FrameCrafter: Novel View Synthesis as Video Completion

Qi Wu, Khiem Vuong, Minsik Jeon, Srinivasa Narasimhan, and Deva Ramanan

Carnegie Mellon University, USA
https://frame-crafter.github.io

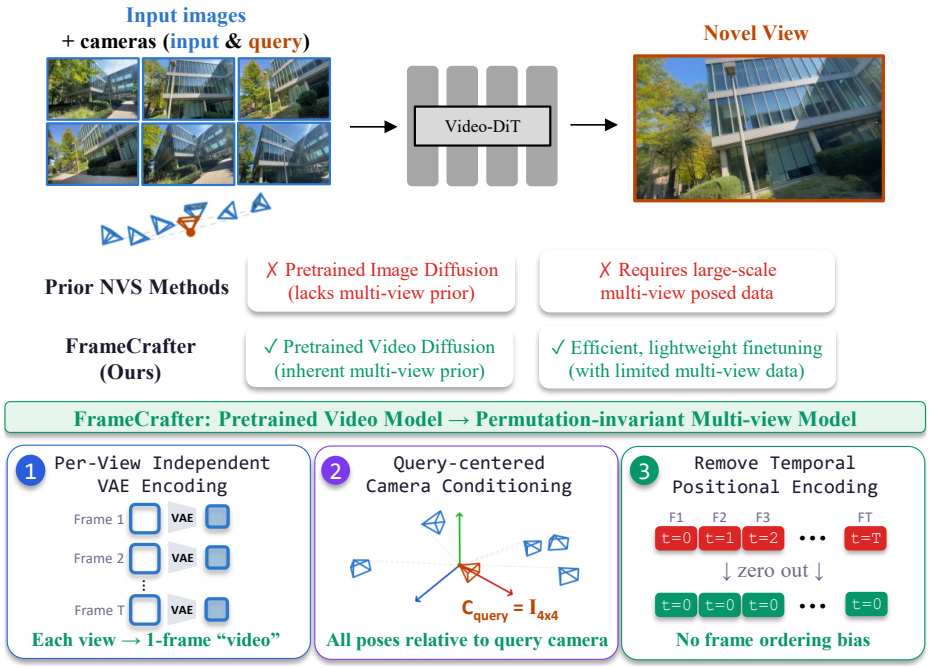

Fig. 1: Video Diffusion as Multi-view Prior. While prior generative NVS methods typically initialize from image diffusion models and rely on large pose-annotated multi-view datasets, videos naturally capture viewpoint changes and cross-view consistency, making them a more *scalable* source of supervision for NVS. We present **FrameCrafter**, which adapts pretrained video diffusion models to permutation-invariant multi-view model with lightweight modifications trained using LoRA on only 1,000 DL3DV-10K scenes, achieving strong performance and demonstrating a more data-efficient path to high-fidelity novel view synthesis.

Abstract. We tackle the problem of sparse novel view synthesis (NVS) using video diffusion models; given K (≈ 5) posed images of a scene, we predict the view from a target camera pose. Many prior approaches leverage generative image priors encoded via diffusion models. However, models trained on single images lack multi-view knowledge, and we instead argue that video models contain implicit multi-view knowledge and are therefore easier to adapt for NVS. Our key insight is to formulate

sparse NVS as a low-frame-rate video completion task. However, one challenge is that sparse NVS is defined over an unordered set of input images, often too sparse to admit a meaningful order, so the models should be *invariant* to permutations of that input image set. To this end, we present ***FrameCrafter***, which adapts video models (naturally trained with coherent frame orderings) to permutation-invariant NVS through simple but effective architectural modifications, including per-frame latent encoding, query-centered camera conditioning and removal of temporal positional embeddings. Our results suggest that video models can be easily modified to "forget" about time with minimal supervision, producing state-of-the-art performance on sparse-view NVS benchmarks.

Keywords: Novel View Synthesis · Video Diffusion Models

1 Introduction

Video diffusion models have advanced rapidly in recent years [3, 4, 6, 19, 39, 46], demonstrating remarkable capabilities in high-fidelity video generation and strong spatio-temporal consistency. Beyond visual realism, these models exhibit increasingly sophisticated world understanding, suggesting an implicit awareness of 3D structure and viewpoint transformation [14]. Motivated by this geometric consistency, prior works have attempted to repurpose video diffusion models for 3D/4D reconstruction [14, 16] and view-conditioned generation [1, 7, 31, 45, 57]. However, their potential for sparse-view novel view synthesis (NVS) — particularly under large viewpoint changes — remains largely underexplored.

Sparse-view novel view synthesis requires a model to understand multi-view relationships and infer geometry that is only partially-observed. Regression methods [18,26] learn deterministic mappings under geometric constraints from input views, but often struggle when geometry is sparsely observed or unseen. Image diffusion [10,20,24,55], in contrast, benefits from large-scale generative pretraining and is better equipped to synthesize plausible content under large viewpoint changes. However, videos, as another scalable source to train diffusion models, are inherently more aligned with the NVS objective: they naturally capture multi-view consistency and viewpoint transitions over time. Despite this, most prior diffusion-based methods for sparse-view NVS rely on image diffusion backbones and require substantial multi-view supervision to induce 3D awareness.

Our key thesis is as follows: *video* diffusion models provide a stronger prior for sparse-view NVS than *image* diffusion models. We leverage geometric reasoning that has already emerged from large-scale video pretraining, instead of learning multi-view geometry "from scratch" on multi-view training data. While perhaps obvious in retrospect, our approach is nontrivial to operationalize because one must modify a video model to accept unordered collections of views (and cameras) as inputs.

