360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model

Qian Wang[1,2], Weiqi Li[1], Chong Mou[1,2], Xinhua Cheng[1,2], Jian Zhang[1,2] ✉
[1]School of Electronic and Computer Engineering, Peking University
[2]Peking University Shenzhen Graduate School-Rabbitpre AIGC Joint Research Laboratory
{qianwang,liweiqi,eechongm,chengxinhua}@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

Abstract

*Panorama video recently attracts more interest in both study and application, courtesy of its immersive experience. Due to the expensive cost of capturing 360° panoramic videos, generating desirable panorama videos by prompts is urgently required. Lately, the emerging text-to-video (T2V) diffusion methods demonstrate notable effectiveness in standard video generation. However, due to the significant gap in content and motion patterns between panoramic and standard videos, these methods encounter challenges in yielding satisfactory 360° panoramic videos. In this paper, we propose a pipeline named **360-D**egree **V**ideo **D**iffusion model (**360DVD**) for generating 360° panoramic videos based on the given prompts and motion conditions. Specifically, we introduce a lightweight 360-Adapter accompanied by 360 Enhancement Techniques to transform pre-trained T2V models for panorama video generation. We further propose a new panorama dataset named WEB360 consisting of panoramic video-text pairs for training 360DVD, addressing the absence of captioned panoramic video datasets. Extensive experiments demonstrate the superiority and effectiveness of 360DVD for panorama video generation. Our project page is at https://akaneqwq.github.io/360DVD/.*

1. Introduction

With the recent advancements in VR technology, 360-degree panoramic videos have been gaining increasing popularity. This video format which offers audiences an immersive experience, is helpful for various applications, including entertainment, education, and communication. To capture details of the entire scene, 360° videos are typically recorded using an array of high-resolution fisheye cameras that yields a 360° × 180° field-of-view (FoV) [1], which is quite costly in both time and resources. Therefore, the generation of 360° panoramic videos is urgently required for border applications, while panoramic video generation receives little attention in studies to date.

Thanks to the emerging theory and training strategies, text-to-image (T2I) diffusion models [26, 27, 31, 32, 35] demonstrate remarkable image generation capacity from prompts given by users, and such impressive achievement in image generation is further extended to text-to-video (T2V) generation. Various T2V diffusion models [3, 16, 37, 46, 52, 60] are recently proposed with adopting space-time separable architectures, wherein spatial operations are inherited from the pre-trained T2I models to reduce the complexity of constructing space-time models from scratch. Among these, AnimateDiff [16] enables the capability to generate animated images for various personalized T2I models, which alleviates the requirement for model-specific tuning and achieves compelling content consistency over time.

Although T2V methods on standard videos are widely studied, there is no method proposed for panorama video generation. One potential approach is to leverage existing powerful T2V models, *e.g.*, AnimateDiff to directly generate the equirectangular projection (ERP) of panoramic videos. Since ERP is a commonly adopted format for storing and transmitting panoramic videos, each frame is treated by ERP as a rectangular image with an aspect ratio of 1:2, which aligns well with the output format of existing standard T2V models. However, due to the significant differences between panoramic videos and standard videos, existing methods suffer challenges in directly producing satisfactory 360° panoramic videos. Concretely, the main challenges include three aspects: **(1)** The content distribution of ERPs differs from standard videos. ERPs require a wider FoV, reaching 360° × 180°. **(2)** The motion patterns of ERPs are different from standard videos, with movements often following curves rather than straight lines. **(3)** The left and right ends of ERPs should exhibit continuity since they correspond to the same meridian on the Earth.

Therefore, we propose a specifically designed method

This work was supported by National Natural Science Foundation of China under Grant 62372016. (✉ Corresponding author: Jian Zhang)

