MV-Forcing: Long Multi-View Video Generation via 4D-Grounded Spatio-Temporal Self-Forcing

Gal Fiebelman[1], Hadar Averbuch-Elor[2], and Sagie Benaim[1]

[1] The Hebrew University of Jerusalem, Jerusalem, Israel
galf251@gmail.com
[2] Cornell University, New York, USA

Abstract. Recent advances in video diffusion models have enabled either long single-view generation through temporal autoregression, or short multi-view synthesis through bidirectional attention. However, generating long, multi-view consistent videos of dynamic scenes remains unsolved. In this work, we present MV-Forcing, a framework that composes temporal and view-wise autoregression within a single diffusion model by introducing a 4D geometric bridge between sequentially generated views. Our key insight is that an autoregressive 3D reconstruction model naturally interfaces between autoregressively generated views. Given a completed source view, we reconstruct its 3D structure and render a geometric prior of the next target viewpoint, which the diffusion model refines into a high-quality video. To extend generation beyond the teacher's fixed temporal window, we introduce a joint denoising regime where both view slots are initialized from noise during training, enabling temporally unbounded generation. We distill the model via Distribution Matching Distillation with Spatio-Temporal Self-Forcing, closing the train-inference exposure bias gap for both temporal and view-sequential autoregression. Extensive experiments on both synthetic and real-world data demonstrate that MV-Forcing produces geometrically consistent multi-view videos of dynamic scenes at arbitrary lengths and viewpoint counts using a single few-step student model.

Keywords: Video Diffusion Models · Multi-View Generation · Autoregressive Generation

1 Introduction

Generating a long, temporally coherent video from multiple viewpoints simultaneously is a long-standing challenge in computer vision. This task, long multi-view video generation, requires a model to synthesize the continuous temporal dynamics of a scene while maintaining strict 3D geometric consistency across arbitrary camera trajectories. Unlocking this capability is critical for a wide array of downstream applications, ranging from immersive virtual reality and advanced cinematic content creation to building dynamic, interactive simulations. However, jointly modeling the complex distribution of photorealistic video across

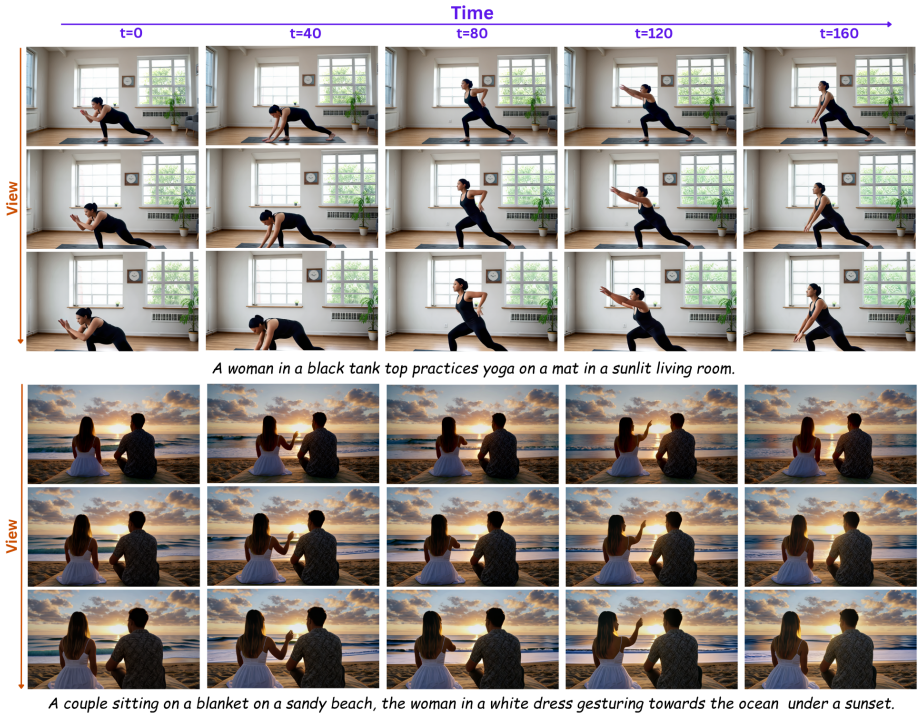

Fig. 1: Given a text prompt and camera sequences, MV-Forcing generates coherent video across an arbitrary number of viewpoints and unbounded temporal horizons. For each prompt, we show 3 views at increasing camera displacements (rows) across 160 timesteps (columns). Appearance, motion, and scene geometry remain consistent across all views and timesteps, demonstrating the effectiveness of our 4D-grounded geometric prior in maintaining cross-view consistency over long-horizon generation.

both unbounded temporal horizons and an arbitrary number of viewpoints demands an intricate balance of spatial, temporal, and geometric reasoning.

Recent years have witnessed remarkable progress along two orthogonal directions in video synthesis. On one hand, continuous advances in autoregressive video diffusion have pushed the boundaries of temporal duration, enabling single-view generation over minute-long horizons by sequentially conditioning on previously generated context [16,23,33]. On the other hand, multi-view generation has seen exciting progress, with models now capable of synchronizing dynamic open-world scenes across multiple camera viewpoints [2, 20]. However, despite these parallel successes, unifying these paradigms to achieve long, multi-view generation remains a formidable bottleneck. Current multi-view approaches heavily rely on bidirectional attention across the entire time-view grid to enforce 3D consistency. While effective for short clips, this dense all-to-all attention scales quadratically, completely preventing streaming inference and severely restricting outputs to short, fixed temporal windows. Conversely, directly adapting temporal autoregressive models to the multi-view setting introduces severe exposure bias; without an explicit geometric anchor, sequential view generation quickly drifts,

failing to maintain rigid spatial consistency. Consequently, generating videos that are simultaneously unbounded in time and geometrically consistent across arbitrary viewpoints remains an important, unsolved challenge.

