DreamDrone: Text-to-Image Diffusion Models are Zero-shot Perpetual View Generators — *Supplement Material* —

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

In this supplementary material, we provide more comprehensive ablation studies, comparisons of visual results with text-to-video methods, and additional visual results of our method. Finally, we discuss the limitations and social impact of this approach.

1 Implementation details

We take Stable Diffusion [7] with the pre-trained weights from version 2.1[3] as the basic text-to-image diffusion and MiDas [6] 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 high-pass filter is 20 and the hyper-parameter λ for feature-correspondence guidance is 300. We conducted the experiments on Titan-RTX GPU. The generated speed is roughly 15 seconds per image.

2 Ablation studies

In this section, we first provide additional ablation results for Fig. 1 as referenced from Fig. 3 in the main text. To enhance the robustness of our ablations, we include one example for each experimental setup. The comprehensive results of the ablation study for each component are illustrated in Fig. 1. For quantitative results, please refer to Tab. 1 in the main text. The simplest method for tasks involving infinite scene generation is frame-by-frame image warping. However, this approach is impractical, as is the direct warping of the latent code. Warping images results in non-integer pixel coordinates, which leads to interpolation-induced blurring and distortion. Furthermore, these errors accumulate with each generated frame, causing a significant degradation in quality, with the images becoming progressively blurred, as shown in `warp image` and `warp latent` in Fig. 1.

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

Fig. 1: 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.

To address these challenges, we introduce DDPM to increase the degrees of freedom for the diffusion model, facilitating the generation of high-quality images (`warp latent + DDPM`). With the incorporation of DDPM, the CLIP score improves from 0.125 to 0.308 for a series of 32 images. Nevertheless, the introduction of DDPM inadvertently affects the semantic consistency between adjacent frames.

To maintain the integrity of image content, and inspired by previous methods [2, 8], we integrate cross-view attention modules into our framework. As demonstrated in `warp latent + DDPM + cross attn.` in Fig. 1, the consistency of geometries across views is significantly improved compared to `warp latent + DDPM`. To ensure high-quality image generation while also maintaining consistency across adjacent views, we propose a feature-correspondence guidance strategy. Comparing `warp latent + DDPM + guidance` with `warp latent + DDPM + guidance + cross view attn.`, it is evident that semantic consistency between adjacent frames is significantly enhanced after incorporating guidance, as indicated by the noticeable improvements in both PSNR and SSIM scores in Table 1 of the main text. To further improve the cross-view consistency of high-frequency details, we have employed high-pass filtering. This approach aids in preserving the high-frequency details of the current frame, thus enhancing the semantic consistency of high-frequency details between consecutive frames. For instance, the enhanced cross-view consistency, particularly of the house on the left side in Fig. 1, illustrates the effectiveness of adding the proposed modules.

Then, we conduct more detailed ablation studies on each proposed module in a Q&A manner.

Q1: Does DDIM inversion limit the reconstruction fidelity?

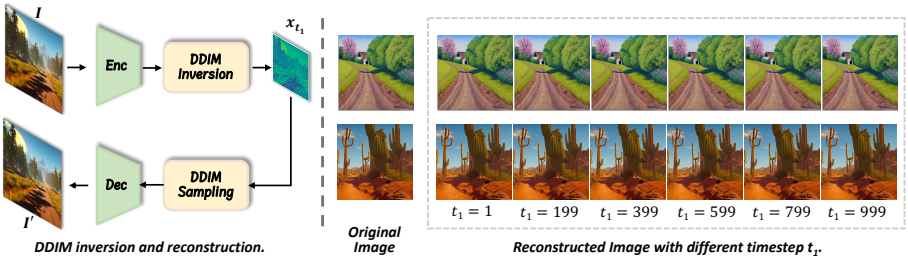


Fig. 2: DDIM inversion and image reconstruction. We illustrate the pipeline for evaluating the reconstruction performance using DDIM inversion. The left side is the pipeline and the right side is the reconstructed results at different timesteps t_1. We visualize two different result samples.

Before warping latent, the most important thing is to ensure we can reconstruct the original image without any editing of the intermediate latent code. To this end, we establish a simple experiment. We obtain the intermediate latent

code x_{t_1} at different timestep t_1, denoising the noise, and decode the reconstructed image. In Fig. 2, the left side is the pipeline of this experiment and the right side is the reconstructed images at different timestep t_1. We take two sample images as examples. From the results, we can figure out that the images can be reconstructed at different timestep t_1.

