This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

NVComposer: Boosting Generative Novel View Synthesis with Multiple Sparse and Unposed Images

Lingen Li[1*] Zhaoyang Zhang[2†‡] Yaowei Li[3] Jiale Xu[2] Wenbo Hu[2] Xiaoyu Li[2]
Weihao Cheng[2] Jinwei Gu[1] Tianfan Xue[1‡] Ying Shan[2]
[1]The Chinese University of Hong Kong [2]ARC Lab, Tencent PCG [3]Peking University

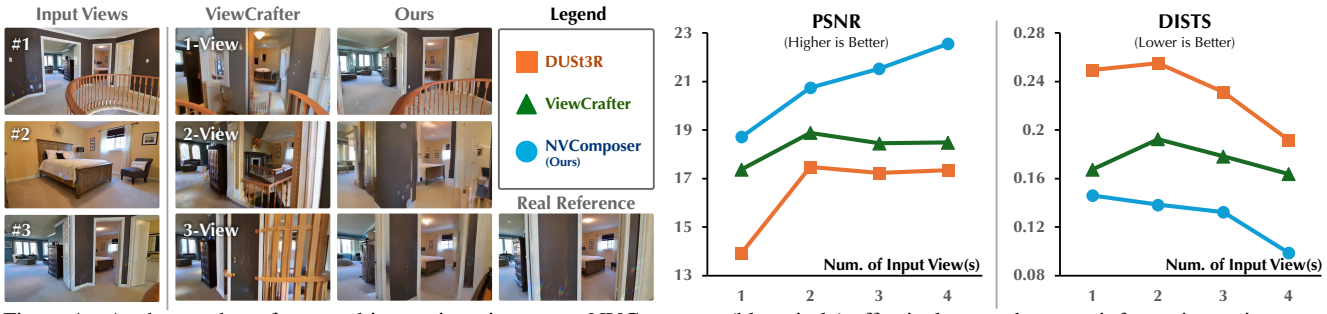


Figure 1. As the number of unposed input views increases, NVComposer (blue circle) effectively uses the extra information to improve NVS quality. In contrast, ViewCrafter [49] (green triangle), which relies on external multi-view alignment (via pre-reconstruction from DUSt3R [38]), suffers performance degradation as the number of views grows due to instability of the external alignment. This result contradicts the common expectation that "more views lead to better performance." Please refer to Sec. 4.2 for full results.

Abstract

Recent advancements in generative models have significantly improved novel view synthesis (NVS) from multi-view data. However, existing methods depend on external multi-view alignment processes, such as explicit pose estimation or pre-reconstruction, which limits their flexibility and accessibility, especially when alignment is unstable due to insufficient overlap or occlusions between views. In this paper, we propose NVComposer, a novel approach that eliminates the need for explicit external alignment. NVComposer enables the generative model to implicitly infer spatial and geometric relationships between multiple conditional views by introducing two key components: 1) an image-pose dual-stream diffusion model that simultaneously generates target novel views and condition camera poses, and 2) a geometry-aware feature alignment module that distills geometric priors from dense stereo models during training. Extensive experiments demonstrate that NVComposer achieves state-of-the-art performance in generative multi-view NVS tasks, removing the reliance on external alignment and thus improving model accessibility. Our approach shows substantial improvements in synthesis quality as the number of unposed input views increases, highlighting its potential for more flexible and accessible generative NVS systems.

*Work done during the internship at ARC Lab. Project Page.
†Project lead. ‡Corresponding authors.

1. Introduction

With recent advances in generative models, generative novel view synthesis (NVS) methods have drawn considerable attention [8, 23, 24, 30, 43, 49] due to its ability to synthesize novel views with only one or a few images. Unlike reconstruction-based NVS methods, where dense-view images with full coverage of the scene are required [17, 19, 25, 38], generative NVS methods could take only one or a few views as inputs, completing unseen parts of a scene with plausible content [28, 48]. This capability is particularly useful in applications where capturing extensive views is impractical, offering greater flexibility and efficiency for virtual scene exploration and content creation.

In addition to generating novel views from a single input image, generative NVS methods [8, 22, 49] have demonstrated more flexible utility by reducing ambiguity through giving additional input images [45]. To leverage multi-view images in the generative NVS tasks, existing methods [8, 43, 49] all rely on external multi-view alignment processes before generation. For example, assuming accurate poses of condition images are given (through *explicit pose estimation*) [8] or generate novel views conditioned on results extracted from reconstructive NVS methods (through *pre-reconstruction*) [22, 43, 49]. However, in the case of the overlap region being small and hard to do stereo matching, external multi-view alignment processes

