PanoWan: Lifting Diffusion Video Generation Models to 360° with Latitude/Longitude-aware Mechanisms

Yifei Xia[1,2,3*] **Shuchen Weng**[4*] **Siqi Yang**[5] **Jingqi Liu**[1,2] **Chengxuan Zhu**[6]
Minggui Teng[1,2] **Zijian Jia**[7] **Han Jiang**[3] **Boxin Shi**[1,2†]
[1]State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[2]Nat'l Eng. Research Ctr. of Visual Tech., School of Computer Science, Peking University
[3]OpenBayes Information Technology Co., Ltd. [4]Beijing Academy of Artificial Intelligence
[5]Institute for Artificial Intelligence, Peking University
[6]Nat'l Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[7]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{yfxia,shuchenweng,yousiki,peterzhu,minggui_teng,shiboxin}@pku.edu.cn
liujingqi@stu.pku.edu.cn jiazijian@bupt.edu.cn hahn@openbayes.com

Abstract

Panoramic video generation enables immersive 360° content creation, valuable in applications that demand scene-consistent world exploration. However, existing panoramic video generation models struggle to leverage pre-trained generative priors from conventional text-to-video models for high-quality and diverse panoramic videos generation, due to limited dataset scale and the gap in spatial feature representations. In this paper, we introduce PanoWan to effectively lift pre-trained text-to-video models to the panoramic domain, equipped with minimal modules. PanoWan employs latitude-aware sampling to avoid latitudinal distortion, while its rotated semantic denoising and padded pixel-wise decoding ensure seamless transitions at longitude boundaries. To provide sufficient panoramic videos for learning these lifted representations, we contribute PANOVID, a high-quality panoramic video dataset with captions and diverse scenarios. Consequently, PanoWan achieves state-of-the-art performance in panoramic video generation and demonstrates robustness for zero-shot downstream tasks. Our project page is available at https://panowan.variantconst.com.

1 Introduction

Text-based panoramic video generation aims to produce a complete 360° view, ensuring coherent spatial and visual relationships between elements within the scene. Such inherent property is highly valuable for conventional VR content, the construction of interactive game worlds [35, 8], and the simulation of environments for embodied AI [18].

The remarkable capabilities of conventional text-to-video models [28, 33, 4] motivate researchers to leverage their generative priors to panoramic video generation. One intuitive strategy generates local perspective conventional videos and integrates them during inference. While these training-free methods [12, 21] entirely preserve generative priors, they sacrifice overall consistency as they struggle to establish cross-view long-range dependencies. Alternatively, fine-tuning conventional text-to-video models [32, 41] also faces challenges. On one hand, existing panoramic video datasets are limited in scope and scale compared to conventional ones. On the other hand, the gap in spatial

*Equal contribution.
†Corresponding author.

