Appendix

A. Implementation Details

A.1. Training Details

Our video generation model is built upon SVD [3], an image-video diffusion model based on UNet. The training process incorporates a continuous-time noise scheduler (Karras et al., 2022), enabling the model to learn to gradually denoise high-variance Gaussian noise. Specifically, we define the noisy data as $x_t = x_0 + n(t)$, where the added Gaussian noise $n(t) \sim \mathcal{N}(0, \sigma^2(t)I)$. The iterative denoising process corresponds to the probability flow ODE:

$$dx = -\dot{\sigma}(t)\sigma(t)\nabla_x \log p(x; \sigma(t))dt,$$

where $\nabla_x \log p(x; \sigma(t))$ is the score function parameterized by a neural network D_θ. The training objective focuses on denoising score matching: $E\left[\|D_\theta(x_0 + n; \sigma, c) - x_0\|_2^2\right]$, with c representing the conditioning information. We follow the EDM framework, parameterizing the learnable denoiser as $D_\theta(x; \sigma) = c_{\text{skip}}(\sigma)x + c_{\text{out}}(\sigma)F_\theta(c_{\text{in}}(\sigma)x; c_{\text{noise}}(\sigma))$.

In our work, we use a relative camera system to convert all camera poses relative to the first frame located at the world origin. First, we train the model at a resolution of 320×512 with a frame length of 25, for 50,000 iterations, and we perform center cropping on the input images to standardize the resolution. Subsequently, we adjust the model to a higher resolution of 576×1024 and continue training for another 10,000 iterations, setting the learning rate to $1e-5$ and using the Adam optimizer with a 1,000-step warm-up. The training is conducted on 8 NVIDIA A800 GPUs with a batch size of 16. It is important to note that we only train the camera encoder, epipolar attention modules, and temporal attention while keeping the weights of other parts frozen. We first train on the Re10K and ACID datasets, with the sampling interval of video frames linearly increasing from 2 to 10. After that, we incorporate the DL3DV dataset, randomly sampling the video frame intervals between $[2, 10]$ to enable the model to adapt to different motion speeds. We choose Lightning as the training framework, utilizing mixed precision fp16 and DeepSpeed ZeRO-2 to enhance efficiency. During inference, we implement the DDIM sampler [51] alongside classifier-free guidance [21] to enhance performance.

A.2. Evaluation Details

Video Generation. When using COLMAP to extract the camera pose of the generated video, the reliability of the obtained pose is often low. Therefore, we employ DUSt3R for a more robust pose estimation. To compare with previous methods, we transform the camera coordinates of the estimated poses to be relative to the first frame and normalize the scale using the furthest frame. We then calculate the rotation distance relative to the ground truth rotation matrices of each generated novel view sequence, expressed as $R_{\text{dist}} = \sum_{i=1}^{n} \arccos\left(\frac{\text{tr}(R^i_{\text{gen}} R^{i^T}_{\text{gt}}) - 1}{2}\right)$. We also compute the $T_{dist} = \sum_{i=1}^{n} \|T^i_{gt} - T^i_{gen}\|_2$, which measures the Euclidean distance between ground truth and estimated translations across n video frames.

Sparse View Reconstruction. To ensure a fair comparison with feed-forward baselines [7, 8], we conducted experiments on 256×256 resolutions. To comprehensively evaluate our method, we compared the results of input views under small and large overlap ratios. We employed RoMa to determine the overlap ratio of image pairs. First, we obtained dense matches from the first image to the second and vice versa, considering pixels with matching scores exceeding 0.005 as valid. Next, we calculated the overlap ratio for each image pair by dividing the number of valid matched pixels by the total number of pixels, with the final overlap ratio defined as the minimum of the two directional calculations. Additionally, to compare with optimization-based reconstruction baselines [13, 29], we used only 2 ground truth training images per scene and evaluated using 12 views. The evaluation metrics included PSNR, SSIM, and LPIPS.