In this work, we present MV-Forcing, a novel framework that uniquely composes temporal and view-wise autoregression within a single generative model, enabling the generation of long multi-view videos; see Figure 1. Our core insight is that an autoregressive 4D reconstruction model (like CUT3R [35]) naturally interfaces between autoregressively generated views, serving as a continuous geometric bridge that circumvents the reliance on dense bidirectional attention. Given a completed source view, we reconstruct its 3D structure and render a geometric prior of the next target viewpoint. The generative model integrates this structural prior via a 3D convolution layer, refining it into a video alongside temporal context provided by the previously generated frames' KV cache. By leveraging a recurrent reconstruction model (CUT3R), as each view is generated, it is incrementally integrated into a persistent latent state. This allows accumulating an increasingly complete 4D reconstruction of the scene, aligning reconstruction and generation within a unified autoregressive loop.

To extend generation beyond the teacher's fixed temporal window, we introduce a joint denoising regime where all view slots are initialized from noise during training, enabling temporally unbounded generation. Specifically, during training, we stochastically initialize both view slots from pure noise, which unifies first-view text-to-video generation and view-sequential conditioning within a single architecture. Furthermore, to close the train/inference gap for both temporal and view-sequential autoregression, we distill our model via Distribution Matching Distillation (DMD) paired with Self-Forcing. By unrolling autoregressive generation along both axes and applying the DMD loss to the self-generated outputs, our Spatio-Temporal Self-Forcing mechanism effectively mitigates exposure bias, ensuring robust geometric and temporal coherence over long horizons.

To the best of our knowledge, we provide the first framework enabling long multi-view video generation. As such, we compare our method to baselines tackling either single-view long video generation or short multi-view generation. Specifically, we evaluate on both synthetic and real-world data, comparing against the bidirectional SynCamMaster teacher for short sequences, and state-of-the-art composed baselines (Self-Forcing and ReCamMaster) for long sequences. Quantitatively, we show that our framework maintains robust cross-view synchronization and strict camera accuracy even when scaling generation to arbitrary viewpoint counts and temporal lengths, effectively eliminating the severe geometric drift seen in prior approaches.

Our main contributions are: **(1).** We introduce MV-Forcing, which is, to the best of our knowledge, the first generative framework to enable long multi-view video generation by uniquely composing temporal and view-wise autoregression. **(2).** We propose using an autoregressive 4D reconstruction model as a continuous geometric bridge between sequentially generated views, effectively bypassing the scaling limitations of dense bidirectional attention. **(3).** We introduce a joint denoising regime combined with Distribution Matching Distillation

(DMD) and Spatio-Temporal Self-Forcing, enabling temporally unbounded, geometrically consistent synthesis closing the train-inference exposure bias gap.

2 Related Work

Autoregressive Video Generation. While video diffusion models achieve remarkable quality [19, 25, 33, 40], most rely on bidirectional attention, preventing streaming and scaling. Autoregressive formulations alternatively generate frames sequentially. Recent works combine autoregressive temporal structure with continuous diffusion to generate frame chunks. Teacher Forcing [9, 15, 18, 49] conditions on clean context, whereas Diffusion Forcing [4, 5, 11, 28, 30] uses independently sampled noise levels per frame. CausVid [44] introduces asymmetric distillation, training a causal student against a bidirectional teacher via DMD [42, 43]. Self-Forcing [16] identifies that both paradigms still suffer from exposure bias and addresses this by unrolling autoregressive generation during training. Several concurrent works build on this paradigm [6, 23, 39] to push long-horizon quality further. All of these methods operate in the single-view regime. Our work extends Self-Forcing to the spatio-temporal domain, composing temporal autoregression with view-sequential generation grounded by a 4D geometric prior.

Multi-View Video Generation. Object-centric methods such as SV4D [38], SV4D 2.0 [41], and CAT4D [37] generate dense orbital views of dynamic objects but are restricted to single-object scenes and fixed camera configurations. For open-world scenes, SynCamMaster [2] introduces a plug-and-play multi-view synchronization module for a pretrained text-to-video DiT, enabling multi-camera generation with 6-DoF camera control. CVD [20] adds epipolar attention across views to enforce 3D consistency. ReCamMaster [1] and related approaches [17,32] condition on a reference video to re-render it from novel trajectories, but are video-to-video methods and do not generate multi-view content from text. All these methods rely on bidirectional attention over both time and views, restricting generation to short, fixed-length clips and a fixed number of views. Our method extends Self-Forcing to the multi-view setting, enabling generation over unbounded temporal horizons and an arbitrary number of viewpoints.

Reconstruction for Generation. A growing body of work leverages 3D reconstruction as a structural prior for generation by reconstructing geometry from available observations, rendering it from target viewpoints, and conditioning a generative model on these renders to fill in appearance details [3,7,8,12,22,27,45, 46]. Underlying many of these methods are feed-forward reconstruction models such as DUSt3R [36], MASt3R [21], VGGT [34], and CUT3R [35], which predict dense pointmaps and camera poses without per-scene optimization. Among these, CUT3R is uniquely suited to sequential generation due to its recurrent architecture, it maintains a persistent latent state updated incrementally with each new image, and supports querying from arbitrary virtual cameras. Our work leverages CUT3R as a geometric bridge between sequentially generated views, aligning reconstruction and generation within a unified autoregressive loop.