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

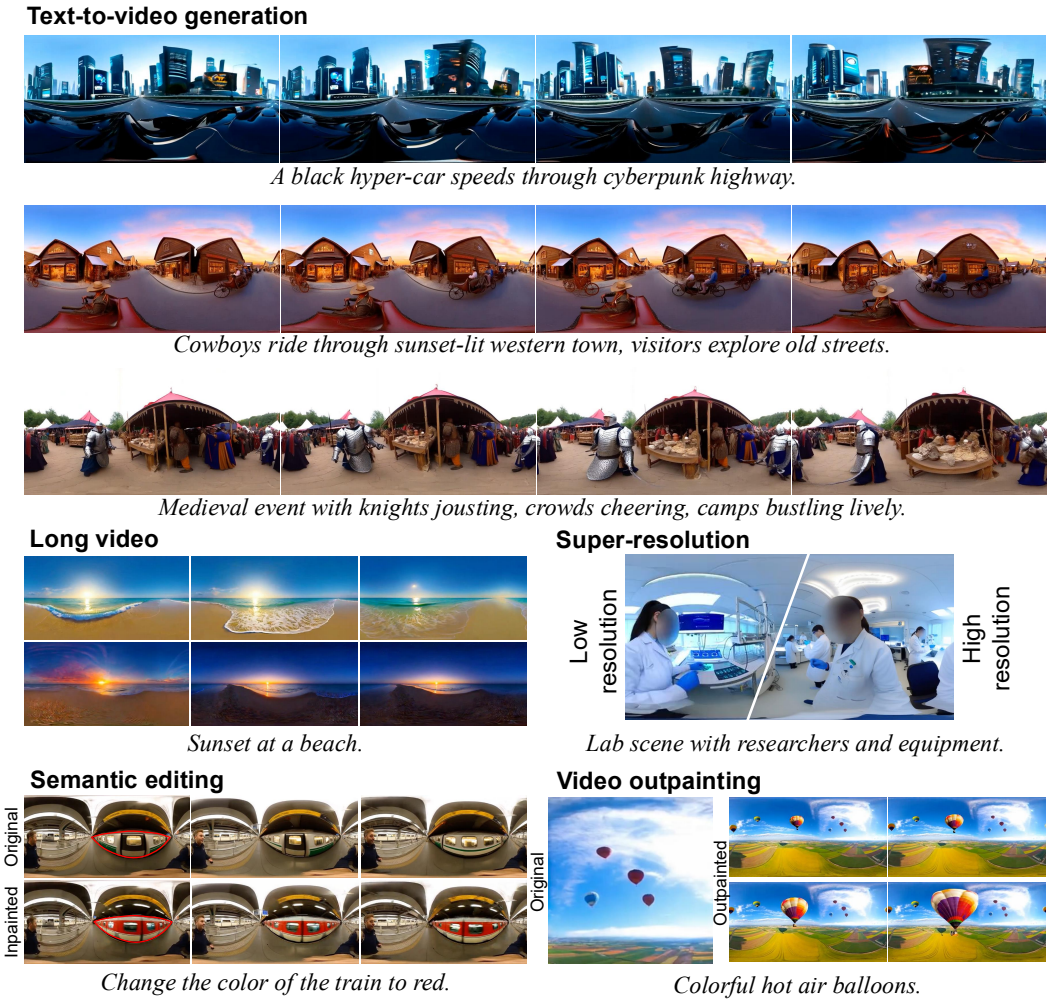


Figure 1: PanoWan is a text-based panoramic video generation framework. It lifts pre-trained generative priors from a conventional text-to-video model to the panorama, and enables generating diverse scenarios for long videos. Equipped with training-free techniques, PanoWan supports zero-shot editing of panoramic videos, including super-resolution, semantic editing, and video outpainting.

representations (*e.g.*, latitudinal distortions and seam longitudes) between panoramic and conventional videos potentially hinder the effective prior leverage from pre-trained conventional models.

In this paper, we pursue a training-based approach to overcome current bottlenecks. We propose **PanoWan**, a framework for **Pano**ramic video generation based on **Wan** 2.1 [33]. Equipped with minimal modules, PanoWan effectively lifts generative priors from a pre-trained conventional text-to-video model to the panorama. To bridge the gap in spatial feature representations between panoramic and conventional videos, we design latitude-aware sampling to address latitudinal distortions caused by equirectangular projection. Since text-to-video models lack continuity awareness for left and right boundaries, we achieve seamless longitude transitions using the rotated semantic denoising to address semantic inconsistency and the padded pixel-wise decoding to resolve pixel-wise disharmony.

As learning to lift spatial representations from conventional videos to the panorama requires large-scale data, we further introduce PANOVID. This **Pano**ramic **Vid**eo dataset offers diverse scenarios (*e.g.*, landscape, streetscape, and humanscape), includes over 13K captioned video clips totaling 944 hours, and features data processing tailored for panoramic video generation. As shown in Fig. 1, our framework produces panoramic videos from text descriptions, enabling long video generation. Additionally, it enables training-free editing of user-provided panoramic videos, including super-resolution, semantic inpainting, and video outpainting. Extensive experiments demonstrate that

PanoWan achieves state-of-the-art panoramic video generation performance across seven metrics, alongside robust zero-shot capabilities for various downstream tasks.

Our contributions can be summarized as follows:

- We contribute PANOVID, a large-scale and high-quality panoramic video dataset with captioned video clips, tailored for text-based panoramic video generation.
- We propose PanoWan, lifting generative priors from a pre-trained model to show state-of-the-art performance on panoramic video generation and robustness for downstream tasks.
- We integrate the latitude-aware sampling, rotated semantic denoising, and padded pixel-wise decoding to bridge the spatial difference between panoramic and conventional videos.

2 Related Works

2.1 Video Diffusion Models

Diffusion models have demonstrated impressive results in image generation, but extending them to videos introduces additional challenges (*e.g.*, temporal consistency and computational efficiency). Early models adapt image diffusion models with cascaded architecture [10], temporal layers [25], and attention mechanisms [3]. To further improve efficiency, LVDM [9] introduces a lightweight latent video diffusion model with a 3D latent space and hierarchical structure for long video generation. Meanwhile, SVD [4] improves video diffusion using large and high-quality datasets. Recent Diffusion Transformers (DiTs) [22] effectively model complex spatio-temporal dynamics for video generation, operating on latent space compressed by 3D VAE [39]. After that, more large-scale models (*e.g.*, HunyuanVideo [14], CogVideoX [38], and Seaweed [24]) emerge and demonstrate the benefits of scaling both model and data size. This motivates us to adopt Wan 2.1 [33] as the backbone to leverage its strong generative priors and temporal modeling capabilities for panoramic video generation.

2.2 Panoramic Video Generation

Text conditioned generation. 360DVD [32] is the pioneer to introduce stable video generation techniques to panoramic video generation by proposing a 360-Adapter and a set of 360 enhancement techniques. Training-free methods (*e.g.*, DynamicScaler [12] and SphereDiff [21]) create panoramic videos by generating local patches and then composing them together into a complete panorama, which inherently break global consistency. With the development of video diffusion models, VideoPanda [36] augments diffusion models with multi-view attention. PanoDiT [41] uses a DiT backbone with global-temporal attention and panoramic-specific losses for coherent long-range generation. Despite these advancements, existing methods still suffer from observable latitude distortions and issues with seam-free longitude transitions. In this work, we introduce PanoWan, a framework that addresses these challenges by lifting generative priors from pre-trained text-to-video models to panorama.

Image or video conditioned generation. Imagine360 [26] adopts antipodal-aware motion modeling to convert perspective videos to panoramic views. Building on static panoramas, 4K4DGen [16] lifts them into dynamic 4D scenes via spatial-temporal denoising. HoloTime [42] further leverages Gaussian splatting and a two-stage diffusion process for high-fidelity 4D reconstruction. VidPanos [20] treats panorama generation as a space-time outpainting task from panning video inputs, while Argus [19] integrates motion and geometry cues for enhanced video-to-360° synthesis. These explorations highlight a trend toward unifying spatial, temporal, and geometric reasoning. We demonstrate that PanoWan possesses these capabilities, with robust zero-shot capabilities for downstream tasks.