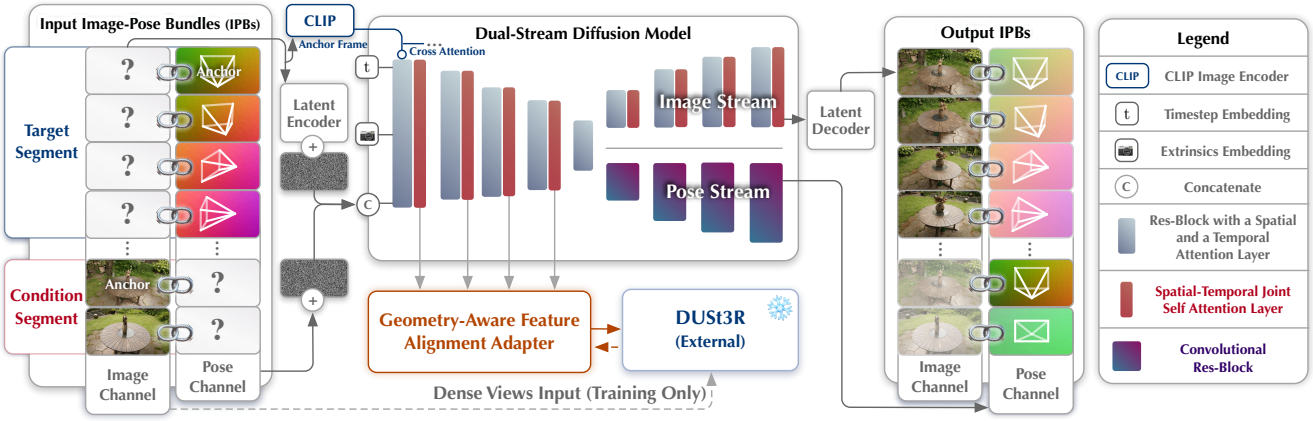


Figure 2. Framework illustration of NVComposer. It contains an image-pose dual-stream diffusion model that generates novel views while implicitly estimating camera poses for conditional images and a geometry-aware feature alignment adapter that uses geometric priors distilled from pretrained dense stereo models [38].

like camera pose estimation become unreliable [45]. As a result, multi-view generative NVS methods that heavily rely on the external alignment also tend to fail as shown in Fig. 1.

To overcome this limitation, we explore the possibility of removing the dependency on the external alignment process and propose **N**ovel **V**iew **Composer** (NVComposer) which could generate novel views from spare unposed images without relying on any external alignment process. Our method is able to generate reasonable results in sparse views with small overlaps and large occlusions.

Firstly, to leverage the powerful generation ability of the video diffusion model, we use a pre-trained video diffusion model as the backbone of our NVComposer with unposed images as the condition in synthesis. Without external alignment of these images, we introduce a novel dual-stream diffusion model in NVComposer to learn the relative poses of condition images during generation. The dual-stream diffusion model not only generates novel views but also implicitly predicts the correct pose relationships between the condition images, ensuring that the model understands the relative positioning of the condition images in the scene and uses them to synthesize novel views correctly.

Moreover, to generate more view-consistent results, we employ features produced by a pretrained dense stereo model [38] to train our model with geometry awareness. Unlike previous methods [22, 49] that directly rely on the reconstruction results from the dense stereo model as input, we propose a more flexible and accessible geometry-aware feature alignment adapter. This adapter aligns our model's features with the predicted 3D features of the dense stereo model and requires no explicit reconstruction during inference. This strategy allows us to distill 3D knowledge implicitly from the dense stereo model [38]. Experiments (Sec. 4.2) demonstrate that this implicit geometry-aware learning achieves competitive performance compared to explicit reconstruction-relied methods [38, 49]. It provides enhanced accessibility and flexibility, as it operates in an end-to-end manner and eliminates the need for an extra step of explicit pose estimation or pre-reconstruction during inference.

To train our model, we construct a mixed dataset from different sources such as video [21, 27, 53] data and 3D [5] data with real indoor and outdoor scenes as well as synthetic 3D objects. NVComposer is trained on this diverse dataset using as few as one to four randomly sampled unposed condition images. Extensive experiments demonstrate that, when provided with multiple unposed input views, NVComposer outperforms state-of-the-art controllable video diffusion models and generative NVS models.

Our key contributions are summarized as follows:

- We introduce the first pose-free multi-view generative NVS model for both scenes and objects, without the requirement for explicit multi-view alignment processes on input images.
- Our proposed design which includes image-pose dual-stream diffusion and geometry-aware feature alignment adapter, highlights a promising direction for creating more flexible and accessible generative NVS systems.
- Our NVComposer achieves state-of-the-art performance on generative NVS tasks for both scenes and objects when given multiple unposed input views.

2. Related Work

Single-View Generative Novel View Synthesis. Early NVS methods using feed-forward networks map a single input image to new views [34, 37], but are limited to small rotations and translations due to the restricted information from one input image. Generative NVS methods effectively hallucinate unseen views given limited input [4, 28, 41]. Recent advances in diffusion models [13, 32] leverage rich image priors for NVS to synthesize more reasonable multi-view content [16, 23, 24, 36] by utilizing pretrained im-

age diffusion models [29]. Zero-1-to-3[23] fine-tunes a latent diffusion model [29] with image pairs and their relative poses for novel view synthesis from a single image. Wonder3D[24] incorporates image-normal joint training and view-wise attention to enhance generative quality. SV3D [36] fine-tunes a video diffusion model [2] for NVS of synthetic objects.

