MVTRACK4GEN: MULTI-VIEW POINT TRACKING FOR PERSISTENT 4D VIDEO GENERATION

Anonymous authors
Paper under double-blind review

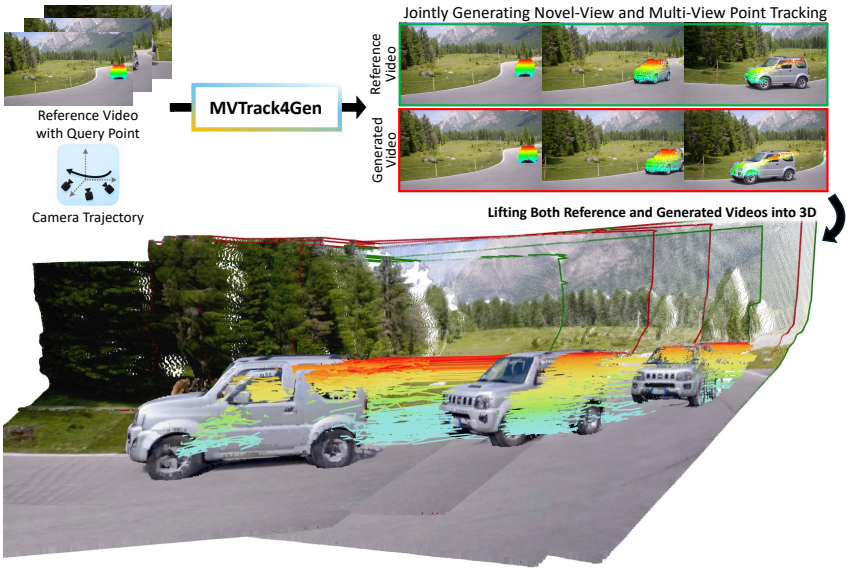


Figure 1: **MVTrack4Gen** jointly generates a novel-view video and multi-view point tracks, given a monocular reference video with query points and a user-specified camera trajectory. Lifting both the **reference** and **generated** frames into 3D space using Depth Anything 3 (Lin et al., 2025) shows that the target views and multi-view point tracks faithfully preserve the dynamic motion of the reference video while remaining geometrically consistent.

ABSTRACT

Synthesizing a novel-view video from a monocular reference video along a target camera trajectory requires geometric consistency and motion fidelity, both with respect to the reference video. Existing methods based on explicit 3D representations are limited by the accuracy of their reconstruction modules, which often produce inaccurate geometry for dynamic objects in monocular videos. In contrast, camera-conditioning-only methods can achieve high visual quality but often struggle to preserve geometric and motion consistency. In this work, we introduce **MVTrack4Gen** (**M**ulti-**V**iew Point **Track**ing **for** 4D Video **Gen**eration), a novel framework that leverages multi-view point tracking as an additional geometric and motion supervision signal for camera-conditioning-only 4D video diffusion models. Our key finding is that specific attention layers in 4D video diffusion models encode strong correspondence cues, where query features attend to key features at geometrically corresponding locations across views and over time, and that misalignment of these correspondences causes motion inconsistency. Based on this observation, we route these features into an auxiliary multi-view tracking head and jointly train the 4D video diffusion model with a point-tracking objective. By explicitly strengthening these motion-aware correspondences, **MVTrack4Gen** improves existing models' ability to follow the motion in the reference view and maintain inter-view geometric consistency. Across diverse benchmarks, our method achieves state-of-the-art geometric and motion consistency and competitive camera accuracy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Novel-view video generation from a monocular reference video aims to synthesize a target-view video along a user-specified camera trajectory, with applications in virtual cinematography, robotics, and immersive AR/VR. To be useful in these real-world scenarios, the generated video should satisfy three requirements: (1) geometric consistency preserving reference scene structure across views, (2) motion consistency preserving reference dynamics, (3) accurate camera control that faithfully follows the user-specified trajectory. The central challenge is achieving all three at once.

Recent works leverage generative priors from pretrained video diffusion models (Agarwal et al., 2025; Bar-Tal et al., 2024; Kong et al., 2024; Wan et al., 2025; Yang et al., 2025) to address this task, achieving notable progress in dynamic scenes (Ren et al., 2025; Lee et al., 2025; Jeong et al., 2025a; Yu et al., 2025; Chen et al., 2025b; Yang et al., 2026; Bai et al., 2025a; Park et al., 2025; Huang et al., 2025b). These approaches largely fall into two paradigms. The first paradigm follows a *reconstruct-then-generate* pipeline (Ren et al., 2025; Lee et al., 2025; Yu et al., 2025; Chen et al., 2025b; Yang et al., 2026), which first reconstructs an explicit 3D representation from the reference video, projects this geometry along the target camera trajectory, and feeds the projected appearance as a spatial condition. By supplying explicit geometric guidance, this design relieves the model from having to jointly infer geometry and synthesize photorealistic frames. However, its quality hinges on the accuracy of the reconstructed geometry, which becomes challenging for dynamic scenes, where inaccurate dynamic-object geometry introduces distortions and flying-pixel artifacts near object boundaries. These artifacts corrupt the spatial condition and prevent the diffusion model from faithfully capturing the geometry and motion of dynamic objects.

To avoid this dependence on error-prone reconstruction, a second and more recent line of work performs *camera-conditioning-only* generation (Bai et al., 2025a; Park et al., 2025; Huang et al., 2025b; Li et al., 2026b), without any explicit 3D primitives. These methods feed the reference video as additional input tokens and give the target camera trajectory through Plücker coordinates (Sitzmann et al., 2021) or camera RoPE embedding within 3D attention (Wu et al., 2026; Li et al., 2026a). Because the reference and target views are processed jointly within the same 3D attention modules, inter-view and intra-view information is exchanged implicitly at every layer. This implicit design yields markedly more photorealistic results, but the lack of explicit geometric grounding leaves the model with a weak understanding of scene structure. As a result, dynamic objects are often placed at incorrect locations or assigned inconsistent motions, producing inter-view geometric and motion inconsistencies.

In this work, we start from the observation that camera-conditioning-only methods already excel at photorealistic synthesis, and ask whether their geometric understanding can be improved without reintroducing explicit 3D primitives. To this end, we analyze the 3D attention maps of these diffusion models (Bai et al., 2025a; Park et al., 2025) across layers and denoising timesteps, where information is implicitly exchanged across views and frames. Inspired by (Nam et al., 2025), which reveals emergent correspondence in the attention maps of video diffusion models, our analysis yields three key observations. First, query-key matching within the 3D attention blocks provides clear correspondence cues, capturing both *intra-view correspondences* within each reference and target view and *inter-view correspondences* between the two views. Second, these intra-view and inter-view correspondences become simultaneously prominent at specific intermediate layers of the diffusion model, revealing which regions the model primarily attends to when synthesizing that region. Third, in regions where dynamic objects exhibit geometric or motion inconsistencies, the attention maps at these layers exhibit incorrect intra-view and inter-view correspondences—indicating that the quality of these correspondences directly governs the consistency of the output.

Building on the above finding, we explore whether geometric and motion consistency can be improved without using explicit 3D primitives by directly supervising the correspondences in these dominant attention layers. A point in 3D space can be geometrically specified by its correspondence between two views with the corresponding camera parameters, and its motion in 4D space can be characterized by multi-view point tracking in each view. Motivated by this, we introduce **MVTrack4Gen** (**M**ulti-**V**iew Point **Track**ing **for** 4D Video **Gen**eration), a framework that leverages ground-truth multi-view point tracks as auxiliary supervision to improve geometric and motion consistency.

Specifically, we build a multi-view tracking head on top of local 4D correlation volumes computed from the query and key features of the selected attention layer, and jointly train it with the diffusion model. This encourages the model to encode motion-aware correspondences in its attention features, so that the target view more faithfully reflects the motion of the reference video. In addition, we introduce a multi-view correspondence loss that applies a cross-entropy objective directly to the attention map, encouraging each query token in the target view to attend to its corresponding ground-truth location. Together, we observe that these objectives improve both inter-view geometric consistency and intra-view temporal consistency of the generated novel-view videos.

To validate its generality, we apply our method to two camera-conditioning-only backbones, ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025), and evaluate on the DAVIS (Perazzi et al., 2016), iPhone (Gao et al., 2022) and Neural3D (Li et al., 2022) benchmarks. Our method consistently improves both backbones, achieving the best scores on most VBench (Huang et al., 2024) visual-quality metrics while reaching almost state-of-the-art geometric consistency and camera accuracy comparable to both reconstruction-based and camera-conditioning-only baselines. Notably, it markedly improves geometric consistency and visual quality for dynamic objects, without requiring any explicit 3D reconstruction at inference time.

2 Related Work

In this section, we focus on prior work most directly related to multi-view generation, while a more comprehensive discussion of related work is provided in the Appendix E.

Novel-View Generation Using Explicit 3D Representation. Recent video diffusion models (Blattmann et al., 2023; Yang et al., 2025; Wan et al., 2025) have demonstrated strong generation capability, ensuring visually plausible results. One line reconstructs explicit 3D representations from the input video and conditions a video diffusion model to synthesize newly visible regions from the target camera viewpoint. Geometry-guided approaches (Yu et al., 2025; Ren et al., 2025; Huang et al., 2025a; Wang et al., 2025; Zhao et al., 2026; Jeong et al., 2025b; Kang et al., 2026; Chen et al., 2025b; Cao et al., 2025) first estimate intermediate geometric cues from the input video, such as depth maps, point cloud, or 3D representations, and use them as spatial conditions to guide the synthesis of disoccluded target-view regions. TrajectoryCrafter (Yu et al., 2025) warps point clouds, GEN3C (Ren et al., 2025) operates in latent space to generate novel views, and ChronosObserver (Wang et al., 2025) synchronizes multi-view diffusion sampling through a hyperspace representation to produce time-synchronized, 3D-consistent multi-view videos. More recently, PostCam (Chen et al., 2025b) fuses pose and visual signals through cross-attention shared by queries, while Infinite-Homography (Kim et al., 2025) conditions the generation of homography transformations as a lightweight geometric proxy that avoids explicit depth estimation. Richer representations have also been explored, including pseudo-4D Gaussian fields from dense point tracking (Bian et al., 2025) and feed-forward 4DGS reconstructors (Yang et al., 2026). Despite their strong spatial grounding, these approaches remain bottlenecked by the accuracy and completeness of off-the-shelf reconstruction models.

Novel-View Generation Using Camera Conditioning Only. An orthogonal line of work bypasses explicit geometry entirely, relying on camera pose conditioning and the generative prior of video diffusion models to implicitly reason about 3D structure (Van Hoorick et al., 2024; Bai et al., 2025a;b; Park et al., 2025; Fan et al., 2025; Van Hoorick et al., 2026). GCD (Van Hoorick et al., 2024) is one of the earliest works to introduce camera-controlled dynamic novel-view generation by training a video generation model on Kubric (Greff et al., 2022), a multi-view synthetic video dataset. To further improve generation quality, ReCamMaster (Bai et al., 2025a) is trained on a realistic synthetic dataset with more diverse camera trajectories. Redirector (Park et al., 2025) and Camera Anything (Li et al., 2026b) designs an additional camera rotational embedding for more accurate camera-controllable video generation. A related line pursues joint space–time controllability: CAT4D (Wu et al., 2025) trains a multi-view video diffusion model to synthesize novel-views at arbitrary camera poses and timestamps, enabling 4D reconstruction via deformable 3D Gaussians, while SpaceTimePilot (Huang et al., 2025b) jointly conditions on camera trajectory and time to achieve generative rendering of dynamic scenes across both space and time.