3 PANOVID Dataset

The absence of paired datasets has long been regarded as one of the primary barriers to advancing the performance of panoramic video generation models [32]. Existing text to panoramic video generation methods [32, 36, 41] mainly rely on WEB360 dataset [32], which contains only 2114 video clips of 10 seconds each. Although Argus [19] filters out over 283K video clips from the 360-1M dataset [29], it is not built for the text-based panoramic video generation task, providing no paired captions, and showing significant distribution bias for the scenario semantics.

To address these limitations, we present PANOVID, a large-scale and high-quality dataset with diverse scenarios and balanced semantics, tailored for text-based panoramic video generation. Our data collection process begins by aggregating videos from existing panoramic sources, including 360-1M [29], 360+x [5], Imagine360 [26], WEB360 [32], Panonut360 [37], the Miraikan 360-degree Video Dataset [1], and a public dataset of immersive VR videos [15]. These sources cover both large-scale web collections (with rich but noisy YouTube content) and more curated institutional datasets (with higher quality but limited size).

Vision-language-based filtering pipeline. To transform these heterogeneous sources into a high-quality paired dataset, we design a scalable five-stage filtering pipeline guided by a vision-language model. *(i) Initial filtering by popularity:* We first filter large-scale collections such as 360-1M [29], retaining only videos with at least 1000 views to discard low-quality or trivial content. *(ii) Shot segmentation:* Each raw video is segmented into 10-second continuous clips using PySceneDetect, ensuring scene consistency and temporal coherence. *(iii) Vision-language annotation:* We employ Qwen-2.5-VL [2] to process each clip, generating a descriptive caption and predicting the associated POI (Point-of-Interest) category in a structured JSON format. Notably, we retain only clips that the model identifies as true ERP (equirectangular projection) panoramic videos. *(iv) Motion score filtering:* Following [7], we compute a normalized optical-flow magnitude and remove clips with a motion score below 0.4 to exclude static or nearly still scenes. *(v) Aesthetic score filtering:* Finally, we utilize Q-Align [34] to evaluate the aesthetic quality of each frame, discarding clips with a mean aesthetic score below 3.0. This ensures that the dataset maintains a consistent level of visual quality. This hierarchical filtering process effectively removes noise while preserving diversity and realism, yielding panoramic video clips that are semantically meaningful, visually appealing, and motion-rich. It also enables large-scale automatic caption generation without manual annotation cost.

Semantic balancing. We observe that the automatically annotated POI categories exhibit strong long-tail distribution: natural scenes such as *Mountains* and *Parks* dominate, while indoor and human-centric environments (*e.g.*, *Libraries*, *Shops*, *Theaters*) are underrepresented. To alleviate this imbalance and enhance the generalization of models trained on PANOVID, we cap the maximum number of clips per POI category to 200, selecting the highest-ranking clips based on a combination of aesthetic and motion scores. All clips from underrepresented categories are fully retained. This strategy prevents overfitting to dominant scene types and ensures a diverse representation across environments.

Dataset statistics and characteristics. After the filtering, balancing, and a final deduplication step based on caption similarity, PANOVID comprises over 13K high-quality panoramic video clips, totaling approximately 944 hours of content. Compared with prior datasets, PANOVID offers not only substantially larger scale but also structured text–video pairs with fine-grained POI categories and balanced semantic coverage. It thereby provides a robust foundation for training and evaluating text-conditioned panoramic video generation models.

4 Method

Panoramic videos have a different spatial feature representation compared to conventional ones. Inspired by GEN3C [23], we effectively preserve the generative prior of pre-trained models by equipping minimal modules and fine-tuning a small subset of parameters via LoRA [11]. We firstly introduce our video diffusion backbone and formulate the spherical coordinate mapping (Sec. 4.1). Next, we propose the latitude-aware sampling to avoid latitude distortion, along with its corresponding analysis (Sec. 4.2). Finally, we present the rotated semantic denoising and the padded pixel-wise decoding to achieve the seamless longitude transitions (Sec. 4.3).

4.1 Preliminaries

Video diffusion models. We employ Wan 2.1 [33] as the video generation backbone, with spatial-temporal Variational AutoEncoders (VAEs) to map high-dimensional videos into compact latent codes. The flow matching framework [17] is used to model a unified denoising diffusion process. Specifically, given a clear video x, a VAE encoder $\mathrm{E}(\cdot)$ first projects the video into the latent space $z_1 = \mathrm{E}(x)$. During training, a noise $z_0 \sim \mathcal{N}(0, I)$ is sampled, and an intermediate latent code $z_t = tz_1 + (1-t)z_0$ is constructed by linearly interpolating between z_1 and z_0 at timestep $t \in [0, 1]$.

The training goal is to predict the ground truth velocity $v_t = \mathrm{d}z_t/\mathrm{d}t = z_1 - z_0$, and the loss function is formulated as:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, c_{\text{txt}}, t}||u(z_t, c_{\text{txt}}, t; \theta) - v_t||^2, \tag{1}$$

where c_{txt} is the text embedding, θ is the parameters of the prediction model, and $u(z_t, c_{\text{txt}}, t; \theta)$ is the predicted velocity of the model.

