DreamDrone: Text-to-Image Diffusion Models are Zero-shot Perpetual View Generators

Hanyang Kong[1], Dongze Lian[1], Michael Bi Mi[2], and Xinchao Wang[1⋆]

[1] National University of Singapore, Singapore
[2] Huawei International Pte. Ltd., Singapore
hanyang.k@u.nus.edu, dzlianx@gmail.com, xinchao@nus.edu.sg

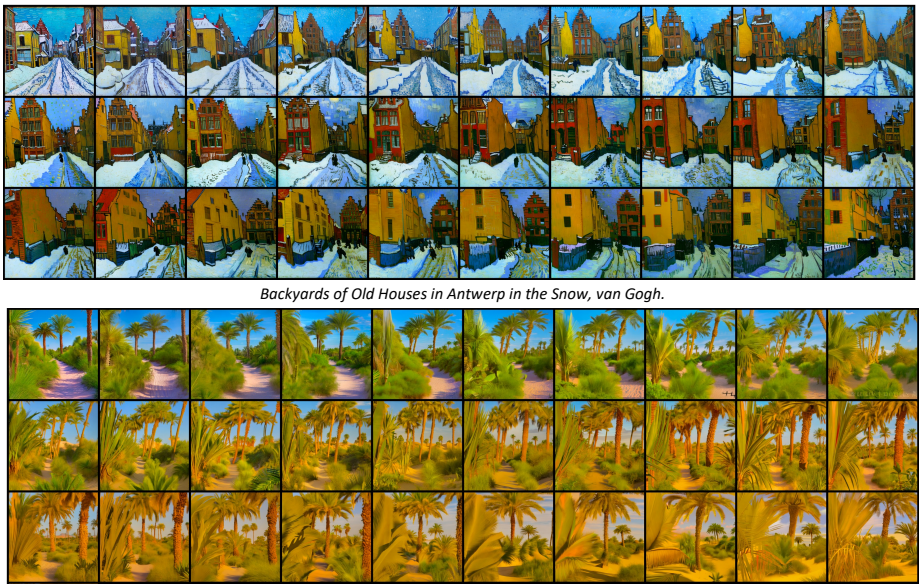

Fig. 1: Visualization results of DreamDrone. Given a single scene image and the textual description, our approach generates novel views corresponding to user-defined camera trajectory, without fine-tuning on any dataset or reconstructing the 3D point cloud in advance.

Abstract. We introduce *DreamDrone*, a novel zero-shot and training-free pipeline for generating unbounded flythrough scenes from textual prompts. Different from other methods that focus on warping images frame by frame, we advocate explicitly warping the intermediate latent code of the pre-trained text-to-image diffusion model for high-quality image generation and generalization ability. To further enhance the fidelity of the generated images, we also propose a feature-correspondence-guidance diffusion process and a high-pass filtering strategy to promote geometric consistency and high-frequency detail consistency, respectively.

⋆ Corresponding author.

Extensive experiments reveal that *DreamDrone* significantly surpasses existing methods, delivering highly authentic scene generation with exceptional visual quality, without training or fine-tuning on datasets or reconstructing 3D point clouds in advance.

1 Introduction

Recent advances in vision and graphics have enabled the synthesis of multi-view consistent 3D scenes along extended camera trajectories [3, 7, 18, 20]. This emerging task, termed *perpetual view generation* [20], involves synthesizing views from a flying camera along an arbitrarily long trajectory, starting from a single RGBD image.

Previous methodologies predominantly engage in warping images frame by frame with traditional 3D geometric knowledge when given RGBD images and subsequent camera extrinsic. However, this operation often leads to blurriness and distortion in images, which arises from inaccurate interpolation, the mismatch between discrete pixels and continuous transformations, and inaccurate depth data. Moreover, such blurriness and distortion tend to amplify with the accumulation of warp operations.

To further alleviate the errors caused by frame-by-frame warp operations, two primary paths have been proposed. i) Some methods [3, 18, 20] try to train a refiner on natural scene datasets. The advantage of this frame-by-frame approach is that it allows for arbitrary changes in camera trajectory during the scene generation process, offering users a higher degree of freedom and enabling infinite generation. However, this training-based method can only be used in natural scenes and cannot be generalized to arbitrary indoor/outdoor scenes or scenes of various styles. ii) Another solution is to first reconstruct the 3D scene model using text prompts, then render 2D RGB images according to the camera trajectory [7, 10, 52]. Although this solution yields more coherent 2D image sequences, the quality of the rendered images highly depends on the quality of the 3D scene model. This method cannot guarantee good rendering effects from every viewpoint. Additionally, since this method requires the reconstruction of 3D point clouds, it cannot achieve "infinite" scene generation in the same way as the frame-by-frame strategy.

In this paper, we advocate that a more general and flexible perpetual view generation pipeline should possess the following capabilities:

i) are versatile across diverse scenes, including indoor and outdoor scenes, as well as scenes depicted in various styles; ii) allow users to interactively control the camera trajectory during the process of scene generation, while ensuring the high quality of the generated images and the semantic consistency between adjacent frames; and iii) enable seamless transitions from one scene to another.