Based on the former discussion, regarding the top branch in Fig. 2 in the main text, the image at the current view can be reconstructed. The reconstruction of the top branch is the foundation of the feature-correspondence guidance and cross-view attention.

Q2: Why do the image sequences become blurrier when generating more images, no matter whether warping the image or warping the latent code?

The 1^{st} and 2^{nd} rows in Fig. 3 in the main text show that the images become blurrier when generating more images. Besides, the reconstruction results in Fig. 2 show that DDIM inversion can reconstruct original images if there is no editing for the intermediate latent code, *i.e.*, warping. It is straightforward that when we fly through, in other words, zoom in, the images, the images will become much blurrier. This is because the warping operation leads to non-integer pixel coordinates. Previous SOTA methods [1, 3, 4] train a refiner to add the details and inpainting or outpainting the missing region when the camera is moving. In this paper, we serve the pre-trained text-to-image diffusion model as a 'refiner' due to its powerful generation capacity.

Q3: Why DDPM forward is needed?

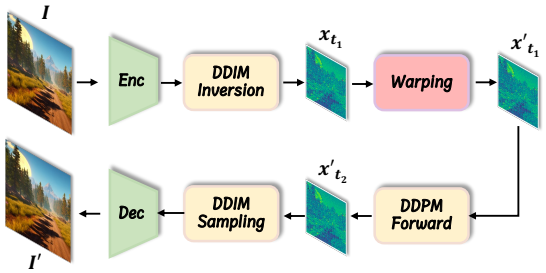


Fig. 3: Ablation studies for DDPM forward without high-pass filtering. To evaluate the necessity of DDPM forward operation, we conduct ablation experiments based on the simple pipeline. The corresponding ablation results are shown in Fig. 4

Now we analyze the necessity of the DDPM forward process. As illustrated in Fig. 2 in the main text, we further apply DDPM forward after warping the latent code. Comparing the 5^{th} and 6^{th} rows in Fig. 1, we can figure out that the image quality improves a lot after DDPM is applied. The details are enhanced and there is no distortion. The side effect of DDPM forward is that the correlation between adjacent views degrades because more degrees of freedom are introduced by DDPM.

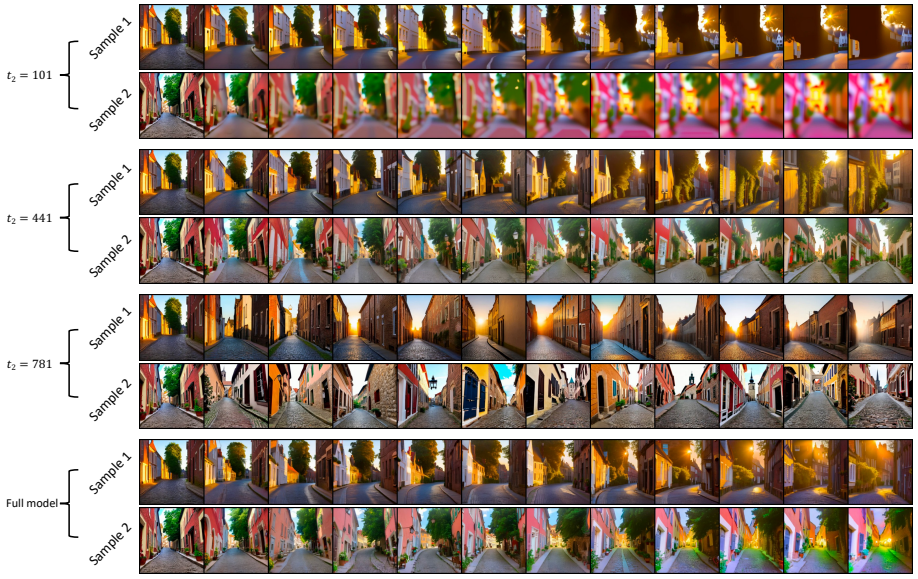


Fig. 4: The visualization results for evaluating DDPM forward without high-pass filtering. We illustrate the results based on the pipeline shown in Fig. 3. We fix the warping timestep t_1 and generate image sequences with different DDPM forward timesteps t_2. To facilitate the comparisons, we further illustrate the generation results using our overall pipeline at last. We visualize every five frames per sample.