However, single-view NVS models struggle to infer occluded or missing details due to the limited information from one viewpoint, making them less practical for real-world applications requiring complete scene understanding.

Multi-View Generative Novel View Synthesis. To overcome single-view limitations, multi-view conditioned generative NVS utilizes images from multiple viewpoints [8, 22, 42, 49], enhancing the fidelity of generated views by capturing finer details and accurate spatial relationships. iFusion [42] employs a pretrained Zero-1-to-3 [23] as an inverse pose estimator and tunes a LoRA [15] adapter for each object to support multi-view NVS. CAT3D [8] uses Plücker ray embeddings [31] as pose representations and masks the target view for inpainting, allowing flexibility in the number of conditioning images. ViewCrafter [49] reconstructs an initial point cloud using a dense stereo model [38] and then employs a video diffusion model to inpaint missing regions in rendered novel views.

These methods, however, rely on accurate pre-computed poses of conditional images. Sparse views that lead to inaccurate poses can significantly degrade the quality of generated views, limiting their robustness in practical scenarios.

Video Diffusion Models. Advancements in diffusion models have extended their capabilities from static images to dynamic videos, enabling temporally coherent video generation conditioned on various inputs [1–3, 7, 10, 14, 18, 40, 44, 46]. Ho *et al*. [14] first introduced diffusion models for video generation. Video LDM [3] operates in the latent space [29] to reduce computational demands. Subsequent works enhance controllability by incorporating additional conditions. AnimateDiff [9] extends text-to-image diffusion models to video by attaching motion modules while keeping the original model frozen. DynamiCrafter [44] introduces an image adapter for image-conditioned video generation. MotionCtrl [40] and CameraCtrl [10] incorporate camera trajectory control using pose matrices and Plücker embeddings, respectively. ReCapture [50] generates new camera trajectory views based on a given video.

Building upon the video diffusion model, our method leverages temporal coherence to synthesize unseen areas not included in the input images. Compared to previous controllable video diffusion models, our approach achieves better accuracy in camera controllability, offering robust performance for generative NVS tasks.

3. Methodology

The objective of NVComposer is to develop a model capable of generating novel views at specified target camera poses, using multiple unposed conditional images without requiring external multi-view alignment (*e.g.*, explicit pose estimation). To achieve this, we propose to enable the model itself to infer the spatial relationships of the conditional views during generation. We introduce this capability through two key strategies: (1) instead of explicitly solving for camera poses, we model pose estimation as a generative task that jointly happens with the image generation, and (2) we distill effective geometric knowledge from expert models into our generative model.

This leads to two main components of our NVComposer, as shown in Fig. 2: an **image-pose dual-stream diffusion** model that generates novel target views while implicitly estimating camera poses for conditional images and a **geometry-aware feature alignment** adapter that uses geometric priors distilled from pretrained dense stereo models [38]. The design and implementation of these components are detailed below.

3.1. Image-Pose Dual-Stream Diffusion

Assume the model accepts T elements as input and produces T elements as output, where each element corresponds to an image captured within the current scene, accompanied by its pose annotation. We refer to these elements as *image-pose bundles*. We partition these bundles into two segments: the first N bundles constitute the *target segment*, and the remaining M bundles form the *condition segment*, as illustrated in Fig. 2.

Image-Pose Bundles. Specifically, let $I_t \in \mathbb{R}^{3\times H\times W}$ denote the t-th RGB image, and $P_t \in \mathbb{R}^{6\times H\times W}$ be the corresponding Plücker ray embedding [31] representing the camera pose. The conditional input and generated output are sequences of T image-pose bundles for a specific scene, denoted as $\mathcal{B} = \{[I'_t, P'_t]\}_{t=1}^{T}$, where the t-th image-pose bundle consists of the concatenation (along the channel dimension) of the latent image $I'_t = \mathcal{E}(I_t) \in \mathbb{R}^{4\times \frac{H}{8}\times \frac{W}{8}}$ and the resized Plücker ray $P'_t \in \mathbb{R}^{6\times \frac{H}{8}\times \frac{W}{8}}$. Here, $[\cdot]$ denotes concatenation and $\mathcal{E}$ represents the VAE encoder of latent diffusion.

The main difference between the target output $\mathcal{B}$ and the conditional input $\mathcal{B}c$ is that $\mathcal{B}$ is complete, while $\mathcal{B}c$ is partially masked. Specifically, for $\mathcal{B}_c$, image latents in the target segment (*i.e.*, the first N elements) and pose embeddings in the condition segment are set to zero. Additionally, we utilize relative camera poses in our model and designate the first elements of both the target and condition segments as the *anchor view* for this relative coordinate system. This implies that these two elements are expected to have the

same image content, and their camera extrinsic matrices are identity transformations. The image for the anchor view is always provided during training and inference, corresponding to the scenario where at least one conditional image is available. With this design, our image-pose dual-stream diffusion model accepts M unposed conditional images and N target poses as input and outputs N novel view images at target poses along with M predicted poses for conditional images, where $N, M \geq 1$ and $N + M = T$.