To do so, we introduce **FrameCrafter** with three key architectural modifications. The first and most important is to modify the video VAE to encode a K-frame video as a collection of K independent 1-frame videos, allowing multi-

view inputs to be treated as an unordered set. This is actually a small architectural change, as current image-to-video diffusion models already treat the input image as a 1-frame video during encoding. Secondly, many NVS solutions break permutation invariance by defining camera poses in a world coordinate system aligned to the first input image. Instead, we define world coordinates aligned to the novel query view (to be synthesized). In practice, we find that video diffusion models can be easily taught to condition on cameras by channel-concatenating latent (1-frame) video encodings with Plücker raymaps. Thirdly, we remove the temporal component of 3D Rotary Positional Embeddings (RoPE) [37] so that the model cannot rely on frame indices when processing input views. We demonstrate that our lightweight architectural modifications can turn state-of-the-art video models into state-of-the-art multi-view models that are invariant to permutations of their inputs.

Our results show that, with only a small number of (LoRA [13]) trainable parameters and minimal training data (1K scenes from DL3DV [22]), we can finetune video diffusion models to outperform prior image diffusion-based models. We further observe a consistent scaling trend across backbones, suggesting that progress in large-scale video foundation models can be directly transferred to NVS without requiring massive curated multi-view datasets. To our knowledge, FrameCrafter is the first approach to turn video diffusion models into permutation-invariant multi-view generators, leveraging their strong multi-view priors for sparse-view NVS while removing their reliance on temporal ordering.

2 Related Work

Novel View Synthesis. Early NVS methods are primarily regression-based, learning radiance fields or renderable scene representations from multi-view inputs using frameworks such as NeRF [26] and its variants [2, 5, 41, 47], as well as explicit representations such as 3D Gaussian Splatting [18]. Subsequent works improve robustness via regularization [27] or feed-forward scene prediction [15, 17, 50, 52]. More recently, diffusion-based approaches have emerged as a powerful alternative by leveraging the geometric and appearance priors of pretrained diffusion backbones [10, 20, 24, 25, 34, 44, 55]. By incorporating large-scale generative priors, these methods synthesize more realistic novel views and generalize better under large viewpoint changes. Nevertheless, achieving reliable multi-view consistency often relies on substantial pose-annotated multi-view datasets [8, 22, 30, 56]. This reliance becomes particularly important in sparse-view settings, where limited observations increase geometric ambiguity and place greater demands on cross-view reasoning. Recent state-of-the-art models such as SEVA [55] demonstrate strong cross-domain performance by combining large pretrained image diffusion backbones with extensive multi-view supervision. In contrast, we explore video diffusion models as scalable priors for sparse-view NVS, leveraging cross-frame consistency from web-scale video pretraining and directly repurposing pretrained video backbones for geometric reasoning without relying on large curated multi-view datasets.

Video Diffusion Models and 3D-Aware Video Generation. Modern video diffusion architectures [39, 46] typically adopt latent diffusion frameworks [32] with spatio-temporal VAEs [4] and transformer-based denoisers [28], achieving strong temporal consistency and scalability through web-scale video pretraining. Beyond generic video generation, several works incorporate camera control or 3D-aware conditioning into video models. One line of research [7,31,48,54] first reconstructs a scene and then applies video diffusion to inpaint missing regions during rendering. These approaches provide precise camera trajectory control but rely on off-the-shelf reconstruction modules and pre-rendered intermediate representations. Another line of work directly conditions video diffusion models on camera poses by injecting pose parameters into the model's latent representations [38,43,45,57] or integrating them into attention mechanisms [11,53]. These approaches enable camera-controllable generation and demonstrate strong cross-frame coherence without explicit reconstruction. However, despite their advances in controllable video synthesis and ability to generate viewpoint changes, the output of such models remains a continuous, temporally ordered video sequence rather than discrete novel view predictions derived from sparse observations. The use of video diffusion models for sparse-view NVS—where both the inputs and outputs are discrete views rather than temporal sequences—remains largely unexplored. In this work, we directly repurpose pretrained video diffusion models for sparse-view NVS without relying on pre-reconstruction or predefined camera trajectories. By removing temporal compression and enforcing permutation-invariant per-view encoding, we transform a video generation model into an effective multi-view synthesis backbone with minimal additional supervision.

3 Method

We present FrameCrafter, a method for sparse-view novel view synthesis (NVS) by repurposing a pretrained video diffusion model as a multi-view generator. Given K posed input images of a scene, our model synthesizes a novel target view at a specified camera pose. The key insight is that a video diffusion transformer, trained on large-scale video data, already possesses strong priors for 3D-consistent scene understanding. We unlock this capability for NVS by (i) encoding each input view *independently* through the VAE to ensure permutation invariance (Sec. 3.2), (ii) injecting camera geometry via Plücker ray conditioning while removing the temporal component of 3D Rotary Positional Embeddings to avoid reliance on frame order (Sec. 3.3), and (iii) fine-tuning with LoRA on multi-view data (Sec. 3.4). An overview of FrameCrafter is shown in Fig. 2.