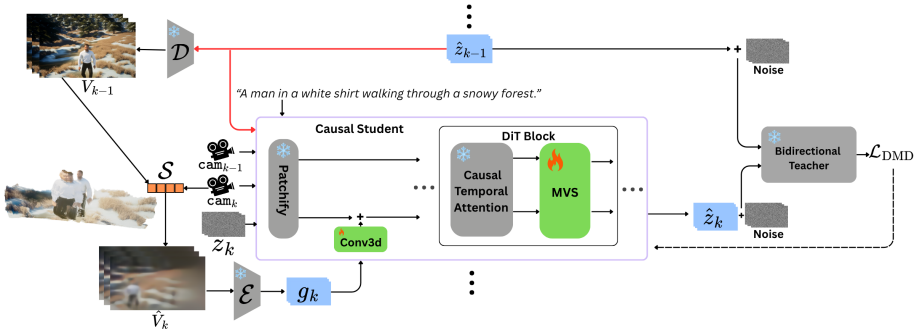

Fig. 2: Overview of MV-Forcing. Given a text prompt, camera sequences $\texttt{cam}_{k-1}$ and $\texttt{cam}_k$, and a previously generated view $\hat{z}_{k-1}$, the causal student generates the next view $\hat{z}_k$. The decoded previous view $V_{k-1} = \mathcal{D}(\hat{z}_{k-1})$ is integrated into CUT3R's persistent state $\mathcal{S}$, which is queried using $\texttt{cam}_k$ to produce a geometric rendering $\hat{V}_k$. This rendering is encoded and projected via a `Conv3d` layer into a conditioning tensor g_k that is added to the patchified noisy latent z_k. Both $\hat{z}_{k-1}$ (clean) and z_k (noisy) are forwarded through DiT blocks with causal temporal attention and MVS cross-view attention. The student's output $\hat{z}_k$ and the preceding $\hat{z}_{k-1}$ are re-noised and scored by the bidirectional SynCamMaster teacher via $\mathcal{L}_{\text{DMD}}$. The red arrow illustrates the view-sequential unrolling, $\hat{z}_k$ is decoded and integrated into CUT3R's state, and fed back as the conditioning view for the next generation step, mirroring the inference-time autoregressive loop. The 3D reconstruction visualized alongside $\mathcal{S}$ illustrates the scene geometry accumulated in the state from all previously generated views.

3 Method

Given an input text prompt and a set of N camera sequences, each defining a desired viewpoint, our goal is to generate a long, multi-view consistent video that scales to an arbitrary number of viewpoints and extends over unbounded temporal horizons. As illustrated in Fig. 2, we achieve this by composing temporal and view-sequential autoregression within a unified self-forcing framework, using a feed-forward dynamic 3D reconstruction model as a geometric bridge between the two axes. We begin by reviewing the multi-view video generation and self-forcing paradigms that form the basis of our approach (Sec. 3.1). We then introduce spatio-temporal self-forcing, which extends the self-forcing paradigm from the temporal dimension to the view dimension (Sec. 3.2). Next, we describe the 4D-grounded geometric prior that bridges temporal and view-sequential generation with an explicit geometric signal (Sec. 3.3). Finally, we describe how these components compose at inference to produce long multi-view videos (Sec. 3.4) and how the framework generalizes to real-world data (Sec. 3.5).

3.1 Preliminaries

Multi-View Video Generation. SynCamMaster [2] extends a pretrained text-to-video diffusion transformer (DiT) [33] to multi-view generation by introducing a plug-and-play multi-view synchronization (MVS) module. Given a pair of

videos V_1, V_2 from two viewpoints, a 3D VAE encoder $\mathcal{E}$ compresses each into a latent representation $z_j = \mathcal{E}(V_j)$, which is then converted into a sequence of tokens via patchification:

$$x_1 = \texttt{patchify}(z_1), \quad x_2 = \texttt{patchify}(z_2), \tag{1}$$

where $x_1, x_2 \in \mathbb{R}^{T \times S \times D}$, with T the number of latent frames, S the spatial tokens per frame, and D the token dimension. Each transformer block applies bidirectional temporal self-attention across all frames within a single view. An MVS module is additionally inserted into each transformer block. Given the intermediate features $F_1^i, F_2^i \in \mathbb{R}^{S \times D}$ at frame i within a transformer block, the module first adds camera embeddings produced by a camera encoder $\mathcal{E}_c$ that maps the 6-DoF extrinsic parameters into the token space, and then applies cross-view self-attention, where tokens from both views at the same frame attend to each other:

$$\overline{F}_1^i, \overline{F}_2^i = \texttt{MVS}(F_1^i + \mathcal{E}_c(\texttt{cam}_1),\ F_2^i + \mathcal{E}_c(\texttt{cam}_2)). \tag{2}$$

The output is projected back to the feature domain by a zero-initialized linear layer and added as a residual to the original features. Only the MVS parameters are trained, while the base DiT weights remain frozen. In SynCamMaster, both temporal and cross-view attention are bidirectional, and generation follows a many-step denoising schedule that requires all frames and views to be denoised jointly within a fixed-length window.

Self-Forcing. A common strategy to accelerate video diffusion is to distill the model into a few-step generator via Distribution Matching Distillation (DMD) [42, 43]. CausVid [44] applies asymmetric DMD to distill a bidirectional video DiT into a causal few-step student. Self-Forcing [16] identifies that the student still suffers from exposure bias, as it trains on ground-truth context but must condition on its own imperfect outputs at inference. It addresses this by unrolling autoregressive generation during training and applying the DMD loss over the self-generated output, closing the train-inference gap and enabling streaming generation of arbitrarily long single-view videos.

