mou, C.; Yu, F.; Wang, X.; Zheng, Y.; Zhang, J.; Dong, C.; Li, G.; Shan, Y.; Timofte, R.; et al. 2023. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1731–1745.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Deng, Z.; He, X.; Peng, Y.; Zhu, X.; and Cheng, L. 2023. MV-Diffusion: Motion-aware video diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7255–7263.

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

Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, 363–370. Springer.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *arXiv e-prints*, arXiv–2307.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6629–6640.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Hornung, R.; Adam, H.; Akbari, H.; Alon, Y.; Birodkar, V.; et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36: 50648–50660.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4296–4304.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, C.; Xu, R.; Lu, K.; Xu, S.; Meng, W.; Zhang, Y.; Fan, B.; and Zhang, X. 2023a. Attention weighted local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10632–10649.

Wang, C.; Xu, R.; Xu, S.; Meng, W.; Xiao, J.; and Zhang, X. 2023b. Accurate lung nodule segmentation with detailed representation transfer and soft mask supervision. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, J.; Chen, Z.; Ling, J.; Xie, R.; and Song, L. 2023c. 360-Degree Panorama Generation from Few Unregistered NFoV Images. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6811–6821.

Wang, Q.; Li, W.; Mou, C.; Cheng, X.; and Zhang, J. 2024. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6913–6923.

Wang, W.; Yang, H.; Tuo, Z.; He, H.; Zhu, J.; Fu, J.; and Liu, J. 2023d. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation.

Wu, J.; Xu, R.; Wood-Doughty, Z.; and Wang, C. 2023. Segment anything model is a good teacher for local feature learning. *arXiv preprint arXiv:2309.16992*.

Xing, J.; Xia, M.; Liu, Y.; Zhang, Y.; Zhang, Y.; He, Y.; Liu, H.; Chen, H.; Cun, X.; Wang, X.; et al. 2024. Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance. *IEEE Transactions on Visualization & Computer Graphics*, 1–15.

Xu, J.; Zou, X.; Huang, K.; Chen, Y.; Liu, B.; Cheng, M.; Shi, X.; and Huang, J. 2024. EasyAnimate: A High-Performance Long Video Generation Method based on Transformer Architecture. *arXiv preprint arXiv:2405.18991*.

Xu, R.; Li, Y.; Wang, C.; Xu, S.; Meng, W.; and Zhang, X. 2022. Instance segmentation of biological images using graph convolutional network. *Engineering Applications of Artificial Intelligence*, 110: 104739.

Xu, R.; Wang, C.; Sun, J.; Xu, S.; Meng, W.; and Zhang, X. 2023a. Self Correspondence Distillation For End-to-End Weakly-Supervised Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, R.; Wang, C.; Xu, S.; Meng, W.; and Zhang, X. 2023b. Wave-Like Class Activation Map With Representation Fusion for Weakly-Supervised Semantic Segmentation. *IEEE Transactions on Multimedia*.

Yang, M.; Guo, J.; Chen, Y.; Chen, L.; Li, P.; Cheng, Z.; Zhang, X.; and Huang, H. 2024. InstanceTex: Instance-level Controllable Texture Synthesis for 3D Scenes via Diffusion Priors. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Zhang, Q.; Song, J.; Huang, X.; Chen, Y.; and Liu, M.-Y. 2023a. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10188–10198. IEEE.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023b. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.