SpatialCrafter: Unleashing the Imagination of Video Diffusion Models for Scene Reconstruction from Limited Observations

Songchun Zhang[1,2] Huiyao Xu[1] Sitong Guo[1] Zhongwei Xie[2]
Hujun Bao[1] Weiwei Xu[1] Changqing Zou[1,3]*
[1]ZJU [2]HKUST [3]Zhejiang Lab

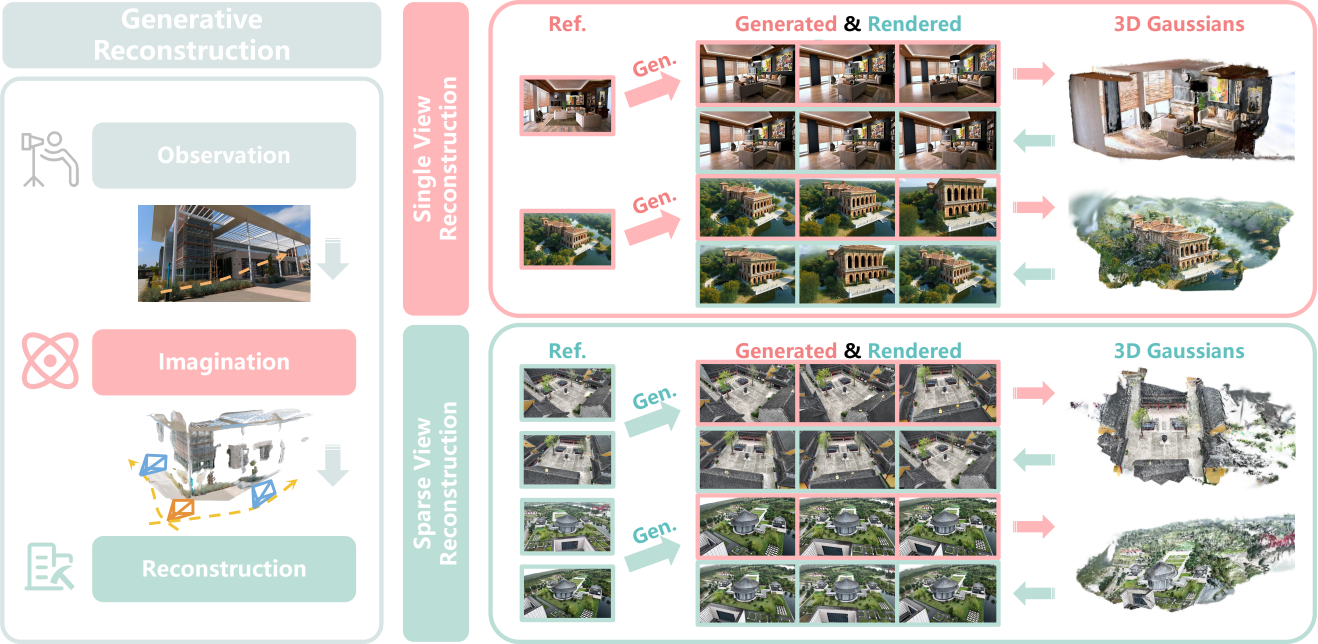
Generative Reconstruction
Observation
Imagination
Reconstruction
Single View Reconstruction
Sparse View Reconstruction
Ref.
Generated & Rendered
3D Gaussians
Gen.

Figure 1. **Overview.** Our method, *SpatialCrafter*, generates additional novel views of a scene from few inputs by leveraging camera trajectory-guided video diffusion models, which alleviates the ambiguity of sparse view reconstruction and robustly reconstructs the scene from the generated video frames. It demonstrates impressive performance in both indoor and outdoor scenes, while also exhibiting generalization capabilities in real scenes and synthetic images. More visualization results can be found in the supplementary material.

Abstract

Novel view synthesis (NVS) boosts immersive experiences in computer vision and graphics. Existing techniques, though progressed, rely on dense multi-view observations, restricting their application. We tackle the task of reconstructing photorealistic 3D scenes from only one or a few input views. We introduce SpatialCrafter, a framework that leverages the rich knowledge in video diffusion models to generate plausible additional observations, thereby alleviating reconstruction ambiguity. Through a trainable camera encoder and an epipolar attention mechanism for explicit geometric constraints, we achieve precise camera control and 3D consistency, further reinforced by a unified scale estimation strategy to handle scale discrepancies across datasets. Furthermore, by integrating monocular depth priors with semantic features in the video latent space, our framework directly regresses 3D Gaussian primitives and efficiently processes long-sequence features using a hybrid network structure. Extensive experiments show our method enhances sparse view reconstruction and restores the realistic appearance of 3D scenes. Project page: `https://franklinz233.github.io/projects/spatialcrafter/`.

* Corresponding author

1. Introduction

Novel View Synthesis (NVS), as a key technology in computer vision and graphics, provides crucial support for immersive experiences in fields such as video games and mixed reality. Although neural reconstruction techniques have made significant progress in recent years [25, 39, 42, 62, 66], these methods typically rely on dense multi-view observation data, which face numerous limitations in practical applications. Therefore, this paper focuses on a more practically valuable challenge: how to achieve high-quality 3D scene reconstruction and synthesize realistic novel views from sparse or even single-view observation.

This problem is hard because (1) real-world scenes are diverse and complex, (2) occluded regions are unseen in sparse captures, and (3) geometric cues are too few to constrain reconstruction. To address these challenges, some methods [17, 26, 29, 40, 84, 88] constrain the optimization process by introducing regularization terms, but these methods need to be optimized for each scene, are not generalizable, and are difficult to cope with complex scenarios. Recently, some feed-forward methods [7, 8, 22, 50, 55, 60, 79, 80] have achieved generalizable sparse view scene reconstruction by directly predicting 3D Gaussian primitives for each pixel. However, such methods produce severe artifacts when reconstructing occluded regions and applying them to the extrapolation setting. To address this challenge, some methods [35, 45, 64, 65] have introduced the priors from the diffusion model to achieve sparse or single-view NVS, but they perform poorly on scene-level data and lack precise pose control and 3D consistency. This is due to the fact that low-dimensional camera information such as Euler angles and extrinsic struggle to provide comprehensive control signals to the generative models.