3.2 Spatio-Temporal Self-Forcing

While Self-Forcing [16] enables temporal autoregression for long single-view videos, extending this scalability across arbitrary viewpoints requires view-wise autoregression. We propose *spatio-temporal self-forcing*, which extends self-forcing to the view dimension by introducing view-sequential autoregressive unrolling during training. Following Self-Forcing [16], we distill SynCamMaster into a causal few-step student G_ϕ via asymmetric DMD [42]. The student inherits the teacher's MVS module (Eq. 2), but replaces the bidirectional temporal attention with causal attention under a blockwise mask:

$$M_{i,j} = \begin{cases} 1, & \text{if } \lfloor \frac{j}{K} \rfloor \leq \lfloor \frac{i}{K} \rfloor, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where i and j index latent frames in the sequence and K is the temporal block size in latent frames. Under this mask, tokens at frame i attend only to frames within the same or earlier temporal blocks.

The student is trained via asymmetric DMD, where the frozen bidirectional SynCamMaster teacher serves as the data score function s_{data} and a trainable copy s_{gen} tracks the student's output distribution. At each training step, the student generates an output $\hat{z}_0 = G_\phi(\{z^i_{\tau^i}\}, \{\tau^i\})$ via its few-step schedule, where $z^i_{\tau^i}$ is the i-th temporal block corrupted at independently sampled noise level τ^i. Noise is re-injected at a shared level τ to obtain $\hat{z}_\tau$, and the student is updated via:

$$\nabla_\phi \mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_\tau \left[\left(s_{\text{data}}(\hat{z}_\tau, \tau) - s_{\text{gen}}(\hat{z}_\tau, \tau) \right) \frac{\partial \hat{z}_\tau}{\partial \phi} \right]. \tag{4}$$

Following CausVid [44], we first initialize the student by training on a small set of ODE solution pairs generated by the bidirectional teacher before applying the DMD objective. The student's temporal layers are initialized from a pretrained Self-Forcing model and kept frozen, while the MVS layers are initialized from SynCamMaster and finetuned during training. Additional training details are provided in the supplementary material.

View-Sequential Generation. Given N target viewpoints, we generate views $0, 1, \ldots, N{-}1$ sequentially. To produce view k, we denoise its latent starting from pure noise $z_k^{\tau_Q} \sim \mathcal{N}(0, I)$ through a Q-step schedule $\tau_Q > \cdots > \tau_1 > \tau_0 = 0$, conditioned on the fully denoised preceding view and the text prompt P. At each denoising step $q = Q, \ldots, 1$, the student takes the noisy latent $z_k^{\tau_q}$ along with the clean preceding latent z_{k-1}, text prompt P, and camera parameters $\texttt{cam}_k, \texttt{cam}_{k-1}$, predicts a clean estimate, and re-noises it to the next level:

$$z_k^{\tau_{q-1}} = \Psi\big(G_\phi(z_k^{\tau_q},\, \tau_q,\, z_{k-1},\, P,\, \texttt{cam}_k, \texttt{cam}_{k-1}),\, \tau_{q-1}\big)\,, \tag{5}$$

where G_ϕ is the student network and $\Psi(\cdot, \tau_{q-1})$ re-injects noise at level τ_{q-1}. The final output $z_k = z_k^{\tau_0}$ then serves as clean conditioning for view $k{+}1$ via the cross-view attention in Eq. 2, naturally extending to arbitrary-length view chains. Since each view only attends to the single preceding view, the process is fully autoregressive across viewpoints. Note that this clean-noisy asymmetry across views is absent from the SynCamMaster teacher, which jointly denoises both views at the same noise level. The student learns this clean cross-view conditioning entirely through distillation.

View-Sequential Unrolling. This view-sequential generation introduces a view-level analogue of exposure bias. During DMD training, the cross-view attention conditions on ground-truth latents, but at inference, view k is conditioned on the model's previously generated z_{k-1}. We close this gap by unrolling autoregressive generation along the view dimension during training. Starting from the ground-truth first view z_0^{gt}, the student sequentially generates views $\hat{z}_1, \hat{z}_2, \ldots$, conditioned on its previously generated output, via its few-step schedule (Eq. 5):

$$\hat{z}_{k-1} = G_\phi(\epsilon_{k-1},\, \{\tau_q\},\, \hat{z}_{k-2},\, P,\, \texttt{cam}_{k-1}, \texttt{cam}_{k-2}), \tag{6}$$

$$\hat{z}_k = G_\phi(\epsilon_k,\, \{\tau_q\},\, \hat{z}_{k-1},\, P,\, \texttt{cam}_k, \texttt{cam}_{k-1}), \tag{7}$$

where $\hat{z}_0 = z_0^{\text{gt}}$, $\epsilon_k \sim \mathcal{N}(0, I)$, $\{\tau_q\}$ denotes the few-step denoising schedule, and we abbreviate the full denoising chain (Eq. 5) for brevity. Each generated view beyond the first is conditioned on the student's own imperfect output rather than ground-truth, mirroring the inference-time regime and forcing the student to learn robustness to its own errors. The bidirectional SynCamMaster teacher then scores each consecutive generated pair $(\hat{z}_{k-1}, \hat{z}_k)$ via the DMD loss (Eq. 4). Since the teacher evaluates view pairs independently, the distillation objective naturally decomposes over both the temporal and view axes, jointly mitigating the exposure-bias gap along both dimensions. In practice, due to GPU memory constraints, we unroll from z_0^{gt} to generate $\hat{z}_1$ and $\hat{z}_2$.