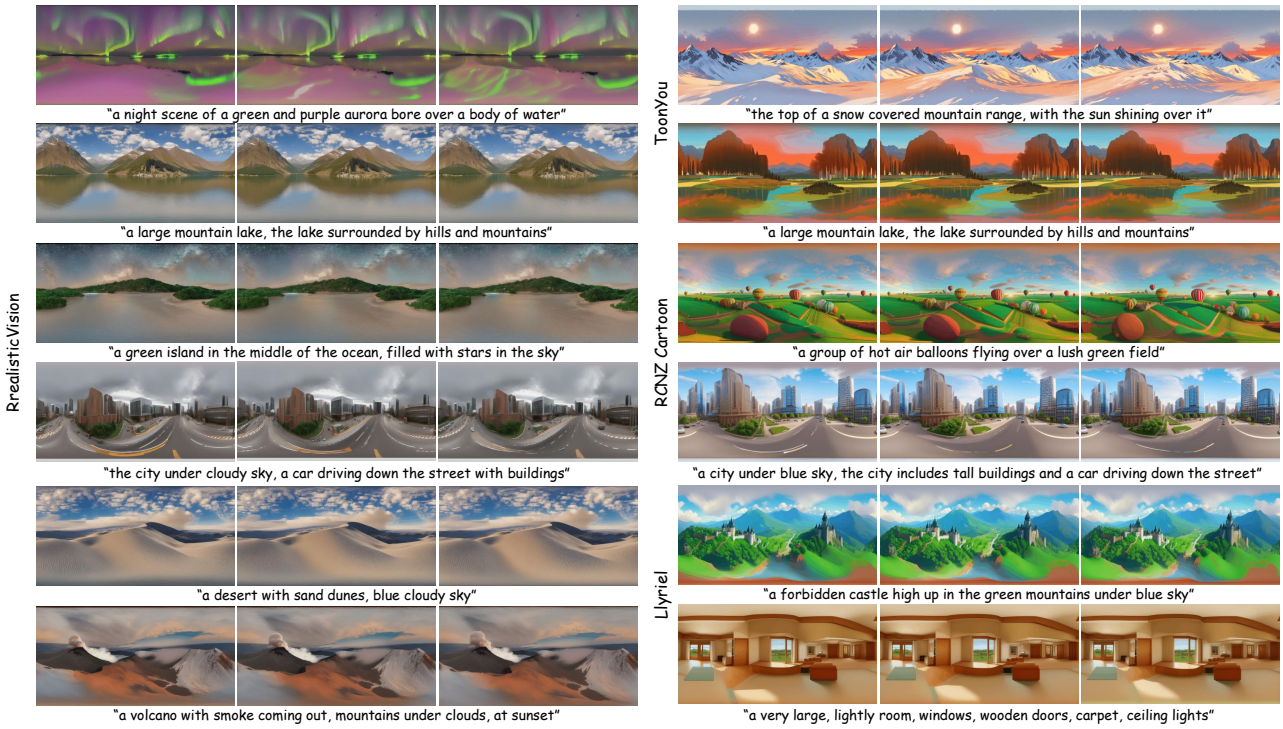

Figure 1. **Main results.** Our 360DVD creates text-aligned, coherent, and high-quality 360° panorama videos. Furthermore, 360DVD can cooperate with multiple personalized text-to-image models and consistently generate stylized panorama videos.

named **360**-**D**egree **V**ideo **D**iffusion (360DVD) for generating panorama videos. We first introduce a plug-and-play module named 360-Adapter to address challenge mentioned above. Our 360-Adapter receives zero values or motion conditions (*e.g.,* optical flow) as input and outputs motion features, which are fed into the frozen denoising U-Net at different levels of the encoder. This transformation is aimed at converting the T2V model into a panoramic video generation without altering the foundational generative capabilities. In addition, we introduce 360 Enhancement Techniques including two mechanisms to enhance continuity at both ends of ERPs from both macro and micro perspectives, and a latitude-aware loss function for encouraging the model to focus more on low-latitude regions. Cooperated with carefully designed techniques, our 360DVD generates text-aligned, coherent, high-quality, 360° panorama videos with various styles, as shown in Fig. 1.

Furthermore, we collect a panorama dataset named WEB360 including ERP-formatted videos from the internet and games for training our method. WEB360 involves approximately 2,000 video clips with each clip consisting of 100 frames. Considering the domain gap between panoramic and standard images, to enhance the accuracy and granularity of captions, we introduce a GPT-based 360 Text Fusion module for obtaining detailed captions. Our contributions can be summarized as follows:

- We introduce a controllable 360° panorama video generation diffusion model named 360DVD, achieved by adopting a controllable standard T2V model with a trainable lightweight 360-Adapter. Our model can generate text-guided panorama videos conditioned on desired motions.
- We design 360 Enhancement Techniques including a latitude-aware loss and two mechanisms to enhance the content and motion quality of generated panorama videos.
- We propose a new high-quality dataset named WEB360 comprising approximately 2,000 panoramic videos, with each video accompanied by a detailed caption enhanced through 360 Text Fusion.
- Experiments demonstrate that our 360DVD is capable of generating high-quality, high-diversity, and more consistent 360° panorama videos.

2. Related Works

2.1. Text-to-Image Diffusion Model

The Denoising Diffusion Probabilistic Model [9, 17, 39] has proven to be highly successful in generating high-quality images, outperforming previous approaches such as generative adversarial networks (GANs)[11, 57], variational autoencoders (VAEs)[20, 38], and flow-based methods [5]. With text guidance during training, users can generate images based on textual input. Noteworthy examples include

GLIDE [27], DALLE-2 [31], Imagen [35]. To address the computational burden of the iterative denoising process, LDM [32] conducts the diffusion process on a compressed latent space rather than the original pixel space. This accomplishment has prompted further exploration in extending customization [14, 34], image guidance [53, 55], precise control [25, 26, 58] and protection [56].

2.2. Text-to-Video Diffusion Model

Despite significant advancements in Text-to-Image (T2I) generation, Text-to-Video (T2V) generation faces challenges, including the absence of large-scale, high-quality paired text-video datasets, the inherent complexity in modeling temporal consistency, and the resource-intensive nature of training. To address these challenges, many works leverage the knowledge from pre-trained T2I models, and they manage training costs by executing the diffusion process in the latent space. Some methods [15, 29, 48, 49, 54] utilize T2I models in zero-shot or few-shot ways. However, these methods often suffer from suboptimal frame consistency due to insufficient training. To address this limitation, another category of T2V diffusion models typically adopts space-time separable architectures. These models [3, 37, 46, 60] inherit spatial operations from pre-trained T2I models, reducing the complexity of constructing space-time models from scratch. Given that most personalized T2I models are derived from the same base one (e.g. Stable Diffusion [32]), AnimateDiff [16] designs a motion modeling module that trained with a base T2I model and could animate most of derived personalized T2I models once for all. There are also efforts focused on enhancing control in T2V models. Gen-1 [13], MCDiff [6], LaMD [18] and VideoComposer [47] introduce diverse conditions to T2V models. Despite these advancements, the aforementioned methods demand extensive training and lack a plug-and-play nature, making it challenging to apply them to a diverse range of personalized T2I models.