Video Prior. Video priors from the video diffusion model have been validated to be useful for generative NVS tasks [8, 22, 49]. To fully leverage the generative priors in NVComposer, we initialize the dual-stream diffusion model using the pretrained weights of the video diffusion model DynamiCrafter [44]. We omit the frame rate and text conditions from the original video diffusion model, focusing solely on the relevant components for our task. We retain the image CLIP [26] feature conditioning with its Q-Former-like [20] image adapter in the cross-attention layers, using the anchor view as the conditioning image. Additionally, we enhance the model's capacity to capture cross-view correspondences at different spatial locations by adding extra spatial-temporal self-attention layers after each Res-Block of the original video diffusion model. Although camera poses information is provided in $\mathcal{B}_c$ as Plücker ray embeddings, we further incorporate the sequence of target camera extrinsics embedding encoded by a learnable multi-layer perceptron layer from the corresponding 3×4 camera-to-world matrices $R_c \in \mathbb{R}^{T \times 3 \times 4}$ of all image-pose bundles, where the last M elements on the temporal dimension are masked with zeros like what we have done with $\mathcal{B}_c$. These embeddings are added to the time-step embedding and serve as supplemental signals indicating the poses for each image-pose bundle in the target segment.

Pose Decoding Head Separation. We observed that training the model to generate image-pose bundles directly makes it hard to converge. This is caused by the difference of the two modalities: images latents I'_t contain complex latent features representing diverse content, while the pose embeddings P'_t are dominated by low-frequency components. This disparity can lead to interference when jointly denoising $[I'_t, P'_t]$ using tightly coupled network layers.

To address this issue, we design an additional decoding head specifically for denoising P'_t. As shown in Fig. 2, this *pose decoding head* operates in parallel with the original decoding part of the diffusion denoising U-Net and follows a fully convolutional architecture similar to the original decoder. It takes as input the bottleneck features, residual connections from the encoding part, and the denoising time-step embeddings. Since the Plücker ray embeddings of poses are predominantly low-frequency and relatively straightforward to denoise, we empirically reduce the base channel number of the pose decoding head to one-tenth of that of the original decoder for images and remove its attention layers. The outputs of the pose decoding head are concatenated along the channel dimension with those of the original decoding part of the diffusion U-Net to form the final output.

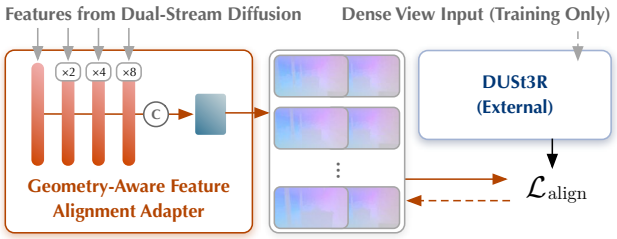


Figure 3. Structure of the geometry-aware feature alignment adapter in NVComposer, which aligns the internal features of the dual-stream diffusion models with the 3D point maps produced by DUSt3R [38] during training. Block with notation "×2", "×4", and "×8" refer to bilinear upsampling on spatial dimensions. The four red bars refer to the channel-wise MLPs.

3.2. Geometry-aware Feature Alignment

Since the dual-stream diffusion model is initialized from a video diffusion model that is not inherently trained with geometric constraints across views, we introduce a geometry-aware feature alignment mechanism in NVComposer. This mechanism distills effective geometric knowledge from an external model with strong geometry priors during training. Specifically, we leverage the dense stereo model DUSt3R [38], which performs well with dense views (both target and condition images), to compute T pointmaps across all views (T_t, for $t = 1, 2, \ldots, T$) relative to the anchor view T_1.

We then align the internal features of our diffusion model with these point maps through a geometry-aware feature alignment adapter during training. Specifically, as illustrated in Fig. 3, the alignment adapter (the red block on the left) takes features from the encoding part of the dual-stream diffusion U-Net, immediately after each spatial-temporal self-attention layer. These features are resized to match the spatial dimensions of the image latent inputs, $\frac{H}{8} \times \frac{W}{8}$ (the white squares in Fig. 3). The resized features retain their temporal dimension of length T, and all operations within the geometry-aware feature alignment adapter are temporally independent. The features are then processed by channel-wise MLPs (four red bars in Fig. 3) to reduce them to 320 channels, followed by a convolutional residual block that outputs a 4-D tensor $F \in \mathbb{R}^{T \times 6 \times \frac{H}{8} \times \frac{W}{8}}$.