In this paper, we propose a framework named *Spatial-Crafter* for scene reconstruction and novel view synthesis based on sparse or single-view inputs. Our innovation lies in leveraging the rich physical world knowledge embedded in video diffusion models to provide plausible additional observations for scene reconstruction, thereby effectively reducing the problem complexity. Specifically, we first focus on enhancing precise camera control and 3D consistency in generated videos. To achieve this, we parameterize camera settings using ray embeddings [78] or metric depth-warped images [77] and incorporate them into the video diffusion model via a trainable camera encoder. Furthermore, we propose an epipolar attention mechanism that improves 3D consistency between video frames through explicit geometric constraints. Previous methods [19, 64] often experience performance degradation and difficulty generating large-motion videos when trained on multiple scene datasets. We identified that these issues primarily stem from scale ambiguity in scene datasets. To overcome this, we introduce a unified scale estimator to calibrate scene dataset, enabling effective joint training across multiple datasets.

Although we can generate visually coherent video sequences, relying solely on generated frames to reconstruct general scenes often leads to suboptimal solutions, particularly in large-scale outdoor environments and stylized scenes. To address this, we propose incorporating monocular depth priors with rich semantic features extracted from the video latent space. Leveraging these latent features, our method directly regresses 3D Gaussian primitives of the scene via a feed-forward network. Furthermore, to efficiently handle long-sequence feature interactions, we design a hybrid architecture that integrates Mamba blocks with Transformer blocks. Experiments show that our method not only improves the sparse-view reconstruction, but also accurately restores the appearance of 3D scenes, especially when extrapolating from a single view and when there are few sparse view overlaps. In conclusion, our key contributions can be summarized as follows:

- We introduce a framework that effectively utilizes the physical-world knowledge embedded in video diffusion models to provide additional plausible observations for sparse-view scene reconstruction, thus reducing the ambiguity of sparse view scene reconstruction.
- To address the scale ambiguity problem that occurs in joint training across datasets, we develop a unified scale estimation approach for trajectory calibration. This solves the performance degradation problem, thus enabling effective multi-dataset training.
- We combine monocular depth priors with semantic features extracted from the video latent space, and directly regress 3D Gaussian primitives through a feed-forward manner. Meanwhile, we propose a hybrid architecture integrating Mamba blocks with Transformer blocks to efficiently handle long-sequence feature interactions.

2. Related Work

2.1. Sparse-View Scene Reconstruction

NeRF [62] and Gaussian Splatting [25] have achieved photorealistic representations of 3D scenes, but require optimization on densely collected image sets for each scene. To address this problem, some approaches [17, 29, 71, 88] focus on constraining the optimization process of 3D representations via manually designed regularization. Some other methods [7, 8, 23, 54, 58, 69, 72, 83] are trained on large-scale datasets and directly predict the 3D representation in a feed-forward manner. Recently, some approaches [13, 50] based on end-to-end 3D reconstruction models [7, 28, 60], have achieved efficient and high-quality 3D reconstruction. However, these methods cannot handle occluded regions, and reconstruction failures (e.g., under large viewpoint changes) can severely affect the quality of novel view synthesis. More relevant to our work, some

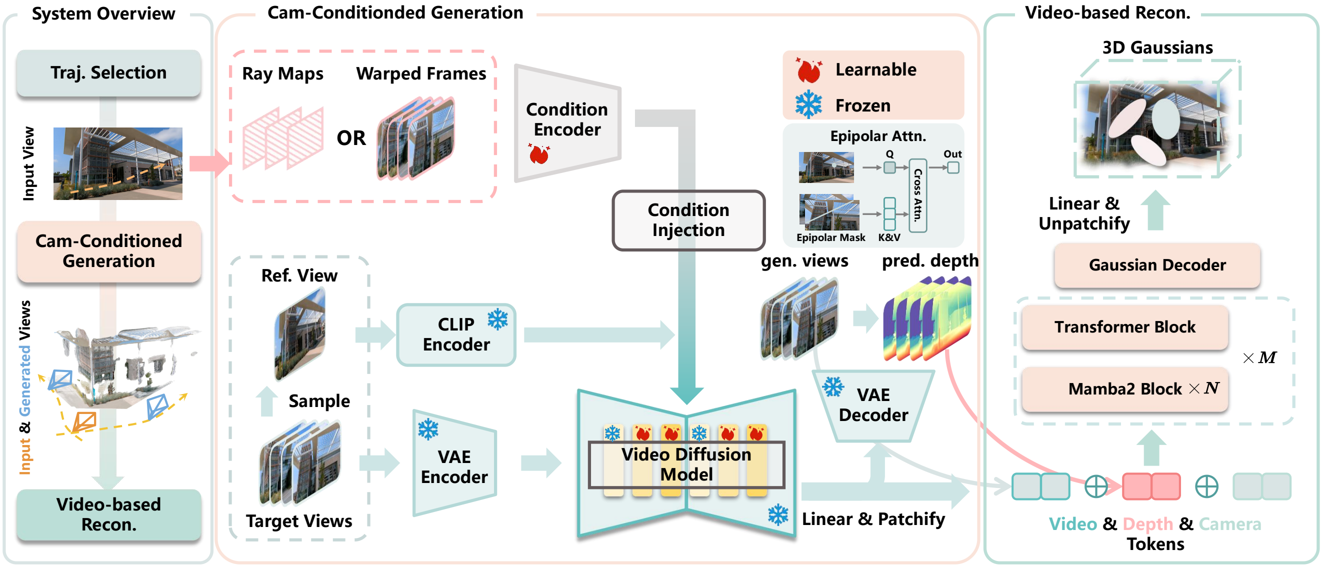

Figure 2. **Overview of our pipeline.** Our generative reconstruction pipeline consists of two parts: camera-controlled video generation, and then video-based reconstruction. First, we set the exploration path based on the input views. In the video generation module, ray embeddings are used to parameterize the camera trajectories, and a cross-attention mechanism with epipolar constraints is introduced to improve the 3D consistency of the generated video. The video-based scene reconstruction pipeline integrates monocular depth priors with semantic features extracted from the video latent space, and directly regress 3D Gaussian primitives via a feed-forward manner.

works [31, 34, 57] have employed priors from video diffusion models to enhance the performance of sparse-view reconstruction.