2.3. Panorama Generation

GAN-based methods for generating panoramic images have been widely studied [2, 4, 7, 10, 12, 23, 24, 28, 40, 41, 43, 50]. For instance, OmniDreamer [2] accepts a single NFoV image as an input condition and introduces a cyclic inference scheme to meet the inherent horizontal cyclicity of 360-degree images. ImmenseGAN [12] fine-tunes the generative model using a large-scale private text-image pair dataset, making the generation more controllable. Text2Light [7] introduces a zero-shot text-guided 360-image synthesis pipeline by utilizing the CLIP model. Very recently, diffusion models have achieved promising results in panoramic image generation. DiffCollage [59] uses semantic maps as conditions and generates images based on complex factor graphs using retrained diffusion models. PanoGen [21] employs a latent diffusion model and synthesizes new indoor panoramic images through recursive image drawing techniques based on multiple text descriptions. PanoDiff [45] achieves a multi-NFoV synthesis of panoramic images through a two-stage pose estimation module. IPO-LDM [51] uses a dual-modal diffusion structure of RGB-D to better learn the spatial distribution and patterns of panoramic images. StitchDiffusion [44] employs a T2I diffusion model, ensuring continuity at both ends through stitching. However, to date, panoramic video generation has received limited attention. To the best of our knowledge, we are the first to leverage diffusion models for panoramic video generation.

3. Method

In this section, we begin with a concise review of the latent diffusion fusion model and AnimateDiff [16]. Following that, we introduce the construction method of the WEB360 dataset. We then provide an overview of 360DVD and elaborate on the implementation details of 360-Adapter. Finally, we describe the 360 enhancement techniques aimed at enriching the panoramic nature of the video.

3.1. Preliminaries

Latent Diffusion Model. Given an input signal $\mathbf{x}_0$, a diffusion forward process in DDPM [17] is defined as:

$$p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

for $t = 1, \ldots, T$, where T is the total timestep of the diffusion process. A noise depending on the variance β_t is gradually added to $\mathbf{x}_{t-1}$ to obtain $\mathbf{x}_t$ at the next timestep and finally reach $\mathbf{x}_T \in \mathcal{N}(0, \mathbf{I})$. The goal of the diffusion model is to learn to reverse the diffusion process (denoising). Given a random noise $\mathbf{x}_t$, the model predicts the added noise at the next timestep $\mathbf{x}_{t-1}$ until the origin signal $\mathbf{x}_0$:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \quad (2)$$

for $t = T, \ldots, 1$. We fix the variance $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ and utilize the diffusion model with parameter θ to predict the mean of the inverse process $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$. The model can be simplified as denoising models $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, which are trained to predict the noise of $\mathbf{x}_t$ with a noise prediction loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \mathbf{y}, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}), t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \boldsymbol{\tau}_\theta(\mathbf{y}))\|_2^2], \quad (3)$$

where $\boldsymbol{\epsilon}$ is the added noise to the input image $\mathbf{x}_0$, $\mathbf{y}$ is the corresponding textual description, $\boldsymbol{\tau}_\theta(\cdot)$ is a text encoder mapping the string to a sequence of vectors.

Latent Diffusion Model (LDM) [32] executes the denoising process in the latent space of an autoencoder, namely $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$, implemented as VQ-GAN [19] or VQ-VAE [42] pre-trained on large image datasets. During the

training of the latent diffusion networks, an input image $\mathbf{x}_0$ is initially mapped to the latent space by the frozen encoder, yielding $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. Thus, the training objective can be formulated as follows:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0),\mathbf{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \boldsymbol{\tau}_\theta(\mathbf{y}))\|_2^2]. \quad (4)$$

In widely-used LDM Stable Diffusion (SD), which our method is based on, $\boldsymbol{\epsilon}_\theta(\cdot)$ is implemented with a modified UNet [33] that incorporates four downsample/upsample blocks and one middle block, resulting in four resolution levels within the networks' latent space. Each resolution level integrates 2D convolution layers as well as self- and cross-attention mechanisms. Text model $\boldsymbol{\tau}_\theta(\cdot)$ is implemented using the CLIP [30] ViT-L/14 text encoder.

AnimateDiff. AnimateDiff inflates base SD by adding temporal-aware structures and learning reasonable motion priors from large-scale video datasets. Since the original SD can only process 4D image data batches, while T2V task takes a 5D video tensor as input. It transforms each 2D convolution and attention layer in the original image model into spatial-only pseudo-3D layers. The motion module is inserted at every resolution level of the U-shaped diffusion network, using vanilla temporal transformers consisting of several self-attention blocks operating along the temporal axis. The training objective of AnimateDiff can be written as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0^{1:N}),\mathbf{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{1:N}, t, \boldsymbol{\tau}_\theta(\mathbf{y}))\|_2^2], \quad (5)$$

where $\mathbf{x}_0^{1:N}$ is the sampled video data, $\mathbf{z}_0^{1:N}$ is the latent code which $\mathbf{x}_0^{1:N}$ are encoded into via the pre-trained autoencoder, $\mathbf{z}_t^{1:N}$ is the latent code obtained by perturbing the initial latent code $\mathbf{z}_0^{1:N}$ with noise at timestep t. During training, the pre-trained weights of the base T2I model are frozen to keep its feature space unchanged.

3.2. WEB360 Dataset

Diverse text-video pairs datasets are essential for training open-domain text-to-video generation models. However, existing 360° panorama video datasets lack corresponding textual annotations. Moreover, these datasets are often constrained either in scale or quality, thereby impeding the upper limit of high-quality video generation.