Spherical coordinate mapping. Panorama captures a 360° view, inherently representing signals in spherical coordinates (φ, θ). To leverage generative priors from conventional images and videos that operate in Cartesian coordinates (x, y), we employ the equirectangular projection (ERP) $\mathcal{P}_{\text{ERP}}$ to map between these coordinate systems for panoramic videos:

$$\mathcal{P}_{\text{ERP}} : [2R] \times [R] \to [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}], \quad (x, y) \mapsto (\varphi, \theta) = \Big(\frac{2x+1}{2R}\pi, \frac{2y+1-R}{2R}\pi\Big), \tag{2}$$

where R is the radius of the sphere. φ and θ are longitude and latitude respectively. While ERP enables the direct application of pre-trained VAEs to encode panoramic videos into latent codes for diffusion processes, it introduces extreme horizontal stretching in polar regions. This horizontal stretching phenomenon arises from the altered representation of distances during projection, and is recognized by changes in horizontal signal frequency. Let $\mathrm{d}s_\varphi$ and $\mathrm{d}s_\theta$ represent the infinitesimal arc lengths along lines of constant latitude and longitude, respectively. They are formulated as:

$$\mathrm{d}s_\varphi = 2R \arcsin(\cos\theta \cdot \sin\frac{\mathrm{d}\varphi}{2}) = R\cos\theta\,\mathrm{d}\varphi, \quad \mathrm{d}s_\theta = R\,\mathrm{d}\theta, \tag{3}$$

where θ is the latitude. We further consider the spherical frequency f_{sph} (cycles per unit physical distance) and the Cartesian frequency f_{car} (cycles per pixel in the image). Assuming that warping preserves content, their relationship in frequency is scaled by the change in distance:

$$f_{\text{car},y}(x) = f_{\text{sph},\theta}(\varphi)\frac{\mathrm{d}s_\theta}{\mathrm{d}\theta} = Rf_{\text{sph},\theta}(\varphi), \quad f_{\text{car},x}(y) = f_{\text{sph},\varphi}(\theta)\frac{\mathrm{d}s_\varphi}{\mathrm{d}\varphi} = Rf_{\text{sph},\varphi}(\theta)\cos\theta. \tag{4}$$

Consequently, in polar regions ($|\theta| \approx \frac{\pi}{2}$, namely $y \approx 0$ or $y \approx R$), the horizontal frequency in the Cartesian coordinate becomes near-zero ($f_{\text{car},x}(y) \approx 0$). Such distortion in the horizontal frequency distribution significantly degrades the effectiveness of transferring priors.

4.2 Latitude-Aware Mechanisms

Latitude-aware sampling. Conventional text-to-video models typically assume independent and identically distributed (i.i.d.) Gaussian noise vectors for each Cartesian coordinate (x, y). To avoid latitudinal distortion in polar regions of ERP, we propose the latitude-aware sampling to better align the initial noise with the spherical frequency distribution for panoramic video generation. As illustrated in the top-left of Fig. 2, our latitude-aware sampling remaps the horizontal sampling coordinates based on latitude to preserve frequency consistency across the sphere. Specifically, after initializing the latent map with i.i.d. Gaussian noise vectors, we calculate the sampling noise by remapping the horizontal sampling coordinate x based on the latitude corresponding to row y, and then applying the interpolation:

$$P'(x, y) = \text{Interp}_P\Big(R + (x - R)\cos(\frac{2y+1-R}{2R}\pi), y\Big), \tag{5}$$

where $P'(x, y)$ is the interpolated noise vector at coordinate (x, y). Interp$(\cdot)$ is formulated as the interpolation function for normalization:

$$\text{Interp}_P(x, y) = \text{sgn}(\text{BI}(P, x, y))\sqrt{\text{BI}(P^2, x, y)}, \tag{6}$$

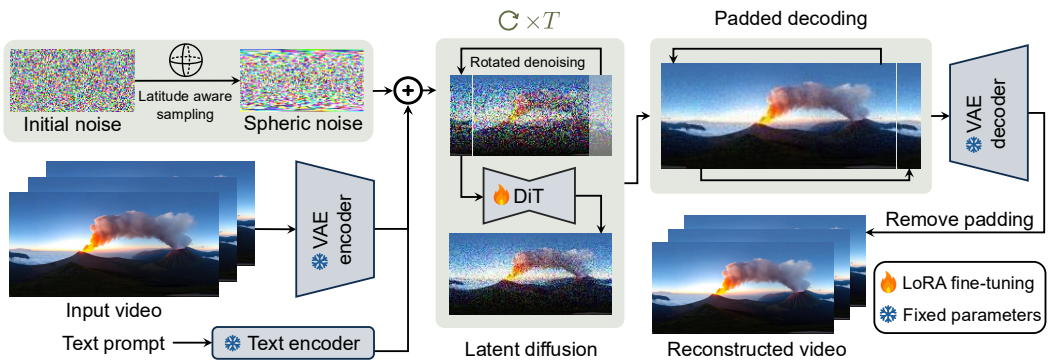


Figure 2: The pipeline of our proposed PanoWan, aware of spherical coordinates. To avoid latitudinal distortion, initial random Gaussian noise is remapped to align with the spherical frequency distribution using the latitude-aware sampling (Sec. 4.2). Next, this remapped noise serves as the latent code within the VAE-encoded latent space. A DiT-based denoising network then iteratively refines this latent representation, where rotated denoising is applied by rolling the latent grid to ensure semantic consistency across longitudinal boundaries. After that, padded pixel-wise decoding provides the VAE decoder with extended context, enabling the mapping of the denoised latent code back into seamless panoramic videos (Sec. 4.3). The DiT backbone within PanoWan is efficiently fine-tuned using LoRA, where most parameters of the pre-trained text-to-video model remain frozen to preserve its strong generative priors.