2.2. Diffusion-based Novel View Synthesis

Since text-to-image diffusion models [44] contain rich natural image priors, some methods [41, 63] directly optimize 3D representations [25, 62] via score distillation sampling [41]. Several approaches [6, 35, 45, 56] enable zero-shot novel view synthesis by fine-tuning the image and pose-conditioned generative models on large-scale 3D object [10] or scene [56, 87] datasets. However, they still have difficulty synthesizing consistent novel views. To solve this, some approaches [5, 15, 36, 38, 47, 65, 85] generate multi-views simultaneously and model the correlation between multiple views with ground-truth poses. Some recent methods [30, 37] proposed to use large-scale unposed images to train NVS methods. In addition, some methods [9, 12, 14, 48, 76, 82] utilize depth-based warping to synthesize novel views and employ the T2I model to inpaint the warped images. However, the novel views generated by these methods tend to suffer from artifacts and cumulative errors, limiting their use in generalized scenes.

2.3. Controllable Video Generation

While recent text-to-video diffusion models have achieved remarkable progress, they inherit the controllability limitations of their text-to-image counterparts, often requiring additional conditioning mechanisms to align generated videos with user intent. Several works [18, 24, 61, 70] have been carried out to introduce a variety of conditions into video generation models. Recently, attention has focused on controlling the camera motion of the video. Some approaches [18, 53] employs motion-specific training for predefined camera movements, though this framework struggles with complex trajectory synthesis. Some methods [1, 19, 20, 64, 67] enable more complex camera control by parameterizing the camera trajectory and injecting it into a pre-trained video diffusion model via a trainable camera encoder. Another line of work [34, 43, 74, 77] advances this direction by enabling pixel-accurate view compositing through a point-based rendering approach.

3. Method

This section begins by explaining our camera-controlled video generation model (Section 3.1), which employs ray embeddings for precise pose control. Then, we describe how to use epipolar geometry constraints to enhance 3D consistency in the video frames. Next, we introduce ourscene reconstruction pipeline (Section 3.2), which aims to perform stable 3D reconstruction based on the generated video latents.

3.1. Camera-Conditioned Video Generation

Scale Alignment. To address this challenge, our approach first employs VGGT [58] to estimate initial camera parameters, represented by the extrinsic matrix $\mathbf{P} = [\mathbf{R}|\mathbf{T}]$, and a corresponding depth map d_v for each video frame. To establish a consistent scale across these datasets, we leverage the Metric3D [73] to infer a canonical metric depth map, denoted d_m. We then align our initial depth prediction, d_v, to the canonical scale by calculating a scale factor s, defined as the ratio of the inter-percentile ranges (IPR) of the two depth maps: $s = \mathrm{IPR}_{0.8,0.2}(d_v) / \mathrm{IPR}_{0.8,0.2}(d_m)$. This scale factor is then used to adjust the translation component of the camera extrinsics, resulting in $\mathbf{P}_{scaled} = [\mathbf{R}|s \cdot \mathbf{T}]$. This procedure yields camera extrinsics with a consistent absolute scale, effectively resolving alignment issues when fusing data from different sources.

Camera Injection. Inspired by recent methods [19, 49, 78], we chose to use ray embeddings [19] or the depth warping frames [77] to represent the camera information. Specifically, for a ray defined by the origin $o \in \mathbb{R}^3$ and the normalized direction $d \in \mathbb{R}^3$, we can represent it as $(o \times d, d) \in \mathbb{R}^6$. Given the camera parameters, the direction of the ray d' corresponding to the pixel coordinates (u, v) can be calculated as $d' = \mathbf{R}\mathbf{K}^{-1}(u, v, 1)^\top + \mathbf{T}$. For depth warping frames, we leverage the pretrained depth estimation model [59] to map the reference image into a 3D point cloud. Subsequently, we render novel views at specified camera poses by projecting this point cloud via the target camera parameters. Afterwards, we use a trainable conditional encoder to inject the camera information into the video model instead of fine-tuning all model parameters. The training objective is:

$$\mathcal{L} = \mathbb{E}_{z,z_0,\epsilon,C,t}\left[\|\epsilon - \epsilon_\theta(z_t; z_0, t, \phi(C))\|_2^2\right] \qquad (1)$$

where $\phi(C)$ represents the camera conditional encodings, z_0 denotes the latent embeddings of the reference frame, and t indicates the timestamps.

Epipolar Feature Aggregation. Camera embedding enhances control but 3D video consistency remains challenging due to dense self-attention allowing unrestricted cross-frame pixel interactions. We address this by incorporating epipolar geometric constraints, and aggregating features along epipolar lines. For a pixel $\boldsymbol{p} = (u, v)$ in frame i, its corresponding epipolar line $\boldsymbol{l}_{ik}$ in frame k is computed as $\boldsymbol{l}_{ik}(\boldsymbol{p}) = \boldsymbol{F}_{ik} \cdot \tilde{\boldsymbol{p}}$, where $\tilde{\boldsymbol{p}} = (u, v, 1)^\top$ are the homogeneous coordinates, and $\boldsymbol{F}_{ik} \in \mathbb{R}^{3\times3}$ is the fundamental matrix. The fundamental matrix is decomposed as $\boldsymbol{F}_{ik} = \boldsymbol{K}_k^{-\top}\boldsymbol{E}_{ik}\boldsymbol{K}_i^{-1}$, where $\boldsymbol{K}_i, \boldsymbol{K}_k \in \mathbb{R}^{3\times3}$ are the camera intrinsics, and $\boldsymbol{E}_{ik}$ is the essential matrix. We then convert epipolar lines into attention masks by computing per-pixel distances and applying a threshold, restricting attention to geometrically valid regions.