3.1 Preliminaries

Novel View Synthesis. Given a set of K input images $\{\mathbf{I}_k\}_{k=1}^{K}$ of a scene captured from known camera poses $\{\boldsymbol{\pi}_k\}_{k=1}^{K} \in \mathrm{SE}(3)$, novel view synthesis (NVS) aims to generate a photorealistic image $\hat{\mathbf{I}}_{\mathrm{tgt}}$ at a novel target pose $\boldsymbol{\pi}_{\mathrm{tgt}}$. Classical methods [18, 26] reconstruct explicit or implicit 3D representations before

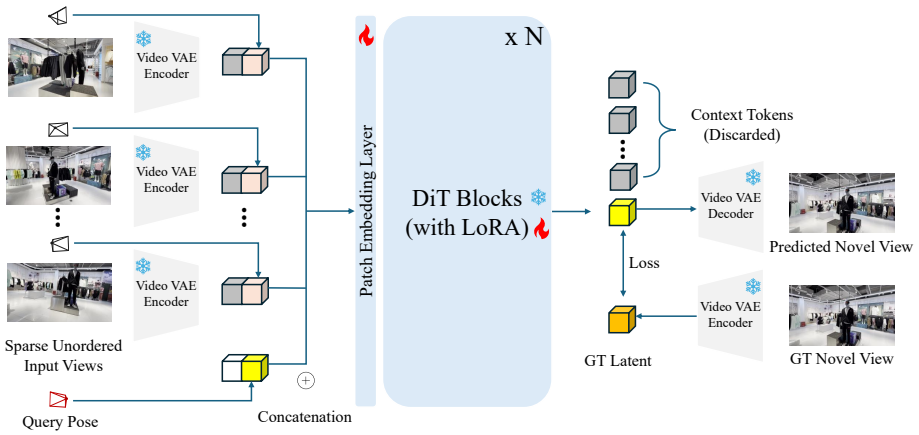

Fig. 2: Each input image is encoded by a frozen video VAE, and its camera pose is converted into a Plücker ray map that is concatenated with the image latent along the channel dimension. The resulting view tokens are then concatenated along the temporal dimension, where the first K tokens correspond to input views and the final token represents the query view. Only the patch embedding layer and the LoRA modules in the DiT backbone are trained. During both training and inference, only the predicted query latent is used for decoding and supervision.

rendering novel views, but degrade with sparse inputs due to limited geometric constraints. Recent methods [10,24,55] instead directly predict target views from input images, bypassing 3D reconstruction and relying more on learned priors.

Diffusion Models for Video Generation. Diffusion models [12, 36] generate data by learning to reverse a gradual noising process. In video generation, a latent diffusion framework [32] is commonly adopted: a variational autoencoder (VAE) maps video frames into a compact latent space, and a denoising network–typically a Diffusion Transformer (DiT) [28]–operates in this latent domain.

Formally, a video $\mathbf{V} \in \mathbb{R}^{3\times T\times H\times W}$ is encoded by a VAE encoder $\mathcal{E}$ into latent representations: $\mathbf{z}_0 = \mathcal{E}(\mathbf{V}) \in \mathbb{R}^{d_z\times T'\times h\times w}$, where d_z is the latent channel dimension, $h = H/f_s$, $w = W/f_s$ denote spatially downsampled resolutions with factor f_s, and T' is the temporally compressed frame count.

During training, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added using a flow-matching schedule [23]: $\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}$, where $t \in [0, 1]$, and the DiT $\boldsymbol{\epsilon}_\theta$ is trained to predict the velocity field $\mathbf{v}_t = \boldsymbol{\epsilon} - \mathbf{z}_0$.

Causal 3D VAE in Modern Video Diffusion. Modern video diffusion models such as CogVideoX [46] and Wan [39] employ temporally-causal 3D convolutions in the VAE to ensure future frames do not influence past frames. The first frame is processed independently as a single-frame "video", while each subsequent group of four frames contributes one additional latent timestep, resulting in an effective temporal length of $T' = 1 + (T-1)/4$. Due to the temporal receptive field of causal convolutions, each frame encoding depends on all preceding frames,

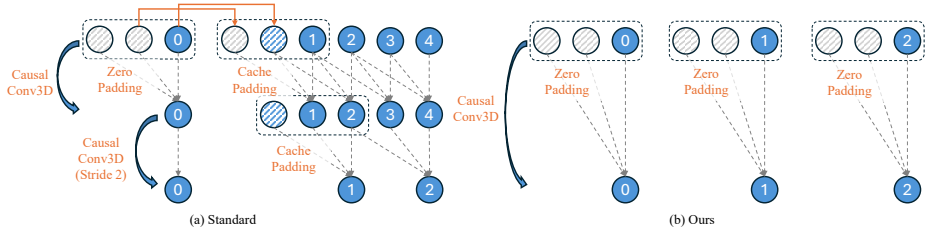

Fig. 3: Comparison between standard causal video VAE encoding and our per-view independent encoding. (a) Standard causal encoding mixes frame information through feature caching and temporal stride, leading to inter-frame dependency. (b) Our per-view encoding eliminates temporal interaction, ensuring independent view-to-latent mapping suitable for unordered sparse-view inputs.

introducing a dependency on frame order. While desirable for video generation, this is undesirable for processing unordered sparse-view inputs.

3.2 Per-View VAE Encoding for Sparse Views

A naive approach to multi-view NVS with a video diffusion model would concatenate the K input views into a video sequence and encode them jointly through the temporal VAE. However, this is problematic for sparse-view NVS:

(1) Temporal compression degrades per-view information. The causal video VAE reduces temporal resolution via strided 3D convolutions (typically by a factor of $\sim$4), mapping multiple frames to fewer latent timesteps. For a video, this may be desirable—it removes redundant information while capturing motion dynamics. For sparse multi-view images that depict the scene from arbitrarily-ordered viewpoints, this compression is harmful: potentially far away viewpoints can be merged into a shared encoding, losing fine-grained per-view information.