To evaluate how DDPM affects the generation results, we fix the warping timestep t_1 and illustrate the generated image sequences with DDPM at various timesteps t_2. The pipeline of this ablation experiment is shown in Fig. 3. The results are shown in Fig. 4. A smaller t_2 means less degree of freedom for the diffusion model, which can result in blurring and distortion. As shown in the first two rows in Fig. 4, the generated images become blurrier when generating more images. A proper t_2 makes the geometry between adjacent views more consistent. In the 3^{rd} and 4^{rd} rows in Fig. 4, the geometry layout in the image sequences becomes consistent. For instance, the geometry of the house on the left side of the image looks roughly consistent. Moreover, the image quality is satisfied. As t_2 becomes larger, more random noise are added to the warped latent code. Though the image quality is promising, the consistency degrades a lot. As shown in the 5^{th} and 6^{th} rows, the consistency across adjacent views is much worse than 3^{th} and 4^{th} rows.

This ablation demonstrates that the DDPM forward module with a proper timestep t_2 improves the image quality. But the consistency is still not satisfied. That's the reason why we further propose the feature-correspondence guidance strategy.

Moreover, since the high-pass filter preserves details from the previous view, is it possible to remove DDPM forward and only use the high-pass filter? To this end, we further conduct an ablation experiment. The pipeline for this ablation

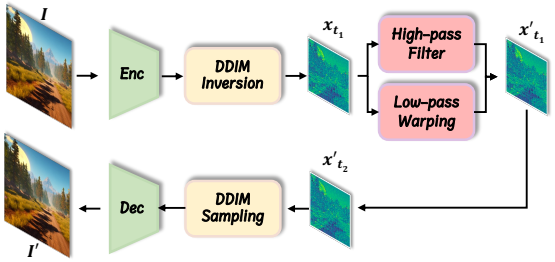


Fig. 5: Ablation studies for high-pass filtering without DDPM forward. To analyze if the DDPM can be replaced by the high-pass filter, we conduct a simple ablation based on this pipeline. The corresponding ablation results are shown in Fig. 6.

is shown in Fig. 5. In this pipeline, we remove DDPM forward operation and add the high-pass filter when warping the latent code. The experimental results are shown in Fig. 6. We show two results for each σ. No matter how large the threshold σ is, the high-pass filter cannot help to preserve the details from previous view. The reason is that the high-pass filter can only preserve high-frequency details from the previous view, rather than the low-frequency content. Combined with the results shown in Fig. 7, the low-frequency information dominants the content when combining frequencies from different images. As discussed before, the content would become much blurrier when warping the latent code, which motivates us to propose the feature-correspondence guidance strategy.

Q4: Will the combination of low and high frequencies from different images break up the correlation of the frequency of the original image and introduce more errors?

To evaluate the feasibility of the frequency combination, we conduct a toy experiment, which is shown in Fig. 7. In this experiment, we first obtain the frequency from two different images and combine the frequencies given different threshold σ. As shown on the right side in Fig. 7, the content of Elon Musk does not change much with different σ, which demonstrates the feasibility of frequency combination. An extremely small σ, for instance, $\sigma = 10$, introduces excessive details from van Gogh's portrait.

In this toy experiment, the content of the two images is extremely different. However, regarding the perpetual view generation task, the content of the adjacent view would not be so different. Now we analyze how different σ affects the generation results. We apply all the proposed modules in this experiment and change the σ value. The comparison results are shown in Fig. 8. As shown in Fig. 8, a small σ neglects more low-frequency content from the previous view, which results in an inconsistency between images. A large σ introduces less high-frequency details from the previous view. Though looks consistent, the generated images look not as realistic as $\sigma = 20$.