Sparse-View Setting. To adapt to the sparse view input setting, we formulate the task as a video frame interpolation problem conditioned on the given start and end frames. To maximally preserve the priors from the pre-trained models [4], we inject boundary frame conditions in both latent and semantic space. Specifically, we combine latent features of the first and last frames with their noisy latent representations, and then concatenate the extracted CLIP embeddings from both frames for cross-attentional feature injection. The training objective is:

$$\mathcal{L} = \mathbb{E}_{z,z_0,z_n\epsilon,C,t}\left[\|\epsilon - \epsilon_\theta(z_t; z_0, z_n, t, \phi(C))\|_2^2\right] \qquad (2)$$

where z_n and z_0 are the latent embeddings of the final and initial frames, respectively. During the inference phase, we use VGGT [58] to estimate the extrinsic and intrinsic parameters for the input views.

3.2. Video-based Scene Reconstruction

Motivation. There are several challenges in reconstructing scenes directly from generated videos. First, the limited number of generated frames makes it difficult to capture complete scene information. In addition, the diverse styles of video frames can pose a challenge to traditional reconstruction techniques. Furthermore, the generated video frames may contain imperfect or low-quality areas, making the reconstruction process unstable. Moreover, In stylized scenes, previous methods often yield poor results, as recovering poses and sparse point clouds from video frames is difficult, leading to many artifacts. To address this, a powerful reconstruction module specifically designed to handle the generated videos is required.

Latent Feature Fusion. Given input video latents $\mathbf{z} \in \mathbb{R}^{T\times H\times W\times C}$ and camera poses encoded as ray embeddings $\mathbf{p} \in \mathbb{R}^{T\times H\times W\times 6}$, we transform them into token sequences through a patchification process. Specifically, we apply spatial patchification to latent features to obtain latent tokens $\mathbf{z}_t \in \mathbb{R}^{N\times d_z}$, while ray embeddings undergo 3D-patchification to produce pose tokens $\mathbf{p}_t \in \mathbb{R}^{N\times d_p}$. To incorporate explicit geometric guidance, we leverage monocular depth estimation within our pipeline. For each input RGB video frame, we estimate dense depth maps $\mathbf{D} \in \mathbb{R}^{T\times H\times W\times 1}$, which are processed by a dedicated depth encoder to yield depth tokens $\mathbf{d}_t \in \mathbb{R}^{N\times d_d}$. The three token sets are channel-wise concatenated to form a unified representation $\mathbf{x} = [\mathbf{z}_t; \mathbf{p}_t; \mathbf{d}_t] \in \mathbb{R}^{N\times(d_z+d_p+d_d)}$, which is then linearly projected to a lower-dimensional space $\mathbf{x}' \in \mathbb{R}^{N\times d}$ before being fed into a sequence of mamba and transformer blocks. Mamba [16] achieves the same token sequence processing functionality as Transformer, but reduces computational complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L)$, making it particularly suitable for dense reconstruction tasks. In our Mamba blocks, we implement bi-directional scanning across the token sequence. First, we compute state parameters $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{N\times d}$ from the input tokens using a linear projection. Then, we execute the State Space Model in both forward $(\mathbf{y}_f)$ and backward $(\mathbf{y}_b)$ directions. The final output of each Mamba block is computed as $\mathbf{y} = \mathbf{y}_f + \mathbf{y}_b$ followed by a final linear transformation.

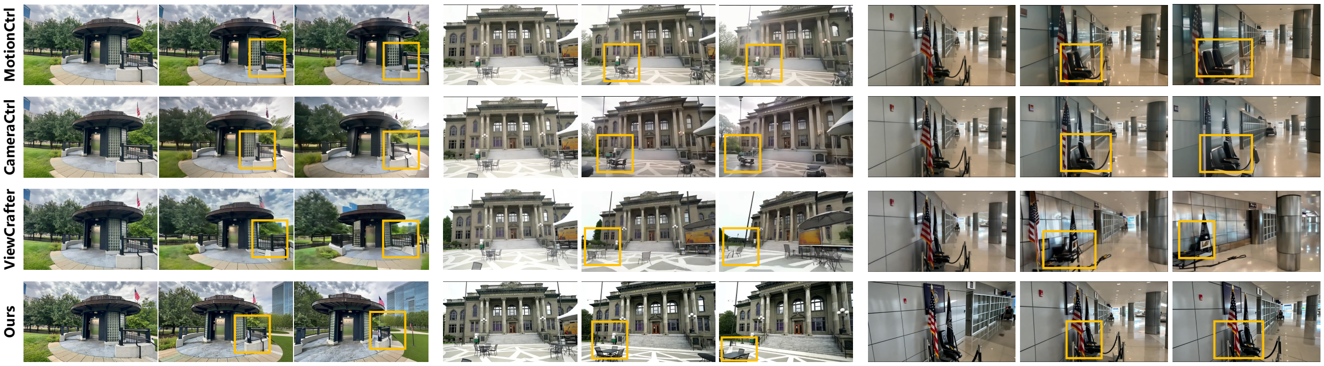

Figure 3. **Qualitative comparison of generated videos.** Compared to other benchmark methods, the videos generated by our method have better foreground object-background consistency, while also being able to generate videos with larger motion amplitude.

Method	**FVD**↓	**FID**↓	**R$_{err}$** ↓	**T$_{err}$** ↓	**LPIPS** ↓	**PSNR** ↑	**SSIM** ↑
RealEstate10K							
MotionCtrl [64]	22.65	230.12	0.234	0.798	0.299	14.72	0.404
CameraCtrl [19]	21.48	188.21	0.054	0.127	0.230	17.33	0.516
ViewCrafter [77]	20.96	204.18	0.055	0.153	0.215	18.95	0.503
Ours	**18.25**	**183.25**	**0.052**	**0.103**	**0.207**	**19.21**	**0.523**
Tanks-and-Temples							
MotionCtrl [64]	30.25	290.38	0.838	1.505	0.315	14.64	0.388
CameraCtrl [19]	24.41	244.82	0.118	0.294	0.285	15.38	0.469
ViewCrafter [77]	22.50	231.35	0.126	0.307	0.247	**16.24**	**0.508**
Ours	**20.17**	**192.52**	**0.097**	**0.175**	**0.228**	16.12	0.501
DL3DV							
MotionCtrl [64]	25.65	249.51	0.473	1.118	0.312	14.38	0.386
CameraCtrl [19]	22.76	233.54	0.095	0.238	0.262	16.33	0.489
ViewCrafter [77]	20.59	211.24	0.093	0.244	0.242	**17.12**	0.521
Ours	**18.21**	**171.52**	**0.063**	**0.134**	**0.225**	17.02	**0.537**

Table 1. **Quantitative comparison** to the camera conditioned video generation method on RealEstate10K [86], DL3DV [32], and Tanks-and-Temples [27] dataset.