where $\mathrm{sgn}(\cdot)$ is the sign function, and $\mathrm{BI}(P, x, y)$ is the standard bilinear interpolation for vector P at coordinate (x, y). Consequently, the resulting noise vectors sampled by our strategy preserve $\mathbb{E}[P'(x, y)] = 0$ and $\mathbb{E}[\mathrm{Var}\, P'(x, y)] = 1$, approaching the distribution on which the diffusion models are pre-trained. The proof is given in the supplementary materials.

Frequency domain analysis. We aim to prove that the horizontal frequency properties of the proposed sampling correctly represent the inherent properties of the spherical coordinate, following the methodology of 1-D Discrete Fourier Transform (DFT). Denote the maximal frequency of the original signal in the spherical space as $f_{\max}$, and the maximal frequency in the Cartesian coordinates at the latitude of θ is determined by:

$$\max f_{\mathrm{car},x}(y) = \max R\cos\theta \cdot f_{\mathrm{sph},\varphi}(\theta) \leq R f_{\max}, \tag{7}$$

where the equation only holds when $\theta = 0$, namely on the equator. For the proposed design, it guarantees $\max f_{\mathrm{car},x} = 2R$ along the equator as $\cos(\theta) = 1$ and the original pixels along $P(\cdot, \frac{R-1}{2})$ is taken. Therefore, the maximal spherical frequency is $f_{\max} = 2$. According to Eq. (5), the warped results only depends on the values between $(R - R\cos(\theta), y)$ and $(R + (R-1)\cos(\theta), y)$ in the original Cartesian grid P. According to DFT, the support of the spectrum is reduced to:

$$\lceil R\cos(\theta)\rceil + \lceil (R-1)\cos(\theta)\rceil \approx 2R\cos\theta = f_{\max} R(\theta)\cos\theta = \max f_{\mathrm{car},x}(y),\ \forall\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]. \tag{8}$$

This matches the inherent property of the panorama in the frequency domain in every latitude.

4.3 Longitude Continuity Mechanisms

Seamless longitude transitions. Pre-trained conventional text-to-video models lack continuity awareness required between the columns of left and right boundaries. Consequently, applying their generative priors directly for panoramic video generation leads to seam artifacts, resulting in an observable transition where the easternmost and westernmost longitudes meet. To achieve seamless longitude transitions, we recognize that these artifacts arise from both semantic inconsistency and pixel-wise disharmony. This motivates us to propose the rotated denoising and the padded decoding to significantly remove the artifacts.

Rotated semantic denoising. The video generation backbone inherently introduces semantic inconsistency at each denoising step. Since the pre-trained generative priors lack the continuity awareness, the semantic in leftmost and rightmost longitudes is typically inconsistent, which are further accumulated during the iterative denoising steps and finally produce an obvious transition.

Our proposed rotated semantic denoising aims to spread the transition error evenly to different longitudes. Let $\mathcal{R}_{s_t}(\cdot)$ be the circular-shift operator and W denote the width of the latent code. As shown in Fig. 2, we horizontally roll the latent code Z_t by $\{s_t = t \bmod W\}$ columns at denoising step t and then undo the shift:

$$Z_{t+1} = \mathcal{R}_{-s_t}\Big(\phi_\theta\big(\mathcal{R}_{s_t}(Z_t)\big)\Big), \tag{9}$$

where $\phi_\theta(\cdot)$ is the noise predictor. As a result, the inherent accumulative error for horizontal coordinate x after T denoising steps is:

$$E_T(x) = \sum_{t=1}^{T} \varepsilon_t\big((x + s_t) \bmod W\big), \tag{10}$$

where $\varepsilon_t(\cdot)$ is prediction error for the transition and step t, which would concentrate at a fixed seam if no rotation are applied. Due to the rotation strategy, this error at physical coordinate x at step t is determined by the logical position $\{(x + s_t) \bmod W\}$. Over T steps, these logical coordinates $\{(x + s_t) \bmod W\}_{t=1}^{T}$ ideally approach a uniform permutation for all longitudes. This effectively suppresses seam artifacts by a factor approaching $1/W$.

Padded pixel-wise decoding. When decoding latent codes back to the pixel space, the pre-trained VAE decoder D often introduces pixel-wise inconsistencies, as it is trained on conventional videos and lacks awareness of the spatial continuity required across the left-right seam of panoramic videos [19]. Inspired by previous works [30, 40], we present the padded pixel-wise decoding. Let Z_0 be the denoised latent code. We first create a padded latent code $Z_0' = P_r(Z_0)$, where $P_r(\cdot)$ is a circular padding operator that extends Z_0 by r columns of context on the side, and the content at horizontal coordinate x is $\{x \bmod W\}$ in Z_0. Finally, we center crop the decoded panoramic videos after the decoding $V = \mathrm{Crop}\big(\mathrm{D}(Z_0')\big)$, as illustrated in Fig. 2. This approach ensures that pixels near the original seam boundaries are decoded with r columns of horizontal panoramic context. Consequently, the VAE decoder can effectively leverage its generative priors learned from conventional videos to avoid the seam artifacts.