To this end, we introduce DreamDrone, a novel zero-shot, training-free, infinite scene generation pipeline from text prompts, which does not require any optimization or fine-tuning on any dataset. A core principle of our approach is to warp the latent code of a pre-trained text-to-image diffusion model rather than

the frames, enriching it with temporal and geometric consistency. To be specific, given RGBD image I of the current view and camera rotation $\mathcal{R}$ and translation $\mathcal{T}$ for the next view (which is interactively defined by users), we first obtain the latent code x_{t_1} of the diffusion model at timestep t_1, warp it to latent code of the next view x'_{t_1} based on $\mathcal{R}$ and $\mathcal{T}$, and denoise x'_{t_1} to the image I' of the next view. To ensure geometry consistency across adjacent views, we propose a novel feature-correspondence-guidance diffusion process when denoising from x'_{t_1} to image at the next view I'. Moreover, we propose a novel high-pass filter mechanism when warping the latent code x_{t_1}, for preserving high-frequency details across adjacent views.

Our experiments demonstrate that the proposed DreamDrone effectively leads to high-quality and geometry-consistent scene generation. Quantitative and qualitative results demonstrate our comparable, even superior performance compared with other training-based and training-free methods from the aspects of temporal consistency and image quality. Moreover, the significant advantage of DreamDrone is its versatility: it is adept not only at generating real-world scenarios but also shows promising capabilities in creating imaginative scenes. Additionally, users can interactively control the camera trajectory (Fig. 5) and shuttle from one scene to another (Fig. 4). Our contributions are summarized as follows:

- To our best knowledge, we are the first attempt to generate novel views by explicitly warping the latent code of the pre-trained diffusion model.
- A novel feature-correspondence-guidance diffusion process is proposed to enforce geometry consistency across adjacent views. Moreover, a high-pass filtering strategy is introduced to preserve high-frequency details for novel views.
- Extensive experiments demonstrate that our method generates high-quality and geometry-consistent novel views for any scene, from realistic to fantastical. More interestingly, our method realizes the scene shuttle, *i.e.*, travels from one scene to another when the user controls the camera trajectory.

2 Related Works

Perpetual view generation. Perpetual view generation extrapolates unseen content outside a single image. InfNat [20], InfNat-0 [18], and DiffDreamer [3] use iterative training for long-trajectory perpetual view extrapolation. InfNat [20] pioneered the *perpetual view generation* task with a database for infinite 2D landscapes. InfNat-0 [18] adapted this to 3D, introducing a render-refine-repeat phase for novel views. DiffDreamer [3] improved consistency with image-conditioned diffusion models. However, these methods lack robustness in complex and urban environments. In very recent concurrent work, SceneScape [7] and WonderJourney [52] firstly generate 3D point cloud for scene by zoom-out and inpainting strategy. 2D image sequences are further rendered based on the reconstructed 3D point cloud. However, the accuracy of the 3D model critically impacts performance, particularly with novel camera trajectories.

Text-to-3D generation. Several text-to-3D generation methods [1,4,15,28,30,54] apply text-3D pair databases to learning a mapping function. However, supervised strategies remain challenging due to the lack of large-scale aligned text-3D pairs. CLIP-based [33] 3D generation methods [12, 13, 16, 29, 56] apply pre-trained CLIP model to create 3D objects by formulating the generation as an optimization problem in the image domain. Recent text-to-3D methods like [19, 25, 26, 31, 38, 47, 50] blend text-to-image diffusion models [36] with neural radiance fields [27] for training-free 3D object generation. Other approaches [21, 22, 35, 41] focus on novel view synthesis from a single image, often limited to single objects or small camera motion ranges. Text2room [10] generates 3D indoor scenes from text prompts, but is confined to room meshes.

Text-to-video generation. Generating videos from textual descriptions [2, 9, 11, 24, 39, 40, 46, 55] poses significant challenges, primarily due to the scarcity of high-quality, large-scale text-video datasets and the inherent complexity in modeling temporal consistency and coherence. CogVideo [11] addresses this by incorporating temporal attention modules into the pre-trained text-to-image model CogView2 [6]. The video diffusion model [9] employs a space-time factorized U-Net, utilizing combined image and video data for training. Video LDM [2] adopts a latent diffusion approach for generating high-resolution videos. However, these methods typically do not account for the underlying 3D scene geometry in scene-related video generation, nor do they offer explicit control over camera movement. Additionally, their reliance on extensive training with large datasets can be prohibitively costly. While T2V-0 [14] introduced the concept of zero-shot text-to-video generation, its capability is limited to generating a small number of novel frames, with diminished quality in longer video sequences.

3 Method

We formulate the task of perpetual view generation as follows: given a starting image I, we generate the next view image I' corresponding to an arbitrary camera pose $\{\mathcal{R}, \mathcal{T}\}$, where the camera pose can be specified or via user's control.