3 ANALYSIS

Since camera-conditioning-only models (Bai et al., 2025a; Park et al., 2025) implicitly exchange information through 3D attention both across the reference and target views and over time within each view, we analyze their attention maps to examine how geometry and motion are encoded inside the models. Following (Nam et al., 2025), we conduct our main analysis on ReCamMaster (Bai et al., 2025a), and show that its 3D attention maps contain emergent token-level correspondences across layers and denoising timesteps. By extracting matches from attention weights and comparing them with pseudo ground-truth point tracks, we identify three types of emergent correspondences: *intra-view correspondences* within each reference and target view and *inter-view correspondences* between the two views. The intra-view correspondences capture motion-related temporal consistency within each view, whereas the inter-view correspondences are responsible for geometric alignment across views. Our analysis is also applicable to other camera-conditioning-only architectures, as demonstrated by our analysis of Redirector (Park et al., 2025) in Appendix B.3.

3.1 ANALYSIS SETUP

We conduct our analysis on the MultiCamVideo dataset (Bai et al., 2025a), which provides time-synchronized multi-view recordings of dynamic scenes with ground-truth camera trajectories, naturally yielding paired reference–target videos. We sample 40 scenes and randomly select two views per scene as the reference and target.

To obtain pseudo ground-truth correspondences that span both temporal and cross-view axes, we use MV-TAP (Koo et al., 2026), an off-the-shelf multi-view point tracker that jointly tracks points across reference and target video frames. This produces multi-view pseudo ground-truth point tracks $\mathcal{T} = \{p_i^{v,\mathrm{GT}}\}$ with visibility $\mathcal{O} = \{o_i^{v,\mathrm{GT}}\}$, where v and i index the view and frame, respectively. More details are in Appendix B.1.

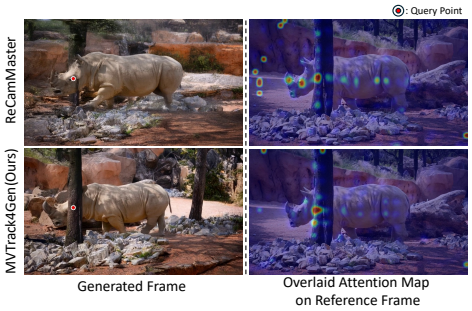


Figure 2: **Cross-View Attention Visualization.** For the same query point in the generated frame, ReCamMaster attends to incorrect regions in the reference frame, whereas MVTrack4Gen localizes attention on the corresponding object, enabling more consistent motion across views.

3.2 EVALUATING ATTENTION-BASED CORRESPONDENCES

Correspondence in 3D attention map. We investigate token-level correspondences from the 3D attention map defined in Eq. 4. For a pair of latent frames, the attention matrix $C_{i,j}^{v_1,v_2}$ measures how each query features in the i-th frame of view v_1 attends to key features in the j-th frame of view v_2, where $v_1, v_2 \in \{\mathrm{ref}, \mathrm{tgt}\}$. Given a query feature in the i-th frame of view v_1, we take the key feature with the highest attention weight in the j-th frame of view v_2 as its forward match.

To enable a more accurate analysis, we apply a cycle-consistency check to filter out unreliable matches. A forward match is accepted as a *reliable correspondence* only if matching backward from the destination latent frame to the source latent frame returns to the original query token. Formal definitions are provided in Appendix B.2.

Matching accuracy and harmonic mean. We measure matching accuracy using the Percentage of Correct Keypoints (PCK), which counts a query point as a positive only if (i) it is reliable under the cycle-consistency check and (ii) its forward match coincides with the corresponding co-visible point in the pseudo ground-truth $\mathcal{T}$ on the latent grid. PCK is reported separately for *intra-view correspondence* in the reference and target views, and for *inter-view correspondences*, averaged over all query points across latent frames and scenes. Beyond matching accuracy alone, we further analyze two attention-weight-based quantities to characterize not only whether the selected match is correct, but also how the attention distribution supports information exchange across layers and denoising timesteps: an attention score, which measures how much attention weight is assigned to the correct match, and a confidence score, which measures how sharply the attention distribution is localized. We then compute the harmonic mean of three normalized metrics—matching accuracy, attention

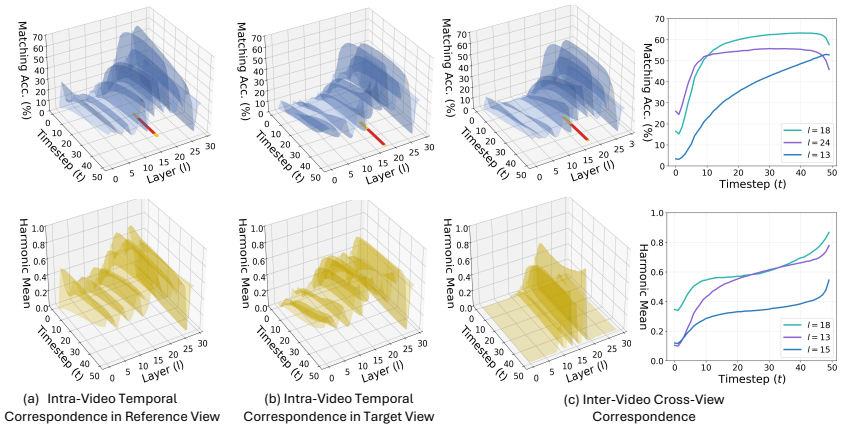


Figure 3: **Matching Accuracy and Harmonic Mean in ReCamMaster (Bai et al., 2025a).** We visualize matching accuracy (top) and the harmonic mean of matching accuracy, attention score, and confidence score (bottom) across diffusion layers and denoising timesteps. Results are shown for intra-video temporal correspondence in the reference and target views in (a) and (b), respectively, and for inter-view correspondence in (c). The rightmost column plots the top three layers with the highest cross-view correspondence scores. This figure shows that accurate cross-view and temporal matching emerge at specific intermediate layers.

score, and confidence score—to identify layers in which all three metrics are simultaneously high, thereby assessing how the correspondences emerging inside the attention layers contribute to generation; further details are provided in Appendix B.2.

3.3 Results

Fig. 3 shows the matching accuracy and harmonic mean by comparing correspondences from the attention maps and the pseudo ground-truth matches $\mathcal{T}$ across denoising timesteps and layers. We observe correspondence cues both within each view over time and across the reference and target views, which we refer to as *intra-view correspondence* and *inter-view correspondences*, respectively.

We observe that specific intermediate layers exhibit strong matching performance for both intra-view correspondence and inter-view correspondences, indicating that the same layer range supports motion tracking within each view and geometric alignment across views, as visualized in Fig. 3 (c), with a dominant peak around the 18th diffusion layer. The harmonic mean follows the same layer–timestep trend as the matching accuracy, suggesting that cross-view matching capability is encapsulated within only specific layers. Moreover, in failure cases of generation as visualized in Fig. 2, the attention map at the 18th layer fails to attend to the corresponding regions in the reference video. Together, these results motivate our design choice of explicitly reusing features from this layer range as a unified correspondence signal for both temporal and cross-view consistency.

4 Methodology

Given a reference video X_{ref} captured from a known camera trajectory cam_{ref} and a target camera trajectory cam_{tgt}, our goal is to synthesize a geometrically consistent target video X_{tgt} that faithfully depicts the same dynamic scene from the new viewpoint. By jointly providing camera and multi-view matching information as supervision, we encourage the model to reason about both the spatial structure and motion of the scene in 4D space. We achieve this by jointly training a camera-controlled video diffusion model and a multi-view tracking module that shares the query and key features from the diffusion model's 3D attention layers. We further directly supervise the 3D attention maps with a multi-view correspondence loss, guiding each query feature to attend to the correct corresponding region across views and time. Together, these objectives provide geometric and motion supervision to the generation process. An overview of our framework is illustrated in Fig. 4.

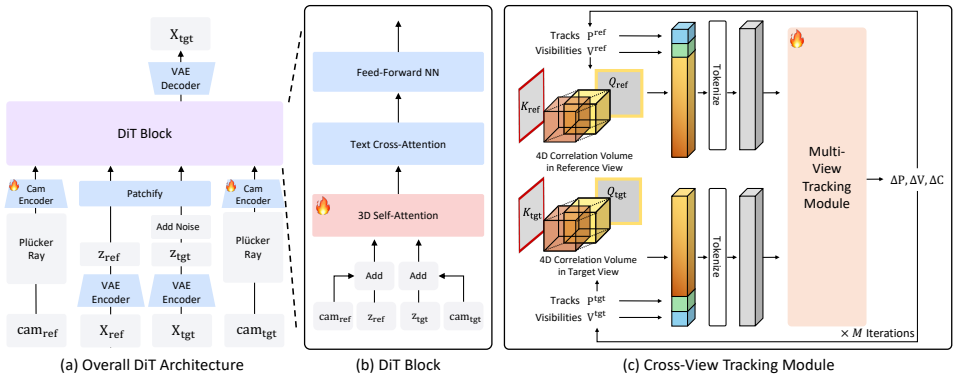


Figure 4: **Main architecture.** We jointly train a camera-controlled DiT and a multi-view tracking module that shares query and key features from the input of the DiT's 3D self-attention layers. From these shared features, the tracking module constructs *intra-video temporal correlation* for temporal consistency and *inter-video cross-view correlation* for geometric correspondence.

4.1 Improved Camera Encoding

We adopt ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025) as our backbones, but condition on both the extrinsics and intrinsics of the reference and target views, rather than the target extrinsics alone. This richer conditioning improves geometric consistency in novel-view generation. Each camera is encoded as a dense Plücker ray map (Sitzmann et al., 2021), in which every pixel is represented by a 6D ray derived from the extrinsics and intrinsics. Plücker maps from both views are injected into each DiT layer, where ReCamMaster (Bai et al., 2025a) adds the camera features to the DiT features, while Redirector (Park et al., 2025) incorporates them through its original Rotary Camera Encoding (RoCE). Additional details are provided in Appendix C.1.

4.2 Multi-View Point Tracking as Geometric Supervision

Multi-Scale 4D Correlation Volume in Each View. We construct a multi-scale local 4D correlation volume (Cho et al., 2024) in each view, leveraging the query–key similarity in the 3D attention map. This is motivated by our analysis in Sec. 3, which shows that the *intra-view correspondence* in the 3D attention map inherently encode motion across time. Local query and key features q_i^v and k_j^v are bilinearly sampled within a Δ-sized neighborhood centered at the query points $p_i^v = (x_i^v, y_i^v)$ of the i-th frame in each view $v \in \{\text{ref, tgt}\}$ and the estimated points $\hat{p}_j^v = (\hat{x}_j^v, \hat{y}_j^v)$ of the j-th frame. The local 4D correlation volume is then computed via softmax:

$$\text{Corr}_{i,j}^v = \text{Softmax}\left(\frac{q_i^v \left(k_j^v\right)^\top}{\sqrt{d_{\text{head}}}}\right) \in \mathbb{R}^{(2\Delta+1)^4}, \quad v \in \{\text{ref, tgt}\}. \tag{1}$$

Each temporal correlation volume within each view, Corr^{ref} and Corr^{tgt}, encodes the motion within its corresponding view and serves as the input to the multi-view tracking module. Query points are sampled among all video frames, since any point should be trackable and its motion detectable wherever it newly appears in the video. This encourages the query feature to be more similar to its corresponding ground-truth key feature, which we empirically find makes the generated motion more faithful to the reference. More details are described in Appendix C.2.

Multi-View Point Tracking Head. Built on the correlation volumes described above, we adopt a transformer-based multi-view tracking head following (Koo et al., 2026). At each iteration, the tracking head Ψ, equipped with factorized temporal and multi-view attention, predicts residuals $(\Delta P, \Delta V, \Delta C) = \Psi(G)$, where input tokens G are formed by concatenating current visibility V, confidence C, and local 4D correlation volume $\text{Corr}_{i,j}^v$ in each view. Visibility and confidence are initialized to zero for both views, i.e., $V = C = 0$.

4.3 Training Objectives

We supervise the correspondences in the 3D attention maps with two additional objectives: a multi-view point-tracking loss $\mathcal{L}_{\text{track}}$ and a multi-view correspondence loss $\mathcal{L}_{\text{corr}}$. The tracking objective

Table 1: **Quantitative comparison on the DAVIS dataset Perazzi et al. (2016).** The best score for each metric is in **bold**, and the colored numbers denote the change relative to each backbone, with (green) indicating a gain and (red) a loss.