Joint View Denoising. As our goal is text to multi-view video generation, the model must also generate the first view from text prompt P alone, without any cross-view conditioning. To this end, during training we set both views to start from noise with probability p, training the model to generate each view independently from text. When only view k starts from noise (probability $1-p$), the model trains the view-sequential conditioning path described above. This training regime unifies first-view generation and view-sequential generation within a single architecture. At inference, the model generates z_0 from text via the joint denoising path with both view latents initialized from noise, and then generates each subsequent view autoregressively conditioned on the preceding one.

3.3 4D-Grounded Geometric Prior

As views are generated sequentially over long temporal horizons and across multiple viewpoints, geometric consistency degrades as each view is conditioned only on the single preceding view through the cross-view attention in Eq. 2, and errors in 3D structure accumulate along the view chain. To mitigate this, we introduce an explicit geometric conditioning signal that bridges temporal and view-sequential generation, grounding each new viewpoint in a persistent 3D reconstruction of the scene built from all previously generated views and timesteps.

Recurrent 3D Reconstruction. Our geometric prior requires a 3D reconstruction model that satisfies two properties: it must operate autoregressively since views and temporal blocks are generated sequentially, and it must maintain a persistent state that can be queried from arbitrary virtual cameras, enabling rendering of geometric priors for viewpoints and timesteps not yet generated. CUT3R [35] is a feed-forward dynamic 3D reconstruction model that meets both requirements. It operates recurrently over a stream of images, maintaining a persistent latent state $\mathcal{S}$ that encodes the accumulated 4D structure of the scene. Given a new image I, the state is updated:

$$\mathcal{S}' = \texttt{CUT3R\_update}(\mathcal{S}, I), \tag{8}$$

where $\mathcal{S}$ is the state prior to observing I and $\mathcal{S}'$ is the updated state. This state can also be queried from arbitrary virtual cameras without updating it. Given

a target camera parameterized as a raymap $\mathcal{R}$ [10, 37, 47], CUT3R decodes a rendered image and a per-pixel confidence map:

$$\hat{I},\ \hat{C} = \texttt{CUT3R\_query}(\mathcal{S}, \mathcal{R}), \tag{9}$$

where $\hat{I}$ is the RGB rendering of the scene from the queried viewpoint, $\hat{C}$ is a confidence map indicating reconstruction certainty at each pixel, and $\mathcal{R}$ is a 6-channel image encoding ray origins and directions at each pixel.

Geometric Conditioning. To condition the generation of view k, we first decode all preceding views to pixel space, $V_j = \mathcal{D}(z_j)$ for $j = 0, \ldots, k-1$, where $\mathcal{D}$ is the 3D VAE decoder, and integrate them into CUT3R's state via sequential updates. We then query the accumulated state from view k's camera sequence, for each frame at time t, we construct a raymap $\mathcal{R}_k^t$ from the camera extrinsics and intrinsics and obtain a rendering $\hat{I}_k^t$ and confidence map $\hat{C}_k^t$ via Eq. 9. The sequence of rendered frames $\hat{V}_k = \{\hat{I}_k^1, \ldots, \hat{I}_k^T\}$ is encoded into video latents $\mathcal{E}(\hat{V}_k)$ and concatenated channel-wise with the confidence maps to form a conditioning tensor:

$$g_k = \text{Concat}\Big(\mathcal{E}(\hat{V}_k),\ \hat{C}_k\Big) \in \mathbb{R}^{T \times D' \times H \times W}, \tag{10}$$

where $D' = C_{\text{latent}} + 1$ combines the VAE latent channels with the single-channel confidence map. A 3D convolution layer projects the conditioning tensor into the token space and adds it to the patch embeddings via a residual connection:

$$\tilde{x}_k = x_k + \texttt{Conv3d}(g_k), \tag{11}$$

where $x_k = \texttt{patchify}(z_k)$ are the patch embeddings of view k. Following ControlNet [48], the Conv3d layer is zero-initialized, ensuring the geometric prior has no effect when training begins and allowing the model to gradually learn to incorporate it. Together with the clean cross-view conditioning described in Sec. 3.2, the geometric prior constitutes a second conditioning pathway that the student acquires entirely through distillation, as neither is present in the original teacher model.

State Accumulation. As each temporal block of a view is generated, its latent z_k is decoded to pixel space $V_k = \mathcal{D}(z_k)$ and fed into CUT3R, updating $\mathcal{S}$ incrementally along both time and views. When generating block (k, t), the state already incorporates all decoded frames from earlier views $0, \ldots, k-1$ and preceding temporal blocks $(k, < t)$ of the current view. Additional details of the accumulated geometric prior are provided in the supplementary material.

3.4 Long Multi-View Video Generation

The autoregressive structure along both axes, and the accumulated geometric prior, yields a key property at inference, generation can advance along *either* axis. At any block (k, t), the model can proceed temporally to $(k, t+1)$ or across views to $(k+1, t)$, since both transitions require only the KV cache (for

temporal context) and a clean preceding view with its geometric prior (for cross-view context). This allows arbitrary traversal orders over the time×view grid, rather than requiring one axis to be fully completed before advancing the other.

Concretely, the first block $(0, 0)$ is generated from the text prompt alone using joint view denoising (Sec. 3.2). For each subsequent block (k, t), the generated latent is decoded to pixel space $V_k = \mathcal{D}(z_k)$ and integrated into CUT3R's state $\mathcal{S}$. When advancing temporally within a view, the KV cache provides temporal context for the next block. When advancing to a new view k, the accumulated state is queried from view k's cameras to produce the geometric conditioning tensor g_k (Eq. 10), which along with the clean preceding latent z_{k-1} provides cross-view conditioning through the MVS module (Eq. 2). In both cases, the decoded frames are fed back into CUT3R, progressively enriching the state along both axes. By the time generation reaches view k+1, the accumulated state captures the full 4D structure observed so far, providing increasingly rich geometric context for each successive viewpoint.