For all item $f_t, t = 1, 2, .., T$ in F along the temporal dimension, we minimize the mean squared error (MSE) with the concatenated point maps produced by DUSt3R [38] D

given the t-th view I_t and the anchor view I_1:

$$\mathcal{L}_{\text{align}} = \frac{1}{T}\sum_{t=1}^{T} \|f_t - D(I_1, I_t)\|_2^2. \quad (1)$$

3.3. Training Objectives

We train the image-pose dual-stream diffusion model h_θ in NVComposer to predict noises ϵ given the uniformly sampled denoising time step k, the noisy version of complete image-pose bundles $\mathcal{B}^{\{k\}}$ at time step k, the conditional image-pose bundles S_{cond}, and CLIP [26] image feature of the anchor view $\Phi(I_1)$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathcal{B},\epsilon,\mathcal{B}_c,\Phi(I_0),R_c,k}[\|\epsilon - h_\theta(\mathcal{B}^{\{k\}}, \mathcal{B}_c, \Phi(I_1), R_c, k)\|], \quad (2)$$

where θ is trainable parameters of the image-pose dual-stream diffusion model, Φ is the CLIP [26] image feature encoder. Our total loss combines the diffusion loss $\mathcal{L}_{\text{diff}}$ and the feature alignment loss $\mathcal{L}_{\text{align}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda\mathcal{L}_{\text{align}}, \quad (3)$$

where λ is a loss re-weighting factor.

4. Experiments

In this section, we evaluate the performance of NVComposer on generative NVS tasks for real-world scenes and synthetic 3D objects, followed by an analysis of the model's sub-components. More results are in the supplementary.

4.1. Training Details

We train our model on a large-scale mixed dataset containing around 200K samples built from Objaverse [5], RealEstate10K [53], CO3D [27], and DL3DV [21]. The sequence length of image-pose bundles T is set to 16. To get samples from video datasets (RealEstate10K [53], CO3D [27], and DL3DV [21]), we randomly select a frame interval between 1 and $\lfloor T_{max}/N \rfloor$ to sample T consecutive frames, where T_{max} is the total number of frames in the scene. The value of T_{max} ranges from 100 to 400 depending on the data sample. We randomly sample N condition views and shuffle them within the image-pose bundle. For samples from the 3D dataset (Objaverse [5]), we render each 3D object in two versions: one with 36 orbit views and another with 32 random views. Target views are sampled from the orbit renderings, and condition views are sampled from the random renderings, following the procedure described above.

Since different parts of our training dataset have varying available resolutions, we first train the model at a resolution of 512×512 for $10,000$ steps across all datasets. Then, we train the model at a resolution of 576×1024 on RealEstate10K and DL3DV for $20,000$ steps, where a higher resolution is available. We use a learning rate of 1×10^{-5} and perform all training on a cluster with 64 NVIDIA V100 GPUs, with an effective batch size of 128. For more details, please refer to the supplementary material.

4.2. Results

4.2.1. Generative NVS in Scenes

Benchmark Settings. We evaluate the performance of NVComposer on generative NVS tasks for scenes, comparing it with four state-of-the-art models: MotionCtrl [40], CameraCtrl [10], DUSt3R [38], and ViewCrafter [49]. MotionCtrl [40] and CameraCtrl [10] are controllable video generation models that work with a single input image. DUSt3R [38] is a dense stereo model that can be used for multi-view reconstructive NVS. ViewCrafter [49] is a multi-view generative NVS method that relies on explicit pose estimation and point cloud guidance.

For the RealEstate10K [53] dataset, we categorize scenes into three difficulty levels: easy, medium, and hard. The categorization is based on the angular distances between views. Specifically, we assess the rotation angle between the anchor view and the furthest target view (θ_{target}), as well as the rotation angle between the anchor and the furthest condition view (θ_{cond}) when more than one condition images are used. Samples are classified as follows: (1) Easy: $\theta_{\text{cond}} < 10$ and $\theta_{\text{target}} < 10$; (2) Medium: $10 \leq \theta_{\text{cond}} < 30$ and $10 \leq \theta_{\text{target}} < 30$; (3) Hard: $60 \leq \theta_{\text{cond}} < 120$ and $30 \leq \theta_{\text{target}} < 60$. We then randomly select 20 samples from the easy set, 60 from the medium set, and 20 from the hard set for evaluation. For the DL3DV [21] dataset, we randomly select 20 test scenes.

Results. We measure performance by comparing generated novel views to reference images using several metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM)[39], and perceptual distance metrics including LPIPS[52] and DISTS [6]. Tab. 1 shows the numerical results on RealEstate10K and Tab. 2 shows the results on the DL3DV test set. As seen, NVComposer outperforms other methods across both datasets. Fig. 4 further demonstrates the visualized comparison among all these methods. For MotionCtrl [40] and CameraCtrl [10], pose controllability is limited. When the target camera poses involve large rotations or translations, these models generate sequences with minimal motion, failing to accurately follow the given instructions. These visual results align with the poor numerical performance observed in Tab. 1.