3.1 Preliminaries

We implement our method based on the recent state-of-the-art text-to-image diffusion model (*i.e.* Stable Diffusion [36]). Stable diffusion is a latent diffusion model (LDM), which contains an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$ and a U-Net [37] denoiser. Diffusion models are founded on two complementary random processes. The *DDPM forward* process, in which Gaussian noise is progressively added to the latent code of a clean image: $\boldsymbol{x}_0$:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{z}, \tag{1}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$ and $\{\alpha_t\}$ are the noise schedule.

The *backward* process is aimed at gradually denoising $\boldsymbol{x_T}$, where at each step a cleaner image is obtained. This process is achieved by a U-Net $\boldsymbol{\epsilon}_\theta$ that predicts the added noise $\boldsymbol{z}$. Each step of the backward process consists of applying $\boldsymbol{\epsilon}_\theta$ to the current $\boldsymbol{x}_t$, and adding a Gaussian noise perturbation to obtain a cleaner $\boldsymbol{x}_{t-1}$.

Classifier-guided DDIM sampling [5] aims to generate images from noise conditioned on the class label. Given the diffusion model $\boldsymbol{\epsilon}_\theta$, the latent code $\boldsymbol{x}_t$ at timestep t, the classifier $p_\theta(y|x_t)$, and the gradient scale s, the sampling process for obtaining $\boldsymbol{x}_{t-1}$ is formulated as:

$$\hat{\epsilon} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t) - \sqrt{\bar{\alpha}_{t-1}} \nabla_{\boldsymbol{x}_t} \log p_\phi(y|\boldsymbol{x}_t), \tag{2}$$

and

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}, \tag{3}$$

where α is the denoise schedule.

In the self-attention block of the U-Net, features are projected into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. The output of the block $\boldsymbol{o}$ is obtained by:

$$\boldsymbol{o} = \mathbf{AV}, \quad \text{where } \mathbf{A} = \text{Softmax}(\mathbf{QK}^\top) \tag{4}$$

The self-attention operation allows for long-range interactions between image tokens.

3.2 Overview

Perpetual view generation as the camera moves presents a complex challenge. This process involves seamlessly filling in unseen regions caused by image warping, adding details to objects as they come closer, while ensuring the imagery remains realistic and diverse. Prior works [3, 18, 20] have focused on training a refiner to enhance details and create new content for areas requiring inpainting or outpainting. These efforts have shown promising outcomes, yet the effectiveness of the refiner is generally limited to scenarios that align with the training dataset.

Since diffusion models can generate high-quality large-variety images from random latent code, a direct solution arises: can we modify the powerful pre-trained text-to-image diffusion model as a refiner? Empirically, DDIM inversion strategy [14, 43] can obtain the intermediate latent code at each timestep and the image can be reconstructed by those latent codes. To this end, we attempt to explicitly warp the latent code of the current view and generate the novel view by the pre-trained text-to-image diffusion model.

Our overall pipeline is illustrated in Fig. 2. Initially, we obtain the latent code x_{t_1} of the current view's RGB image I at timestep t_1 through the DDIM inversion process. We then warp the current frame's latent code x_{t_1} to the next view x'_{t_1} using depth information and camera extrinsic parameters. However, directly denoising from x'_{t_1} to the image also suffers from blurry, which results in the non-integer pixel coordinates and the interpolation operation. More

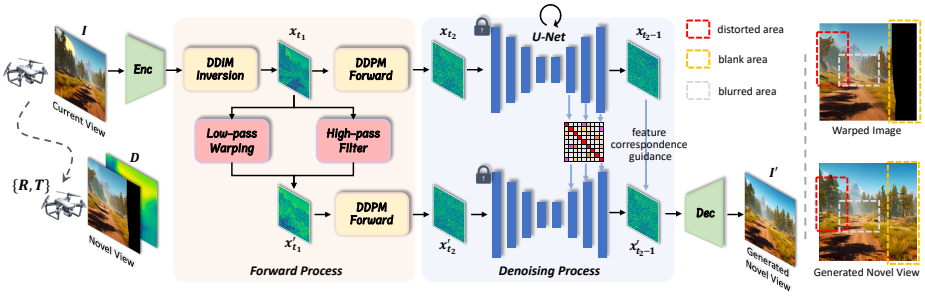

Fig. 2: Overview of our proposed pipeline. Starting from a real or generated RGBD (I, D) image at the current view, we apply DDIM inversion to obtain intermediate latent code x_{t_1} at timestep t_1 using a pre-trained U-Net model. A warping with the high-pass filter strategy is applied to generate latent code for the next novel view. A few more DDPM forward steps from timestep t_1 to t_2 are applied for enlarging the degree of freedom *w.r.t.* the warped latent code. In the denoising process, we apply pre-trained U-Net to generate the novel view from x'_{t_2}. The cross-view self-attention module and feature-correspondence guidance are applied to maintain the geometry correspondence between x_{t_2} and x'_{t_2}. The right side shows the warped image and our generated novel view I'. Our method greatly alleviates blurring, inconsistency, and distortion. The overall pipeline is zero-shot and training-free.

noise is added from x'_{t_1} at timestep t_1 to x'_{t_2} at timestep t_2 by DDPM forward operation, for generating high-quality images. The side effect of DDPM is the geometry inconsistency between adjacent views. To this end, we propose a feature-correspondence-guidance denoising strategy to enforce geometry consistency. Moreover, a high-pass filtering strategy is proposed to maintain the consistency of high-frequency details between adjacent views. Please refer to Fig. 3 for the motivation of our proposed modules. Our overall pipeline requires only a pre-trained text-to-image diffusion model and a depth estimation model, eliminating the need for any additional training or fine-tuning.