(2) Causal ordering violates permutation invariance. Sparse input views have no natural temporal ordering—the set $\{\mathbf{I}_1, \ldots, \mathbf{I}_K\}$ should be treated as unordered. The causal 3D convolutions, however, encode a strict left-to-right temporal dependency: the latent of frame k is conditioned on frames $1, \ldots, k-1$ through feature caching. This means permuting the input order produces different latent representations, introducing an undesirable inductive bias. As illustrated in Fig. 3(a), standard causal VAE encoding propagates multi-frame information through feature caching and temporal downsampling, resulting in ordering dependency and latent compression.

To address these issues, we adopt *per-view independent encoding* (Fig. 3(b)). Notably, in image-to-video (I2V) generation, the first frame is processed independently without temporal compression, serving as a clean conditioning signal. Inspired by this design, we similarly encode each input view individually as a single-frame "video":

$$\mathbf{z}_k = \mathcal{E}(\mathbf{I}_k) \in \mathbb{R}^{d_z \times 1 \times h \times w}, \quad k = 1, \ldots, K, \tag{1}$$

and concatenate the results along the temporal axis:

$$\mathbf{z}_{\text{ctx}} = [\mathbf{z}_1; \mathbf{z}_2; \ldots; \mathbf{z}_K] \in \mathbb{R}^{d_z \times K \times h \times w}. \tag{2}$$

For the target view, we encode a zero placeholder $\mathbf{z}_{\text{tgt}} = \mathcal{E}(\mathbf{0})$ and append it:

$$\mathbf{z}_{\text{all}} = [\mathbf{z}_{\text{ctx}};\ \mathbf{z}_{\text{tgt}}] \in \mathbb{R}^{d_z \times (K+1) \times h \times w}. \tag{3}$$

Since each frame is encoded as a standalone single-frame input, the causal temporal convolutions have no preceding frames to condition on—each view's latent is computed identically and independently of its position. This guarantees:

- **Permutation-invariant encoding**: $\mathcal{E}(\mathbf{I}_k)$ is independent of the ordering of the other views.
- **No temporal compression**: each view maps to exactly one latent frame ($T' = 1$ per view), preserving full spatial detail and every view receives the same single-frame encoding treatment, regardless of K.

During decoding, we similarly decode each latent frame independently:

$$\hat{\mathbf{I}}_k = \mathcal{D}(\mathbf{z}_k), \quad k = 1, \ldots, K+1, \tag{4}$$

where $\mathcal{D}$ is the VAE decoder applied per-frame. This avoids temporal blending artifacts in the decoded output.

3.3 Camera-Conditioned Diffusion Transformer

To enable geometric control, we condition the DiT on camera pose information via Plücker ray maps [29, 49].

Plücker ray maps. For each camera with intrinsic matrix $\mathbf{K}_k$ and camera-to-world extrinsic $\mathbf{T}_k \in \text{SE}(3)$, we compute a dense ray map. For each pixel (u, v), the ray origin $\mathbf{o}$ and direction $\mathbf{d}$ are computed via back-projection, and the Plücker representation is: $\mathbf{p}(u, v) = \left[\hat{\mathbf{d}};\ \mathbf{o} \times \hat{\mathbf{d}}\right] \in \mathbb{R}^6$, where $\hat{\mathbf{d}} = \mathbf{d}/\|\mathbf{d}\|$ is the normalized ray direction. Per-pixel Plücker coordinates are downsampled to the latent resolution via *pixel unshuffle* [35] with factor f_s, converting 6 channels at resolution $H \times W$ into $6f_s^2$ channels at resolution $h \times w$. Unlike interpolation-based downsampling, pixel unshuffle preserves per-pixel ray information by rearranging spatial details into channels without loss, maintaining precise geometric correspondence with VAE latents and retaining fine-grained camera geometry. Much prior work defined cameras (and rays) with respect to a world coordinate system aligned to the first input view [40, 42], but this introduces a dependence on the ordering of input views. Instead, we define the world coordinate system with respect to the query view, preserving invariance (to permutations of input views). Finally, we define the scale of the world coordinate system by normalizing all input cameras to have a mean distance of unit length (as in past work [33]).

Input channel expansion. The DiT's input is formed by channel-wise concatenation of three components:

$$\mathbf{x} = [\mathbf{z}_t;\ \mathbf{y};\ \mathbf{r}] \in \mathbb{R}^{D_{\text{in}} \times (K+1) \times h \times w}, \tag{5}$$

where $\mathbf{z}_t$ is the noisy latent, $\mathbf{y}$ is the VAE-encoded embedding concatenated with a binary mask indicating input vs. target frames, and $\mathbf{r}$ is the Plücker ray map. The patch embedding layer (a 3D convolution) is re-initialized to accommodate the expanded channel count, while all other pretrained weights are preserved.

Zero-Temporal Positional Encoding. Standard video diffusion transformers employ Rotary Positional Embeddings (RoPE) [37] to encode temporal order. While appropriate for video generation, such encodings implicitly assume a sequential structure and introduce ordering bias across frames. In sparse-view NVS, the input views form an unordered set; injecting temporal position embeddings would therefore violate permutation invariance.

To preserve invariance, we remove temporal positional encoding entirely and treat the $(K+1)$ latent tokens as an unordered set. As a result, geometric identity is provided solely through camera conditioning via Plücker ray maps. This design encourages the transformer to reason over views based on camera geometry rather than frame index. In practice, we observe that removing temporal positional encoding slightly slows early training convergence, but leads to more stable multi-view generalization.

CLIP and text Embedding. To preserve permutation invariance, we additionally disable the first-frame CLIP image conditioning used in the original Wan2.1-I2V model, while using an empty text prompt for global text conditioning.