To address the aforementioned challenges and achieve high-quality 360 panorama video generation, we introduce a novel text-video dataset named WEB360. This dataset comprises 2114 text-video pairs sourced from open-domain content, presented in high-definition (720p) ERP format. Our dataset creation process involved extracting 210 high-resolution panoramic video clips from the ODV360 [4] training set. Additionally, we collected over 400 original videos from YouTube. Due to the complex scene transitions present in the original videos, which pose challenges for models in learning temporal correlations, we perform a manual screening process to split the original videos into 1904 single-scene video clips. We employ BLIP [22] to annotate the first frame of the 2104 video clips. However, we observed that direct application of BLIP to ERP images often resulted in bad captions. Therefore, we propose a panoramic image caption method named 360 Text Fusion, based on ChatGPT.

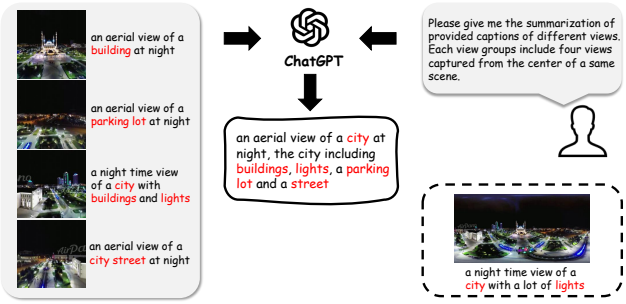

Figure 2. **360 Text Fusion.** The captions of four images with a FoV of 90 are fed into ChatGPT to generate a new 360° summarization. Compared to the caption of ERP at the bottom right, 360 Text Fusion allows for more fine-grained captions.

360 Text Fusion. We find that directly using BLIP [22] to label ERP has drawbacks. On one hand, errors may arise due to the distortion caused by the polarities, leading to misidentifications such as labeling "person" as "dog". On the other hand, the captions generated by BLIP lack granularity, making them insufficient for providing a detailed description of the current scene. Thus, we propose 360 Text Fusion (360TF) method, as shown in Fig. 2. To deal with the irregular distortion of ERP, we turn to less-distorted perspective images. We first project the original ERP image to four non-overlapping perspective images at 0 degrees longitude, with a FoV of 90. The four images are then fed into BLIP to be captioned. By pre-informing ChatGPT about the task and providing examples, these four captions are collectively input to ChatGPT, which then generates a summary of the scene as our final caption. In comparison to directly using BLIP to label the entire image, our 360TF demonstrates a significant advantage in granularity.

3.3. 360-degree Video Diffusion Model

An overview of the 360-degree Video Diffusion Model (360 DVD) is presented in Fig. 3, which is composed of a pre-trained denoising U-Net and 360-Adapter. The pre-trained denoising U-Net adopts a structure identical to that of AnimateDiff. In every resolution level of the U-Net, the spatial layer unfolds pre-trained weights from SD, while the temporal layer incorporates the motion module of AnimateDiff trained on a large-scale text-video dataset.

During the training process, we first sample a video $\mathbf{x}_0^{1:N}$

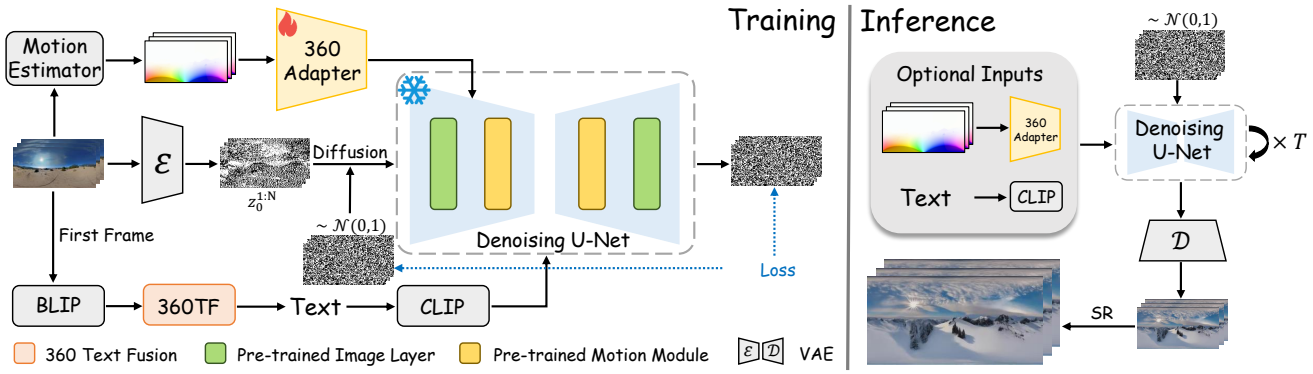

Figure 3. **Overview of 360DVD.** 360DVD leverages a trainable 360-Adapter to extend standard T2V models to the panorama domain and is able to generate high-quality panorama videos with given prompts and optional motion conditions. In addition, 360 Enhancement Techniques are proposed for quality improvement in the panorama perspective.