3.3 Warping latent codes

Algorithm 1 Warping latent code with high-pass filter

Require: $\boldsymbol{x}_t$ ▷ latent code at timestep t of current view c
$F(\boldsymbol{x}_t) \leftarrow FFT(\boldsymbol{x}_t)$ ▷ Apply Fast Fourier Transform
Split $F(\boldsymbol{x}_t)$ into F_{low} and F_{high} using threshold σ
$\boldsymbol{x}_t^{low} \leftarrow IFFT(F_{low})$ ▷ Inverse FFT on low-frequency component
$\boldsymbol{x}_t^{low-warped} \leftarrow warp(\boldsymbol{x}_t^{low})$ ▷ warp the low-frequency content
$F_{warped} \leftarrow FFT(\boldsymbol{x}_t^{low-warped})$ ▷ FFT on warped content
$F' \leftarrow F_{warped} + F_{high}$ ▷ Combine low-frequency of warped content with high-frequency of original content
$\boldsymbol{x}'_t \leftarrow IFFT(F')$ ▷ Inverse FFT to get latent code for next view c'
return $\boldsymbol{x}'_t$ ▷ warped latent code at timestep t for next view c'

The results in the right side of Fig. 2 reveal that directly warping images based on camera intrinsics K, extrinsics $\{\mathcal{R}, \mathcal{T}\}$, and depth information leads to regions of distortion in the images. Additionally, the use of inpainting [23, 36, 53] and outpainting [17, 49, 51] models to fill these gaps does not achieve satisfactory outcomes. In pursuit of photo-realistic images, we opt to edit the latent code corresponding to timestep t. PnP [44] and DIFT [42] have shown that the features of diffusion possess strong semantic information, with semantic parts being shared across images at each step. The simplest method for warping the latent code follows the same approach as warping the image. The only difference between warping the latent code and warping the image is a slight modification in the camera intrinsics; this entails scaling the camera intrinsics proportionally based on the different resolutions of the image and latent code.

The overall procedure for warping the latent code is illustrated in Alg. 1. Initially, a latent code x_t is obtained and transformed via Fast Fourier Transform (FFT) to $F(\boldsymbol{x}_t)$. This is divided into low-frequency F_{low} and high-frequency F_{high} components, segregated at threshold σ. The key step involves warping the Inverse FFT (IFFT) processed low-frequency component $\boldsymbol{x}_t^{\text{low}} = \text{IFFT}(F_{\text{low}})$, warping to the next view $\boldsymbol{x}_t^{low-warped}$. Merging $\text{FFT}(\boldsymbol{x}_t^{low-warped})$ with F_{high}, we obtain F', from which the final latent code $\boldsymbol{x}_t' = \text{IFFT}(F')$ is reconstructed. This approach efficiently preserves high-frequency details, enabling high-fidelity scene generation aligned with text prompts.

3.4 Feature-correspondence-guidance design

After obtaining the latent code x_t' corresponding to the next frame, we employ the DDPM (Denoising Diffusion Probabilistic Models) method to increase the degrees of freedom of the latent code, enabling the generation of richer image details. However, increasing freedom introduces a challenge: the correlation between frames. An unconstrained diffusion denoising process can result in poor semantic correlation between adjacent frames. To address this, we propose a feature-correspondence guidance strategy with a cross-view self-attention mechanism. We introduce these approaches in detail below.

Cross-view self-attention. To maintain consistency between the generated result and the original image, inspired by recent image and video editing works [8, 44, 45, 48], we modify the process of the self-attention module of U-Net when denoising the latent code $\boldsymbol{x}_t'$. Specifically, we denoise the views for the current and next view together. The key and value of the self-attention modules from the next view are replaced by that of the current view. To be specific, for obtaining the original view, the self-attention module is defined the same as Eq. (4). The modified cross-view self-attention for generating a novel view is defined as:

$$\boldsymbol{o}' = \mathbf{A}'\mathbf{V}, \quad \text{where } \mathbf{A}' = \text{Softmax}(\mathbf{Q}'\mathbf{K}^\top), \tag{5}$$

where $\mathbf{Q}'$, $\mathbf{A}'$, and $\boldsymbol{o}'$ are query, attention matrix, and output features for the novel views. $\mathbf{K}$ and $\mathbf{V}$ are injected keys and values obtained from the self-

attention module for generating the original view. Please note that the **K** and **V** are also warped before injection.

Feature-correspondence guidance. Maintaining geometry consistency between adjacent views using the cross-view self-attention mechanism presents challenges, especially in preserving high-frequency details as the camera moves forward. The recent DIFT [42] highlights the potential of using intermediate features of diffusion models for accurate point-to-point image matching [32]. Additionally, the concept of vanilla classifier guidance [5] steers the diffusion sampling process using pre-trained classifier gradients towards specific class labels. Building on these ideas, we integrate feature correspondence guidance into the DDIM sampling process to enhance consistency between adjacent views, addressing the challenge of detail preservation in dynamic scenes.