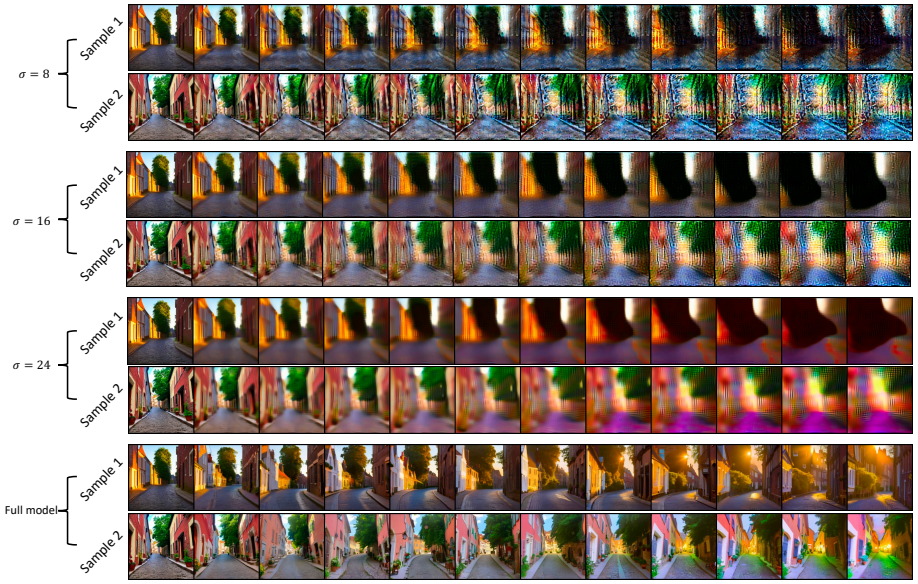


Fig. 6: The visualization results for high-pass filter without DDPM forward. To evaluate if we can preserve high quality from the previous view using the high-pass filter, we remove the DDPM forward module and visualize the image sequences with different threshold σ. To facilitate the comparisons, we further illustrate the generation results using our overall pipeline at last. We visualize every five frames per sample.

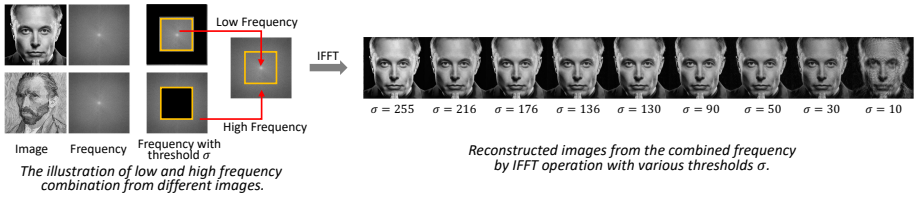


Fig. 7: Toy experiments of low and high-frequency combination from different images. Our toy experiments are illustrated on the left side. We combine the low frequency of Elon Musk and the high frequency of Vincent van Gogh's self-portrait with various threshold σ. The higher σ, the more low-frequency of Elon Musk is used. The results with various σ are illustrated on the right side. Please zoom in for comparisons.

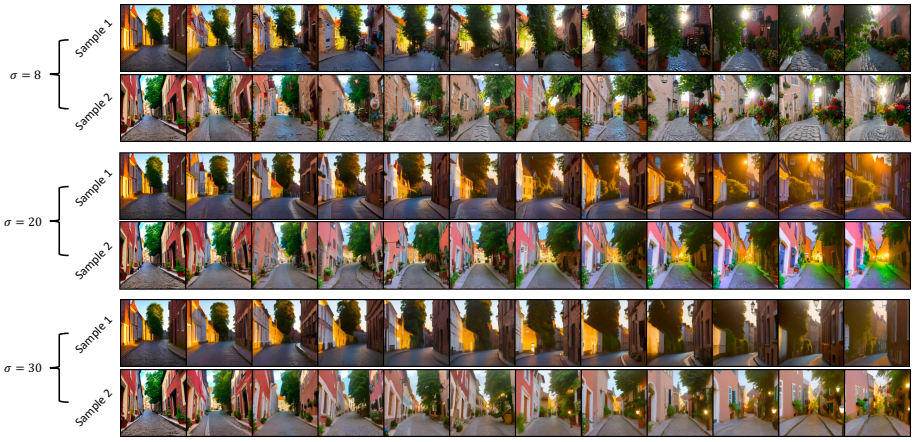


Fig. 8: Ablation studies on high-pass filter. We apply all the proposed modules with various thresholds σ of the high-pass filter.