Method	Visual Quality↑						Geo. Consist.		Camera Accuracy		
	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness	MEt3R↓	MEt3R$_{\text{dynamic}}$ ↓	mRotErr (°)↓	mTransErr↓	mCamMC↓
Explicit 3D Lifting											
GEN3C Ren et al. (2025)	0.856	0.894	0.461	0.582	0.950	0.980	0.290	0.328	2.538	0.127	0.163
TrajectoryCrafter Yu et al. (2025)	0.847	0.880	0.438	0.550	0.942	0.970	0.291	0.306	10.126	0.190	0.355
CogNVS Chen et al. (2025a)	0.811	0.862	0.417	0.530	0.959	0.978	0.333	0.346	10.439	0.228	0.400
NeoVerse Yang et al. (2026)	0.858	0.878	0.446	0.591	0.954	0.983	0.302	0.323	4.705	0.159	0.228
Camera Conditioning Only											
ReCamMaster Bai et al. (2025a)	0.904	0.907	0.502	0.652	**0.962**	0.985	0.337	0.369	3.660	0.113	0.169
Redirector Park et al. (2025)	0.897	0.911	0.506	0.680	0.952	0.985	0.318	0.395	**1.714**	0.086	0.109
Camera Anything (Li et al., 2026b)	0.883	0.904	0.503	0.684	0.948	0.981	0.276	0.320	2.577	0.115	0.152
MVTrack4Gen $_{\text{ReCamMaster}}$	0.892 (-.012)	0.909 (+.002)	0.507 (+.005)	0.685 (+.033)	0.953 (-.009)	0.984 (-.001)	0.274 (-.063)	**0.287** (-0.082)	1.858 (-1.802)	0.100 (-.013)	0.125 (-.044)
MVTrack4Gen $_{\text{Redirector}}$	**0.905** (+.008)	**0.919** (+.008)	**0.508** (+.002)	**0.687** (+.007)	0.956 (+.004)	**0.986** (+.001)	**0.267** (-.051)	0.349 (-0.036)	1.718 (+.004)	**0.073** (-.013)	**0.097** (-.012)

makes the model follow the dynamic motion of the reference video by jointly training the diffusion model with the multi-view point tracking head. The multi-view correspondence loss enforces consistency—not only between the reference view and the target view, but also within the target view itself—by strengthening matching information directly into the 3D attention map to guide where the model should attend.

The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{track}}\,\mathcal{L}_{\text{track}} + \lambda_{\text{corr}}\,\mathcal{L}_{\text{corr}}$, where $\lambda_{\text{track}} = \lambda_{\text{corr}} = 0.01$. The diffusion loss follows the rectified-flow formulation, $\mathcal{L}_{\text{diff}} = w(t)\,\big\|\mathbf{v}_\theta - (\epsilon - x_0)\big\|^2_{\text{tgt}}$, where $w(t)$ is a timestep-dependent weighting and the squared error is evaluated only on the target-view tokens. The tracking loss jointly supervises the multi-view point-tracking head and diffusion model with ground-truth trajectories and decomposes into $\mathcal{L}_{\text{track}} = \lambda_{\text{seq}}\,\mathcal{L}_{\text{seq}} + \lambda_{\text{conf}}\,\mathcal{L}_{\text{conf}} + \lambda_{\text{vis}}\,\mathcal{L}_{\text{vis}}$, where $\mathcal{L}_{\text{seq}}$ is a visibility-weighted Huber loss on the predicted 2D point coordinates, $\mathcal{L}_{\text{conf}}$ is a probabilistic confidence loss that calibrates the predicted per-point confidence against the regression error, and $\mathcal{L}_{\text{vis}}$ is a binary cross-entropy on the predicted visibility logits.

The multi-view correspondence loss $\mathcal{L}_{\text{corr}}$ directly supervises the attention weight matrix defined in Eq. 4 with a cross-entropy objective. For each query point, the ground-truth multi-view point tracks provide its corresponding locations and visibility across all frames and views. Since each frame contains a single ground-truth correspondence for a given query point when the target point is visible, we formulate this as a single-label classification problem. Specifically, we supervise the attention weight matrix to assign the highest probability to the matching token in each visible reference or target frame. More details are provided in Appendix C.4.

5 Experiments

5.1 Experimental Setup

Implementation Details. We build on ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025) as our backbones, fine-tuning the 3D attention layers and camera encoder while keeping all other parameters frozen. The multi-view point tracking head estimates point tracks using query-key features from the 18th layer of the DiT. We train on 4 NVIDIA H100 GPUs using mixed-precision training and the AdamW optimizer with a learning rate of 1×10^{-4} for 20,000 iterations with a batch size of 16. All videos are used at 480×832 resolution with 81 frames for training.

Table 2: **Quantitative comparison on the iPhone (Gao et al., 2022) and Neural3D (Li et al., 2022) datasets.**

Method	iPhone				Neural3D			
	Visual Quality			Geo. Consist.	Visual Quality			Geo. Consist.
	PSNR↑	SSIM↑	LPIPS↓	MEt3R↓	PSNR↑	SSIM↑	LPIPS↓	MEt3R↓
Explicit 3D Lifting								
GEN3C (Ren et al., 2025)	10.419	0.270	0.699	0.382	16.072	0.519	0.466	0.222
TrajectoryCrafter (Yu et al., 2025)	10.271	0.276	0.766	0.379	16.380	0.523	0.496	0.257
CogNVS (Chen et al., 2025a)	10.105	0.280	0.774	0.406	16.466	0.514	0.470	0.243
NeoVerse (Yang et al., 2026)	11.886	**0.394**	**0.579**	0.493	15.268	0.489	0.511	0.260
Camera Conditioning Only								
Redirector (Park et al., 2025)	11.447	0.329	0.720	0.539	16.170	0.511	0.462	0.213
Camera Anything (Li et al., 2026b)	10.455	0.287	0.663	0.425	14.307	0.473	0.517	0.244
MVTrack4Gen $_{\text{ReCamMaster}}$	11.521	0.340	0.640	**0.325**	16.078	0.500	0.461	0.210
MVTrack4Gen $_{\text{Redirector}}$	**11.969** (+.522)	0.338 (+.009)	0.633 (-.087)	0.347 (-.192)	**16.694** (+.524)	**0.527** (+.015)	**0.447** (-.015)	**0.199** (-.013)

Datasets. We train our model on the combined Kubric and MultiCamVideo dataset. For evaluation, we adopt two benchmarks covering complementary regimes: (i) the DAVIS dataset (Perazzi et al., 2016) for in-the-wild monocular videos with diverse object and camera motion, and (ii) the iPhone dataset (Gao et al., 2022) for casual hand-held captures of dynamic scenes.

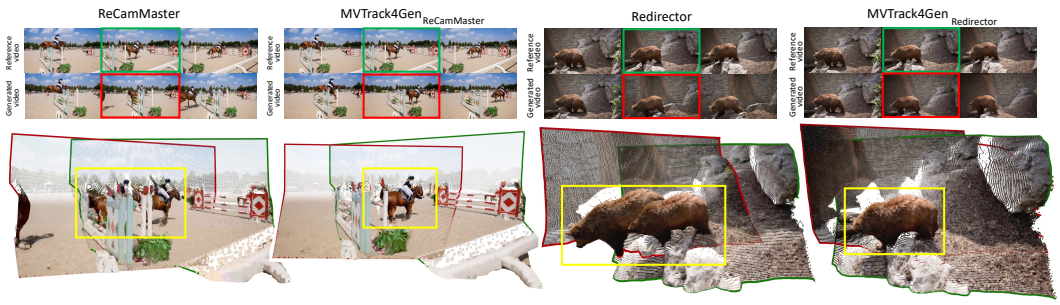


Figure 5: **Qualitative results on the DAVIS dataset (Perazzi et al., 2016).** We attach our MVTrack4Gen to two backbones, ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025), and compare against each baseline. For every example, we show the reference and generated frames, together with their point-cloud renderings reconstructed by Depth Anything3 (Lin et al., 2025) for the **reference view** and **target view**. By strengthening cross-view correspondences, MVTrack4Gen generates target frames that remain faithfully aligned with the reference view, even for dynamic objects, while preserving scene structure, appearance, and coherent motion across views.

Baselines. We compare against two families of camera-controlled video generation methods. *Explicit 3D Lifting* approaches first reconstruct geometry and re-render under the target camera; in this category we include GEN3C (Ren et al., 2025), TrajectoryCrafter (Yu et al., 2025), CogNVS (Chen et al., 2025a), and NeoVerse (Yang et al., 2026). *Camera Conditioning Only* approaches rely solely on the diffusion prior together with camera conditioning, without explicit 3D reconstruction; here we compare with ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025).

Evaluation Metrics. We evaluate along three complementary axes: generation quality, geometric consistency, and camera accuracy. For generation quality, we report PSNR, SSIM, and LPIPS against ground-truth target views on the iPhone dataset (Gao et al., 2022) and Neural3D dataset (Li et al., 2022); on DAVIS, where ground-truth views are unavailable, we follow VBench (Huang et al., 2024) and report Subject Consistency, Background Consistency, Aesthetic Quality, Imaging Quality, Temporal Flickering, and Motion Smoothness. For geometric consistency, we report MEt3R (Asim et al., 2025), which measures multi-view geometric consistency via feed-forward 3D reconstruction, and additionally report $\text{MEt3R}_{\text{dynamic}}$, which restricts this measure to dynamic regions. For camera accuracy, evaluated on DAVIS, we estimate camera trajectories from the generated videos using Depth Anything3 (Lin et al., 2025) and report mean rotation error (mRotErr), mean translation error (mTransErr), and mean camera-motion consistency (mCamMC) against the target trajectory.

5.2 Quantitative Results

We compare MVTrack4Gen against state-of-the-art baselines from both the *Explicit 3D Lifting* and *Camera Conditioning Only* paradigms on the DAVIS (Perazzi et al., 2016), iPhone (Gao et al., 2022), and Neural3D (Li et al., 2022) datasets, as summarized in Tab. 1 and Tab. 2. For a fair comparison, we exclude ReCamMaster from the iPhone and Neural3D evaluations, since it is trained under a setting where the first frames of the reference and target views are aligned, whereas these benchmarks involve reference and target views with different initial camera poses, resulting in a mismatch with its training setup.

On DAVIS, MVTrack4Gen achieves the best overall performance across all three evaluation axes. For visual quality, it ranks first on Subject Consistency, Background Consistency, Aesthetic Quality, Imaging Quality, and Motion Smoothness. For geometric consistency, it attains the lowest MEt3R and $\text{MEt3R}_{\text{dynamic}}$, outperforming all baselines. For camera accuracy, it achieves the lowest mTransErr and mCamMC while remaining competitive on mRotErr. On iPhone and Neural3D, MVTrack4Gen consistently outperforms all *Camera Conditioning Only* baselines on PSNR, LPIPS, and MEt3R, demonstrating that strengthening cross-view correspondence is highly effective even without any explicit 3D representation. Although *Explicit 3D Lifting* methods attain higher pixel-level fidelity on iPhone owing to their direct reliance on reconstructed geometry, MVTrack4Gen achieves comparable or even better geometric consistency without using any explicit 3D structure at

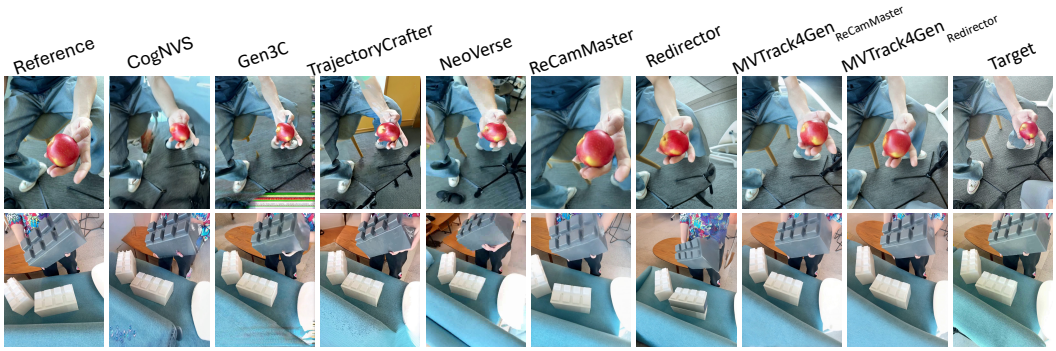


Figure 6: **Qualitative results on the iPhone dataset (Gao et al., 2022).** We visualize novel-views generated from the reference video by each method, alongside the ground-truth target view. Our MVTrack4Gen, built on both ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025), produces geometrically consistent novel-views that faithfully preserve the scene structure and appearance of the reference, while prior methods exhibit viewpoint inaccuracies, geometric distortions, or texture degradation.

inference. This confirms that strengthening cross-view correspondence in a video diffusion model effectively bridges the gap between the two paradigms.