B. Discussions

B.1. Comparison with Recent Methods

ViewCrafter. ViewCrafter [77] leverages point cloud rendering as a condition for fine-tuning video diffusion models. However, constrained by Dust3R's capabilities, the approach may encounter challenges in synthesizing novel views with extensive view range transformations, particularly when generating front-view images from limited back-view inputs. Moreover, in scenarios where DUSt3R fails—such as incorrectly predicting planar geometries—ViewCrafter's performance can be significantly compromised. In contrast, our proposed generation module decouples from the reconstruction module, effectively

mitigating these limitations and enhancing view synthesis robustness.

DimensionX. DimensionX [53] achieves promising results by fine-tuning respective LORA models for different motion patterns. However, this method has some limitations: first, it cannot generate more complex motion patterns; second, it lacks effective control over the speed and amplitude of motion. In contrast, our method introduces camera embedding as the conditional signal, which effectively solves the above problems, thereby enabling the generation of more complex motion patterns and providing precise control over the speed and amplitude of motion.

ReconX. ReconX [34] initially constructs a global point cloud representation and encodes it into a contextual space, serving as 3D geometric conditional information for video diffusion models. However, the method's performance critically depends on the geometric reconstruction accuracy of the underlying DUST3R algorithm. When DUST3R fails to precisely align and reconstruct geometric structures, the point cloud information injected into the video diffusion model may introduce significant geometric distortions, leading to suboptimal generation results. More importantly, the current approach is limited to multi-view interpolation scenarios, lacking an effective solution for single-view extrapolation tasks

CAT3D. CAT3D [15] extends the text-to-image (T2I) generation model by a large-scale data-driven method to learn a 3D prior of the scene, which usually requires a large amount of training time. In addition, the generation stage requires rendering a large number of frames (about 100 frames) to reconstruct the scene. We recognize that video generation models trained on large video datasets embed inherent scene prior information, and therefore propose a new method based on fine-tuning with a video diffusion model. Our method speeds up convergence and retains the generative prior of the pre-trained video model, thereby enhancing the model's generalization ability. By introducing explicit geometric constraints, we further improve the 3D consistency of the generated scenes. Notably, during the inference stage, we utilize the geometric prior for initialization and reconstruction, greatly reducing the number of images required and the optimization time.

C. More Qualitative Results

C.1. More Generated Video Results

As shown in Figure 14, we present a series of generated video sequences covering different scenarios, including indoor environments, outdoor landscapes, and object-centered scenes. Our method shows excellent stability in generating video frame sequences with high 3D consistency, and can effectively predict the plausible appearance details of previously unseen regions. In addition, Figure 12 demonstrates the view interpolation ability of our video generation model using sparse input views, and the performance remains stable even when the input view angles differ greatly.

C.2. Basic Camera Trajectories

Our method demonstrates promising generalization ability in camera-controllable video generation, especially for common motion patterns such as zoom-in, zoom-out, and panning. As shown in Figure 10, our method can generate stable videos with basic camera trajectories using a large-scale training dataset that contains a diverse and complex distribution of camera motion patterns. In addition, in Figure 11, we also show the control ability of our method for different motion amplitudes.

C.3. Single View Exploration Results

Figure 13 illustrates that scene reconstruction from sparse input views often suffers from severe occlusions and geometric gaps due to limited observation angles. By implementing a multi-view observation strategy using video diffusion models, we can progressively fill in these missing regions, thereby enhancing the reconstruction's completeness and geometric detail. As we continuously increase and synthesize observation perspectives, the scene's geometric structure becomes more complete, significantly improving the quality and accuracy of the 3D reconstruction. In Figure 11, we show that by controlling the camera parameters, we can directly control the magnitude of the movement, which is general in different scenarios.

C.4. More 3D Reconstruction Results

Figure 16 presents additional reconstruction and rendering results based on generated video frames, demonstrating that our method effectively achieves high-quality reconstructions in indoor, outdoor, and object-centered scenes.

D. Limitations

Our method encounters several challenges in video generation and scene reconstruction. First, existing video diffusion models suffer from significant inference efficiency bottlenecks, particularly when processing large-scale scene reconstruction tasks requiring numerous video frames, leading to prohibitively high computational costs. Second, current video diffusion models are constrained in the number of generated video frames, which impedes comprehensive scene reconstruction for complex scenarios. Lastly, the lack of effective user interaction mechanisms represents a substantial limitation, as users cannot exert fine-grained control over the generation process and results, restricting the model's adaptability in practical applications.