3.5 Generalization to Real-World Data

As the DMD objective (Eq. 4) distills the teacher's distribution through the data score function s_{data}, the student inherits the domain of its teacher, which in our case is the synthetic SynCamMaster. Extending the framework to real-world content requires replacing s_{data} with a teacher whose score function captures a real-world distribution. As no such multi-view teacher exists, we adopt ReCamMaster [1], a video-to-video model that re-renders a source video from a novel camera trajectory, and finetune the student for N_{ft} iterations on real-world videos to bridge the distribution shift. As ReCamMaster conditions on an existing video rather than generating from text, it cannot supervise the joint denoising path (Sec. 3.2), and we set p=0 during this stage. At inference, we compose Self-Forcing [16] for first-view generation with our finetuned student for subsequent views. Additional details are provided in the supplementary material.

4 Experiments

We evaluate MV-Forcing on the task of long multi-view video generation. We present comparisons against baselines in Sec. 4.1, ablate our design choices and analyze scaling behavior along both the view and temporal axes in Sec. 4.2. Finally, we discuss limitations in Sec. 4.3.

Evaluation Setup. We train and evaluate our model in two settings. For the synthetic setting, we use the SynCamVideo Dataset [2], which contains 3,400 synthetic multi-view video scenes, each captured from 10 synchronized cameras, and hold out 100 scenes for evaluation. For the real-world setting, we follow the extension described in Sec. 3.5. We finetune our model on real-world videos from the Mixkit subset of the Open-Sora Dataset [24], which contains diverse open-domain video content, and hold out 100 videos for evaluation. All metrics are

Table 1: Short Sequence Evaluation. All methods generate 2 views at 81 frames. We compare MV-Forcing against the bidirectional SynCamMaster [2] teacher.

	Visual Quality				Camera Accuracy		View Synchronization		
Method	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
SynCamMaster	**166.57**	**1451.44**	**30.02**	**99.32**	3.83	8.83	236.72	1697.37	90.13
MV-Forcing (Ours)	167.90	1468.84	29.67	99.21	**3.64**	**8.26**	**251.13**	**1691.05**	**91.81**

averaged across the evaluation set. Additional evaluation details are provided in the supplementary material.

Evaluation Metrics. Following SynCamMaster [2], we evaluate our method along three axes, visual quality, camera accuracy, and cross-view synchronization. For visual quality, we report Fréchet Inception Distance (FID) [14], Fréchet Video Distance (FVD) [31], CLIP-T, which measures average CLIP [26] similarity between each frame and its text prompt and CLIP-F, which measures average CLIP similarity between consecutive frames. For camera accuracy, we use GIM [29] to estimate per-frame camera poses from dense correspondences and report the average rotation and translation error against the ground truth, denoted as RotErr and TransErr, respectively [13]. For cross-view synchronization, we report Mat. Pix, which measures the number of matching pixels with confidence above a threshold from the computed dense correspondences between view pairs from GIM, CLIP-V [20], which measures average CLIP similarity between views at the same timestep, and FVD-V [38], which measures FVD of stacked frames across views at each timestep.

4.1 Comparison to Baselines

Baselines. No existing method addresses long multi-view video generation. We therefore construct baselines by composing single-view, multi-view, and video-to-video methods. SynCamMaster [2] is the bidirectional many-step teacher, generating 2 views in a fixed 81-frame window, serving as a strong short sequence baseline, but cannot scale beyond this setting. SF+ReCamMaster uses Self-Forcing [16] to generate an arbitrarily long first view, then applies ReCamMaster [1] to re-render additional views in independent 81-frame segments. SF+ReCamMaster+SF uses Self-Forcing to generate an 81-frame first view, then applies ReCamMaster and uses Self-Forcing to temporally extend each re-rendered view beyond its first segment. For the synthetic evaluation, we additionally finetune Self-Forcing on SynCamVideo, replacing its real-world teacher with the synthetic SynCamMaster, denoted SF(ft), to ensure a fair in-domain comparison. Additional baseline details are provided in the supplementary material.

Comparisons. We evaluate under two settings. Tab. 1 compares against SynCamMaster at 2 views and 81 frames. Tab. 2 extends the evaluation to 3 views and 162 frames, beyond SynCamMaster's capacity. We compare against SF+ReCamMaster and SF+ReCamMaster+SF in the real-world setting, and against their SF(ft) variants in the synthetic setting.

Table 2: Long Sequence Evaluation. All methods generate 3 views at 162 frames. SF+ReCamMaster generates each view independently from a Self-Forcing [16] generated first view then re-renders with ReCamMaster [1]. SF+ReCamMaster+SF additionally extends each view temporally using Self-Forcing after the initial ReCamMaster re-rendering. We evaluate in two settings, real-world (Real), where all methods operate on real-world data, and synthetic (Synth.), where we replace Self-Forcing with SF(ft) finetuned on SynCamVideo for a fair in-domain comparison.