5.3 Qualitative Results

Fig. 5 presents qualitative comparisons on the DAVIS dataset, where we attach MVTrack4Gen to both ReCamMaster (Bai et al., 2025a) and Redirector (Park et al., 2025) backbones and show, for each method, the generated frames together with their point-cloud renderings reconstructed by DepthAnything3 (Lin et al., 2025) for the reference and target views. The baselines produce plausible-looking frames but lack geometric grounding under viewpoint change, particularly on dynamic objects: their reference and generated point clouds fail to align on the moving foreground, indicating inconsistent scene geometry across views. In contrast, MVTrack4Gen produces point clouds that align faithfully between the reference and target views even for dynamic objects, preserving scene structure and keeping the motion of moving foreground objects coherent across viewpoints.

Fig. 6 presents comparisons on challenging real-world scenes from the iPhone dataset. Camera Conditioning Only baselines such as ReCamMaster and Redirector produce plausible frames but exhibit geometric inconsistencies under viewpoint changes, including incorrect object depths and non-rigid warping. Explicit 3D Lifting methods such as CogNVS, GEN3C, and TrajectoryCrafter better preserve coarse scene structure but suffer from artifacts in regions with unreliable depth, particularly for dynamic foreground objects. In contrast, MVTrack4Gen preserves scene structure and appearance while maintaining geometric consistency under significant viewpoint changes and complex motion. These results are consistent with our quantitative gains in MEt3R and camera accuracy. More generation results are provided in Appendix. F.

6 Conclusion

We introduce MVTrack4Gen, which strengthens both motion and geometric correspondence in a video diffusion model for dynamic novel-view generation. Our key insight is that correspondence-specialized attention layers naturally emerge in novel-view diffusion models, jointly encoding intra-video temporal and inter-view correspondences — a property largely overlooked in prior works. We exploit this property with two complementary objectives that share the DiT's attention features: a multi-view point tracking objective that makes the generated view follow the reference motion, and a multi-view correspondence loss that enforces geometric consistency across and within views. Without any explicit 3D reconstruction at inference, our method achieves state-of-the-art geometric consistency over both reconstruction-based and reconstruction-free baselines.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation and providing feedback on research methodology and experimental design. We have not used generative AI tools for developing the core research concepts or hypotheses, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for drafting and editing portions of the manuscript to improve readability. We have reviewed all AI-assisted work. All AI-assisted text was reviewed and revised by the authors, and any technical claims, citations, and literature suggestions were independently verified against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6034–6044, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, 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, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with dynamic 3d gaussian fields through efficient dense 3d point tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21717–21727, 2025.

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

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. *arXiv preprint arXiv:2507.12646*, 2025a.

Yipeng Chen, Zhichao Ye, Zhenzhou Fang, Xinyu Chen, Xiaoyu Zhang, Jialing Liu, Nan Wang, Haomin Liu, and Guofeng Zhang. Postcam: Camera-controllable novel-view video generation with query-shared cross-attention. *arXiv preprint arXiv:2511.17185*, 2025b.

Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. In *European conference on computer vision*, pp. 306–325. Springer, 2024.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10061–10072, 2023.

Xiang Fan, Sharath Girish, Vivek Ramanujan, Chaoyang Wang, Ashkan Mirzaei, Petr Sushko, Aliaksandr Siarohin, Sergey Tulyakov, and Ranjay Krishna. Omniview: An all-seeing diffusion model for 3d and 4d view synthesis. *arXiv preprint arXiv:2512.10940*, 2025.

Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35: 33768–33780, 2022.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–12. IEEE, 2025.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pp. 59–75. Springer, 2022.

Jiaxin Huang, Sheng Miao, BangBang Yang, Yuewen Ma, and Yiyi Liao. Vivid4d: Improving 4d reconstruction from monocular video by video inpainting, 2025a. URL `https://arxiv.org/abs/2504.11092`.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y Wang, Joan Lasenby, and Chun-Hao Huang. Spacetimepilot: Generative rendering of dynamic scenes across space and time. *arXiv preprint arXiv:2512.25075*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7276–7287, 2025a.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. 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, 2025b.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pp. 3334–3342, 2015.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11116–11126, 2026.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025.

Min-Jung Kim, Jeongho Kim, Hoiyeong Jin, Junha Hyung, and Jaegul Choo. Infinite-homography as robust conditioning for camera-controlled video generation. *arXiv preprint arXiv:2512.17040*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Jahyeok Koo, Inès Hyeonsu Kim, Mungyeom Kim, Junghyun Park, Seohyeon Park, Jaeyeong Kim, Jung Yi, Seokju Cho, and Seungryong Kim. Mv-tap: Tracking any point in multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20932–20941, 2026.

Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6216–6226. IEEE, 2025.

JoungBin Lee, Jaewoo Jung, Jisang Han, Takuya Narihira, Kazumi Fukuda, Junyoung Seo, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. 3d scene prompting for scene-consistent camera-controllable video generation. *arXiv preprint arXiv:2510.14945*, 2025.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026a.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5511–5521. IEEE, 2022.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, et al. Cameraanything: Refilming videos with arbitrary camera control. *arXiv preprint arXiv:2607.24591*, 2026b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers, 2025. URL https://arxiv.org/abs/2506.17220.

Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. Redirector: Creating any-length video retakes with rotary camera encoding. *arXiv preprint arXiv:2511.19827*, 2025.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Frano Rajič, Haofei Xu, Marko Mihajlovic, Siyuan Li, Irem Demir, Emircan Gündoğdu, Lei Ke, Sergey Prokudin, Marc Pollefeys, and Siyu Tang. Multi-view 3d point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 59–68, 2025.

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

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

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

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *European Conference on Computer Vision (ECCV)*, 2024.

Basile Van Hoorick, Dian Chen, Shun Iwase, Pavel Tokmakov, Muhammad Zubair Irshad, Igor Vasiljevic, Swati Gupta, Fangzhou Cheng, Sergey Zakharov, and Vitor Campagnolo Guizilini. Anyview: Synthesizing any novel view in dynamic scenes. *arXiv preprint arXiv:2601.16982*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Qisen Wang, Yifan Zhao, Peisen Shen, Jialu Li, and Jia Li. Chronosobserver: Taming 4d world with hyperspace diffusion sampling. *arXiv preprint arXiv:2512.01481*, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 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.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. Rayrope: Projective ray positional encoding for multi-view attention. In *European Conference on Computer Vision*, pp. 192–210. Springer, 2026.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37:76115–76138, 2024.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *International Conference on Learning Representations*, volume 2025, pp. 102934–102948, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 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.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2063–2073. IEEE, 2025.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026.

Supplementary Material

A Preliminaries

B Correspondence in 3D Attention Map

B.1 Dataset for Analysis and Pseudo Ground-Truth Generation

B.2 Details of Attention-Based Correspondence Evaluation

B.3 Generalization to Another Backbone

C Model Architecture

C.1 Plücker Ray Camera Encoding

C.2 Multi-Scale Local 4D Correlation

C.3 Training Data

C.4 Training Objective Details

D Ablations

D.1 Architecture Ablation

D.2 Layer Ablation

D.3 Effect on Attention Correspondence

D.4 Zero-Shot Multi-View Point Tracking

D.5 Inference Cost

E Related Works

F More Generation Results

G Attention Visualization Results

H Limitations

A PRELIMINARIES: VIDEO DIT FOR NOVEL-VIEW GENERATION

Here, we explain the details of camera-conditioning-only frameworks for novel-view video generation, which aim to synthesize a video from a novel camera viewpoint given a monocular reference video. A typical video generation model for novel-view synthesis consists of two main components: VAE and Diffusion Transformer (DiT). Given a reference video $X_{\text{ref}} \in \mathbb{R}^{F\times H\times W\times 3}$ where F, H, W, and 3 denote the number of frames, spatial height, width, and RGB channels, respectively, the VAE encodes it into a latent representation $z_{\text{ref}} \in \mathbb{R}^{f\times h\times w\times d_{\text{video}}}$ where $(f, h, w, d_{\text{video}})$ are the temporal, spatial, and channel dimensions in the latent space. The DiT then denoises a target-view latent $z_{\text{tgt}} \in \mathbb{R}^{f\times h\times w\times d_{\text{video}}}$ conditioned on z_{ref} to generate the novel-view video.

Specifically, the VAE downsamples the input by $4\times$ temporally and $16\times$ spatially. The DiT $\mathbf{v}_\theta$ takes the concatenation of z_{ref} and a noisy z_{tgt} as input, and predict the velocity field:

$$\hat{\mathbf{v}} = \mathbf{v}_\theta([\, z_{\text{ref}},\, z_{\text{tgt}}\,],\, t,\, c,\, \text{cam}_{\text{tgt}})\,, \tag{2}$$

where $[\,\cdot\,,\,\cdot\,]$ denotes concatenation along the token dimension, t is the flow matching timestep, c is the text caption, and cam_{tgt} is the target camera trajectory. The model is trained via flow matching (Lipman et al., 2023).

3D Attention Across Reference and Target Latent Tokens. Since z_{ref} and z_{tgt} are concatenated as input to DiT, they jointly participate in 3D attention. At each transformer layer l and flow matching timestep t, the i-th latent frame ($i \in \{1, \ldots, f\}$) of each view, either reference or target, is projected into query and key matrices:

$$Q_i^{\text{ref}},\, K_i^{\text{ref}},\, Q_i^{\text{tgt}},\, K_i^{\text{tgt}} \in \mathbb{R}^{hw\times d_{\text{head}}}, \tag{3}$$

where d_{head} is the per-head channel dimension and we omit the attention head for brevity. The attention weight matrix $\mathcal{C}_{i,j}^{v_1,v_2}$, where each entry measures the similarity between the i-th frame of view v_1 and the j-th frame in view v_2, is computed as:

$$\mathcal{C}_{i,j}^{v_1,v_2} = \text{Softmax}\left(\frac{Q_i^{v_1}\left(K_j^{v_2}\right)^\top}{\sqrt{d_{\text{head}}}}\right), \quad v_1, v_2 \in \{\text{ref},\, \text{tgt}\}. \tag{4}$$

Fig. 7 illustrates the resulting attention map structure.

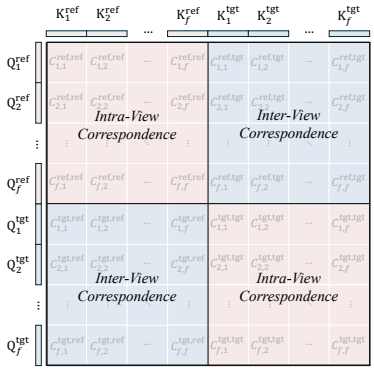


Figure 7: **Illustration of full 3D attention in video DiT for novel-view generation.** The attention jointly captures intra-view and inter-view correspondences.