from the dataset. The video is encoded into latent code $\mathbf{z}_0^{1:N}$ through pre-trained VAE encoder $\mathcal{E}(\cdot)$ and noised to $\mathbf{z}_t^{1:N}$. Simultaneously, the corresponding text $\mathbf{y}$ for the video is encoded using the text encoder $\tau_\theta(\cdot)$ of the CLIP. The video is also input into a motion estimation network to generate corresponding motion conditions $\mathbf{c}$, which are then fed into the 360-Adapter $\mathcal{F}_{360}(\cdot)$. Finally, noised latent code $\mathbf{z}_t^{1:N}$, timestep t, text embedding $\tau_\theta(\mathbf{y})$, and the feature maps $\mathbf{f}_{360}$ generated by 360-Adapter are collectively input into the U-Net $\epsilon(\cdot)$ to predict the noise strength added to the latent code. As we aim to preserve the priors learned by SD and AnimateDiff on large datasets, we freeze their weights during the training process. If we use a simple L2 loss term, the training objective is given as follows:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0^{1:N}),\mathbf{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{1:N}, t, \tau_\theta(\mathbf{y}), \mathbf{f}_{360})||_2^2]. \tag{6}$$

To ensure satisfactory generation of 360° panoramic videos without motion control input, we set the input of the 360-Adapter to zero with a probability P during training. This strategy aims to encourage the model to learn representations that are not solely reliant on motion conditions, enhancing its ability to generate compelling panoramic videos without explicit motion guidance.

In inference, users have the option to selectively provide text prompts and motion guidance to carry out denoising over a total of T steps. Here, we employ DDIM [39] to accelerate the sampling process. The estimated latent code $\hat{\mathbf{z}}_0^{1:N}$ is then input into a pre-trained VAE decoder to decode the desired 360° panoramic videos $\hat{\mathbf{x}}_0^{1:N}$. Due to constraints such as resolution limitations imposed by existing SD and considerations regarding GPU memory usage, the experimental results presented in this paper showcase a resolution of 512 × 1024. In practical applications, super-resolution methods [8, 40] can be employed to upscale the generated results to the desired size.

360-Adapter. Our proposed 360-Adapter is simple and lightweight, as shown in Fig. 4. The original condition input has the the same resolution as the video of $H \times W$. Here, we utilize the pixel unshuffle [36] operation to downsample it to $H/8 \times W/8$. Following that are four 360-Adapter blocks, we depict only one for simplification in Fig. 4. To maintain consistency with the U-Net architecture, the first three 360-Adapter blocks each include a downsampling block. In each 360-Adapter block, one 2D convolution layer and a residual block (RB) with pseudo-3D convolution layers are utilized to extract the condition feature $\mathbf{f}_{360}^k$. Finally, multi-scale condition features $\mathbf{f}_{360} = \{\mathbf{f}_{360}^1, \mathbf{f}_{360}^2, \mathbf{f}_{360}^3, \mathbf{f}_{360}^4\}$ are formed. Suppose the intermediate features in the U-Net encoder block is $\mathbf{f}_{enc} = \{\mathbf{f}_{enc}^1, \mathbf{f}_{enc}^2, \mathbf{f}_{enc}^3, \mathbf{f}_{enc}^4\}$. $\mathbf{f}_{360}$ is then added with $\mathbf{f}_{enc}$ at each scale. In summary, the condition

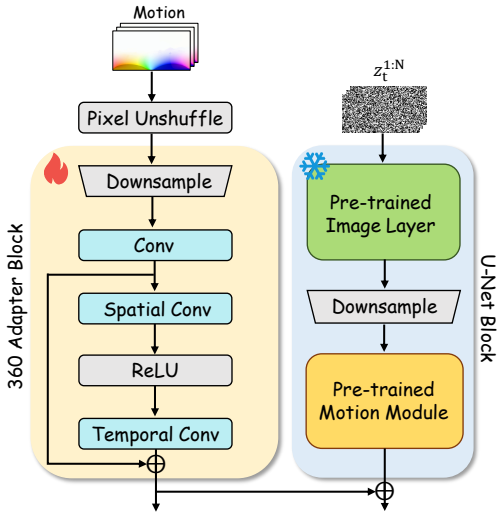

Figure 4. **Overview of 360-Adapter.** 360-Adapter is a simple but effective module in which intermediate features are fed into the U-Net encoder blocks for modulation.

feature extraction and conditioning operation of the 360-Adapter can be defined as the following formulation:

$$\mathbf{f}_{360} = \mathcal{F}_{360}(\mathbf{c}), \tag{7}$$

$$\hat{\mathbf{f}}^i_{enc} = \mathbf{f}^i_{enc} + \mathbf{f}^i_{360}, i \in \{1, 2, 3, 4\}. \tag{8}$$

In the previous description, we omit some details. Our motion condition $\mathbf{c}$ is a 5D tensor, assuming its size is $batch \times channels \times frames \times height \times width$. We first reshape it into a 4D tensor of size $(batch \times frames) \times channels \times height \times width$ to allow it to be fed into the 2D convolution layer and restore it to 5D to go through the RB with pseudo-3D convolution layers. Subsequently, in the RB, we employ a $1 \times 3 \times 3$ pseudo-3D convolution to extract features in the spatial dimension, followed by a $3 \times 1 \times 1$ pseudo-3D convolution to model information along the temporal dimension. The resulting features are reshaped back to $(batch \times frames) \times channels \times height \times width$ to add the output of the skip connection. Finally, condition features are reshaped back into a 5D vector of size $batch \times channels \times frames \times height \times width$ to align with the U-Net encoder intermediate features.