3.4 Training

Dataset and sampling. We train on multi-view images from a subset of DL3DV-10K [22], which provides diverse real-world scenes with known camera parameters. At each iteration, we sample $K+1$ frames from a scene: K input views and 1 target view. We employ a probabilistic sampling strategy: with 80% probability, frames are sampled uniformly in random order from the full video; with 20% probability, frames are sampled from a local temporal window to encourage learning from nearby viewpoints.

LoRA fine-tuning. Rather than full fine-tuning, we apply Low-Rank Adaptation (LoRA) [13] to the attention and feed-forward layers of the pretrained DiT. Due to the change in input dimensionality, we re-initialize the patch embedding layer and train it with full gradients. This strategy preserves the pretrained video generation prior while efficiently adapting the model for camera-conditioned NVS.

Objective. We train with the standard flow-matching loss [23] applied *only* to the target frame: $\mathcal{L} = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{z}_0} \left\| \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{z}_0) \right\|_{\text{tgt}}^2$, where $\mathbf{c}$ denotes all conditioning signals (input latents, mask, ray maps, and text embeddings), and $\|\cdot\|_{\text{tgt}}^2$ indicates that the loss is masked to the target frame only.

4 Experiments

4.1 Experiment Setup

Implementation Details. We adopt Wan2.1-I2V-14B [39] as our video diffusion backbone. Training is conducted in two stages: we first train the model at a resolution of 192×336, and then at a higher resolution of 480×832 to enhance visual fidelity. Low-Rank Adaptation (LoRA) with rank r=32 is applied to the DiT backbone, and only the patch embedding layer and LoRA parameters are updated, while all other pretrained weights remain frozen. This lightweight adaptation strategy ensures that only a small fraction ($\sim$ 1%) of parameters are trained, significantly reducing training cost compared to full fine-tuning baselines. We also extend the model to support multi-target prediction via mixed m-to-n training (see supplementary).

Benchmark and Metrics. We evaluate our method on DL3DV-Benchmark [22] and Mip-NeRF 360 [2], which contain diverse real-world scenes with calibrated camera poses and multi-view images. For fair comparison, we follow the train/test split from SEVA [55]. We report performance using standard image reconstruction and perceptual metrics, including PSNR, SSIM, LPIPS [51], and DreamSim [9]. PSNR measures pixel-level fidelity, SSIM structural similarity, while LPIPS and DreamSim quantify perceptual similarity in learned feature spaces.

Baselines. We compare against representative generative NVS approaches. (1) EscherNet [20], a scalable view synthesis model built upon image diffusion priors, which we fine-tune on the 10K scene-level multi-view dataset (DL3DV [22]) following its original training protocol. (2) Aether [57], a recent action-conditioned video diffusion model designed for camera-controlled generation. (3) SEVA [55], state-of-the-art generative method which adapts an image diffusion model to multi-view synthesis using large-scale pose-annotated data.

4.2 Quantitative Results

We evaluate all methods under a unified protocol. As SEVA produces square images at 576×576 and video diffusion models operate at different aspect ratios, we resize or center-crop all predictions to 480×480 for metric computation, ensuring a fair comparison without modifying the original generation settings.

Tab. 1 reports quantitative results on the 6-view NVS setting.

We first compare with EscherNet [20]. Despite using only 1K training scenes and updating only a small fraction of parameters via LoRA, our model substantially outperforms EscherNet across all metrics. These results indicate that video diffusion priors provide significantly stronger multi-view consistency than image-based generative priors under sparse-view settings.

Compared to Aether [57], our method achieves consistent improvements when using the same CogVideoX-5B backbone. Aether is a camera-conditioned video

Table 1: Quantitative comparison on 6-view NVS on DL3DV-Benchmark and Mip-NeRF 360. FrameCrafter uses pretrained video diffusion backbones and is trained on only 1K multi-view scenes. Scaling the backbone (CogVideoX-5B → Wan2.1-14B) consistently improves performance, achieving competitive results with prior state-of-the-art methods. Best results are highlighted as first, second, third. DS denotes DreamSim [9]. For completeness, we also compare to regression baselines (top two rows) which tend to consistently output blurrier images that produce poor LPIPS and DS but better PSNR and SSIM.

Method	Backbone	#Scenes	DL3DV				Mip360			
			PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓
LVSM [17]	–	–	17.09	0.478	0.333	0.204	15.25	0.317	0.609	0.577
E-RayZer [52]	–	–	16.85	0.442	0.455	0.254	16.56	0.343	0.621	0.340
EscherNet [20]	SD1.5 [32]	10K	12.07	0.251	0.484	0.227	11.14	0.126	0.540	0.315
Aether [57]	CogVideoX-5B [46]	–	12.66	0.258	0.469	0.140	12.60	0.220	0.651	0.334
SEVA [55]	SD2.1 [32]	80K	16.15	0.470	0.253	0.088	14.59	0.294	0.372	0.137
Ours	CogVideoX-5B [46]	1K	13.65	0.272	0.389	0.113	13.09	0.209	0.540	0.204
Ours	Wan2.1-14B [39]	1K	17.18	0.445	0.223	0.066	15.64	0.279	0.365	0.111

diffusion model that also incorporates camera ray maps via channel concatenation and generates frames along a specified camera trajectory, for example by conditioning on the first and last frames and interpolating intermediate poses. While such a formulation can in principle include the query pose, it assumes a coherent temporal sequence and relies on temporally compressed video encodings, which makes precise pose control more difficult. In contrast, our formulation explicitly treats the inputs as an unordered set of views and preserves per-view representations, enabling more accurate geometry-aware generation for sparse-view NVS. When scaling to Wan2.1-14B, performance further improves substantially, highlighting the importance of stronger pretrained video backbones.