B CORRESPONDENCE IN 3D ATTENTION MAP

B.1 DATASET FOR ANALYSIS AND PSEUDO GROUND-TRUTH GENERATION

MultiCamVideo Dataset. We conduct our analysis (Sec. 3) on the MultiCamVideo dataset (Bai et al., 2025a), a multi-camera synchronized video dataset rendered via Unreal Engine 5 that features large camera movements. The dataset provides time-synchronized multi-view recordings of dynamic scenes together with diverse camera trajectories. Since each scene is captured by multiple cameras at the same timestamps, the dataset naturally offers paired reference–target videos suitable for evaluating both temporal and multi-view correspondences. From this dataset, we sample 40 scenes, and for each scene we randomly select two camera views where one serves as the reference view and the other as the target view for novel-view generation.

Pseudo Ground-Truth Multi-View Point Tracking via MV-TAP. To obtain dense point correspondences that span both the temporal axis within each video and the cross-view axis between the reference and target, we leverage MV-TAP (Koo et al., 2026), a recent multi-view point tracker that jointly reasons over synchronized views via cross-view attention. Concretely, we set query points on the first frame in a regular grid manner, and run MV-TAP jointly over the 10 synchronized view



Figure 8: **Pseudo Ground-Truth Multi-View Tracks on the MultiCamVideo Dataset (Bai et al., 2025a).** We run MV-TAP (Koo et al., 2026) on the synchronized multi-view videos of the MultiCamVideo dataset to extract pseudo ground-truth multi-view point tracks. The tracked points are propagated consistently across both frames and views, providing intra-video temporal and inter-view correspondences that serve as supervision for training.

videos to estimate their correspondences across both time and viewpoints. Running MV-TAP on the reference–target pairs from the MultiCamVideo dataset yields pseudo multi-view point tracks $\mathcal{T} = \{p_i^{v,\mathrm{GT}}\}$ and visibility states $\mathcal{O} = \{o_i^{v,\mathrm{GT}}\}$, where $p_i^v \in \mathbb{R}^{N\times 2}$ denotes the 2D locations of N tracked points in view v at the i-th frame, and $o_i^v \in \{0,1\}^N$ indicates whether each point is visible. These tracks serve as our pseudo ground-truth, as shown in Fig. 8.

B.2 Details of Attention-based Correspondence Evaluation

Forward Match via Mapping Operator. For a query point $p_i^{v_1}$ in the i-th latent frame of view v_1, its forward match in the j-th latent frame of view v_2 is obtained via argmax over the attention weight matrix $\mathcal{C}_{i,j}^{v_1,v_2}$ in Eq. 4, which selects the spatial location with the highest attention score within the spatial domain Ω of the j-th latent. We define a mapping operator $\mathcal{F}$ that returns this forward match under a given attention weight matrix:

$$\hat{p}_j^{v_2} = \mathcal{F}\big(\mathcal{C}_{i,j}^{v_1,v_2},\, p_i^{v_1}\big) = \underset{p\in\Omega}{\mathrm{argmax}}\, \mathcal{C}_{i,j}^{v_1,v_2}(p_i^{v_1},\, p), \quad v_1, v_2 \in \{ref, tgt\}. \tag{5}$$

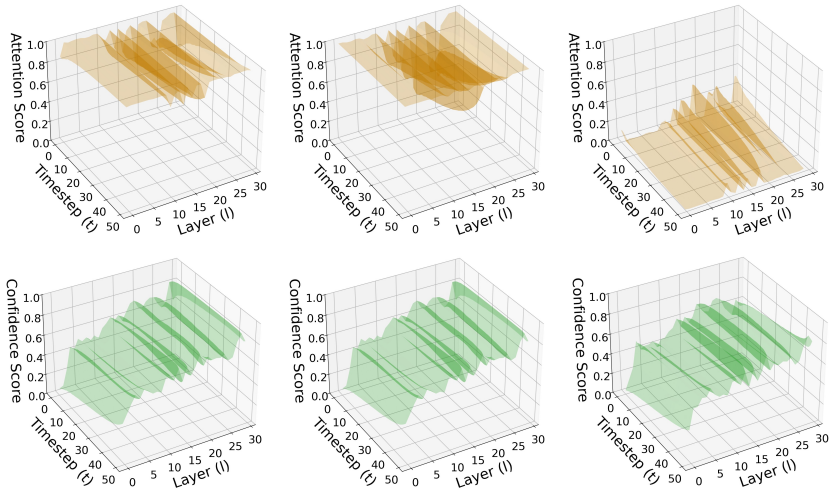
With this mapping, the forward match from view v_1 to view v_2 is $\hat{p}_j^{v_2} = \mathcal{F}(\mathcal{C}_{i,j}^{v_1,v_2}, p_i^{v_1})$, and the backward match from $\hat{p}_j^{v_2}$ back to view v_1 is obtained by applying $\mathcal{F}$ once more with the reverse attention weight matrix $\mathcal{C}_{j,i}^{v_2,v_1}$.

Reliable Correspondence via Cycle Consistency. To identify reliable correspondences in the attention map, we adopt a cycle-consistency check: a query and its forward match form a valid correspondence only when matching backward returns the original query. Formally, $p_i^{v_1}$ is regarded as reliable if

$$\mathrm{Reliable}\big(p_i^{v_1};\, \mathcal{C}_{i,j}^{v_1,v_2}\big) = \mathbb{1}\big[\, \big\|\mathcal{F}\big(\mathcal{C}_{j,i}^{v_2,v_1},\, \mathcal{F}(\mathcal{C}_{i,j}^{v_1,v_2},\, p_i^{v_1})\big) - p_i^{v_1}\big\|_2 \le \delta \,\big] \tag{6}$$

where δ denotes the cycle-consistency threshold on the pixel grid and we set $\delta = 16$. This enforces a mutual best-match relation: both directions of attention should agree on the same point pair. Since the attention is computed in the latent space, the pseudo ground-truth tracks $\mathcal{T}$ are accordingly rescaled to the latent resolution.

Matching Accuracy via PCK. We measure matching accuracy using the Percentage of Correct Keypoints (PCK). Given the pseudo ground-truth tracks $\mathcal{T}$ rescaled to the latent grid, a query point $p_i^{v_1}$ is counted as a positive when both conditions hold: (i) it is reliable under Eq. 6, and (ii) its forward match $\hat{p}_j^{v_2} = \mathcal{F}(\mathcal{C}_{i,j}^{v_1,v_2}, p_i^{v_1})$ lies within δ pixels of the co-visible ground-truth match $p_j^{v_2,\mathrm{GT}}$



(a) Intra-Video Temporal Correspondence in Reference View (b) Intra-Video Temporal Correspondence in Target View (c) Inter-Video Cross-View Correspondence

Figure 9: **Attention Score and Confidence Score in ReCamMaster (Bai et al., 2025a).** We visualize the attention score (top) and confidence score (bottom) of query points across diffusion layers l and timesteps t for the three correspondence types: (a) intra-video temporal in the reference view, (b) intra-video temporal in the target view, and (c) inter-video cross-view.

on the latent grid. Formally,

$$\mathrm{PCK} = \frac{1}{|\mathcal{Q}|} \sum_{p_i^{v_1} \in \mathcal{Q}} \mathbb{1}\Big[\mathrm{Reliable}\big(p_i^{v_1}; \mathcal{C}_{i,j}^{v_1,v_2}\big) = 1 \ \wedge \ \big\| \hat{p}_j^{v_2} - p_j^{v_2,\mathrm{GT}} \big\|_2 \leq \delta \Big], \tag{7}$$

where $\mathcal{Q}$ is the set of all query points whose ground-truth match is co-visible across the relevant frames. Matches are obtained on the latent grid, while distances are measured in the original pixel space with δ set to 16 pixels. PCK is computed separately for each of the three correspondence types and averaged across latent frames and scenes.

Attention Score. A single query point attends to tokens from both the reference and target views through the attention mechanism, allowing us to examine how its attention weights concentrate across layers and denoising timesteps. The resulting attention map encodes three types of correspondence: (a) intra-video temporal correspondence within the reference view, (b) intra-video temporal correspondence within the target view, and (c) inter-view correspondence between the two views. For each type, we visualize the attention weights of individual query points. As shown in the top row of Fig. 9, the temporal correspondences within each view, shown in (a) and (b), are emphasized in the early and final diffusion layers, whereas in the intermediate layers the attention shifts toward the cross-view correspondence in (c).

Confidence Score. The confidence score captures *how sharply* the correspondence is localized within the 3D attention map. For each query point, we re-normalize the attention over only the keys of the ground-truth frame and take the maximum value of the attention map, which approaches one when the distribution collapses onto a single key and stays low when it remains diffuse. As shown in the bottom row of Fig. 9, the temporal correspondences within each view, shown in (a) and (b), tend to become sharply localized in the later diffusion layers, whereas the cross-view correspondence in (c) exhibits higher confidence in the intermediate layers.

B.3 Generalization to Another Backbone

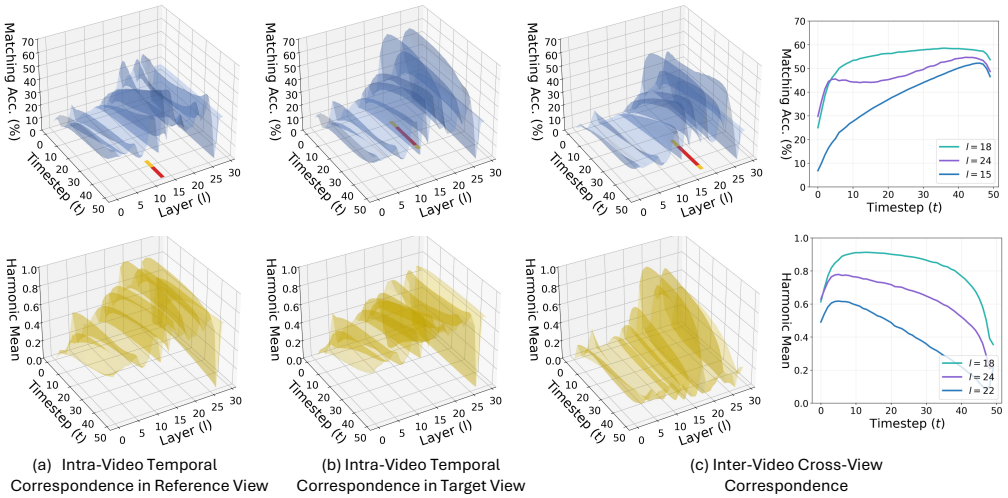
To verify that the correspondence-specialized layer structure is not an artifact of a single model, we repeat the same analysis on Redirector (Park et al., 2025), a camera-controlled video diffusion back-

bone. We compute matching accuracy, attention score, and confidence score, and average the three into a harmonic mean, using identical settings to those used for ReCamMaster (Bai et al., 2025a), and report them over various diffusion layers l and denoising timesteps t for the three correspondence types.

Accuracy and Harmonic Mean. As shown in Fig. 10, the matching accuracy (top) and the harmonic mean (bottom) on Redirector follow the same layer–timestep trend observed on ReCamMaster: intra-view correspondences in both the reference and target views, shown in (a) and (b), peak in a similar intermediate layer range, whereas the inter-view correspondence in (c) is more sharply localized around a specific middle 18th layer. The selected-layer curves in the right column further show that matching becomes accurate as denoising progresses. This consistency indicates that the emergence of a correspondence-specialized layer is a property shared across camera-controlled video diffusion backbones rather than one tied to a particular model.