Specifically, we obtain the features of the current and next view at each timestep t of the DDIM process and calculate the cosine distance between the warped original features and features from the next novel views:

$$\mathcal{L}_{sim}^{t} = \frac{1 - \cos\left[\text{warp}(f_t), f_t'\right]}{2}, \tag{6}$$

where f_t and f_t' are intermediate features extracted from pre-trained U-Net $\boldsymbol{\epsilon}_\theta$ at timestep t and warp is the warping functions. The lower $\mathcal{L}_{sim}^{t}$, the higher the similarity.

We further introduce the similarity score $\mathcal{L}_{sim}^{t}$ to the DDIM sampling process, for generating novel views with geometry consistency. The predicted noise $\hat{\epsilon}$ is formulated as:

$$\hat{\epsilon} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t) - \lambda\sqrt{\bar{\alpha}_{t-1}}\nabla_{\boldsymbol{x}_t}\mathcal{L}_{sim}^{t}, \tag{7}$$

where λ is the constant hyper-parameter and latent code $\boldsymbol{x}_{t-1}$ is calculated by Eq. (3)

4 Experiments

4.1 Implementation details

We take Stable Diffusion [36] with the pre-trained weights from version 2.1[3] as the basic text-to-image diffusion and MiDas [34] with weights dpt_beit_large_512[4]. The overall diffusion timesteps is 1000. We warp the latent code at timestep t_1=21 and add more degrees of noise to timestep t_2=441. The threshold σ for the high-pass filter is 20 and the hyper-parameter λ for feature-correspondence guidance is 300. Due to the page limit, please refer to the supplementary material (*supp.*) for details.

[3] https://huggingface.co/stabilityai/stable-diffusion-2-1-base
[4] https://github.com/isl-org/MiDaS

4.2 Baselines

We compare against 1) two supervised methods for perpetual view generation: *InfNat* [20] and *InfNat-0* [18]. 2) one text-conditioned 3D point cloud-based scene generation: *SceneScape* [7]. 3) two supervised methods for text-to-video generation: *CogVideo* [11] and *VideoFusion* [24]. 4) one method for zero-shot text-to-video generation: *T2V-0* [14].

4.3 Evaluation metrics

We evaluate our zero-shot perpetual scene generation into two aspects: 1) the quality of generated images and text-image alignment, and 2) the temporal consistency of generated image sequences.

Image quality and text-image alignment. We evaluate CLIP score [33], which indicates text-scene alignment for quantitative comparisons. A high average CLIP score indicates not only that the generated images are more aligned with the corresponding prompts but also that they consistently maintain high quality [14]. CogVideo [11], VideoFusion [24], SceneScape [7], and T2V-0 [14] are all engaged in text-conditioned generation tasks. We generated 50 scene-related text prompts using GPT-4[5] and then created videos using each of the three methods. For the InfNat [20] and InfNat-0 [18] methods, we used Stable Diffusion to generate the initial frame, followed by subsequent frame generation based on this initial frame. We calculated the distance between each generated frame and the text embedding, known as the CLIP score. Considering that the InfNat [20] and InfNat-0 [18] methods trained on natural scene datasets, we further provided 10 very general prompts such as 'an image of the landscape' and 'an image of the mountain' for these methods, and then selected the highest CLIP score as the CLIP score for the current frame.

Table 1: Ablations of image quality and temporal coherence of generated image sequences with various lengths. Please refer to Fig. 3 for quality comparisons.

Methods	PSNR ↑			SSIM ↑			CLIP ↑		
	8 frames	16 frames	32 frames	8 frames	16 frames	32 frames	8 frames	16 frames	32 frames
warp image	26.90	22.46	21.62	0.25	0.23	0.24	0.138	0.112	0.106
warp latent	28.35	28.57	28.75	0.27	0.28	0.24	0.135	0.122	0.125
warp latent+DDPM	24.67	23.04	22.59	0.12	0.10	0.06	0.302	0.297	0.308
warp latent+DDPM+guidance	28.27	28.21	28.10	0.34	0.30	0.26	0.317	0.316	0.313
warp latent+DDPM+guidance +cross-view attn.	28.89	28.83	28.75	0.32	0.31	0.27	0.318	0.315	0.315
warp latent+DDPM+guidance +cross-view attn.+high pass filter	29.91	29.86	29.79	0.39	0.38	0.35	0.320	0.318	0.319

Temporal consistency of generated image sequences. We demonstrate our advancements in temporal consistency against other SOTA methods by calculating average PSNR and SSIM scores across adjacent frames for generated videos with different lengths. The higher scores demonstrate the superiority in terms of cross-view consistency.

[5] https://openai.com/gpt-4

4.4 Ablation studies

We perform ablation studies on our three proposed modules: 1) warping latent with high-pass filter, 2) cross-view self-attention module, and 3) feature-correspondence guidance. The quantitative ablation results are shown in Tab. 1 and we visualize the ablation samples in Fig. 3.