		Visual Quality				Camera Accuracy		View Synchronization		
	Method	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
Real	SF+ReCamMaster	157.57	1397.42	32.27	98.22	4.74	10.12	146.81	1759.82	87.26
	SF+ReCamMaster+SF	156.78	1363.23	32.48	98.67	4.89	10.61	127.48	1771.34	86.61
	MV-Forcing (Ours)	**153.27**	**1309.68**	**32.74**	**99.03**	**3.88**	**8.78**	**239.37**	**1554.23**	**90.83**
Synth.	SF(ft)+ReCamMaster	192.34	1576.83	29.07	98.03	4.32	10.27	143.77	1956.38	87.19
	SF(ft)+ReCamMaster+SF	191.26	1592.78	29.26	98.52	4.67	10.73	121.29	1987.43	86.27
	MV-Forcing (Ours)	**186.73**	**1560.54**	**29.54**	**99.17**	**3.72**	**8.41**	**243.63**	**1729.35**	**89.88**

In the short sequence setting, MV-Forcing outperforms the bidirectional teacher on all camera accuracy and view synchronization metrics, demonstrating that explicit geometric grounding through the accumulated CUT3R prior can outperform joint bidirectional denoising even on short sequences. Visual quality metrics remain slightly lower, reflecting the expected cost of distilling a many-step bidirectional model into a few-step causal student. In the long sequence setting (Tab. 2), MV-Forcing outperforms all baselines across all metrics in both evaluation settings. SF+ReCamMaster generates each view as a sequence of independent 81-frame segments with no shared context between them, leading to visible degradation beyond the first segment. SF+ReCamMaster+SF recovers temporal continuity by extending each view autoregressively, but lacks any cross-view conditioning mechanism, and consistency between viewpoints degrades as views are generated independently. The consistent results across both real-world and synthetic settings indicate that the performance gap is architectural rather than driven by domain alignment. These results demonstrate that view-sequential conditioning with a persistent geometric prior provides the 3D grounding necessary for cross-view consistency, while temporal autoregression ensures coherent generation beyond any fixed temporal window.

A qualitative comparison is shown in Fig. 3. SF+ReCamMaster produces plausible results within its first temporal window but degrades at later timesteps due to independent per-segment generation. SF+ReCamMaster+SF maintains temporal continuity but exhibits cross-view drift, with scene structure and appearance varying between viewpoints. In contrast, our method preserves geometric consistency across all views and timesteps. Qualitative comparisons for the short setting and synthetic long setting are provided in the supplementary material.

4.2 Ablation Study

We ablate the key components of our framework: (1) *w/o Accumulation* conditions each view only on the single preceding view's geometry; (2) *Manual Rendering* replaces CUT3R's learned raymap query with classical forward-warping

Fig. 3: Qualitative Comparison on Long Sequences. We generate multiple views across 162 frames for the prompt *"A young man wearing white clothes practicing karate kicks on an urban street."* For each method, we show 3 views (rows) at 5 timesteps (columns). SF+ReCamMaster degrades when operating beyond its training window, producing visible artifacts in the third view and beyond 81 frames. SF+ReCamMaster+SF recovers temporal continuity but lacks cross-view consistency, with scene structure and appearance varying significantly between viewpoints. Our method maintains consistency across both views and time. As highlighted by the red boxes (shown on the last timestep for visibility), the baselines produce inconsistent leg positions across views at the same timestep, while our method preserves a coherent pose across all viewpoints.

Table 3: Quantitative Ablation Results. All variants are evaluated at 3 views and 162 frames. We ablate view-sequential unrolling (w/o View Unrolling), the geometric prior (w/o CUT3R), state accumulation (w/o Accumulation), and CUT3R's latent state query (Manual Rendering).

	Visual Quality				Camera Accuracy		View Synchronization		
Method	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
w/o View Unrolling	212.89	1791.40	27.74	97.66	4.72	10.45	159.73	2318.94	86.41
w/o CUT3R	199.10	1613.21	28.87	98.21	4.56	9.83	173.75	2109.82	87.12
w/o Accumulation	193.66	1576.84	29.26	98.78	3.97	9.04	229.63	1984.35	88.27
Manual Rendering	191.14	1571.44	29.11	99.10	3.93	8.93	232.71	1963.42	88.36
Ours	**186.73**	**1560.54**	**29.54**	**99.17**	**3.72**	**8.41**	**243.63**	**1729.35**	**89.88**

Table 4: View scaling. All variants are evaluated at a fixed length of 81 frames. We increase the number of generated views from 2 to 5.

	Visual Quality				Camera Accuracy		View Synchronization		
Views	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
2	**167.9**	**1468.84**	**29.67**	**99.21**	**3.64**	**8.26**	**251.13**	**1691.05**	**91.81**
3	168.33	1468.92	29.58	99.13	3.65	8.29	251.08	1691.86	91.77
4	168.57	1469.03	29.55	99.06	3.65	8.31	251.02	1691.94	91.75
5	169.13	1469.12	29.43	98.92	3.67	8.35	250.95	1692.07	91.73

via z-buffer splatting; (3) *w/o CUT3R* removes the geometric prior entirely; and (4) *w/o View Unrolling* trains with standard DMD along the view axis without spatio-temporal self-forcing unrolling.

Tab. 3 reports results at 3 views and 162 frames. Removing view-sequential unrolling yields the largest degradation, as the train-inference distribution mismatch propagates along the view chain. Removing CUT3R produces the second-largest drop, validating the geometric prior as critical for cross-view consistency. Removing accumulation and replacing CUT3R's query with classical warping each degrade metrics by a smaller but consistent margin. Qualitative comparisons of the ablation variants are provided in the supplementary material.

Scaling Analysis. We additionally evaluate how generation quality evolves as the number of views and temporal length are independently extended.