Gaussian Decoding. We design a lightweight decoder module that efficiently transforms output feature tokens into per-pixel Gaussian parameters. The decoding module consists of 3D-DeConv layers, which generates refined Gaussian feature map $G \in \mathbb{R}^{(T\times H\times W)\times 12}$. This 12-channel representation precisely encodes the complete set of Gaussian parameters: RGB color (3 channels), scale factors (3 channels), rotation quaternion (4 channels), opacity (1 channel), and ray distance (1 channel). The final output of the model is the merge of 3D Gaussians from all input video frames.

Training Objective. During training, we render images from predicted Gaussians using randomly selected supervision views. Our approach employs a composite loss that integrates three components:

$$\mathcal{L}_{recon} = \lambda_1 \mathcal{L}_{mse} + \lambda_2 \mathcal{L}_{perc} + \lambda_3 \mathcal{L}_{depth}, \qquad (3)$$

where $\mathcal{L}_{mse}$ represents pixel-wise mean squared error, $\mathcal{L}_{perc}$ denotes perceptual loss, and $\mathcal{L}_{depth}$ enforces depth consistency. This formulation jointly optimizes for photometric accuracy, high-level perceptual fidelity, and geometric coherence.

4. Experiment

4.1. Experiment settings

Datasets. To comprehensively capture the underlying distribution of real-world scenarios, we trained our video diffusion model on three diverse datasets: RealEstate-10K [86], ACID [33], and DL3DV-10K [32]. The Re10K dataset from YouTube comprises 67,477 training and 7,289 testing indoor and outdoor camera trajectories. The ACID dataset focuses on natural landscapes, with 11,075 training and 1,972 testing scenes. The DL3DV-10K dataset is a large-scale collection featuring 10,510 scenes captured under controlled, standardized conditions.

Implementation Details. Our video generation model is based on SVD [3], an image-video diffusion model based on UNet. In our experiments, we use a relative camera system in which all camera poses are converted to poses relative to the first frame. The camera in the first frame is located at the world origin. In the first stage, we train the model at a resolution of 320 × 512 for 50,000 iterations with the frame length set to 25. Subsequently, we train on 576 × 1024 for 10,000 iterations to adapt to high resolution. The learning rate is set to $1e-5$ with a warmup of 1,000 steps, using the Adam optimizer. We chose Lightning as the training framework, using mixed-precision fp16 and DeepSpeed ZeRO-2. We trained the proposed method and its variants on 16 NVIDIA A800 GPUs with a batch size of 32. During inference, we adopt DDIM sampler [51] with classifier-free guidance [21].

Progressive Training. Controlling a video generation model to generate arbitrary motion trajectories remains a challenging task. To enable robust arbitrary trajectory generation, we employ a three-stage curriculum learning strategy. The model first trains on smooth camera motions with small temporal intervals, then progressively adapts to more complex motion patterns through linear interval scheduling, and finally incorporates random sampling intervals.

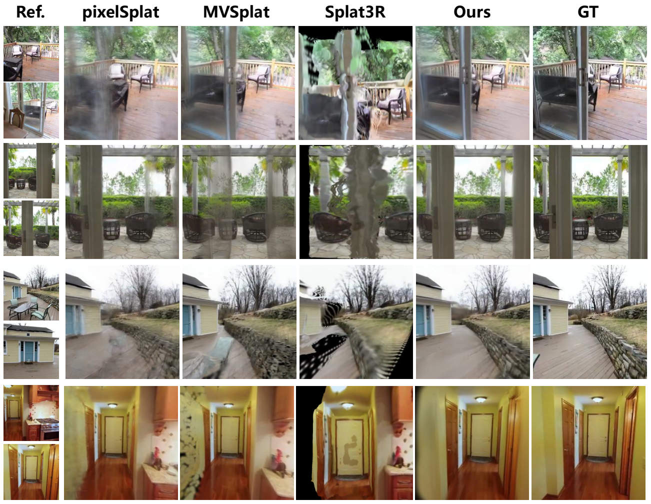

Figure 4. **Qualitative comparison** on Re10K [86]. Compared to baselines, we obtain superior reconstruction from limited overlap, and enhanced geometry reconstruction in non-overlapping regions.

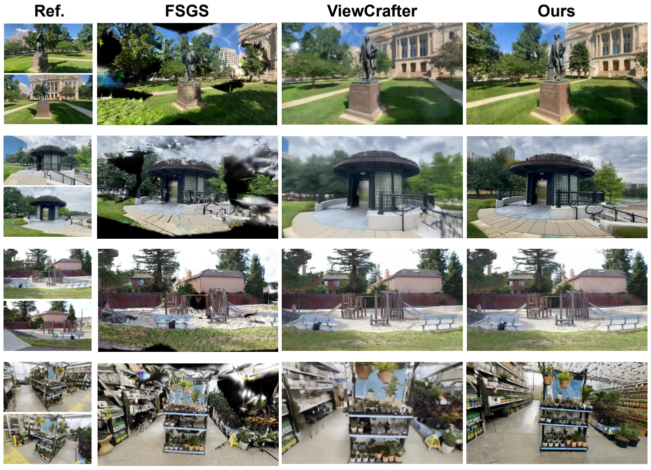

Figure 5. **Qualitative comparison** on DL3DV and Tank-and-Temples dataset. Our method reconstructs better than baselines, even with limited image overlap and in non-overlapping regions.