Most notably, we compare against SEVA [55], the state-of-the-art image diffusion-based method, which adapts image diffusion models for multi-view synthesis using substantially larger supervision, including approximately 80K scene-level samples and 340K object-level samples. Despite being trained on only 1K scene-level samples – a large disparity in supervision scale – our model achieves competitive and often superior performance. With the Wan2.1-14B backbone, our method surpasses SEVA on PSNR, LPIPS, and DreamSim on DL3DV, while also achieving the best DreamSim and PSNR on Mip-NeRF 360. These results suggest that video diffusion priors provide strong geometric and perceptual cues for sparse-view synthesis, enabling high-quality predictions even with significantly less multi-view supervision.

These results strongly support our central claim: large-scale video diffusion models already encode powerful geometric and cross-view priors, and sparse-view NVS can be achieved through lightweight specialization rather than massive multi-view training. Furthermore, the consistent improvement from CogVideoX-5B to Wan2.1-14B reveals a clear backbone scaling trend. As video diffusion models continue to advance, our adaptation framework with lightweight training

Table 2: Ablation study on 192×336 resolution generation (on DL3DV-Benchmark). Removing the per-view independent encoding, pixel unshuffle, or target-only supervision degrades performance, with per-view "single-frame" video encodings being the most crucial design choice.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓
✗ Per-View Encoding	11.50	0.147	0.676	0.656
✗ Pixel Unshuffle	13.53	0.212	0.359	0.129
✗ Target-Only Supervision	15.08	0.287	0.270	0.116
w/ PRoPE	14.91	0.289	0.315	0.174
w/ Original RoPE	15.30	0.300	0.265	0.122
Full model	**15.69**	**0.326**	**0.246**	**0.114**

provides a direct and efficient pathway for transferring these gains to sparse-view novel view synthesis.

We also include two regression-based models as reference. Regression-based models often achieve higher PSNR and SSIM due to direct pixel supervision, but typically produce blurrier or less perceptually realistic results in such cases, which is reflected by worse LPIPS and DreamSim scores. Qualitative comparisons further illustrate this difference (see Sec. 4.3). More quantitative results are provided in the supplementary.

4.3 Qualitative Results.

We present qualitative comparisons in Fig. 4. Under sparse-view inputs, our model is able to recover plausible geometry even when certain structures appear only in small regions of the input views, whereas SEVA occasionally fails to maintain geometric consistency. Compared with Aether, our method produces more accurate viewpoint alignment and sharper visual details, which we attribute to the per-view independent encoding and our formulation of sparse unordered views within the video diffusion backbone. LVSM tends to produce overly smooth predictions with noticeable blur and loss of fine structures, particularly when synthesizing views that require extrapolation from sparse observations. More qualitative comparisons are provided in the supplementary.

4.4 Ablation Study

Design Components Tab. 2 evaluates the contributions of pixel unshuffle, per-view independent encoding, supervision strategy, and positional encoding.

Per-View Independent Encoding. We also conducted an experiment where, instead of encoding each view independently, we applied the standard spatio-temporal encoding of the causal video VAE. Removing individual view encoding and jointly encoding all frames leads to substantial performance degradation. This validates our key design choice to eliminate temporal compression and the ordering bias introduced by causal 3D convolutions, which is particularly important for unordered sparse view inputs.

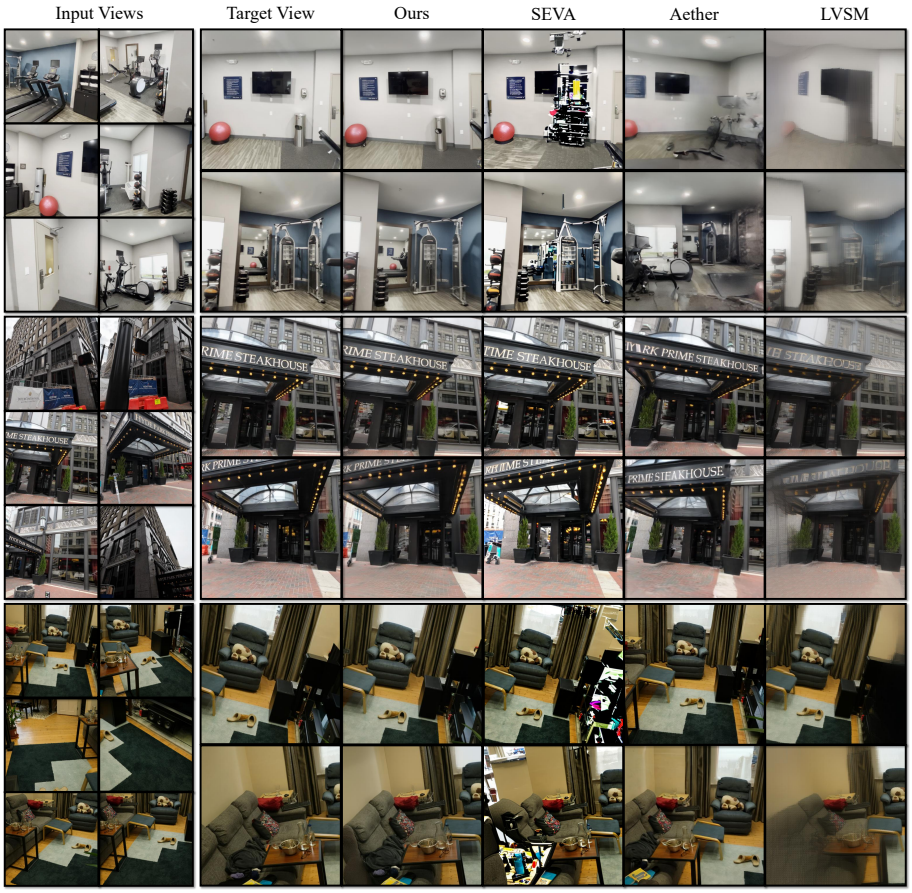

Fig. 4: Qualitative comparison under sparse inputs. FrameCrafter produces more geometrically consistent and sharper novel views compared to SEVA [55], Aether [57], LVSM [17], with improved viewpoint alignment and detail preservation.