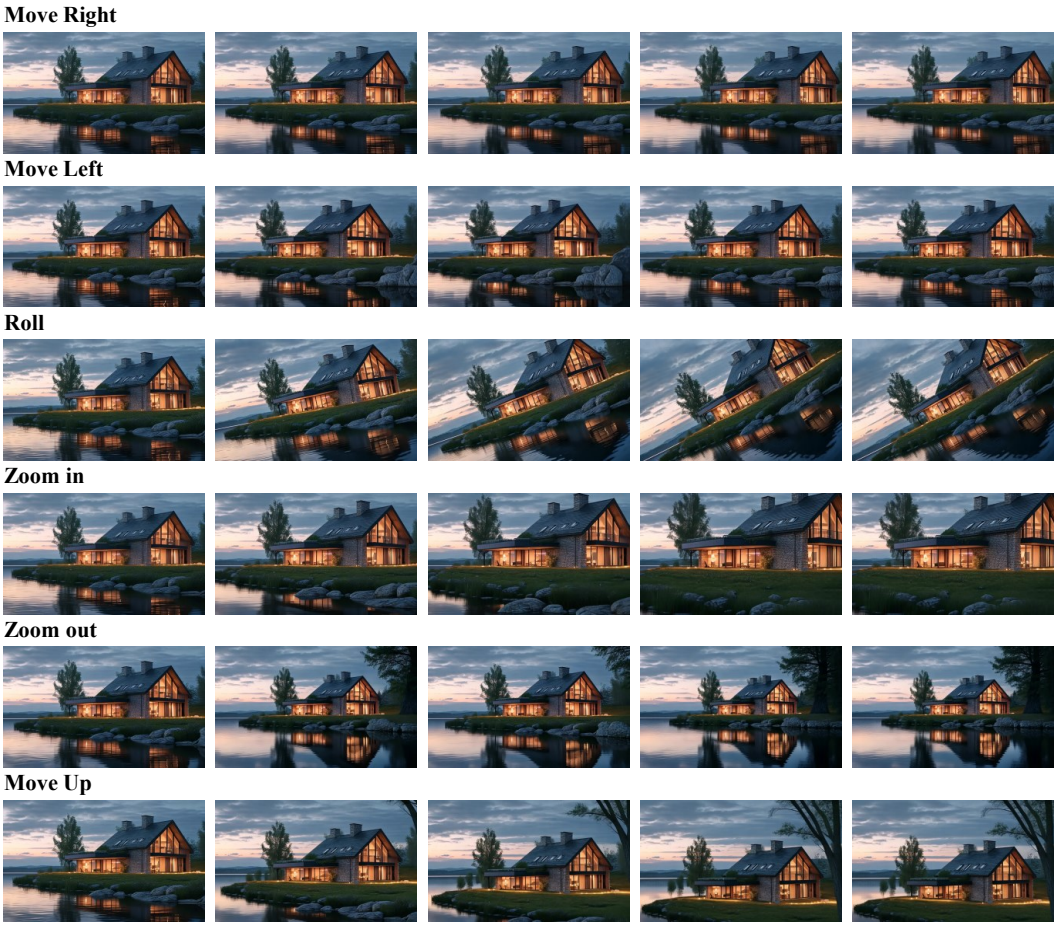

Figure 10. Visualization of our method for generating videos with basic motion trajectories.

Figure 11. Visualization of our method for controlling different motion amplitudes.