It is noteworthy that when there are more given input views, the performance of our method consistently increases, as we also showed in Fig. 1 before. In contrast, ViewCrafter [49] suffers from a performance drop when the number of given views increases from one to two in the hard set. This is because the two conditional views in the hard

Model	Views	Easy ($\theta_{cond} < 10$ and $\theta_{target} < 10$)				Medium ($10 \leq \theta_{cond} < 30$ and $10 \leq \theta_{target} < 30$)				Hard ($60 \leq \theta_{cond} < 120$ and $30 \leq \theta_{target} < 60$)			
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
MotionCtrl [40]	1	15.0741	0.6071	0.3616	0.0999	12.0674	0.5667	0.5439	0.1584	11.6381	0.5276	0.5762	0.1633
CameraCtrl [10]	1	13.6082	0.5050	0.4234	0.1458	11.9639	0.4934	0.5217	0.1957	11.7599	0.4716	0.5478	0.2021
DUSt3R [38]	1	13.9443	0.5582	0.3914	0.1565	11.4854	0.4520	0.5570	0.2294	10.9003	0.4029	0.6089	0.2495
	2	17.4837	0.6148	0.3582	0.1503	13.3077	0.4886	0.5434	0.2126	11.5381	0.4003	0.6407	0.2551
	3	17.2341	0.6097	0.3585	0.1504	13.2212	0.4978	0.5287	0.2056	11.9211	0.4387	0.5942	0.2313
	4	17.3545	0.6193	0.3541	0.1481	14.6845	0.5534	0.4892	0.1870	14.2381	0.5280	0.5295	0.1917
ViewCrafter [49]	1	17.3750	0.6670	0.2849	0.1221	13.6015	0.6016	0.4315	0.1762	14.0781	0.5894	0.4293	0.1676
	2	18.8906	0.6685	0.3079	0.1334	14.2891	0.5947	0.4478	0.1761	13.5859	0.5537	0.5100	0.1925
	3	18.4531	0.6548	0.3024	0.1294	14.1172	0.5913	0.4401	0.1717	13.7031	0.5620	0.4867	0.1784
	4	18.4844	0.6553	0.3068	0.1346	14.7421	0.6011	0.4230	0.1672	15.1875	0.5874	0.4327	0.1638
NVComposer (Ours)	1	18.7227	0.7215	0.2354	0.0996	15.3101	0.6056	0.3445	0.1516	15.2115	0.6408	0.4048	0.1462
	2	20.7395	0.7681	0.1781	0.0793	16.9100	0.6445	0.2742	0.1198	15.3461	0.6638	0.3789	0.1384
	3	21.5278	0.7981	0.1522	0.0716	17.7071	0.7418	0.2759	0.1097	15.3825	0.6822	0.3699	0.1324
	4	22.5519	0.8226	0.1188	0.0537	19.5346	0.7847	0.2030	0.0851	17.8181	0.7359	0.2644	0.0988

Table 1. NVS evaluation with varying numbers of input views on RealEstate10K [53] for controllable video models MotionCtrl [40] and CameraCtrl [10], reconstructive model DUSt3R [38], and generative models ViewCrafter [49] and NVComposer. θ_{target} denotes the rotation angle between the anchor view and the furthest target view, while θ_{cond} indicates the angle between the anchor view and the furthest conditional view (when multiple conditions are used). The model that achieves the best average performance is indicated in bold.

Model	Views	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
MotionCtrl [40]	1	13.4003	0.5539	0.4004	0.1396
CameraCtrl [10]	1	12.2995	0.4692	0.4337	0.1829
DUSt3R [38]	1	11.7650	0.4652	0.4900	0.2295
	2	14.6660	0.5158	0.4531	0.2104
	3	13.9156	0.5010	0.4699	0.2127
	4	14.8716	0.5193	0.4478	0.2072
ViewCrafter [49]	1	15.5625	0.4932	0.4122	0.2125
	2	15.6875	0.4775	0.4417	0.2212
	3	14.8593	0.4670	0.4617	0.2273
	4	15.0625	0.4712	0.4549	0.2301
NVComposer (Ours)	1	15.3101	0.6056	0.3445	0.1516
	2	16.9100	0.6445	0.2742	0.1198
	3	17.3115	0.6687	0.2558	0.1122
	4	17.9248	0.6958	0.2277	0.1023

Table 2. NVS evaluation on DL3DV [21]. Our model (noted in bold) consistently reports higher performance when more unposed input views are provided.

set have large rotation differences, *i.e.*, small overlapping regions, and possibly some occlusion are there between the two given views. This makes the external alignment process (explicit pose estimation and pre-reconstruction) tend to produce unstable results, thus leading to poor generative NVS performance.

Distribution Evaluation. In addition to evaluating per-view NVS performance, we assess the distribution of generated novel view sequences using several metrics: Fréchet Inception Distance (FID) [12], Fréchet Video Distance (FVD) [35], and Kernel Video Distance (KVD) [35]. For FID, we treat the novel views as individual images, while for FVD and KVD, we treat them as video clips. We compute these metrics for each model using 1,000 ground truth

Method	FID↓	FVD↓	KVD↓
MotionCtrl [40]	60.83	509.96	14.26
CameraCtrl [10]	52.33	561.97	24.38
ViewCrafter [49]	**46.08**	485.11	13.06
NVComposer (Ours)	46.19	**425.44**	**8.04**

Table 3. Distribution evaluation on generated views of MotionCtrl [40], CameraCtrl [10], ViewCrafter [49], and our NVComposer using FID [12], FVD [35], and KVD [35] metrics. The best performances for each metric are highlighted in bold.

sequences. To ensure fairness, we report results based on single input view conditions. Results are shown in Tab. 3. Our method achieves a comparable FID to ViewCrafter but outperforms it in both FVD and KVD. This suggests that our method produces more accurate novel views when considering the entire multi-view sequence. Overall, our model generates results that are closer to the ground truth, both in terms of image and video generation.