5 Experiments

5.1 Training Details

PanoWan is built on Wan 2.1-1.3B-T2V [33] as the video generation backbone. We train PanoWan at a resolution of 448×896, closely matching the pre-trained resolution of this backbone model. For parameter-efficient training, LoRA [11] with a rank of 64 is applied to the query, key, value, and output projections of the attention mechanisms, as well as to the feed-forward networks. The model is trained for 200K iterations on our contributed PANOVID dataset. The training process employs the AdamW optimizer [13] with a learning rate of 1×10^{-4} and a batch size of 8. Training is conducted on 8 NVIDIA H100 GPUs for approximately 18 hours. During each iteration, clips of 81 consecutive frames are randomly sampled from the videos. Consequently, only 21.9M parameters are adjusted, constituting approximately 1.6% of the base model's total parameters.

5.2 Panoramic Video Evaluation Metrics

Existing panoramic video generation methods either directly apply conventional video evaluation metrics [12, 41] or rely on subjective user preferences [32], lacking metrics that comprehensively assess both perceptual quality and spherical consistency critical for panoramic video evaluation. This motivates us to adapt general video quality metrics for panoramic videos and to introduce additional panorama-specific metrics for structural properties of 360° content.

General metrics. We apply Frechét Video Distance (FVD) [27] to evaluate overall video quality and VideoCLIP-XL [31] to assess text-video alignment. Following DynamicScaler [12], we also

Table 1: Quantitative comparison results of PanoWan and previous text-based panoramic video generation models. ↑ (↓) means higher (lower) is better. Throughout the paper, best performances are highlighted in **bold**.

Method	General Metrics			Panoramic Metrics		
	FVD ↓	VideoCLIP-XL ↑	Image Quality ↑	End Continuity ↓	Motion Pattern ↑	Scene Richness ↑
360DVD [32]	1750.36	20.27	0.7054	0.0323	5.8%	6.6%
DynamicScaler [12]	2146.04	21.13	0.7188	0.0339	4.0%	2.6%
Ours (W/o LAS)	1520.69	21.20	0.7205	0.0278	16.2%	19.4%
Ours (W/o RSD)	1302.48	21.76	0.7243	0.0327	15.6%	18.8%
Ours (W/o PPD)	1294.03	21.81	0.7239	0.0294	22.0%	17.4%
Ours (full)	**1281.21**	**21.86**	**0.7249**	**0.0270**	**36.4%**	**35.2%**

calculate specific metrics for image quality. To adapt these general metrics for panoramic videos, we project each video onto a cube map and compute metric scores separately on each of the six faces. The final reported score for a video v is a weighted average:

$$\overline{\mathbf{f}}(v) = \sum_{f \in \mathcal{F}} \alpha_f \cdot \Phi\big(\mathcal{P}_f(v)\big), \tag{11}$$

where $\mathcal{F}$ denotes the set of cube map faces, $\mathcal{P}_f(v)$ is the projection of video v onto face $f \in \mathcal{F}$, Φ is the metric function, and α_f is the weight assigned to face f. Following OmniFID [6], we assign weights $\alpha_{\text{top}} = \alpha_{\text{bottom}} = \frac{1}{3}$ and $\alpha_{\text{side}} = \frac{1}{12}$ for each of the four lateral faces.

Panoramic metrics. Following previous works [32, 12], we evaluate motion patterns and scene richness with user preferences. We additionally introduce a quantitative metric for evaluating the end continuity of generated panoramic videos, tailored to capture artifacts across longitude boundaries. Specifically, this metric computes the mean absolute pixel difference across the left and right boundaries, directly capturing discontinuities at the longitude seam.

5.3 Comparison with State-of-the-art Methods

We evaluate PanoWan against existing text-based panoramic video generation methods, including 360DVD [32] and DynamicScaler [12], on the PANOVID test split containing 67 non-overlapping clips. Quantitatively, as shown in Tab. 1, PanoWan achieves state-of-the-art performance across both general and panoramic metrics (detailed in Sec. 5.2). Qualitatively, we present visual results to highlight our advantages. For instance, DynamicScaler [12] falls short in complex scenarios (Fig. 3, first sample), and 360DVD [32] exhibits notable distortion in polar regions (Fig. 3, second sample). In contrast, PanoWan effectively maintains global consistency and visual coherence, achieving superior performance in generating high-fidelity panoramic videos.

Following 360DVD [32], we also conducted a human-preference study to complement automatic metrics. 25 participants compared videos generated from 50 random test prompts among PanoWan, 360DVD [32], and DynamicScaler [12]. PanoWan was preferred in 71.36% of cases, versus 16.72% and 11.92% for 360DVD [32] and DynamicScaler [12], respectively. These results confirm that the perceptual improvements observed quantitatively are also recognized subjectively.

5.4 Ablation Studies

We conduct ablation studies to validate the effectiveness of proposed modules in PanoWan: latitude-aware sampling (LAS), rotated semantic denoising (RSD), and padded pixel-wise decoding (PPD).

Quantitative results. As shown in Tab. 1, removing LAS primarily affects general metrics—FVD increases from 1281.21 to 1520.69, and VideoCLIP-XL drops from 21.86 to 21.20—indicating the model struggles to learn panoramic features in high-latitude regions without frequency-aligned noise initialization. In contrast, removing RSD or PPD mainly degrades panoramic metrics (*e.g.*, end continuity increases from 0.0270 to 0.0327 and 0.0294, respectively), confirming their roles in achieving seamless longitude transitions.