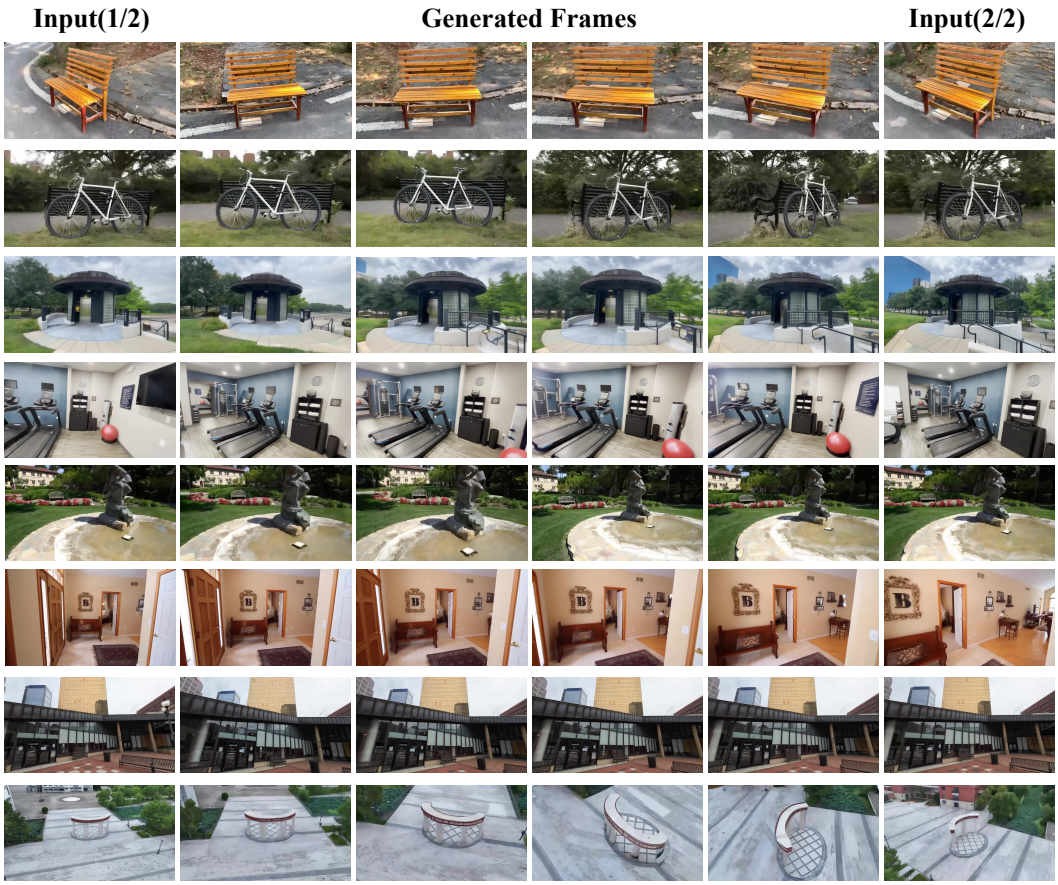

Figure 12. More Visualization of interpolated video based on sparse view input.