Method	3-view			6-view			9-view		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Mip-NeRF360									
Zip-NeRF [2]	12.77	0.271	0.705	13.61	0.284	0.663	14.30	0.312	0.633
ReconFusion [65]	15.50	0.358	0.585	16.93	0.401	0.544	18.19	0.432	0.511
CAT3D [15]	16.62	0.377	0.515	17.72	0.425	0.482	18.67	0.460	0.460
ReconX [34]	17.16	0.435	0.407	19.20	0.473	0.378	20.13	0.482	0.356
ViewCrafter [77]	14.51	0.315	0.682	15.87	0.336	0.665	16.45	0.348	0.655
Ours	**17.32**	**0.439**	**0.418**	**19.42**	**0.510**	**0.371**	**20.19**	**0.536**	**0.345**
DTU									
Zip-NeRF [2]	9.18	0.601	0.383	8.84	0.589	0.370	9.23	0.592	0.364
FSGS [88]	17.34	0.818	0.169	21.55	0.880	0.127	24.33	0.911	0.106
ReconFusion [65]	20.74	0.875	0.124	23.62	**0.904**	0.105	24.62	0.921	0.094
CAT3D [15]	22.02	0.844	0.121	24.28	0.899	0.095	25.92	**0.928**	0.073
ViewCrafter [77]	15.63	0.522	0.383	15.03	0.525	0.408	14.92	0.487	0.452
Ours	**27.92**	**0.879**	**0.103**	**28.82**	0.895	**0.082**	**29.03**	0.905	**0.065**
LLFF									
Zip-NeRF [2]	17.23	0.574	0.373	20.71	0.764	0.221	23.63	0.830	0.166
FSGS [88]	20.31	0.652	0.288	24.20	0.811	0.173	25.32	0.856	0.136
ReconFusion [65]	21.34	0.724	0.203	24.25	0.815	0.152	25.21	0.848	0.134
CAT3D [15]	21.58	0.731	0.181	24.71	**0.833**	**0.121**	25.63	**0.860**	0.107
ViewCrafter [77]	17.73	0.521	0.332	17.43	0.512	0.345	17.33	0.488	0.351
Ours	**22.04**	**0.741**	**0.165**	**25.11**	0.814	0.124	**25.95**	0.838	**0.103**

Table 2. **Quantitative comparison** of sparse view 3D reconstruction on Out-of-Domain Datasets.

Method	RealEstate10K			ACID		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
pixelNeRF [75]	20.43	0.589	0.550	20.97	0.547	0.533
GPNR [52]	24.11	0.793	0.255	25.28	0.764	0.332
AttnRend [11]	24.78	0.820	0.213	26.88	0.799	0.218
MuRF [68]	26.10	0.858	0.143	28.09	0.841	0.155
pixelSplat [7]	25.89	0.858	0.142	28.14	0.839	0.150
MVSplat [8]	26.39	0.869	0.128	28.25	0.843	0.144
GS-LRM [79]	28.10	0.892	0.114	-	-	-
DepthSplat [69]	24.23	0.790	0.217	-	-	-
ReconX [34]	28.31	**0.912**	**0.088**	28.84	0.891	**0.101**
ViewCrafter [77]	24.22	0.788	0.218	23.48	0.660	0.299
Ours	**28.35**	0.862	0.121	**28.93**	**0.899**	0.116

Table 3. **Quantitative comparison** of two-view sparse view reconstruction methods on RealEstate10K [86] and ACID [33] dataset.

This gradual transition from simple to complex trajectories proves crucial for high-quality video generation with arbitrary trajectories.

4.2. Comparisons

4.2.1. Controllable Video Generation

Benchmark and Metrics. We evaluate our video generation model against three baselines [19, 64, 77] on three benchmark datasets: RealEstate10K [86] (500 randomly selected videos with first frame as image condition and subsequent frames at stride 3 for pose guidance), DL3DV-140 [32] (500 video clips at stride 2), and Tanks-and-Temples [27] (100 sequences at stride 4 from 14 scenes with COLMAP-annotated poses) for out-of-domain generalization testing.

Metrics. Our evaluation framework employs multiple metrics to assess performance: visual quality measured through Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID); camera control precision quantified by rotation error (R_{err}) and translation error (T_{err}) computed from camera poses extracted via COLMAP and normalized relative to the first frame; and visual consistency evaluated using PSNR, SSIM, and LPIPS [81] between generated and ground-truth views. To ensure fair comparison across methods with varying output capabilities, we restrict visual similarity assessment to the first 14 frames, as generated content typically diverges progressively from the conditional single-view input as the scene extends.

Comparison. Table 1 shows that our method excels in qualitative results, outperforming other methods in terms of visual quality, pose control, and 3D consistency. Figure 3 further demonstrates our advantage in 3D consistency and visual quality. This performance improvement is attributed to the introduction of a camera parameterization method and an epipolar attention module, which respectively enhance pose control ability and the 3D consistency of generated videos. In addition, training on a diverse scene dataset sig-

Method	Small			Large		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
pixelNeRF [75]	18.417	0.601	0.526	20.869	0.639	0.458
AttnNeRF [11]	19.151	0.663	0.368	25.897	0.845	0.229
pixelSplat [7]	20.263	0.717	0.266	27.151	0.879	0.122
MVSplat [8]	20.353	0.724	0.254	27.408	0.884	0.116
DUSt3R [60]	14.101	0.432	0.468	16.427	0.453	0.402
CoPoNeRF [22]	17.393	0.585	0.462	20.464	0.652	0.358
Ours	**22.514**	**0.784**	**0.213**	**27.411**	**0.913**	**0.109**

Table 4. **Quantitative comparison** on the RealEstate10K [86] dataset. *Small* and *Large* refer to input images with low and high overlap ratios, respectively. Greater overlap means greater temporal coherence between adjacent images.

Dataset	Tanks-and-Temples			DL3DV		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LucidDreamer [9]	14.552	0.364	0.415	15.126	0.452	0.431
ZeroNVS [45]	14.734	0.381	0.480	15.163	0.464	0.466
MotionCtrl [64]	15.322	0.426	0.404	16.889	0.528	0.392
ViewCrafter [77]	21.286	0.654	0.193	21.373	0.686	0.244
Ours	**22.331**	**0.742**	**0.213**	**22.782**	**0.791**	**0.202**

Table 5. **Quantitative comparison** of single-view novel view synthesis on DL3DV [32] and Tank-and-Temples [27] benchmarks.

nificantly improves the generalization ability of the model.