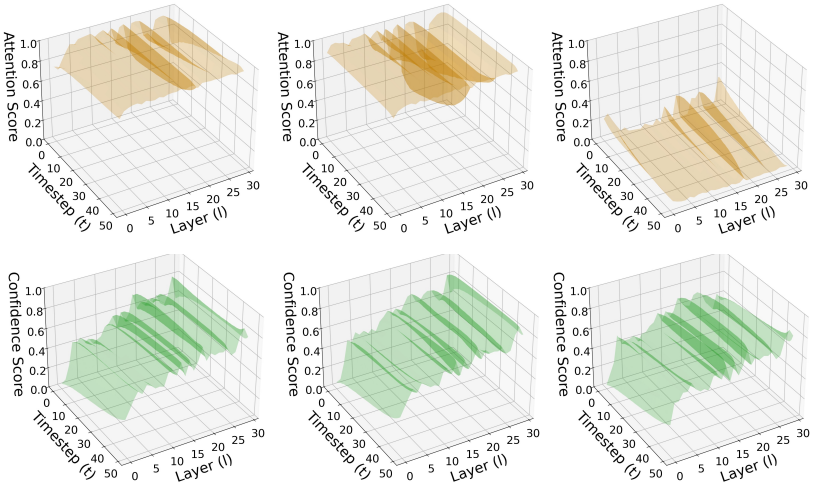
Attention Score. As shown in the top row of Fig. 11, Redirector exhibits a similar attention-routing pattern to ReCamMaster. intra-view correspondences in the reference and target views, shown in (a) and (b), receive stronger attention in the early and later diffusion layers, while inter-view correspondence in (c) becomes more prominent in the intermediate layers. This consistent trend suggests that camera-controlled video diffusion backbones tend to allocate intermediate layers to cross-view information exchange.

Confidence Score. The confidence maps in the bottom row of Fig. 11 further support this observation. Temporal correspondences become more sharply localized toward the later layers, whereas cross-view correspondences show higher confidence around the intermediate layer range. Together with the attention-score analysis, these results indicate that Redirector also develops a correspondence-specialized layer structure, supporting the generality of our choice to supervise attention features from this layer range.



(a) Intra-Video Temporal Correspondence in Reference View (b) Intra-Video Temporal Correspondence in Target View (c) Inter-Video Cross-View Correspondence

Figure 10: **Accuracy and Harmonic Mean in Redirector (Park et al., 2025).** We visualize matching accuracy (top) and the harmonic mean of accuracy, confidence, and attention score (bottom) over various diffusion layers and denoising timesteps, for (a) intra-video temporal correspondence in the reference view, (b) intra-video temporal correspondence in the target view, and (c) inter-view correspondence. The right column shows selected layers from (c) plotted against timestep t. Consistent with our analysis on ReCamMaster, cross-view matching and temporal matching within the reference and target views emerge at specific middle layers and become stronger as denoising progresses, indicating that the correspondence-specialized layer structure is not specific to a single backbone.



(a) Intra-Video Temporal Correspondence in Reference View (b) Intra-Video Temporal Correspondence in Target View (c) Inter-Video Cross-View Correspondence

Figure 11: **Attention Score and Confidence Score in Redirector (Park et al., 2025).** We visualize the attention score (top) and confidence score (bottom) of query points across diffusion layers l and timesteps t, for (a) intra-video temporal correspondence in the reference view, (b) intra-video temporal correspondence in the target view, and (c) inter-view correspondence. As on ReCamMaster, the temporal correspondences (a, b) are emphasized in the early and final layers while the cross-view correspondence (c) concentrates in the intermediate layers, confirming that the same layer-wise behavior holds across backbones.

C Model Architecture

C.1 Plücker Ray Camera Encoding

Plücker Ray Construction. We provide the formal construction of the Plücker ray map. For each pixel, the corresponding ray $\mathbf{r} \in \mathbb{R}^6$ is defined as

$$\mathbf{r} = \begin{bmatrix} \mathbf{d} \\ \mathbf{m} \end{bmatrix}, \quad \text{where } \mathbf{m} = \mathbf{o} \times \mathbf{d}, \tag{8}$$

with the ray direction $\mathbf{d} \in \mathbb{R}^3$ and origin $\mathbf{o} \in \mathbb{R}^3$ computed as

$$\mathbf{d} = \mathbf{R}^\top \mathbf{K}^{-1} \mathbf{x}, \quad \mathbf{o} = -\mathbf{R}^\top \mathbf{t}, \tag{9}$$

where $\mathbf{x} = (u, v, 1)^\top$ is the homogeneous pixel coordinate, $\mathbf{R}$ and $\mathbf{t}$ are the camera rotation and translation, and $\mathbf{K}$ is the intrinsic matrix. The direction $\mathbf{d}$ is normalized to unit length for scale invariance. Aggregating the rays over all pixels and frames yields the dense Plücker maps $\mathbf{r}_{\text{ref}}, \mathbf{r}_{\text{tgt}} \in \mathbb{R}^{6 \times F \times H \times W}$.

Incorporating Plücker Maps into DiT Layers. The Plücker maps are injected into each DiT layer as

$$h' = h + \mathrm{Proj}\big(\mathrm{CamEnc}([\mathbf{r}_{\text{ref}}, \mathbf{r}_{\text{tgt}}])\big), \tag{10}$$

where h is the input token to the 3D attention at each layer, CamEnc is a lightweight convolutional encoder that maps the 6-channel Plücker rays to the token dimension, and Proj is a linear projection. We follow the same injection location as ReCamMaster (Bai et al., 2025a).

C.2 Multi-Scale Local 4D Correlation

We construct a multi-scale local 4D correlation with varying receptive fields to capture multi-scale query–key relationships from the 3D attention. To build a feature pyramid with S scales, we interpolate the DiT query and key features Q_i^v, K_i^v at each transformer layer to resolution $\frac{H}{r \times 2^{s-1}} \times \frac{W}{r \times 2^{s-1}}$, where $r = 4$ denotes the model stride and $s \in \{1, 2, 3, 4\}$ denotes the scale factor:

$$^sQ_i^v = \mathrm{Interpolate}_s(Q_i^v), \tag{11}$$

$$^sK_i^v = \mathrm{Interpolate}_s(K_i^v), \quad v \in \{\text{ref}, \text{tgt}\}. \tag{12}$$

At each scale s, we extract local features around points of interest via bilinear sampling. Local query features $^sq_i^v$ are sampled within a Δ-sized neighborhood centered at the query point $p_i^v = (x_i^v, y_i^v)$ in frame i of view v:

$$^sq_i^v = {}^sQ_i^v\left(\frac{x_i^v}{2^{s-1}} + \delta_x,\ \frac{y_i^v}{2^{s-1}} + \delta_y : \|\delta\|_\infty \le \Delta\right), \tag{13}$$

and local key features $^sk_j^v$ are sampled around the estimated point $\hat{p}_j^v = (\hat{x}_j^v, \hat{y}_j^v)$ across all frame indices j in view v:

$$^sk_j^v = {}^sK_j^v\left(\frac{\hat{x}_j^v}{2^{s-1}} + \delta_x,\ \frac{\hat{y}_j^v}{2^{s-1}} + \delta_y : \|\delta\|_\infty \le \Delta\right), \tag{14}$$

where $\delta \in \mathbb{Z}^2$ and $^sq_i^v, {}^sk_j^v \in \mathbb{R}^{d \times (2\Delta+1)^2}$, with $v \in \{\text{ref}, \text{tgt}\}$. We then construct a local 4D correlation volume $^s\,\mathrm{Corr}_{i,j}^v$ between local features of frame i and frame j in view v via the softmax operation:

$$^s\,\mathrm{Corr}_{i,j}^v = \mathrm{Softmax}\left(\frac{^sq_i^v \left(^sk_j^v\right)^\top}{\sqrt{d_{\text{head}}}}\right) \in \mathbb{R}^{(2\Delta+1)^4}, \quad v \in \{\text{ref}, \text{tgt}\}. \tag{15}$$

We concatenate the local 4D correlation volumes from all S scales along the channel dimension to obtain the multi-scale correlation descriptor $\mathrm{Corr}_{i,j}^v \in \mathbb{R}^{4(2\Delta+1)^4}$.

$$\mathrm{Corr}_{i,j}^v = \mathrm{Concat}\Big(^1\,\mathrm{Corr}_{i,j}^v,\ ^2\,\mathrm{Corr}_{i,j}^v,\ ^3\,\mathrm{Corr}_{i,j}^v,\ ^4\,\mathrm{Corr}_{i,j}^v\Big) \in \mathbb{R}^{4(2\Delta+1)^4}. \tag{16}$$

C.3 TRAINING DATA.

We train on a hybrid dataset to maximize diversity in scene types, camera motions, and visual domains:

Kubric. The Kubric dataset provides synthetic multi-object scenes rendered with Blender, precise camera intrinsics, extrinsics, and dense point tracks with ground-truth visibility labels. Videos are center-cropped from 832×832 to 480×832 at 81 frames.

MultiCamVideo. The MultiCamVideo dataset provides cinematic multi-camera sequences rendered via Unreal Engine 5, featuring 10 synchronized camera viewpoints per scene with known calibration. Dense point tracks are computed on a 30×52 spatial grid (1,560 points per frame).

Reverse MultiCamVideo Augmentation. To improve robustness, we augment training by reversing the temporal order of the video sequences, which increases the diversity of both object motion and camera trajectories seen during training.

C.4 TRAINING OBJECTIVE DETAILS

Diffusion Loss. We adopt the rectified flow formulation,

$$\mathcal{L}_{\text{diff}} = w(t)\,\big\|\mathbf{v}_\theta([\,z_{\text{ref}},\ z_{\text{tgt}}\,], t, c, \text{cam}_{\text{ref}}, \text{cam}_{\text{tgt}}) - (\epsilon - x_0)\big\|^2_{\text{tgt}}, \tag{17}$$

where $\mathbf{v}_\theta(\cdot)$ is the predicted velocity field conditioned on the concatenated reference and target latents $[z_{\text{ref}}, z_{\text{tgt}}]$, the denoising timestep $t \sim \mathcal{U}(0,1)$, the conditioning signal c, and the target camera cam_{tgt}. The training target is the rectified flow velocity $\epsilon - x_0$, where $\epsilon \sim \mathcal{N}(0, I)$ is the sampled noise and x_0 denotes the clean target latent, with the noisy input constructed as $x_t = (1-t)\,x_0 + t\,\epsilon$. The weighting $w(t)$ is a timestep-dependent factor that balances the contribution of different noise levels during training. The squared error is evaluated only on the target-view tokens, as indicated by the $\|\cdot\|^2_{\text{tgt}}$ notation; reference-view tokens serve purely as conditioning context and do not contribute to the diffusion loss.

Tracking Loss. The multi-view point tracking head (Koo et al., 2026) is supervised with ground-truth point trajectories using

$$\mathcal{L}_{\text{track}} = \lambda_{\text{seq}}\,\mathcal{L}_{\text{seq}} + \lambda_{\text{conf}}\,\mathcal{L}_{\text{conf}} + \lambda_{\text{vis}}\,\mathcal{L}_{\text{vis}}, \tag{18}$$

with $\lambda_{\text{seq}} = 0.05$ and $\lambda_{\text{conf}} = \lambda_{\text{vis}} = 1.0$. Given the ground-truth tracks $\mathcal{T} = \{p^{v,\text{GT}}_{n,i}\}$ and visibility $\mathcal{O} = \{o^{v,\text{GT}}_{n,i} \in \{0,1\}\}$ of the N query points indexed by n over the i latent frames of both views $v \in \{\text{ref}, \text{tgt}\}$, together with the head's predictions $\{\hat{p}^v_{n,i}, \hat{c}^v_{n,i}, \hat{o}^v_{n,i}\}$, the three terms are defined as follows. The sequence loss $\mathcal{L}_{\text{seq}}$ is a visibility-weighted Huber loss (Huber, 1992) on the predicted 2D point coordinates,

$$\mathcal{L}_{\text{seq}} = \frac{1}{\sum_{v,n,i} o^{v,\text{GT}}_{n,i}} \sum_{v,n,i} o^{v,\text{GT}}_{n,i}\,\rho\Big(\hat{p}^v_{n,i} - p^{v,\text{GT}}_{n,i}\Big), \tag{19}$$

where $\rho(\cdot)$ is the Huber function; this restricts coordinate regression to frames where the ground-truth point is visible and prevents occluded targets from injecting noisy gradients. The confidence loss $\mathcal{L}_{\text{conf}}$ is a probabilistic loss that calibrates the predicted per-point confidence $\hat{c}^v_{n,i}$ against the realized regression error: confidences are supervised to be high when the predicted location lies within a small radius of the ground-truth and low otherwise, encouraging the head to produce reliability estimates that downstream consumers can use to filter unreliable tracks.