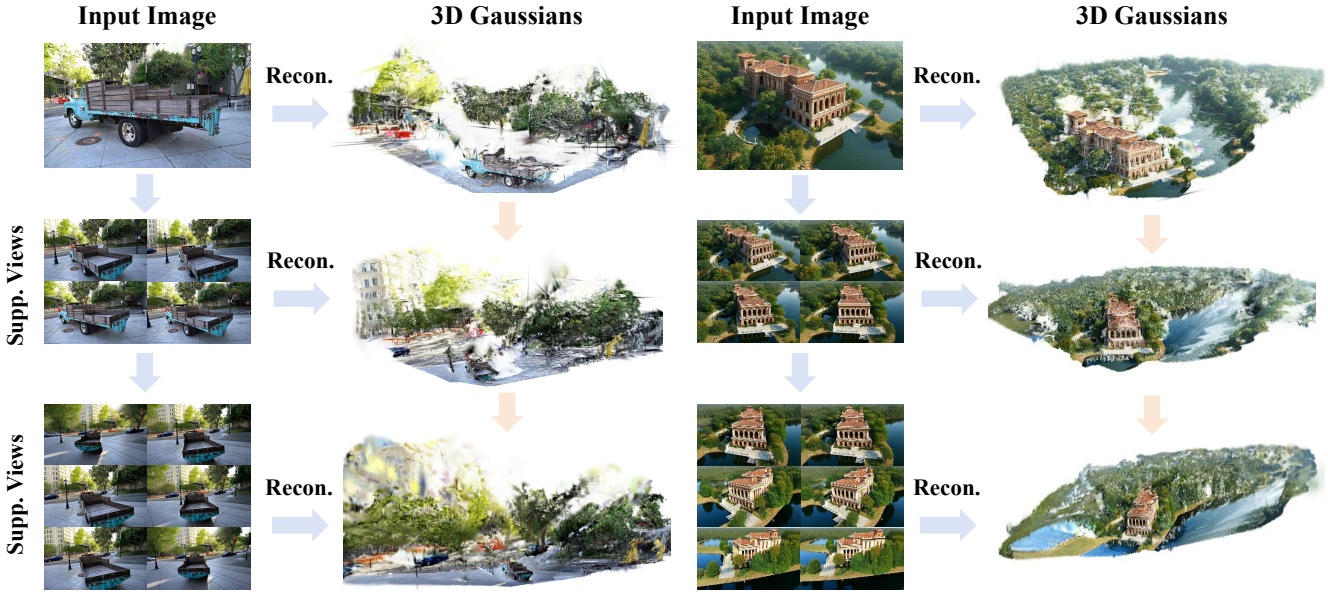

Figure 13. Visualization of exploration and reconstruction based on single-view input.

Figure 14. More video generation results with high 3D consistency based on single-view input.