4.2.2. Sparse View Reconstruction

Benchmark and Metrics. To comprehensively evaluate the performance of our method, we compare it with two categories of reconstruction methods: (1) two-view feed-forward reconstruction methods [7, 8, 11, 52, 69, 79], and (2) optimization-based sparse-view reconstruction methods [13, 15, 34, 77]. The evaluation employs two types of test sets: in-domain test sets (from RealEstate10K [86] and Acid [33]) and out-of-domain test sets. To ensure a fair comparison, the out-of-domain test set is the same as that used in previous work [15, 65]. For evaluating scene reconstruction quality, we adopt three metrics: PSNR, SSIM, and LPIPS [81]. Additionally, we categorize the evaluation into two groups based on the overlap between input views, with detailed overlap estimation methods provided in the supplementary materials. The evaluation protocol follows a consistent procedure: first, we reconstruct the 3D scene based on the input images, then render from novel views, and compute the metrics between the rendered images and the reference images.

Comparison. As shown in Tables 2-4 and Figures 4-5, our method performs well on various evaluation metrics, particularly in challenging scenarios with occlusions and limited view overlap. This improvement can be attributed to our approach that utilizes scene priors and sparse input views to generate helpful additional observations, thereby enhancing visual correlations between different viewpoints and helping to reduce the complexity of sparse-view reconstruction. Compared to the method [77] conditioning on point-based rendering, our approach shows advantages in reconstructing scenes with thin structures, leading to improved 3D reconstruction results.

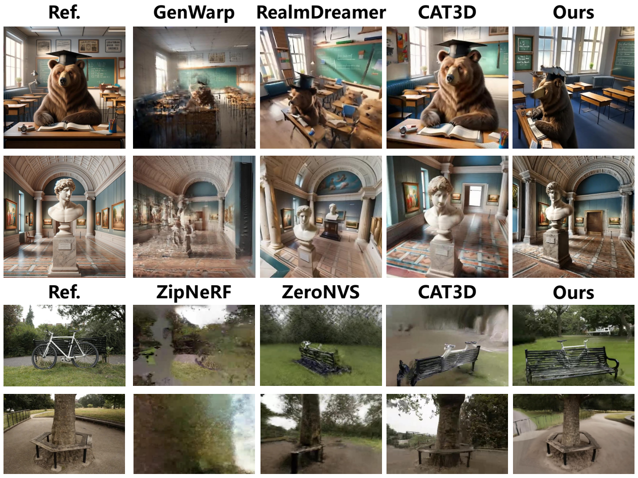

Figure 6. **Qualitative comparisons** of novel views rendered from scenes reconstructed using other 3D generation methods.

4.2.3. Single View 3D Generation

Benchmark and Metrics. In single-view generation, we compare our method with several generative methods [45, 46, 64, 77] on the Tank-and-Temples [27] and DL3DV-140 [32] datasets. Noted that we begin by reconstructing the 3D scene from the given images, followed by calculating the metrics using renderings from novel views. Similarly, we used the PSNR, SSIM, and LPIPS [81] metrics to evaluate our results. Evaluation in this underconstrained setting is challenging, as multiple 3D scenes can produce consistent generations for a given view. Thus, we measure metrics using only temporally adjacent frames to the conditional image, 14 sampled frames and poses captured after the conditional image.

Comparison. As shown in Table 5, our method consistently achieves superior performance on image quality metrics compared to the baselines. For a fair comparison with ZeroNVS [45] and LucidDreamer [9], which are limited to square image inputs, we crop the generated novel views before computing the quantitative metrics. Figure 6 shows that scenes reconstructed using our method have less noise and distortion in occluded areas. For more visualizations, see Figure 7.

4.3. Ablation Study

Ablation on Video Diffusion Model. We evaluate the effectiveness of the design choices. As shown in Table 6, ray embedding achieves significant improvements in camera control accuracy and rendering quality compared to other camera encoding methods by introducing denser position encoding and optimizing geometric correspondences. Specifically, cross-frame consistency shows notable en-

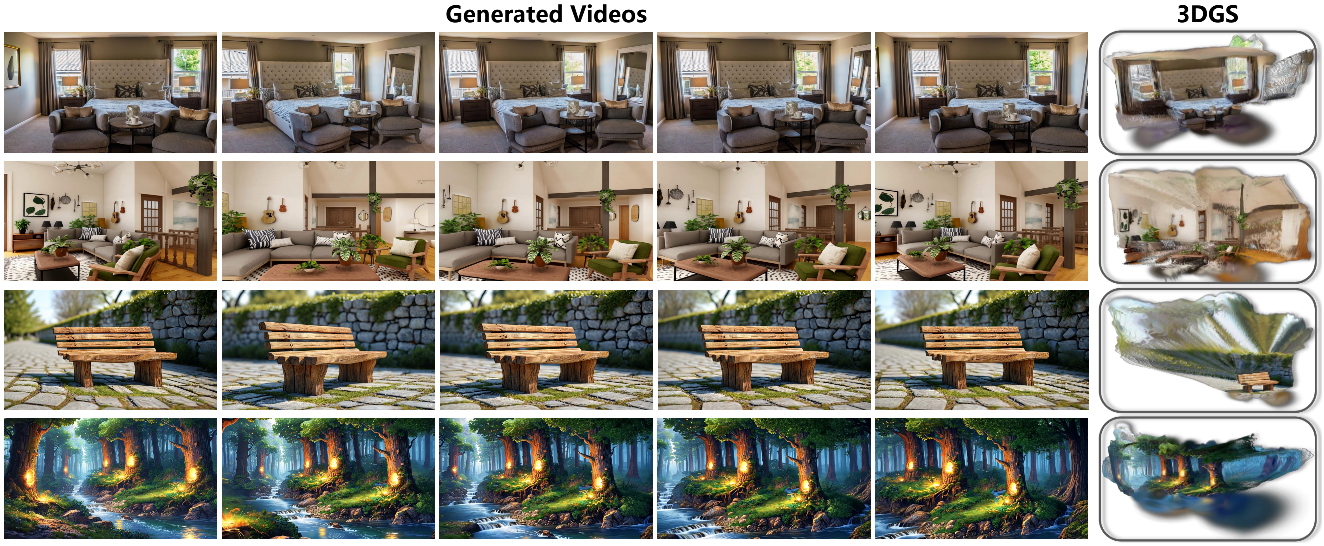

Figure 7. Additional visualizations of generated videos and their corresponding 3DGS representations.