Wide-angle panoramic interior capturing an energetic hot pot restaurant bustling with diners enthusiastically cooking ingredients in bubbling pots. Colorful plates arranged invitingly, steam rising, spirited conversations filling warm, inviting atmosphere emphasizing community and culinary enjoyment.

Expansive panoramic view capturing a vibrant ski resort nestled among towering snowy mountains, as skiers gracefully descend pristine slopes amid cozy alpine chalets. Chairlifts glide leisurely under crisp blue skies while visitors gather around lively outdoor cafes, soaking in the sunny winter scenery.

Figure 3: Visual comparison results with existing text-based panoramic video generation methods.

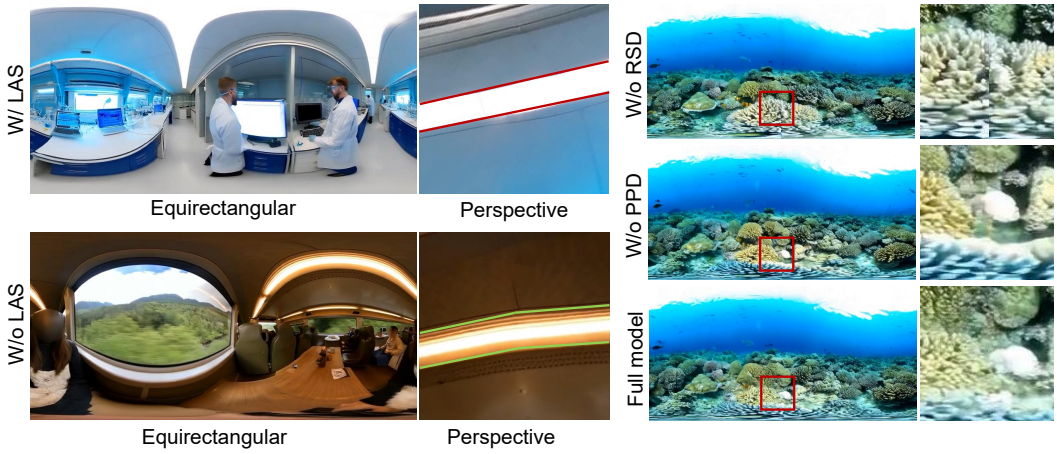
Qualitative results. We further provide qualitative evaluation in Fig. 4. When generating high-latitude elements like LED panels (which should appear straight in perspective views but are inherently distorted in the equirectangular projection), PanoWan without LAS fails to render them with the correct geometric appearance. When RSD is discarded, semantic inconsistencies become apparent at the longitude seam, due to the lack of mechanism for continuity awareness and the error accumulation during the denoising process. When PPD is removed, observable seam artifacts occur because conventional VAE decoder introduces pixel-wise inconsistencies at boundaries. Consequently, the full PanoWan model with all modules enabled achieves the best performance.

5.5 Application

As a text-based panoramic video model, PanoWan shows robust zero-shot capabilities across a wide range of downstream tasks. We present representative examples in Fig. 1 and additional examples in supplementary materials due to the space limitation.

Long video generation. To generate panoramic videos longer than the model's native temporal context, we adopt a sliding-window inference strategy in the latent space. At each denoising step, the latent code is partitioned into temporally overlapping chunks, where each chunk is conditioned on a corresponding segment of text prompts expanded by an LLM. Adjacent chunks share overlapping frames that serve as temporal context, and their outputs are fused through a linear blending function on the overlapping regions.

Super-resolution for panoramic videos. To generate high-resolution panoramic videos from low-resolution ones, we first encode each low-resolution video into its corresponding latent code. After



(a) Visual evaluation of LAS (b) Visual evaluation of RSD and PPD

Figure 4: Qualitative evaluation of proposed latitude/longitude-aware mechanisms. (a) With the proposed Latitude-Aware Sampling (LAS), PanoWan ensures that content generated at high latitudes exhibits an accurate geometry when presented in a perspective view. (b) By combining Rotated Semantic Denoising (RSD) and Padded Pixel-wise Decoding (PPD), PanoWan achieves seamless longitude transitions. For visualization, videos are rolled 180° to center the seam.

injected noise, the latent code is denoised based on user-provided text descriptions, producing results with structural consistency and visual fidelity across the spherical representation.

Inpainting for semantic editing. Given a panoramic video, we identify and mask regions for modification. Next, we apply the denoising process to these regions, guided by user-provided text descriptions. Leveraging its understanding of spherical representations, the inpainted content naturally exhibits the properties of ERP projection.

Outpainting for conventional videos. Similar to the inpainting process, we first map the conventional video to the latent code and then mask the surrounding unseen panoramic regions. With user-provided text descriptions, we denoise the masked regions to generate corresponding content. Our pre-trained model maintains the spatial and temporal consistency for generated panoramic videos.

6 Conclusion

We present PanoWan, a text-based panoramic video generation framework that effectively lifts pre-trained diffusion model to the panorama. By integrating latitude-aware sampling, PanoWan addresses latitudinal distortions caused by equirectangular projection. Equipped with the rotated semantic denoising and the padded pixel-wise decoding, PanoWan achieves the seamless longitude transitions. To provide large-scale data for lifting representations from conventional videos to the panorama, we contribute the PANOVID dataset, offering high-quality and semantically rich 360° video data with annotated text descriptions. Extensive experiments demonstrate that PanoWan achieves state-of-the-art performance on text-based panoramic video generation and strong generalization across diverse zero-shot downstream tasks.