Multi-View Correspondence Loss. While the tracking head is supervised only through the sampled local correlation volumes in each view, the attention weight matrix $\mathcal{C}^{v_1,v_2}_{i,j}$ in Eq. 4 itself encodes three types of correspondence within the attention map: *intra-view correspondence* within the reference view and within the target view, and *inter-view correspondence* between the two views. These correspondences act as a bridge that exchanges information both across views and across time, and visually failed regions in the generated output tend to coincide with correspondences that are

Table 3: **Architecture ablation on the MultiCam Video Dataset Bai et al. (2025a).** We progressively add components to the ReCamMaster baseline to validate each design choice.

Condition Type	Visual Quality			Geo. Consist.	Camera Accuracy		
	PSNR↑	SSIM↑	LPIPS↓	MEt3R↓	mRotErr (°)↓	mTransErr↓	mCamMC↓
(i) Plücker Ray Camera Encoding	13.69	0.546	0.505	0.391	4.92	20.50	23.01
(ii) (i) + Correspondence Loss	14.93	0.554	0.472	0.386	4.57	18.67	21.88
(iii) (i) + Tracking Module	14.49	**0.571**	**0.456**	0.378	**3.47**	17.86	19.43
(iv) (i) + Tracking Module + Correspondence Loss	**15.05**	**0.571**	0.465	**0.363**	3.90	**17.48**	**19.31**

misaligned in the attention map. We therefore supervise it directly with a cross-entropy objective derived from the multi-view point tracks $\mathcal{T}$, so that each query point attends to the location where it physically reappears in the other frames and the other view. The key property we exploit is that the multi-view tracks specify the correct token *exactly*: for a query point $p_i^{v_1}$ in the i-th latent frame of view v_1, its match in any other latent frame j of view v_2 is either a single ground-truth location $p_j^{v_2,\mathrm{GT}}$ or occluded, as indicated by the visibility flag $o_j^{v_2,\mathrm{GT}} \in \{0,1\}$.

We randomly sample a set of query points $\mathcal{Q}$, and for each query we supervise its correspondence against every co-visible target token across all latent frames j of both the reference and target views $v_2 \in \{\mathrm{ref}, \mathrm{tgt}\}$. Since each row $\mathcal{C}_{i,j}^{v_1,v_2}(p_i^{v_1}, \cdot)$ is already a per-frame softmax distribution over the keys of the j-th latent frame, we apply a visibility-weighted cross-entropy loss for each query point $p_i^{v_1}$:

$$\mathcal{L}_{\mathrm{corr}} = \sum_{v_2} \sum_{j} o_j^{v_2,\mathrm{GT}} \, \mathrm{CE}\left(\mathcal{C}_{i,j}^{v_1,v_2}(p_i^{v_1}, \cdot), p_j^{v_2,\mathrm{GT}}\right). \tag{20}$$

Here, $\mathrm{CE}(\cdot,\cdot)$ denotes the cross-entropy loss with the ground-truth matching point $p_j^{v_2,\mathrm{GT}}$ as the target label. By averaging over all co-visible target frames j in both views $v_2 \in \{\mathrm{ref}, \mathrm{tgt}\}$, the loss jointly supervises intra-view correspondence within each view and inter-view correspondences across the two views, directly shaping where each query token attends to enforce geometric consistency both across the reference and generated views and within the generated views themselves.

D Ablations

D.1 Architecture Ablation

We conduct ablation studies on the MultiCamVideo dataset (Bai et al., 2025a) to evaluate the contribution of each component; results are summarized in Tab. 3. All variants are fine-tuned for 10k iterations from ReCamMaster, with "+" indicating additions to the camera-encoding baseline in (i). Configuration (ii) adds the multi-view correspondence loss on the 3D attention map and consistently improves all reported metrics over (i), including visual quality, geometric consistency, and camera accuracy. This shows that directly supervising *where each query attends* strengthens cross-view correspondence and benefits downstream generation. Configuration (iii) instead introduces the multi-view tracking module, yielding larger gains in SSIM, LPIPS, MEt3R, and camera accuracy, which indicates that learning temporally consistent point correspondences provides effective supervision for dynamic novel-view generation. Finally, configuration (iv) combines both objectives and achieves the best PSNR, MEt3R, mTransErr, and mCamMC, while matching the best SSIM. Although (iii) remains slightly better in LPIPS and mRotErr, the combined model provides the strongest overall balance across visual quality, geometric consistency, and camera control, demonstrating that the tracking and correspondence objectives provide complementary supervision.

D.2 Layer Ablation

Our analysis in Sec. 3 identifies the 18th DiT layer as the correspondence-specialized layer whose 3D attention most reliably encodes inter-view correspondence. This observation translates into the best downstream generation, i.e., that the layer carrying the strongest correspondence signal is also the most effective source of features for our multi-view tracking head. We ablate the matching layer l — the DiT layer whose query and key features are routed into the tracking head — across four layers spanning the network, and report generation quality, geometric consistency, and camera accuracy on the MultiCamVideo (Bai et al., 2025a) evaluation set in Table 4. Layer 18 attains the best generation quality on PSNR, SSIM, and LPIPS, the lowest MEt3R, and the most accurate camera control on

Table 4: **Layer ablation on the MultiCamVideo (Bai et al., 2025a) dataset.** We vary the matching layer l that feeds the multi-view tracking head and the best value per metric is shown in **bold**.

Matching Layer	Visual Quality			Geo. Consist.	Camera Accuracy		
	PSNR↑	SSIM↑	LPIPS↓	MEt3R↓	mRotErr (°)↓	mTransErr↓	mCamMC↓
Camera Encoding Only	13.69	0.546	0.505	0.391	4.92	20.50	23.01
$l = 15$	14.63	0.556	0.483	0.375	4.12	17.94	20.48
$l = 18$	**15.05**	**0.571**	**0.465**	**0.363**	**3.90**	**17.48**	**19.31**
$l = 22$	14.53	0.563	**0.465**	0.366	4.33	18.32	21.35
$l = 24$	14.45	0.563	0.473	0.365	4.38	18.44	21.44

Table 5: **Layer-wise attention correspondence accuracy.** We compare ReCamMaster and MVTrack4Gen$_{\text{ReCamMaster}}$ using PCK (%) for inter-view correspondence across views and intra-view correspondence over time.

		DiT Layer										
Matching	Model	0	3	6	9	12	15	18	21	24	27	29
Inter-view	ReCamMaster	12.3	11.7	13.3	11.4	7.5	36.0	55.1	21.2	51.4	1.5	20.8
	MVTrack4Gen$_{\text{ReCamMaster}}$	12.5	15.3	14.7	13.1	9.6	47.3	64.4	24.1	56.2	2.5	22.2
	Δ	+0.2	+3.6	+1.4	+1.7	+2.1	+11.3	+9.3	+2.9	+4.8	+1.0	+1.4
Intra-view (ref.)	ReCamMaster	18.5	23.2	24.3	18.0	12.7	47.3	65.0	28.8	55.8	4.1	27.1
	MVTrack4Gen$_{\text{ReCamMaster}}$	18.1	27.0	26.5	21.7	15.5	60.1	81.2	31.7	63.0	6.7	28.9
	Δ	−0.4	+3.8	+2.2	+3.7	+2.8	+12.8	+16.2	+2.9	+7.2	+2.6	+1.8
Intra-view (tgt.)	ReCamMaster	16.9	16.8	18.2	15.5	12.1	35.7	59.1	28.9	53.1	1.7	33.2
	MVTrack4Gen$_{\text{ReCamMaster}}$	17.4	22.5	19.2	18.7	14.4	49.5	69.8	32.6	57.1	2.9	33.4
	Δ	+0.5	+5.7	+1.0	+3.2	+2.3	+13.8	+10.7	+3.7	+4.0	+1.2	+0.2

all three pose metrics, consistent with our analysis where cross-view correspondence peaks at this layer.

D.3 Effect on Attention Correspondence

We further examine whether our training objectives strengthen the correspondence encoded in the DiT attention maps. Table 5 compares layer-wise PCK between ReCamMaster (Bai et al., 2025a) and MVTrack4Gen$_{\text{ReCamMaster}}$ for both inter-view and intra-view correspondences. Our method improves correspondence accuracy across nearly all layers, with particularly large gains in the intermediate layers where strong correspondence naturally emerges. This result shows that our multi-view supervision directly strengthens both geometric correspondence across views and temporal correspondence within each view.

D.4 Zero-Shot Multi-View Point Tracking

Table 6 evaluates the zero-shot multi-view point tracking capability of MVTrack4Gen on the Panoptic Studio dataset (Joo et al., 2015). For each scene, we randomly select two views and report results averaged over 100 scenes without any task-specific fine-tuning. We report Jaccard and PCK (δ^x) at pixel thresholds $\{1, 2, 4, 8, 16\}$, where PCK measures point localization accuracy and Jaccard additionally accounts for visibility prediction. Although MVTrack4Gen is trained primarily to improve novel-view video generation rather than as a standalone point tracker, it retains non-trivial zero-shot tracking performance across thresholds. In particular, both backbone variants achieve strong localization accuracy at larger thresholds, suggesting that the correspondence supervision learned during generation training transfers to unseen multi-view tracking scenarios.

D.5 Inference Cost

As reported in Tab. 7, we evaluate all methods on a single NVIDIA H100 80 GB under identical generation settings: 81 frames at 832×480, 50 sampling steps, batch size 1, and bfloat16 precision. We report both the number of parameters in the generative denoiser and the total parameters of all networks executed during inference, including the text/image encoders, VAE, denoiser, and auxiliary modules. FLOPs are measured for one complete generation using

Table 6: **Zero-shot multi-view point tracking.** We evaluate zero-shot point tracking on the Panoptic Studio Dataset (Joo et al., 2015). For each scene, we randomly select two views and report threshold-wise Jaccard and PCK, averaged over 100 scenes ($\times 100$).

Method	Jaccard						δ^x (PCK)					
	$< \delta^x_1$	$< \delta^x_2$	$< \delta^x_4$	$< \delta^x_8$	$< \delta^x_{16}$	AJ	$< \delta^x_1$	$< \delta^x_2$	$< \delta^x_4$	$< \delta^x_8$	$< \delta^x_{16}$	$< \delta^x_{\text{avg}}$
MV-TAP (Koo et al., 2026)	5.7	15.8	33.0	53.0	67.4	35.0	11.8	29.3	53.3	75.9	89.7	52.0
MVTrack4Gen$_{\text{ReCamMaster}}$	3.2	10.3	28.4	51.0	62.6	31.1	6.1	18.6	44.6	72.4	88.4	46.0
MVTrack4Gen$_{\text{ReDirector}}$	3.6	11.3	29.5	50.2	60.6	31.0	6.7	19.8	45.5	71.4	86.9	46.1

Table 7: **Inference cost of camera-conditioned video generation.** We evaluate the inference cost of each method under the same generation setting. Rows marked with "+ tracking head" report the additional computational cost introduced by the tracking head.

Method	Params (B)		FLOPs (P)	Time (s)	Peak mem. (GiB)
	Denoiser	Total			
CameraAnything	18.54	24.35	568.0	1135.5	56.7
ReCamMaster	1.49	7.30	98.1	219.1	17.1
ReDirector	1.66	8.08	98.7	321.2	21.7
MVTrack4Gen$_{\text{ReCamMaster}}$	1.50	7.31	98.4	244.7	19.7
+ tracking head	+0.05	+0.05	+0.05	+0.86	+12.9
MVTrack4Gen$_{\text{ReDirector}}$	1.66	7.47	98.4	253.7	21.3
+ tracking head	+0.05	+0.05	+0.05	+0.78	+12.9

`torch.utils.flop_counter` ($1\,\text{PFLOP} = 10^{15}$), and runtime is reported as the median over 10 runs after 3 warm-up runs. Despite introducing an additional tracking head for supervision, MVTrack4Gen incurs negligible inference overhead. The tracking head is lightweight and is not invoked during video generation, so the generation cost remains essentially unchanged; we report the head itself only as a separate incremental cost.