4.2.2. Generative NVS in Objects

In addition to scenes, another important scenario involves generating novel views for synthetic 3D objects. To evaluate the versatility of our proposed pipeline, we compare the generative NVS performance of our model with the object-based generative model SV3D [36] on the Objaverse test set. The numerical and visual results of this comparison are presented in Tab. 4 and Fig. 5. Our model achieves better PSNR and comparable SSIM with SV3D when only a single conditional view is provided. Furthermore, as more unposed input views are added, our model effectively leverages the additional information, producing results that are closer to the real reference.



Figure 4. Visual comparison of NVS results on the RealEstate10K [53] and DL3DV [21] test sets. MotionCtrl [40] and CameraCtrl [10] use the first view as input while other methods use two views as input. MotionCtrl and CameraCtrl produce incorrect camera trajectories. DUSt3R and ViewCrafter exhibit better camera control but introduce artifacts due to occlusions or misaligned multi-view inputs. Our model generates views that are visually closer to the reference. We provide zoomed-in details of the first three scenes in white boxes for a closer look. Additional visual comparisons can be found in the supplementary material.

Model	Views	PSNR↑	SSIM↑	LPIPS↓
SV3D [36]	1	13.8861	0.8130	0.2731
NVComposer (Ours)	1	16.3764	0.8218	0.2286
	2	17.1507	0.8268	0.2067
	4	17.7234	0.8352	0.1889

Table 4. Generative NVS results on the Objaverse [5] test set. When only a single conditional view is provided, NVComposer achieves performance comparable to SV3D [36]. As more random unposed condition views are added, NVComposer 's performance improves significantly.

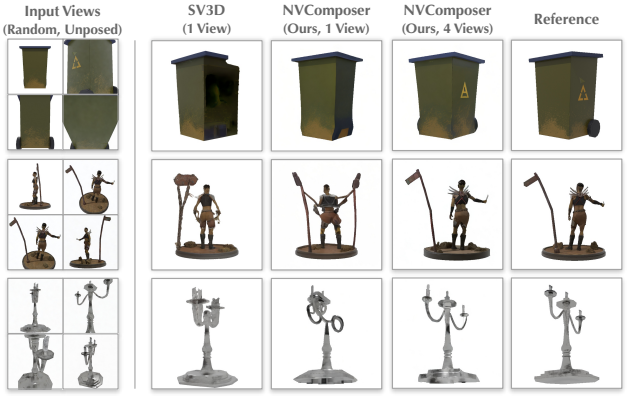


Figure 5. Visual comparison of novel view generation results on the Objaverse [5] test set. All input views are unposed and randomly rendered from the same 3D object.

4.3. Analysis

In this section, we perform several ablation studies and analyses to validate the effectiveness of our model.

Ablation on Image-Pose Dual Stream Diffusion. To ensure both fairness and feasibility, we train two models with and without the dual-stream diffusion design on a subset of Objaverse [5] containing $5,000$ objects for one epoch from the initial weight of the video diffusion model. The evaluation is conducted on a test set with 100 objects. The results presented in Tab. 5 show that the dual-stream diffusion con-

Dual-Stream	PSNR↑	SSIM↑	LPIPS↓
w/	17.0510	0.7501	0.1353
w/o	14.6857	0.7458	0.2095

Table 5. Ablation experiments on dual-stream diffusion on Objaverse [5]. We train the two models (initialized from the same checkpoint) for one epoch on a small subset of Objaverse. The model without dual-stream only generates images instead of the image-pose bundles.

Alignment	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
w/o	14.7218	0.6291	0.3799	0.1494
w/	15.6568	0.6440	0.3284	0.1340

Table 6. Ablation experiments on the geometry-aware feature alignment (*Alignment* in table). We initialize two models with and without the alignment mechanism from the same checkpoint and train the two models for an epoch, then evaluate them on RealEstate10K [53].

siderably enhances the model's performance on generative NVS tasks with multiple unposed condition views.

Ablation on Geometry-Aware Feature Alignment. We further conduct an ablation study on the geometry-aware feature alignment mechanism using the RealEstate10K dataset [53]. In this experiment, we train two models from the same initial checkpoint for one epoch, with and without geometry-aware feature alignment. Tab. 6 demonstrate the numerical results and Fig. 6 shows the visualized results of this ablation study. These results indicate that this feature alignment mechanism helps our model learn the generative NVS task with unposed multiple conditional views.