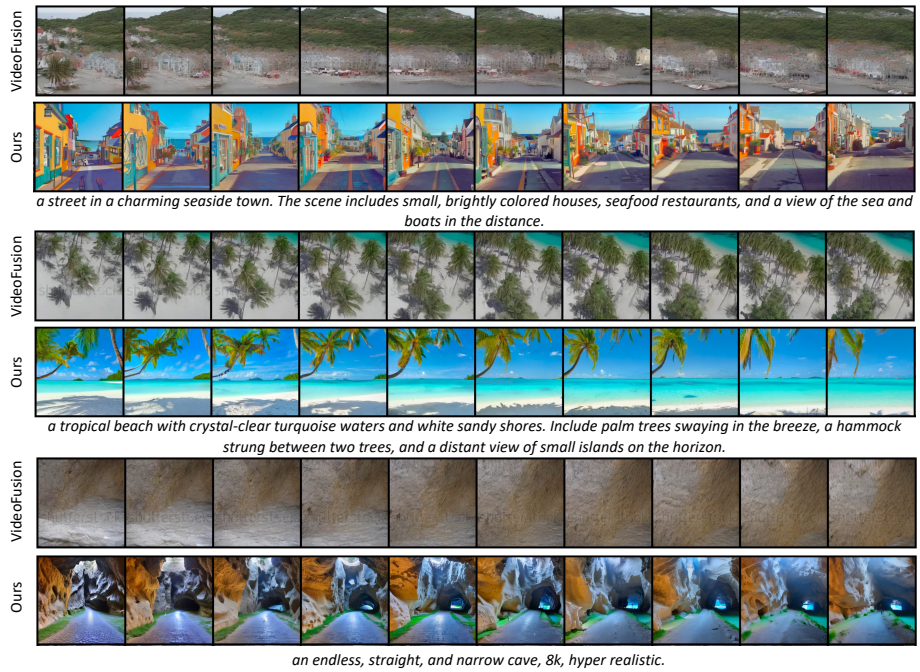


Fig. 9: Qualitative comparisons of VideoFusion [5] and our DreamDrone. We show the visualization results given three prompts and illustrate every five frames for each sample.

3 Additional qualitative comparisons

Our task bears similarities to text-to-video generation, with the key difference being that text-to-video generation cannot be controlled by camera pose, and the quality significantly diminishes as the number of generated frames increases. VideoFusion [5], one of the state-of-the-art methods for video generation tasks, has been visually compared with our method, which is illustrated in Fig. 9. It is evident that VideoFusion's generated results become blurry with an increase in frame count, and the effect of camera movement is less pronounced. In contrast, our method not only generates high-quality continuous scenes but also ensures geometric consistency between frames, clearly conveying the camera's forward movement. Generating scenes in constrained environments like caves is more challenging. VideoFusion does not perform well under such prompts, whereas our method effectively demonstrates the effect of the camera advancing forward.

4 More visualization results

In this section, we provide more visualization results. We generate 120 images for each prompt and visualize one image from every third frame. Please refer to Figs. 10 and 11 for details.

5 Limitation

Given a prompt, our method can infinitely extend a scene without any training or fine-tuning. However, there are some limitations to our approach. Firstly, as our method is zero-shot and training-free, even with the introduction of feature-correspondence guidance and cross-frame self-attention modules, the correspondence of high-frequency details between adjacent frames is not yet perfect. Secondly, our method heavily relies on the accuracy of depth estimation. Although the stable diffusion model exhibits some robustness, for scenes with special styles, the entirely incorrect depth information leads to unsatisfactory generation results. We plan to address these shortcomings in our future work.

6 Social impact

We introduce a new method for creating perpetual scenes from text descriptions, making it easier for people to generate high-quality images without needing complex training or data. This breakthrough can help in various areas, such as making educational content more engaging, aiding in environmental planning, and giving creative professionals new tools to express their ideas. As this technology becomes available, it's important to use it wisely, ensuring it benefits society and does not contribute to misinformation or unethical use. In essence, DreamDrone offers exciting possibilities for innovation while emphasizing the need for responsible use.