Fig. 3: Ablation results for the key components. We perform ablation studies by disabling the key components of our method. We illustrate every five frames for each ablation experiment. Please zoom in for better comparisons.

The simplest method for infinite scene generation tasks is frame-by-frame image warping, but this approach is unfeasible, as is directly warping the latent code. Warping images leads to non-integer pixel coordinates, resulting in interpolation-induced blurring and distortion. Moreover, these errors accumulate with each frame generated, leading to a collapse in quality. The first two rows of Table 1 show that directly warping images or warping latent codes (i.e., removing DDPM) results in very low CLIP scores, indicating poor quality of the generated images. The generated images becoming progressively blurred can also be observed in the first two rows of Fig. 3. We introduce DDPM to increase the degrees of freedom of the diffusion model, thereby generating high-quality images. However, the introduction of DDPM has the side effect of worsening the semantic consistency between adjacent frames (3^{rd} row in Tab. 1 and Fig. 3).With the help of DDPM, the CLIP score increases from 0.125 to 0.308 when generating 32 images. Please refer to the *supp.* for the generated results with different scales of the DDPM forward process.

To ensure the quality of image generation while also maintaining consistency with adjacent views, we propose a feature-correspondence guidance strategy. Comparing the third and fourth rows of Fig. 3, it is evident that the semantic consistency between adjacent frames is significantly enhanced after adding guid-

ance, with noticeable improvements in both PSNR and SSIM scores in Tab. 1. To further enhance cross-view consistency, we adopted cross-view attention modules and high-pass filtering. From the visualized results at the 5^{th} and 6^{th} rows in Fig. 3, it is clear that the semantic consistency of adjacent camera perspectives is further strengthened after incorporating the cross-view attention module. The operation of the high-pass filter further preserves the high-frequency details of the current frame, thereby further enhancing the semantic consistency of high-frequency details between adjacent frames. For instance, comparing the left side house at the 4^{th}, 5^{th}, and 6^{th} rows in Fig. 3, the cross-view consistency is enhanced after adding the proposed modules.

In addition to conducting ablation experiments on the modules we propose, we continue to explore two more questions:

Fig. 4: Ablation study for scene travel. We visualize two image sequences and change the prompt when generating novel views. We illustrate every five images and the prompts are changed when generating 31^{th} image (7^{th} image shown in each row).

Q1: Can DreamDrone shuttle from one scene to another by changing text prompts? During the frame-by-frame process, we changed the textual prompts, with the generated results shown in Fig. 4. The visualized results demonstrate that DreamDrone can smoothly complete the scene travel (from streets in Inazuma City to urban art street) or the transition of scene styles (from realistic to Lego style) while ensuring the semantic consistency of adjacent views, according to the changes in textual prompts.

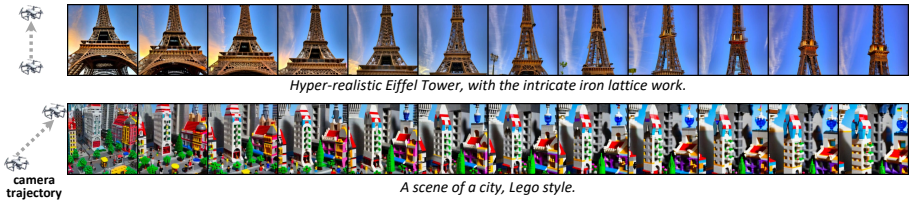

Fig. 5: Ablation study on customized camera trajectory. We generate images with different camera directions. For the sample of the Eiffel Tower, our camera perspective continuously ascends. For the 2^{nd} scene of the Lego city, our camera not only moves forward but also shifts upwards and to the right.

Q2: Can explicitly warping the latent code control the trajectory of camera perspective movement? Since our method generates image sequences frame by frame, we can freely adjust the camera's flight angle by altering the camera's extrinsic parameters. In Fig. 5, we provide sequences of images generated under

different camera trajectories. The results show that our method possesses a high degree of freedom, allowing for the free customization of the camera's trajectory. Other state-of-the-art methods cannot achieve this functionality.

4.5 Qualitative comparison

Fig. 6: Qualitative comparisons of InfNat-0 [18] and ours. We provide four starting scene images with various styles and categories as start points and ask models to fly through the images. 50 frames are generated and we illustrate every five frames for each starting scene image.

In our comparison with InfNat-0 [18] (Fig. 6), focusing on various scenes including coastlines, rivers, Van Gogh-style landscapes, and city streetscapes, we identified four main differences: Firstly, InfNat-0 shows proficiency in coastline scenes, a reflection of its training data, but our training-free *DreamDrone* surpasses it in later frames due to InfNat-0's cumulative errors over time. Secondly, in natural scenes with closer objects, InfNat-0's flawed generation becomes more apparent, whereas our method maintains consistency. Thirdly, InfNat-0's limited approach to gap filling leads to poor performance in stylized scenes, in contrast to *DreamDrone* which preserves high-frequency details and frame correspondence. Finally, in urban environments, InfNat-0 struggles significantly, while *DreamDrone* achieves realistic and geometry-consistent views, demonstrating its versatility across varied scenarios.