Pixel Unshuffle. Replacing pixel unshuffle [35] with standard spatial downsampling significantly degrades performance. We attribute this to the fact that unshuffling produces latent features with a higher effective channel dimension, providing richer geometric context that better aligns with pose conditioning. This confirms that preserving exact per-pixel ray geometry at the latent resolution is critical for accurate view-conditioned generation.

Supervision Strategy. We compare target-only supervision with full-sequence supervision, where input views are also supervised, and find that supervising input views hurts performance. This is likely because the loss is dominated by reconstructing input frames, reducing emphasis on query view prediction. This insight may benefit other conditional generation tasks with clean input frames [38, 45].

Positional Encoding. We compare against retaining full spatio-temporal Rotary Positional Encoding (3D RoPE). Removing the temporal component yields im-

proved performance quality while preserving permutation invariance (see Sec. 4.4). We attribute this to the fact that disabling temporal RoPE encourages the transformer to rely more on explicit camera conditioning rather than implicit frame index cues, which are unnecessary for unordered sparse-view inputs. We also experimented with replacing the original RoPE with Projective Positional Encoding (PRoPE) [21], a relative positional encoding technique that captures complete camera frustums (both intrinsics and extrinsics) in attention conditioning. While PRoPE has been shown to improve performance in feedforward NVS models, we did not observe consistent gains in our setting under minimal training (LoRA with limited multi-view data). We conjecture that PRoPE modifies the attention structure more significantly, and fully adapting the pretrained backbone to this change may require larger-scale retraining.

Overall, each component contributes meaningfully, and their combination achieves the best reconstruction and perceptual performance.

Permutation Invariance We evaluate the sensitivity of different design choices to input ordering. Specifically, we compare: (i) a first-view coordinate formulation (where poses are normalized relative to the first input view), (ii) model using standard spatio-temporal RoPE, and (iii) our design with temporal RoPE removed and target-view coordinate frame. At test time, we evaluate each model under 10 random permutations of the same input set on DL3DV-Benchmark [22]. We report the mean performance across all runs together with the mean standard deviation across permutations (Tab. 3).

Table 3: Performance under 10 random permutations of input views.

Training Setup	PSNR		SSIM		LPIPS		DreamSim	
	Mean ↑	Std ↓	Mean ↑	Std ↓	Mean ↓	Std ↓	Mean ↓	Std ↓
First-view Coordinate	15.67	0.4730	0.309	0.0276	0.262	0.0203	0.112	0.0111
w/ original RoPE	15.61	0.3106	0.310	0.0212	0.254	0.0095	0.115	0.0063
Ours	**16.21**	**0.0081**	**0.348**	**0.0006**	**0.224**	**0.0003**	**0.102**	**0.0005**

Models that rely on a first-view coordinate exhibit degraded performance and high variance, suggesting that this design introduces bias that harms generalization. Using standard temporal RoPE yields similar performance, but still shows non-negligible variance across permutations due to the implicit ordering bias in temporal positional encoding. Removing the temporal component of RoPE and adopting a target-view coordinate frame architecturally eliminate input-order bias, further improving mean performance and significantly reducing permutation variance, demonstrating strong robustness to input reordering.

One may ask why permutation invariance is important for sparse-view NVS, and whether it is possible to simply choose a suitable ordering of the input views. To further investigate this question, we report the performance when the model is trained and evaluated using an "ordered" input sequence (w/ temp RoPE).

As shown in Tab. 4, even when the model is trained and evaluated using ordered inputs, its performance remains worse than the model trained with randomly shuffled inputs under the same ordered evaluation. This compares two

possible problem formulations: if sparse-view NVS is treated as an ordered sequence modeling problem, the ordered model represents a natural upper attempt under this setting; however, if the inputs are instead treated as an unordered set, random permutations become valid training samples. The improvement suggests that sparse-view observations do not naturally correspond to a smooth temporal trajectory: input cameras often exhibit large viewpoint changes and irregular spatial distributions around the scene, making it difficult to define a meaningful global ordering. Therefore, permutation-invariant architecture and training better match the nature of sparse-view NVS. Also this formulation can serves as an effective form of data augmentation, since each scene can be observed under many equivalent input orderings, exposing the model to a much richer set of training configurations. Permutation-invariant training can be viewed as implicitly leveraging a factorially larger set of permuted training samples.

Table 4: Effect of training setup under ordered evaluation on 192×336 resolution 6-view NVS (DL3DV-Benchmark [22]). During evaluation, input views follow their original temporal order in the source video.

Training Setting	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓
Ordered	15.04	0.292	0.276	0.136
Ours (Permutation-Invariant)	**15.69**	**0.326**	**0.246**	**0.114**

Our architectural design naturally supports this behavior. In particular, removing the temporal component of RoPE, using per-view independent VAE encoding, and adopting query-centered camera coordinate normalization together ensure that the model does not rely on a specific ordering of the input views. This allows the network to reason more effectively over unordered sets of views and improves robustness in sparse-view settings.

Data Scaling We further analyze the effect of training data scale on model performance (Fig. 5) to examine the limits of minimal-data adaptation. Remarkably, even when trained on only 20 scenes, our model surpasses EscherNet trained on 10K scenes. This highlights the strength of video diffusion models as geometric priors: the lightweight training primarily adapts the model to the sparse-view setting and camera conditioning, whereas multi-view reasoning appears to already emerge from large-scale video pretraining.