Tab. 4 reports metrics as the number of views increases from 2 to 5 at a fixed temporal length of 81 frames. All metrics show only minor degradation, remaining stable up to 5 views. We attribute this robustness to the accumulated geometric prior, grounding each new view in the full 4D structure. Tab. 5 extends temporal length from 81 to 648 frames at 2 views. Cross-view metrics remain nearly constant across an 8× increase in duration, while CLIP-F shows the most notable decline, reflecting the compounding of small temporal errors inherent to autoregressive temporal generation. Qualitative comparisons for both scaling axes are provided in the supplementary material.

4.3 Limitations

While MV-Forcing enables long multi-view video generation, several limitations remain. First, our model is primarily trained on the synthetic SynCamVideo

Table 5: Temporal scaling. All variants are evaluated at a fixed 2 views. We increase the temporal length from 81 to 648 frames.

	Visual Quality				Camera Accuracy		View Synchronization		
Length	FID ↓	FVD ↓	CLIP-T ↑	CLIP-F ↑	RotErr ↓	TransErr ↓	Mat. Pix.(K) ↑	FVD-V ↓	CLIP-V ↑
81	**167.90**	**1468.84**	**29.67**	**99.21**	**3.64**	**8.26**	**251.13**	**1691.05**	**91.81**
162	169.34	1529.52	29.59	99.19	3.66	**8.26**	249.82	1693.26	91.64
324	170.36	1675.26	29.54	97.82	3.67	8.27	249.72	1694.84	91.57
648	172.91	1859.82	29.52	96.23	3.67	8.29	249.14	1694.98	91.28

dataset. Although we demonstrate generalization to real-world data with minimal finetuning (Sec. 3.5), training on large-scale real multi-view video data could further improve diversity and realism. Second, while Self-Forcing significantly reduces the train-inference exposure bias for temporal autoregression, quality still degrades over long horizons. Recent works [6, 23, 39] propose complementary strategies to mitigate this. These works are complementary to our framework and could be integrated to improve temporal stability. Third, as the SynCamMaster teacher generates only two views at a time, the DMD supervision never directly supervises consistency across more than two simultaneous views. Extending the training window to longer view tuples could further improve quality at high viewpoint counts. Despite this, Tab. 4 shows stable performance up to 5 views at inference, as the model chains pairwise transitions autoregressively. We provide a further failure case analysis in the supplementary material, examining scenarios such as extreme camera displacements, extreme motion, and generation backbone quality.

5 Conclusion

We presented MV-Forcing, a framework for long multi-view video generation that composes temporal and view-sequential autoregression within a single generative model. Our key insight is that an online dynamic 3D reconstruction model can serve as a geometric bridge between sequentially generated views, accumulating a persistent state that grounds each new viewpoint in the geometry of previously generated content. By extending Self-Forcing to the view axis, our approach mitigates the exposure bias gap along both axes, enabling generation over unbounded temporal horizons and arbitrary viewpoint counts with a single few-step model. We believe this takes a step toward spatially consistent, multi-view video generation that captures the full 4D structure of dynamic scenes.

Ethics Statement. Our method builds on publicly available open-source models and datasets. Our contribution is long multi-view video generation, a capability that does not introduce new generative content beyond what existing open-source text-to-video models already produce.

Acknowledgements. This research was supported by The Israel Science Foundation (grant No. 2416/25) and EuroHPC JU (Application ID EHPC-DEV-2025D08-098).

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. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. arXiv preprint arXiv:2412.07760 (2024)
3. Bian, W., Huang, Z., Shi, X., Li, Y., Wang, F.Y., Li, H.: Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. arXiv preprint arXiv:2501.02690 (2025)
4. Chen, B., Monso, D.M., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. arXiv preprint arXiv:2407.01392 (2024)
5. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
6. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
7. Fiebelman, G., Averbuch-Elor, H., Benaim, S.: Let it snow! animating 3d gaussian scenes with dynamic weather effects via physics-guided score distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 37322–37331 (2026)
8. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. Advances in Neural Information Processing Systems **36**, 39897–39914 (2023)
9. Gao, K., Shi, J., Zhang, H., Wang, C., Xiao, J., Chen, L.: Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. arXiv preprint arXiv:2411.16375 (2024)
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. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025)
12. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
13. 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)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
15. Hu, J., Hu, S., Song, Y., Huang, Y., Wang, M., Zhou, H., Liu, Z., Ma, W.Y., Sun, M.: Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. arXiv preprint arXiv:2412.07720 (2024)

16. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
17. Jeong, H., Lee, S., Ye, J.C.: Reangle-a-video: 4d video generation as video-to-video translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11164–11175 (2025)
18. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: ICLR (2025)
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. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. Advances in Neural Information Processing Systems **37**, 16240–16271 (2024)
21. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
22. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25690–25699 (2025)
23. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
24. PKU-YuanGroup: Open-sora-dataset. `https://github.com/PKU-YuanGroup/Open-Sora-Dataset` (2024), accessed: 2026-03-06
25. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. 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)
28. Sand-AI: Magi-1: Autoregressive video generation at scale (2025), `https://static.magi.world/static/files/MAGI_1.pdf`, accessed: 2026-02-24
29. Shen, X., Cai, Z., Yin, W., Müller, M., Li, Z., Wang, K., Chen, X., Wang, C.: Gim: Learning generalizable image matcher from internet videos. arXiv preprint arXiv:2402.11095 (2024)
30. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
31. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. arXiv preprint arXiv:1812.01717 (2019)
32. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
33. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

34. 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)
35. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
36. 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)
37. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26057–26068 (2025)
38. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024)
39. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
40. 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)
41. Yao, C.H., Xie, Y., Voleti, V., Jiang, H., Jampani, V.: Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13248–13258 (2025)
42. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. NeurIPS (2024)
43. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024)
44. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
45. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 100–111 (2025)
46. 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)
47. 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)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
49. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)