3.4. 360 Enhancement Techniques

Latitude-aware Loss. When projecting panoramic videos into ERPs, meridians are mapped as vertically spaced lines with a constant interval, while parallels are mapped as horizontally spaced lines with a constant interval. This projection method establishes a straightforward mapping relationship, but it is neither equal-area nor conformal, introducing significant distortion, particularly in the polar regions. To make the denoiser pay more attention to low-latitude regions with less distortion, which is more crucial for human visual perception, we introduce a latitude-aware loss:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0^{1:N}),\mathbf{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}[||\mathbf{W} \odot (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta)||_2^2], \tag{9}$$

where $\hat{\boldsymbol{\epsilon}}_\theta = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{1:N}, t, \boldsymbol{\tau}_\theta(\mathbf{y}), \mathbf{f}_{360})$, and $\mathbf{W}$ is a weight matrix used to perform element-wise product, defined as:

$$\mathbf{W}_{i,j} = \cos\left(\frac{2i - H/8 + 1}{H/4}\pi\right), \tag{10}$$

where $i \in [0, H/8)$, $j \in [0, W/8)$, $H/8$ and $W/8$ is the height and width of latent code $\mathbf{z}_t^{1:N}$. The visualized result of $\mathbf{W}$ is shown in Fig. 5, where pixels in low and middle latitudes are given more weight during training.

Latent Rotation Mechanism. Because ERPs can be considered as the unfolding of a spherical surface along a meridian, they are meant to be wraparound consistent, implying that their left and right sides are continuous. However, during the process of video generation, the left and right sides are physically separated. Inspired by PanoDiff [45], we employ a latent rotation mechanism to enhance the macroscopic coherence between the left and right ends of the video. During the inference process, we perform a horizontal rotation at an angle of θ on $\mathbf{z}_t^{1:N}$ and motion condition $\mathbf{c}$, at each denoising step. As illustrated in Fig. 5, the content on the far left is shifted to the far right, where we use $\mathbf{x}_0^1$ to replace $\mathbf{z}_t^{1:N}$ for a better visual effect of its continuity. During the training process, we also randomly rotate the training videos along with the motion condition by a random angle as a data augmentation strategy.

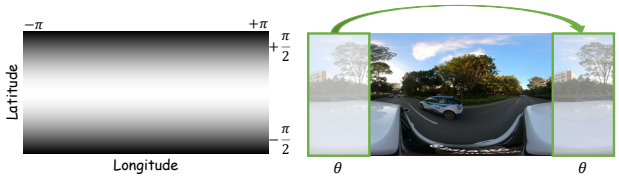

Figure 5. **Left:** the visualization of weight matrix $\mathbf{W}$, brighter colors indicate values closer to 1, while darker colors suggest values closer to 0. **Right:** a schematic diagram of the latent rotation mechanism. In each iteration, the far left portion of angle θ is shifted to the far right.

Circular Padding Mechanism. Although the previous latent rotation mechanism achieves semantic continuity at a macroscopic level, achieving pixel-level continuity is challenging. Therefore, in the inference process, we adopt a mechanism of circular padding by modifying the padding method of the convolution layers. We observe that the early stages of 360° video generation often involve layout modeling, while the later stages focus on detail completion. To maintain the stable video generation quality of 360DVD, we only implement the circular padding mechanism in the late $\lfloor \frac{T}{2} \rfloor$ steps of a total of T denoising steps.

4. Experiment

4.1. Implementation Details

Training Settings. We choose Stable Diffusion v1.5 and Motion Module v14 as our base model. We utilize the panoramic optical flow estimator PanoFlow [45] to generate motion conditions. We train the 360-Adapter using the proposed WEB360 dataset. The resolution is set to 512×1024, the length of frames to 16, the batch size to 1, the learning rate to 1×10^{-5}, and the total number of training steps to $100k$, probability $P = 0.2$. We use a linear beta schedule as AnimateDiff, where $\beta_{start} = 0.00085$ and $\beta_{end} = 0.012$.

Inference Settings. We use DDIM with 25 sampling steps, and the scale for text guidance is 7.5, the angle $\theta = \pi/2$. We collect several personalized Stable Diffusion models from CivitAI to verify the effectiveness and generalizability of our method, including Realistic Vision, Lyriel, ToonYou, and RCNZ Cartoon.