As the number of training scenes increases, performance improves con-

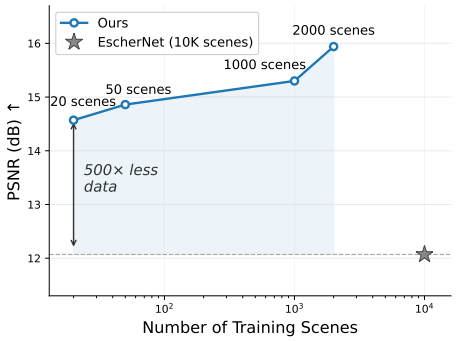

Fig. 5: Performance on different numbers of training scenes. Even with only 20 scenes, our model surpasses EscherNet trained on 10K scenes. Performance improves consistently as training data increases, demonstrating strong data efficiency and scalability.

sistently, revealing a clear scaling trend. These results suggest that video diffusion priors provide a strong geometric foundation, enabling sparse-view NVS to benefit from additional data without requiring large-scale multi-view supervision.

5 Conclusion

We revisit sparse-view novel view synthesis from the perspective of video generation. Instead of adapting image diffusion models with large-scale multi-view supervision, we demonstrate that modern video diffusion models already encode strong geometric and multi-view priors. By introducing per-view independent encoding, query-centered Plücker-based geometric conditioning, temporal RoPE removal and lightweight LoRA adaptation, we transform a pretrained video diffusion backbone into an effective permutation-invariant sparse-view NVS model using only 1K training scenes.

Our results show that this approach achieves competitive performance against state-of-the-art methods trained with substantially larger multi-view datasets. Moreover, performance improves consistently as the video backbone scales, suggesting that advances in large-scale video diffusion models can be directly transferred to novel view synthesis without requiring massive curated multi-view data. These findings point toward a scalable paradigm for NVS: rather than treating multi-view synthesis as a standalone modeling problem, it can be viewed as a lightweight specialization of increasingly powerful video foundation models.

Our analysis also suggests a provocative take on the knowledge emergent in video generative "world models." Indeed, we were surprised by the ease with which such models can be tuned to be permutation-invariant, losing any sense of temporal dynamics. This suggests that current models might capture more "multi-view" knowledge about the world, rather than the dynamics of how it evolves over time. At the very least, such temporal knowledge is surprisingly easy to unlearn. We hope our work leads to more exploration of such representational questions. More broadly, it points to the potential of adapting pretrained video models to non-temporal tasks and using video diffusion priors as alternatives to image diffusion priors.

Acknowledgments: We thank Prof. Shubham Tulsiani for insightful discussions and perspectives, and Nikhil Keetha and other members of Deva's lab at CMU for valuable feedback and suggestions.

This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation `cis250177p` from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work was supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

1. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. 2024. URL https://openai. com/research/video-generation-models-as-world-simulators **3**(1), 3 (2024)
5. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14124–14133 (2021)
6. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
7. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. arXiv preprint arXiv:2507.12646 (2025)
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)
9. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
10. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
11. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
14. Huang, Z., Li, X., Lv, Z., Rehg, J.M.: How much 3d do video foundation models encode? arXiv preprint arXiv:2512.19949 (2025)
15. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., et al.: Rayzer: A self-supervised large view synthesis model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4918–4929 (2025)

16. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4d: Leveraging video generators for geometric 4d scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20658–20671 (2025)
17. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. arXiv preprint arXiv:2410.17242 (2024)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
20. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9503–9513 (2024)
21. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding. arXiv preprint arXiv:2507.10496 (2025)
22. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., 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. pp. 22160–22169 (2024)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
24. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
25. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5480–5490 (2022)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
29. Plucker, J.: Xvii. on a new geometry of space. Philosophical Transactions of the Royal Society of London (155), 725–791 (1865)
30. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: 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. pp. 10901–10911 (2021)
31. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6132 (2025)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

33. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9420–9429 (2024)
34. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
35. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
36. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
37. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
38. Van Hoorick, B., Chen, D., Iwase, S., Tokmakov, P., Irshad, M.Z., Vasiljevic, I., Gupta, S., Cheng, F., Zakharov, S., Guizilini, V.C.: Anyview: Synthesizing any novel view in dynamic scenes. arXiv preprint arXiv:2601.16982 (2026)
39. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
40. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
41. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2021)
42. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
43. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
44. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)
45. Xiao, Z., Zhao, Y., Li, L., Lan, Y., Yu, N., Garg, R., Cooper, R., Taghavi, M.H., Pan, X.: Video4spatial: Towards visuospatial intelligence with context-guided video generation. arXiv preprint arXiv:2512.03040 (2025)
46. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

47. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4578–4587 (2021)
48. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
49. Zhang, J.Y., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. arXiv preprint arXiv:2402.14817 (2024)
50. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: European Conference on Computer Vision. pp. 1–19. Springer (2024)
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
52. Zhao, Q., Tan, H., Wang, Q., Bi, S., Zhang, K., Sunkavalli, K., Tulsiani, S., Jiang, H.: E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. arXiv preprint arXiv:2512.10950 (2025)
53. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024)
54. Zhi, Y., Li, C., Liao, H., Yang, X., Sun, Z., Chang, J., Cun, X., Feng, W., Han, X.: Mv-performer: Taming video diffusion model for faithful and synchronized multi-view performer synthesis. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–14 (2025)
55. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12405–12414 (2025)
56. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
57. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8535–8546 (2025)