Limitation. While PanoWan benefits from the strong priors of its pretrained text-to-video models, it is also inherits a common challenge: the content forgetting problem often seen in such models. This issue is particularly evident when generating long videos due to limited temporal memory. We believe this challenge can be substantially alleviated through future advancements in memory-aware generation techniques (*e.g.*, video caching mechanisms).

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant No.62136001. Authors thank openbayes.com for providing computing resources.

References

[1] Miraikan 360-degree video dataset. https://www.miraikan.jst.go.jp/en/research/AccessibilityLab/dataset360/.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+x: A panoptic multi-modal scene understanding dataset. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

[7] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, 2003.

[8] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

[9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[10] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. of the International Conference on Learning Representations*, 2022.

[12] Liu Jinxiu, Lin Shaoheng, Li Yinxiao, and Yang Ming-Hsuan. DynamicScaler: Seamless and scalable video generation for panoramic scenes. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[13] Diederik P Kingma, J Adam Ba, and J Adam. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2020.

[14] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[15] Benjamin J. Li, Jeremy N. Bailenson, Adam Pines, Walter J. Greenleaf, and Leanne M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in Psychology*, 2017.

[16] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4K4DGen: Panoramic 4D generation at 4K resolution. *arXiv preprint arXiv:2406.13527*, 2024.

[17] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[18] Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille, and Jieneng Chen. GenEx: Generating an explorable world. *Proc. of the International Conference on Learning Representations*, 2025.

[19] Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360° panoramic videos from perspective videos. *arXiv preprint arXiv:2504.07940*, 2025.

[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 *Proc. of ACM SIGGRAPH Asia*, 2024.

[21] Minho Park, Taewoong Kang, Jooyeol Yun, Sungwon Hwang, and Jaegul Choo. SphereDiff: Tuning-free omnidirectional panoramic image and video generation via spherical latent representation. *arXiv preprint arXiv:2504.14396*, 2025.

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[23] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[24] SeaweadTeam, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7B: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.

[26] Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. *arXiv preprint arXiv:2412.03552*, 2024.

[27] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[28] Veo-Team, :, Agrim Gupta, Ali Razavi, Andeep Toor, Ankush Gupta, Dumitru Erhan, Eleni Shaw, Eric Lau, Frank Belletti, Gabe Barth-Maron, Gregory Shaw, Hakan Erdogan, Hakim Sidahmed, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jeff Donahue, José Lezama, Kory Mathewson, Kurtis David, Matthieu Kim Lorrain, Marc van Zee, Medhini Narasimhan, Miaosen Wang, Mohammad Babaeizadeh, Nelly Papalampidi, Nick Pezzotti, Nilpa Jha, Parker Barnes, Pieter-Jan Kindermans, Rachel Hornung, Ruben Villegas, Ryan Poplin, Salah Zaiem, Sander Dieleman, Sayna Ebrahimi, Scott Wisdom, Serena Zhang, Shlomi Fruchter, Signe Nørly, Weizhe Hua, Xinchen Yan, Yuqing Du, and Yutian Chen. Veo 2. 2024. URL `https://deepmind.google/technologies/veo/veo-2/`.

[29] 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. In *Proc. of Neural Information Processing Systems*, 2024.

[30] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[31] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing long description understanding for video clip models. *arXiv preprint arXiv:2410.00741*, 2024.

[32] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360DVD: Controllable panorama video generation with 360-degree video diffusion model. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[33] WanTeam, 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.

[34] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[35] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WORLDMEM: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

[36] Kevin Xie, Amirmojtaba Sabour, Jiahui Huang, Despoina Paschalidou, Greg Klar, Umar Iqbal, Sanja Fidler, and Xiaohui Zeng. VideoPanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*, 2025.

[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*, 2024.

[38] 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. In *ICLR*, 2025.

[39] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *ICLR*, 2024.

[40] 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 *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[41] Muyang Zhang, Yuzhi Chen, Rongtao Xu, Changwei Wang, JinMing Yang, Weiliang Meng, Jianwei Guo, Huihuang Zhao, and Xiaopeng Zhang. PanoDit: Panoramic videos generation with diffusion transformer. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2025.

[42] Haiyang Zhou, Wangbo Yu, Jiawen Guan, Xinhua Cheng, Yonghong Tian, and Li Yuan. Holotime: Taming video diffusion models for panoramic 4D scene generation, 2025. URL `https://arxiv.org/abs/2504.21650`.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction part reflect the main contributions and scope.

   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: The limitations are discussed in Sec. 6.

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: [Yes]

   Justification: Proofs are given in the supplementary materials.

   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: The model framework and experiment settings are described in detail.

   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: [No]

   Justification: Both the dataset and the codes will be released no latter than acceptance.

   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: The experiments are presented in detail.

   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: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   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: These can be found in Sec. 5.

   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: The authors are fully aware of 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: [NA]

    Justification: This paper has no societal impact.

    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: [No]

    Justification: The paper notes potential biases from pretrained models but does not describe safeguards.

    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: They are cited clearly.

    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: Codes and the dataset will be released no latter than acceptance.

    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: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    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: We clearly describe the usage of LLMs to annotate our dataset.

    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.