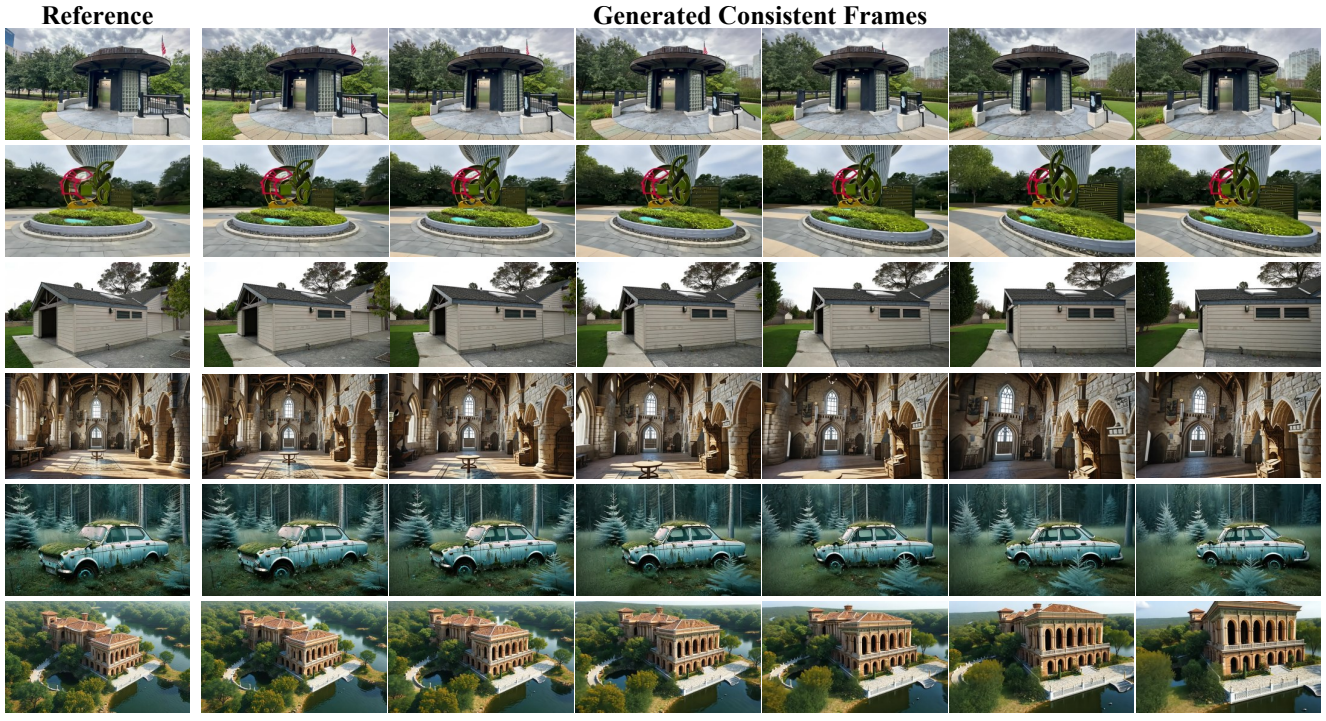

Figure 15. **Consistent frames generated by our approach.** Our method demonstrates strong generalization capability across diverse scenarios, including the DL3DV [32], Tanks-and-Temples benchmark [27], and synthetic images from generative models, while maintaining geometric coherence and structural fidelity.

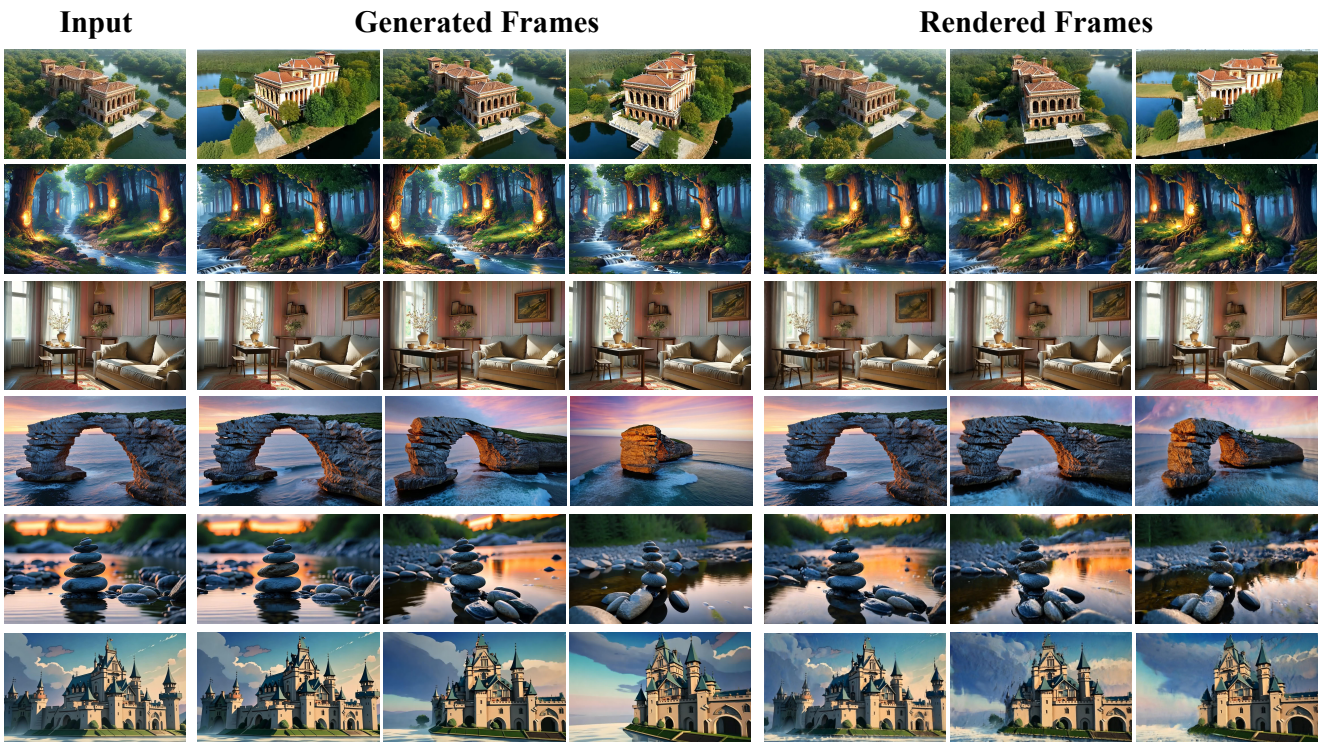

Figure 16. More video generation and reconstruction results with high 3D consistency based on single-view input.