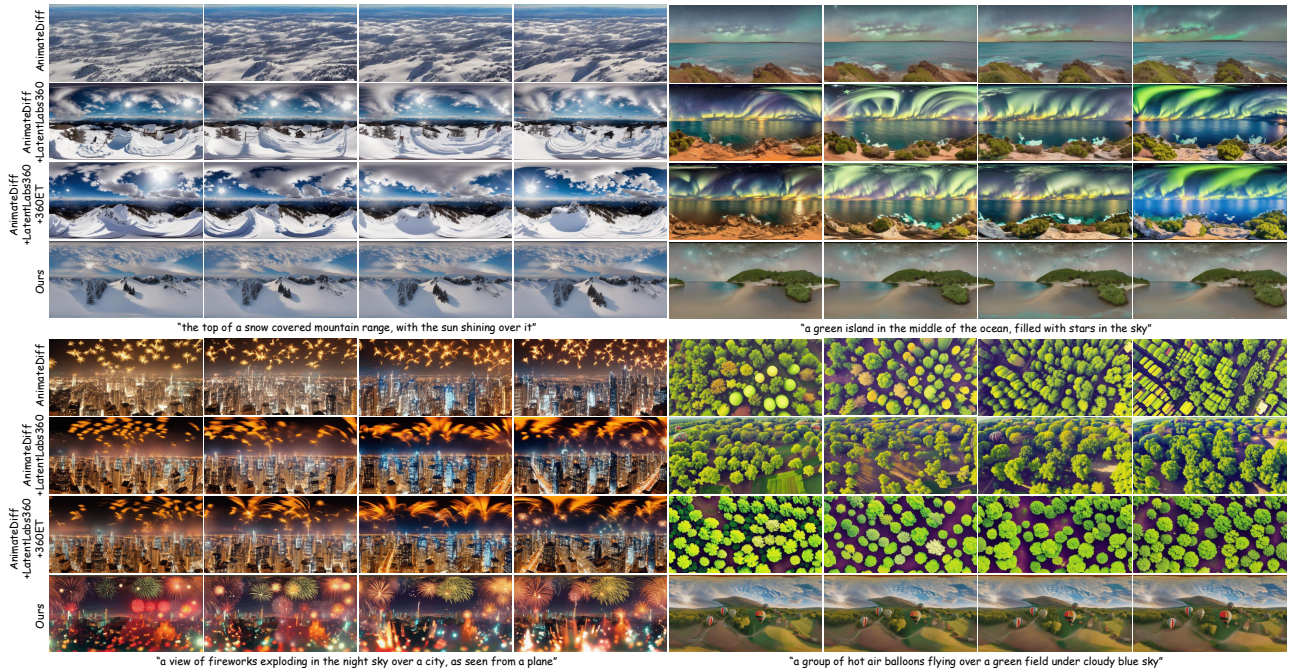

Figure 6. **Qualitative comparisons with baseline methods.** 360DVD successfully produces stable and high-quality panorama video over various prompts while other methods are failed.

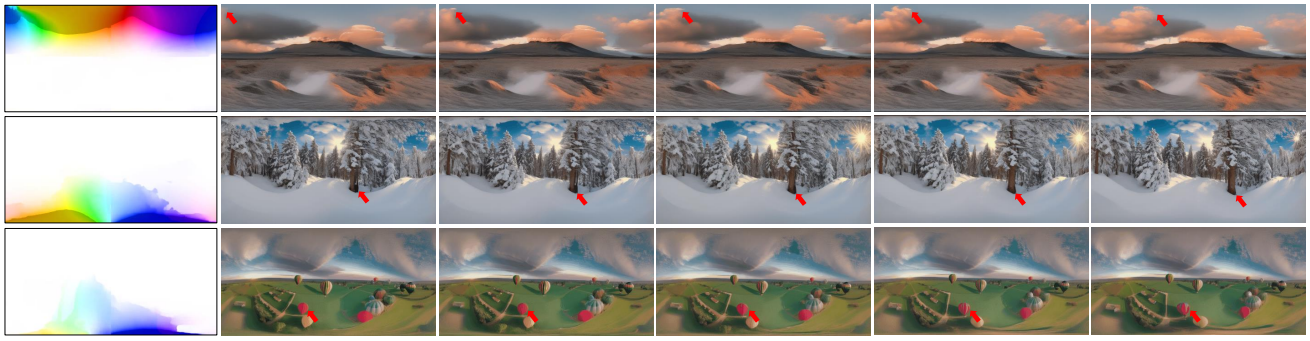

Figure 7. **Qualitative comparisons of optical flow.** 360DVD generates panorama videos with reasonable motion patterns consistent with the conditioned optical flow.

4.2. Qualitative Results

Due to space limitations, we only display several frames of each video. We strongly recommend readers refer to our project page for more results and better visual quality.

Prompt-guided Panorama Video Generation. We present several prompt-guided 360° panorama video generation results across different personalized models in Fig. 1. The figure shows that our method successfully turns personalized T2I models into panorama video generators. Our method can produce impressive generation results ranging from real to cartoon styles, from natural landscapes to cultural scenery. This success is attributed to the fact that our method preserves the image generation priors and temporal modeling priors learned by SD and AnimateDiff on large-scale datasets.

Motion-guided Panorama Video Generation. We showcase panoramic video generation results guided by three typical optical flow maps, as shown in Fig. 7. The optical flow maps in the first row indicate the primary motion areas in the Arctic, where we can observe significant movement of clouds in the sky. The optical flow maps in the second row and third row indicate motion areas primarily in the Antarctic, where we can see the movement of trees and hot air balloons near the Antarctic.

Index	Methods	Video Criteria		Panorama Criteria		
		Graphics Quality	Frame Consistency	End Continuity	Content Distribution	Motion Pattern
A	AnimateDiff	11.3%	15.3%	5.3%	4.8%	4.4%
B	A+LoRA	14.1%	10.5%	6.0%	12.1%	6.5%
C	B+360ET	23.0%	9.7%	16.9%	16.1%	14.5%
D	Ours	51.6%	64.5%	71.8%	67.0%	74.6%

Table 1. **User preference studies.** More raters prefer videos generated by our 360DVD, especially over panorama criteria including if generated videos have left-to-right continuity, the panorama content distribution, and the panorama motion pattern.

4.3. Comparison

We compare our results with native AnimateDiff, AnimateDiff with a LoRA for panorama image generation from CivitAI named LatentLabs360, AnimateDiff with panoramic LoRA, and our proposed 360 Enhancement Techniques (loss excepted). We can observe that the results generated by the native AnimateDiff have a very narrow field of view, which does not align with the content distribution of panoramic videos. When AnimateDiff is augmented with panoramic LoRA, it produces videos with a broader field of view; however, the two ends of videos lack continuity, and object movements are highly random. Our proposed 360ET method significantly enhances the continuity between two ends of the videos but fails to address issues such as non-compliance with panoramic motion patterns and poor cross-frame consistency. Notably, our 360DVD can generate videos that best adhere to the content distribution and motion patterns of panoramic videos. We are pleased to discover that, thanks to the high-quality training data provided by WEB360, the videos generated by 360DVD exhibit more realistic colors and nuanced lighting, providing an immersive experience.