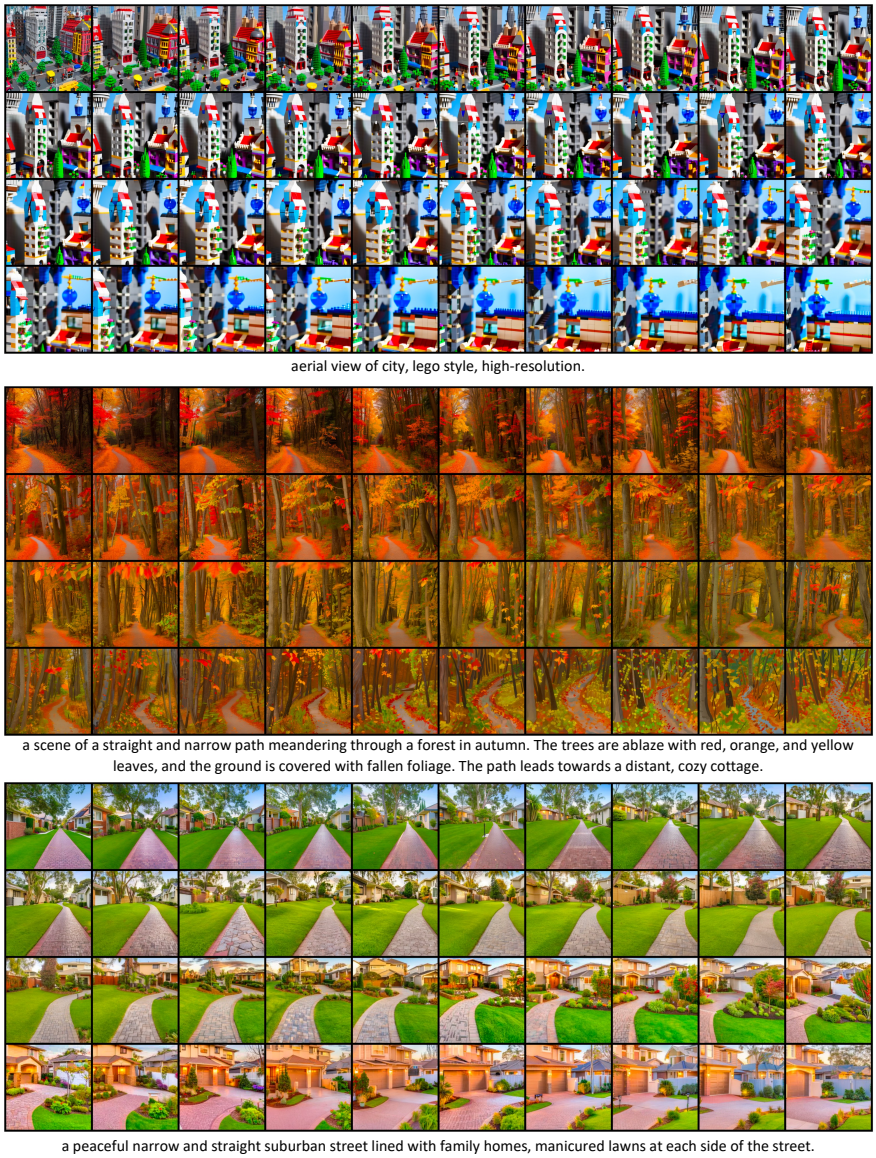
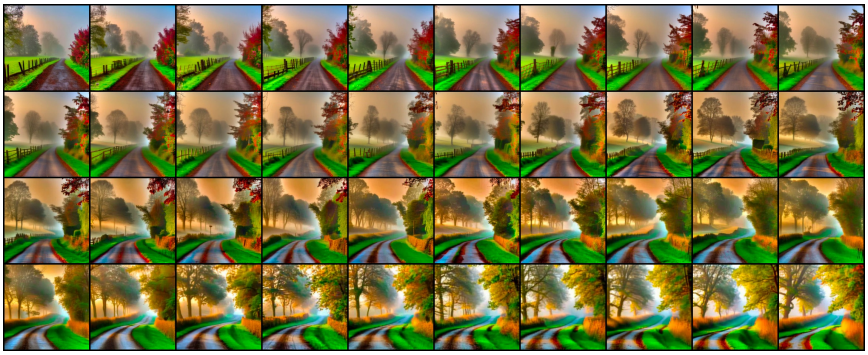


Fig. 10: Visualization results of our DreamDrone. We generated 120 image sequences for each text prompt and visualized one image from every third frame to demonstrate the model's capability in producing diverse and stable visual outputs over time.

an old cobblestone lane winding through the countryside. The lane passes by traditional stone cottages with thatched roofs and well-tended gardens, evoking a sense of nostalgia and timeless beauty.

the vibrant and electric streets of Inazuma City from 'Genshin Impact'. The city combines traditional Japanese elements with a touch of fantasy.



a country lane on a foggy morning. The lane is flanked by old trees and hedges, with the fog adding a mystical quality to the landscape. The early morning light creates a soft, ethereal atmosphere.

Fig. 11: Visualization results of our DreamDrone. We generated 120 image sequences for each text prompt and visualized one image from every third frame to demonstrate the model's capability in producing diverse and stable visual outputs over time.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)