T2V-0 [14] introduces unsupervised text-conditioned video generation using stable diffusion. SceneScape [7] focuses on 'zoom out' effects during backward camera movement. However, as seen in Fig. 7, both methods have limitations.

SceneScape struggles with outdoor scenes and forward camera movement, leading to blurred and distorted results after 8 steps due to its reliance on a pre-trained inpainting model. T2V-0 displays a drop in quality beyond the third frame in complex environments like Lego-style cities, likely from its latent code editing approach that compromises frame continuity and geometric consistency. Conversely, our *DreamDrone* excels across various scenes. It maintains detail, continuity, and quality in advancing camera scenarios, evident in even simpler landscapes like mountains where T2V-0 and SceneScape cannot effectively portray dynamic elements like cloud movement. Our approach ensures the preservation of fine details such as shadows and sunlight, creating a more dynamic and realistic video experience. Please refer to *supp.* for more comparisons.

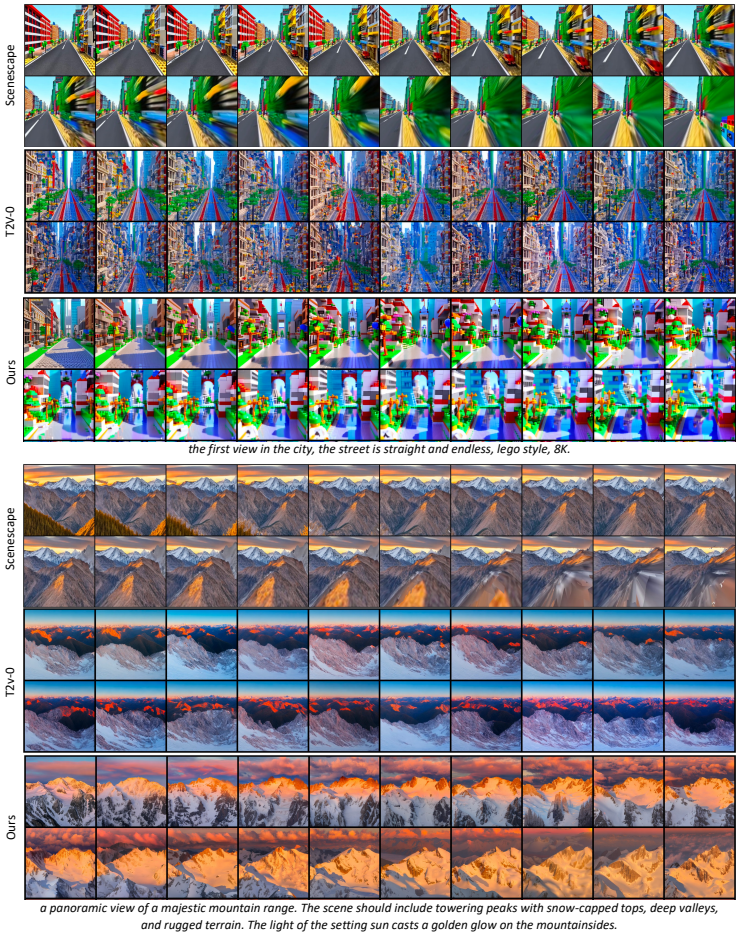

Fig. 7: Qualitative comparisons of SceneScape [7], T2V-0 [14], and our DreamDrone. We visualize 20 continuous frames for each textual prompt. As the camera flies, our method generates geometry-consistent scene sequences.

As our task bears similarities to text-to-video generation, we further provide qualitative comparisons with VideoFusion [24]. Due to the page limit, please refer to *supp.* for detailed comparisons.

4.6 Quantitative comparison

Tab. 2 offers a detailed comparison of various SOTA methods for generating image sequences, including our method, DreamDrone. When compared to other training-based methods, DreamDrone, despite being training-free, consistently achieves higher CLIP scores across all frame lengths (0.320, 0.318, 0.319 for 8, 16, and 32 frames respectively). This is particularly noteworthy as the CLIP scores for training-based methods generally degrade as the number of generated frames increases. For instance, VideoFusion's [24] CLIP scores decrease from 0.281 for 8 frames to 0.272 for 32 frames. This trend suggests a decline in the quality of generated images with an increase in sequence length for training-based methods.

Table 2: Qualitative comparisons with other SOTA methods. We evaluate the quality and temporal coherence of the generated image sequences with various lengths.