4.4. Ablation Study

We primarily conducted ablation studies on the proposed 360 Text Fusion strategy, the pseudo-3D layer in the 360-Adapter, and the latitude-aware loss, as illustrated in Fig. 8. Given the prompt "a car driving down a street next to a forest", the first row without 360TF can not generate the car because of low-quality captions in the training process. The second row without pseudo-3D layer can generate a car, but due to the lack of temporal modeling, the results exhibit flickering. The third row without latitude-aware loss can produce relatively good results, but it still falls slightly short in terms of clarity, field of view, and other aspects compared to the last row with the complete 360DVD.

4.5. User Study

31 participants were surveyed to evaluate the graphics quality, cross-frame consistency, left-right continuity, content distribution, and motion patterns of 8 sets of generated results. For each criterion, they selected the video they deemed most fitting for the theme of high-quality 360-degree panoramic videos. The data presented in Table 1 indicates that our model outperforms the other three methods significantly across all five dimensions. Simultaneously, our proposed 360ET can remarkably improve video quality, and left-right continuity, solely based on the native AnimateDiff and panoramic LoRA.

Figure 8. **Ablation studies** on 360 Text Fusion (360TF), pseudo-3D layer in 360-Adapter (Pseudo-3D), and latitude-aware loss (Lat. Loss).

5. Conclusion

In this paper, we introduce 360DVD, a pipeline for controllable 360° panorama video generation. Our framework leverages text prompts and motion guidance to animate personalized T2I models. Utilizing the proposed WEB360 dataset, 360-Adapter, and 360 Enhancement Techniques, our framework can generate videos that adhere to the content distribution and motion patterns in real captured panoramic videos. Extensive experiments demonstrate our effectiveness in creating high-quality panorama videos with various prompts and styles. We believe that our framework provides a simple but effective solution for panoramic video generation, and leads to inspiration for possible future works.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Lin Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 1

[2] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11441–11450, 2022. 3

[3] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 1, 3

[4] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Xuhan Sheng, Bin Chen, Haoyu Ma, Ming Cheng, Shijie Zhao, Wanwan Cui, Tianyu Xu, Chunyang Li, Long Bao, Heng Sun, Huaibo Huang, Xiaoqiang Zhou, Yuang Ai, Ran He, Renlong Wu, Yi Yang, Zhilu Zhang, Shuohao Zhang, Junyi Li, Yunjin Chen, Dongwei Ren, Wangmeng Zuo, Qian Wang, Hao-Hsiang Yang, Yi-Chung Chen, Zhi-Kai Huang, Wei-Ting Chen, Yuan-Chun Chiang, Hua-En Chang, I-Hsiang Chen, Chia-Hsuan Hsieh, Sy-Yen Kuo, Zebin Zhang, Jiaqi Zhang, Yuhui Wang, Shuhao Cui, Junshi Huang, Li Zhu, Shuman Tian, Wei Yu, and Bingchun Luo. 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 (CVPR) Workshops*, pages 1731–1745, 2023. 3, 4

[5] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[6] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[7] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3

[8] Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Xuhan Sheng, Shijie Zhao, Junlin Li, and Li Zhang. Hybrid transformer and cnn attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1702–1711, 2023. 5

[9] Xinhua Cheng, Nan Zhang, Jiwen Yu, Yinhuai Wang, Ge Li, and Jian Zhang. Null-space diffusion sampling for zero-shot point cloud completion. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. 2

[10] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 3

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 2

[12] Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Jonathan Eisenmann, Siavash Khodadadeh, and Jean-François Lalonde. Guided co-modulated gan for 360° field of view extrapolation. In *2022 International Conference on 3D Vision (3DV)*, pages 475–485. IEEE, 2022. 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[15] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[18] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation. *arXiv preprint arXiv:2304.11603*, 2023. 3

[19] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 3

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[21] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4

[23] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4512–4521, 2019. 3

[24] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: To-

wards infinite-pixel image synthesis. *arXiv preprint arXiv:2104.03963*, 2021. 3

[25] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 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*, pages 4296–4304, 2024. 1, 3

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3

[28] Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, pages 352–371. Springer, 2022. 3

[29] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 3

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, page 234–241. 2015. 4

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[37] 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. *arXiv preprint arXiv:2209.14792*, 2022. 1, 3

[38] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[40] Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, et al. Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1293–1301, 2023. 3, 5

[41] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019. 3

[42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[43] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European Conference on Computer Vision*, pages 477–492. Springer, 2022. 3

[44] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4933–4943, 2024. 3

[45] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. *arXiv preprint arXiv:2308.14686*, 2023. 3, 6

[46] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 1, 3

[47] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning

of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[49] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 3

[50] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 2022. 3

[51] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Ipo-ldm: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. *arXiv preprint arXiv:2307.03177*, 2023. 3

[52] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Y He, H Liu, H Chen, X Cun, X Wang, Y Shan, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[53] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[54] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[55] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023. 3

[56] Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. CRoSS: Diffusion model makes controllable, robust and secure image steganography. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[59] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10188–10198. IEEE, 2023. 3

[60] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1, 3