Sparse-View Pose Estimation Thanks to the unique design of the dual-stream diffusion, NVComposer can implicitly estimate the pose information. We follow the method [51] to solve the camera poses from the Plücker rays generated by the dual-stream diffusion of NVComposer. We perform the evaluation on the RealEstate10K dataset, asking the model to estimate the two sparse condition images given in the easy and hard subsets we discussed in Tab. 1. The accuracy of estimated poses is quantitatively evaluated in the average degrees of rotation angle differences and the average translation difference (with normalization according to the 2-norm of the translation of the furthest view). The results are given in Tab. 7, where we can see that our method is comparable to the performance of DUSt3R [38] in the easy case and outperforms DUSt3R in the hard case. This is because DUSt3R estimates poses using explicit deep feature correspondences, while NVComposer implicitly generates pose information during novel view generation. When given sparse condition views with minimal overlap (*i.e.*, in ill-posed cases), our method's implicit pose estimation proves more robust, delivering accurate pose estimates directly corresponding to the current scene in novel view generation.

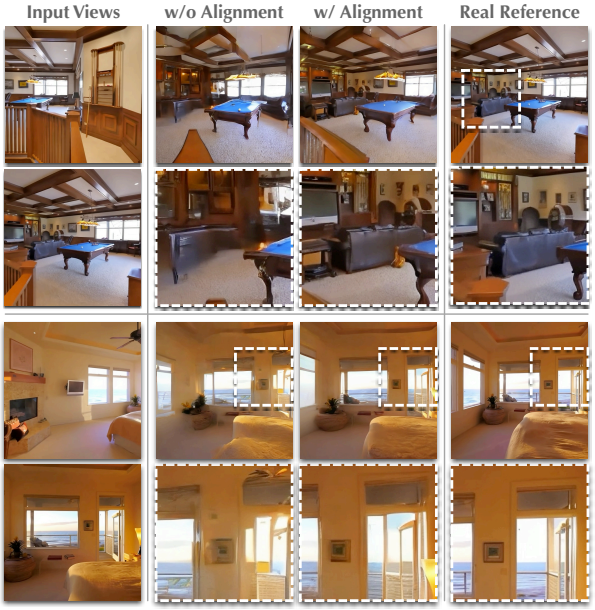


Figure 6. A visual sample in the ablation results of the geometry-aware feature alignment with two input views given. Some patches are zoomed in for a better view. The feature alignment helps NVComposer to properly utilize contents from other views.

Subset	Method	$\Delta\hat{R}\downarrow$	$\Delta\hat{T}\downarrow$
Easy	DUSt3R [38]	9.6968	0.5757
	NVComposer (Ours)	2.7225	0.0257
Hard	DUSt3R [38]	58.3987	0.7603
	NVComposer (Ours)	5.8566	0.0263

Table 7. Comparison with pose estimation accuracy on two spare condition images in our RealEstate10K [53] test sets. Our NVComposer implicitly predicts camera poses by generating ray embeddings of condition views while generating target views.

5. Conclusion

We presented NVComposer, a novel multi-view generative NVS model that eliminates the need for external multi-view alignments, such as explicit camera pose estimation or pre-reconstruction of conditional images. By introducing an image-pose dual-stream diffusion model and a geometry-aware feature alignment module, NVComposer can synthesize novel views from sparse and unposed condition images effectively. Our extensive experiments demonstrate that NVComposer outperforms state-of-the-art methods that rely on external alignment processes. Notably, we show that the model's performance improves as the number of unposed condition views increases, highlighting its ability to implicitly infer spatial relationships and leverage available information from unposed images. This paves the way for more flexible, scalable, and robust generative NVS systems that do not rely on external alignment processes.

Acknowledgments

This work is partially supported by CUHK Direct Grants (RCFUS) No. 4055189.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 3

[2] 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. 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 2

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 5, 7, 8, 1

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5

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

[8] 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. 1, 3, 4

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

[10] 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. 3, 5, 6, 7, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

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

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[16] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 2

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[18] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024. 3

[19] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 5, 6, 7, 1, 3

[22] 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. 1, 2, 3, 4

[23] 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. 1, 2, 3

[24] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 2, 3

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5

[27] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2, 5, 1

[28] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 1, 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[30] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1

[31] 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. 3

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

[33] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2

[34] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 2

[35] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[36] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2, 3, 6, 7, 4

[37] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 2

[38] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1, 2, 3, 4, 5, 6, 8

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[40] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3, 5, 6, 7, 2

[41] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 2

[42] Chin-Hsuan Wu, Yen-Chun Chen, Bolivar Solarte, Lu Yuan, and Min Sun. ifusion: Inverting diffusion for pose-free reconstruction from sparse views. *arXiv preprint arXiv:2312.17250*, 2023. 3

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 1

[44] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3, 4, 1

[45] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2025. 1, 2

[46] 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

[47] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 4

[48] 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. 1

[49] 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. 1, 2, 3, 4, 5, 6

[50] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 3

[51] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 8

[52] 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. 5

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph*, 37, 2018. 2, 5, 6, 7, 8, 1, 3