	Methods	PSNR ↑			SSIM ↑			CLIP ↑		
		8 frames	16 frames	32 frames	8 frames	16 frames	32 frames	8 frames	16 frames	32 frames
training-based	InfNat [20]	28.75	28.67	28.65	0.32	0.30	0.30	0.125	0.123	0.118
	InfNat-0 [18]	28.92	28.89	28.87	0.37	0.35	0.34	0.128	0.125	0.122
	CogVideo [11]	31.03	30.08	29.32	0.45	0.39	0.31	0.255	0.249	0.241
	VideoFusion [24]	29.89	28.36	28.78	0.41	0.37	0.31	0.281	0.283	0.272
training-free	T2V-0 [14]	27.25	26.17	26.03	0.27	0.24	0.23	0.312	0.305	0.287
	Scenescape [7]	29.87	29.75	29.66	0.41	0.38	0.34	0.318	0.282	0.279
	DreamDrone (Ours)	29.91	29.86	29.79	0.39	0.38	0.35	0.320	0.318	0.319

In contrast, DreamDrone maintains high CLIP scores even as the sequence length increases, indicating superior image quality. When compared to other training-free methods, DreamDrone also stands out. For example, while T2V-0's [14] CLIP scores decrease from 0.312 for 8 frames to 0.287 for 32 frames, DreamDrone's CLIP scores remain relatively stable, further demonstrating its robustness in maintaining image quality across varying sequence lengths. This analysis underscores the effectiveness of DreamDrone in generating high-quality, temporally coherent image sequences without the need for training.

5 Conclusion

In this work, we propose *DreamDrone*, a novel approach for generating flythrough scenes from textual prompts without the need for training or fine-tuning. Our method explicitly warps the intermediate latent code of a pre-trained text-to-image diffusion model, enhancing the quality of the generated images and the generalization ability. We propose a feature-correspondence-guidance diffusion process and a high-pass filtering strategy to ensure geometric and high-frequency detail consistency. Experimental results indicate that *DreamDrone* surpasses current methods in terms of visual quality and authenticity of the generated scenes.

Acknowledgement

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

1. Bautista, M.A., Guo, P., Abnar, S., Talbott, W., Toshev, A., Chen, Z., Dinh, L., Zhai, S., Goh, H., Ulbricht, D., et al.: Gaudi: A neural architect for immersive 3d scene generation. Advances in Neural Information Processing Systems **35**, 25102–25116 (2022)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
3. Cai, S., Chan, E.R., Peng, S., Shahbazi, M., Obukhov, A., Van Gool, L., Wetzstein, G.: Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2139–2150 (2023)
4. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 100–116. Springer (2019)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. Advances in Neural Information Processing Systems **35**, 16890–16902 (2022)
7. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. arXiv preprint arXiv:2302.01133 (2023)
8. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
9. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
10. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023)
11. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
12. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
13. Jiang, Z., Lu, G., Liang, X., Zhu, J., Zhang, W., Chang, X., Xu, H.: 3d-togo: Towards text-guided cross-category 3d object generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1051–1059 (2023)
14. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)

15. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14806–14816 (2023)
16. Lee, H.H., Chang, A.X.: Understanding pure clip guidance for voxel grid nerf models. arXiv preprint arXiv:2209.15172 (2022)
17. Li, J., Bansal, M.: Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. arXiv preprint arXiv:2305.19195 (2023)
18. Li, Z., Wang, Q., Snavely, N., Kanazawa, A.: Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In: European Conference on Computer Vision. pp. 515–534. Springer (2022)
19. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
20. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14458–14467 (2021)
21. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
22. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
23. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
24. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)
25. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)
26. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
28. Mo, S., Xie, E., Chu, R., Yao, L., Hong, L., Nießner, M., Li, Z.: Dit-3d: Exploring plain diffusion transformers for 3d shape generation. arXiv preprint arXiv:2307.01831 (2023)
29. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia 2022 conference papers. pp. 1–8 (2022)
30. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)

31. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
32. Qiu, J., Wang, X., Fua, P., Tao, D.: Matching seqlets: An unsupervised approach for locality preserving sequence matching. IEEE transactions on pattern analysis and machine intelligence **43**(2), 745–752 (2019)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
34. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
35. Rockwell, C., Fouhey, D.F., Johnson, J.: Pixelsynth: Generating a 3d-consistent experience from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14104–14113 (2021)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
38. Shen, Q., Yang, X., Wang, X.: Anything-3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261 (2023)
39. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
40. Tan, Z., Yang, X., Liu, S., Wang, X.: Video-infinity: Distributed long video generation. arXiv preprint arXiv:2406.16260 (2024)
41. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023)
42. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881 (2023)
43. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1921–1930 (June 2023)
44. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
45. Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
46. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
47. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)

48. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: 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. pp. 7623–7633 (2023)
49. Yang, C.A., Tan, C.Y., Fan, W.C., Yang, C.F., Wu, M.L., Wang, Y.C.F.: Scene graph expansion for semantics-guided image outpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15617–15626 (2022)
50. Yang, X., Wang, X.: Hash3d: Training-free acceleration for 3d generation. arXiv preprint arXiv:2404.06091 (2024)
51. Yu, H., Li, R., Xie, S., Qiu, J.: Shadow-enlightened image outpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7850–7860 (2024)
52. Yu, H.X., Duan, H., Hur, J., Sargent, K., Rubinstein, M., Freeman, W.T., Cole, F., Sun, D., Snavely, N., Wu, J., et al.: Wonderjourney: Going from anywhere to everywhere. arXiv preprint arXiv:2312.03884 (2023)
53. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)
54. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. arXiv preprint arXiv:2210.06978 (2022)
55. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
56. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8552–8562 (2022)