Method	$\mathbf{T}_{err}\downarrow$	$\mathbf{R}_{err}\downarrow$	COLMAP$_{err}\downarrow$	FVD$\downarrow$
Raw Value Emb.	1.592	2.376	13.5%	81.854
Quaternion Emb.	1.457	2.394	13.1%	80.368
W/o Ray Emb.	2.683	3.145	15.7%	112.843
W/o Epipolar	1.125	2.235	12.2%	79.547
W/o Scale Align.	1.712	2.498	14.3%	81.679
Full model	**1.014**	**2.218**	**4.2%**	**78.132**

Table 6. **Ablation study** on the video diffusion model variants.

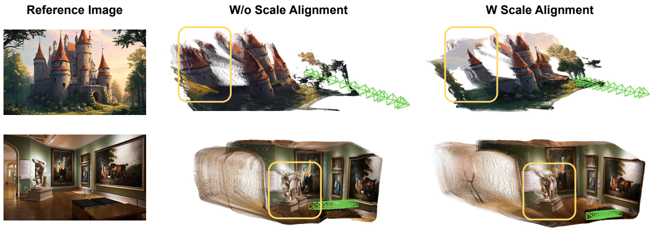

Figure 8. **Ablation on the Scale Alignment.** The video reconstruction results obtained based on scale-alignment training exhibit reduced artifacts while improving the 3D consistency of the scene.

hancement. Furthermore, when incorporating geometric constraints through Epipolar Attention, the system achieves optimal performance in terms of camera controllability and 3D consistency.

Scale Alignment. As illustrated in Table 6 and Figure 8, the incorporation of dataset metric scale alignment significantly improves the pose control accuracy and 3D consistency of the generated video.

Video Latent-based Reconstruction. As demonstrated in Figure 9, incorporating generative priors and depth features in our method significantly enhances the fidelity of fine-grained geometric details in scene reconstruction. By introducing priors from video generative models, we reduced the difficulty of sparse-view reconstruction and achieved reasonable completion of unseen regions.

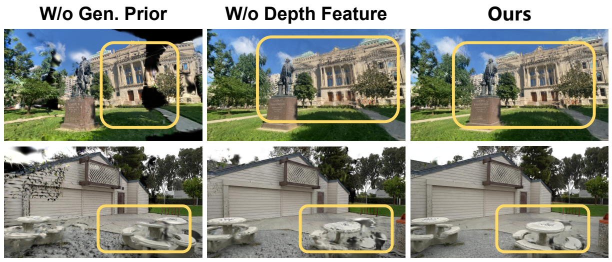

Figure 9. Ablation on the feed-forward Reconstruction Module.

5. Conclusion

We present *SpatialCrafter*, a framework for scene reconstruction and novel view synthesis from sparse or single-view inputs. By leveraging video diffusion models to generate plausible additional observations, we effectively reduce the complexity of sparse-view scene reconstruction. Our key contributions include: (1) precise camera control via ray embeddings or depth-warped images with a trainable condition encoder; (2) a unified scale estimator solving the scale ambiguity in multi-dataset training; and (3) a hybrid Mamba-Transformer architecture that combines monocular depth priors with semantic features from video latent space to directly regress 3D Gaussian primitives. Experiments show our method outperforms existing methods, especially in single-view extrapolation and scenarios with little overlaps. Future work will focus on extending to dynamic scene.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 6

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 5

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

[5] Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. *arXiv preprint arXiv:2411.16157*, 2024. 3

[6] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023. 3

[7] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. 2, 6, 7

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, 2024. 2, 6, 7

[9] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 3, 7

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 3

[11] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4970–4980, 2023. 6, 7

[12] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024. 3

[13] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv:2403.20309*, 2024. 2, 7

[14] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[15] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3, 6, 7

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 4

[17] Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, 2023. 2

[18] 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. 3

[19] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 3, 4, 5, 6

[20] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025. 3

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[22] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20196–20206, 2024. 2, 7

[23] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2024. 2

[24] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2, 3

[26] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. 2

[27] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 5, 6, 7

[28] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2

[29] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *CVPR*, 2024. 2

[30] Lingen Li, Zhaoyang Zhang, Yaowei Li, Jiale Xu, Wenbo Hu, Xiaoyu Li, Weihao Cheng, Jinwei Gu, Tianfan Xue, and Ying Shan. Nvcomposer: Boosting generative novel view synthesis with multiple sparse and unposed images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 777–787, 2025. 3

[31] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 798–810, 2025. 3

[32] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 5, 6, 7

[33] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 5, 6, 7

[34] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 3, 6, 7

[35] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 3

[36] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[37] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. *arXiv preprint arXiv:2412.06699*, 2024. 3

[38] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *CVPR*, 2024. 3

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[40] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 2

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[42] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, 2021. 2

[43] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6121–6132, 2025. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[45] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 2, 3, 7

[46] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *arXiv preprint arXiv:2405.17251*, 2024. 7

[47] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[48] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 3

[49] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34: 19313–19325, 2021. 4

[50] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 5

[52] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 6, 7

[53] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 3

[54] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward gener-

alisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 2

[55] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[56] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 3

[57] Hanyang Wang, Fangfu Liu, Jiawei Chi, and Yueqi Duan. Videoscene: Distilling video diffusion model to generate 3d scenes in one step. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16475–16485, 2025. 3

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 4

[59] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024. 4

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 7

[61] 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

[62] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2, 3

[63] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[64] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH Conference*, 2024. 2, 3, 5, 6, 7

[65] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 2, 3, 6, 7

[66] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *ECCV*, 2022. 2

[67] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[68] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 6

[69] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 2, 6, 7

[70] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[71] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. 2

[72] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2

[73] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 4

[74] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024. 3

[75] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 6, 7

[76] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 3

[77] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3, 4, 5, 6, 7

[78] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 4

[79] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 2, 6, 7

[80] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 2

[81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 7

[82] Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, and Changqing Zou. 3d-scenedreamer: Text-driven 3d-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10170–10180, 2024. 3

[83] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025. 2

[84] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *CVPR*, 2022. 2

[85] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9720–9731, 2024. 3

[86] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph*, 37, 2018. 5, 6, 7

[87] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 2018. 3

[88] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *ECCV*, 2024. 2, 6