E Related Works

Point Tracking. Tracking Any Point (TAP) formulates long-term pixel-level correspondence as a generalization of optical flow that explicitly handles occlusion (Harley et al., 2022; Doersch et al., 2022). Early works such as PIPs (Harley et al., 2022) and TAP-Net (Doersch et al., 2022) suffered from occlusion fragility or per-frame independence, which TAPIR (Doersch et al., 2023) addressed by combining global matching-based initialization with a temporal refinement stage. CoTracker (Karaev et al., 2024) further demonstrated that tracking points jointly via a transformer with cross-track attention substantially improves robustness under occlusion and fast motion, and CoTracker3 (Karaev et al., 2025) unifies these ideas under a simpler architecture trained with pseudo-labels.

In the multi-view setting, MV-TAP (Koo et al., 2026) aggregates spatio-temporal information across views via cross-view attention for robust 2D trajectory estimation, while MVTracker (Rajič et al., 2025) targets feed-forward 3D point tracking by fusing multi-view features. These formulations yield geometrically consistent correspondences beyond what monocular trackers can recover, motivating their use as an auxiliary supervision signal to strengthen the geometric feature learning of video diffusion models.

Motion Aware Video Generation. Recent video generation methods increasingly incorporate explicit motion cues to improve temporal consistency and controllability. Several works condition video diffusion models on optical flow or object trajectories to control camera and object motion (Wang et al., 2024; Yin et al., 2023; Koroglu et al., 2025; Zhang et al., 2025; Geng et al., 2025), while others modify attention mechanisms to propagate features along predicted or user-specified trajectories (Shi et al., 2024; Xiao et al., 2025). Beyond using motion solely as a conditioning signal,

recent studies show that motion information is inherently encoded in intermediate diffusion features and attention maps (Ling et al., 2024; Xiao et al., 2024). In particular, Track4Gen (Jeong et al., 2025a) introduces point-tracking supervision into video diffusion features and demonstrates that explicitly learning temporal correspondences improves generation consistency. These works suggest that establishing reliable motion correspondences within diffusion representations is an effective means of improving temporally coherent video generation.

F More Generation Results

To complement the iPhone results in Fig. 6, we provide additional qualitative comparisons on the DAVIS dataset (Perazzi et al., 2016), which contains in-the-wild monocular videos with diverse object and camera motion. Figs. 12, 13, and 14 show novel-view generations across representative scenes covering both dynamic foreground subjects and wide outdoor environments, each under a different target camera trajectory. *Explicit 3D Lifting* baselines such as CogNVS, GEN3C, and TrajectoryCrafter preserve the coarse scene layout but introduce noticeable artifacts in regions where depth estimation becomes unreliable, including water surfaces, thin structures, and fast-moving subjects. NeoVerse mitigates some of these artifacts but still produces residual distortions on dynamic objects under large camera motion. *Camera Conditioning Only* baselines such as ReCamMaster and Redirector synthesize plausible-looking frames but lack geometric grounding under viewpoint change, leading to inaccurate parallax, drifting backgrounds, and non-rigid warping of foreground objects. In contrast, our MVTrack4Gen generates novel-views that remain visually faithful to the reference video while accurately following the prescribed target camera trajectory, preserving both the structure of dynamic subjects and the geometry of the surrounding scene. These qualitative observations are consistent with the quantitative trends reported in Tab. 1, further confirming that strengthening multi-view correspondence learning into a video diffusion model is effective for in-the-wild novel-view synthesis, even without any explicit 3D representation at inference.

G Attention Visualization Results

We examine the relationship between generation fidelity and the model's internal attention behavior As established there, the 18th layer acts as the *correspondence-specialized layer*, where each query point most strongly attends to its geometrically corresponding location across views; we therefore visualize the attention map at this layer. We find that synthesis errors are tightly coupled with a breakdown of this alignment: when a query point is placed on an incorrectly generated region, a model without multi-view correspondence supervision attends to spurious locations in both the reference view and the generated view, rather than to the physically corresponding point. As shown in Fig. 15, ReCamMaster (Bai et al., 2025a) produces visible geometric inconsistency in the dynamic object, and its attention is dispersed onto irrelevant background structures instead of the queried object. Built upon the same backbone, MVTrack4Gen not only synthesizes markedly higher-fidelity novel-view frames but also keeps the attention sharply concentrated and consistently aligned with the correct correspondences. This visual evidence is consistent with our central finding that explicitly strengthening multi-view correspondence directly translates into geometrically faithful novel-view generation.

H Limitations

While MVTrack4Gen improves geometric and motion consistency for dynamic novel-view generation, several limitations remain. First, our current implementation is trained and evaluated at a fixed resolution of 480×832 with 81-frame video clips. Extending the framework to substantially longer sequences or higher-resolution generation remains unexplored. Second, training requires dense multi-view point correspondence supervision, which limits the available training data compared with unconstrained monocular video collections. Extending MVTrack4Gen to large-scale in-the-wild videos may therefore require reliable pseudo-labeling or self-supervised correspondence learning. Finally, MVTrack4Gen inherits the computational cost of the underlying diffusion-based video generation backbone. Although the auxiliary tracking module is used only during training and does not introduce substantial additional inference cost, the overall video generation process re-

mains too computationally expensive for real-time applications. We leave these directions to future work.



Figure 12: **Additional qualitative comparisons on the DAVIS dataset (Perazzi et al., 2016).** The top Reference Video row shows reference-view frames at three timesteps, each row below renders the same target viewpoint, and the rightmost *Zoom Out* column visualizes the target camera trajectory in 3D. We compare our two models, $\text{MVTrack4Gen}_{\text{ReCamMaster}}$ and $\text{MVTrack4Gen}_{\text{Redirector}}$, against ReCamMaster (Bai et al., 2025a), Redirector (Park et al., 2025) and CogNVS (Chen et al., 2025a), GEN3C (Ren et al., 2025), TrajectoryCrafter (Yu et al., 2025), NeoVerse (Yang et al., 2026). Baselines show geometric distortions on dynamic foreground objects and floating or blurry background artifacts under large viewpoint changes, while MVTrack4Gen produces sharp, geometrically consistent novel-views faithful to the reference video.

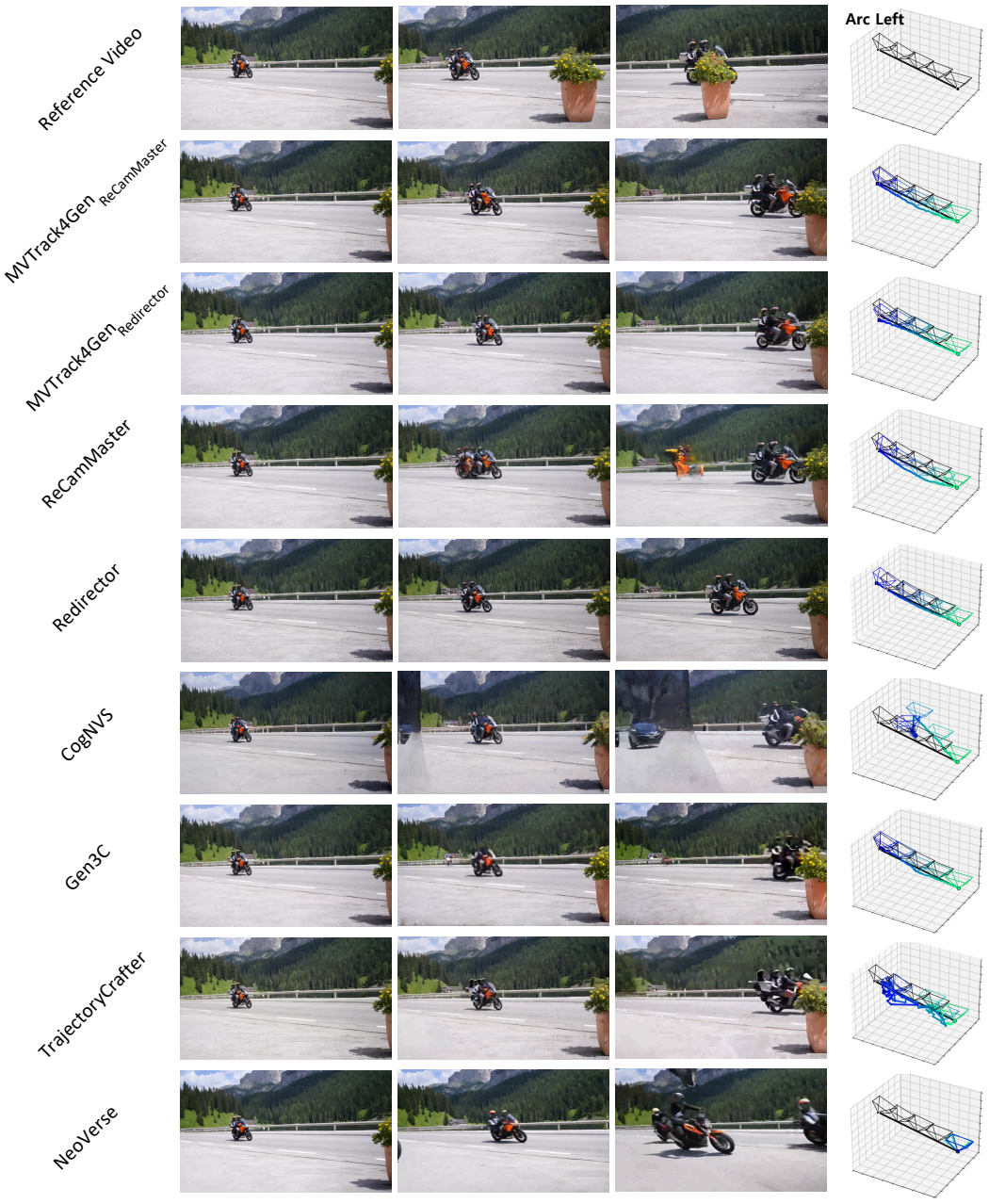


Figure 13: **Additional qualitative comparisons on the DAVIS dataset (Perazzi et al., 2016).** The top Reference Video row shows reference-view frames at three timesteps, each row below renders the same target viewpoint, and the rightmost *Arc Left* column visualizes the target camera trajectory in 3D. We compare our two models, MVTrack4Gen$_{\text{ReCamMaster}}$ and MVTrack4Gen$_{\text{Redirector}}$, against ReCamMaster (Bai et al., 2025a), Redirector (Park et al., 2025) and CogNVS (Chen et al., 2025a), GEN3C (Ren et al., 2025), TrajectoryCrafter (Yu et al., 2025), NeoVerse (Yang et al., 2026). Baselines show geometric distortions on dynamic foreground objects and floating or blurry background artifacts under large viewpoint changes, while MVTrack4Gen produces sharp, geometrically consistent novel-views faithful to the reference video.



Figure 14: **Additional qualitative comparisons on the DAVIS dataset (Perazzi et al., 2016).** The top Reference Video row shows reference-view frames at three timesteps, each row below renders the same target viewpoint, and the rightmost *Translate Up* column visualizes the target camera trajectory in 3D. We compare our two models, $\text{MVTrack4Gen}_{\text{ReCamMaster}}$ and $\text{MVTrack4Gen}_{\text{Redirector}}$, against ReCamMaster (Bai et al., 2025a), Redirector (Park et al., 2025) and CogNVS (Chen et al., 2025a), GEN3C (Ren et al., 2025), TrajectoryCrafter (Yu et al., 2025), NeoVerse (Yang et al., 2026). Baselines show geometric distortions on dynamic foreground objects and floating or blurry background artifacts under large viewpoint changes, while MVTrack4Gen produces sharp, geometrically consistent novel-views faithful to the reference video.

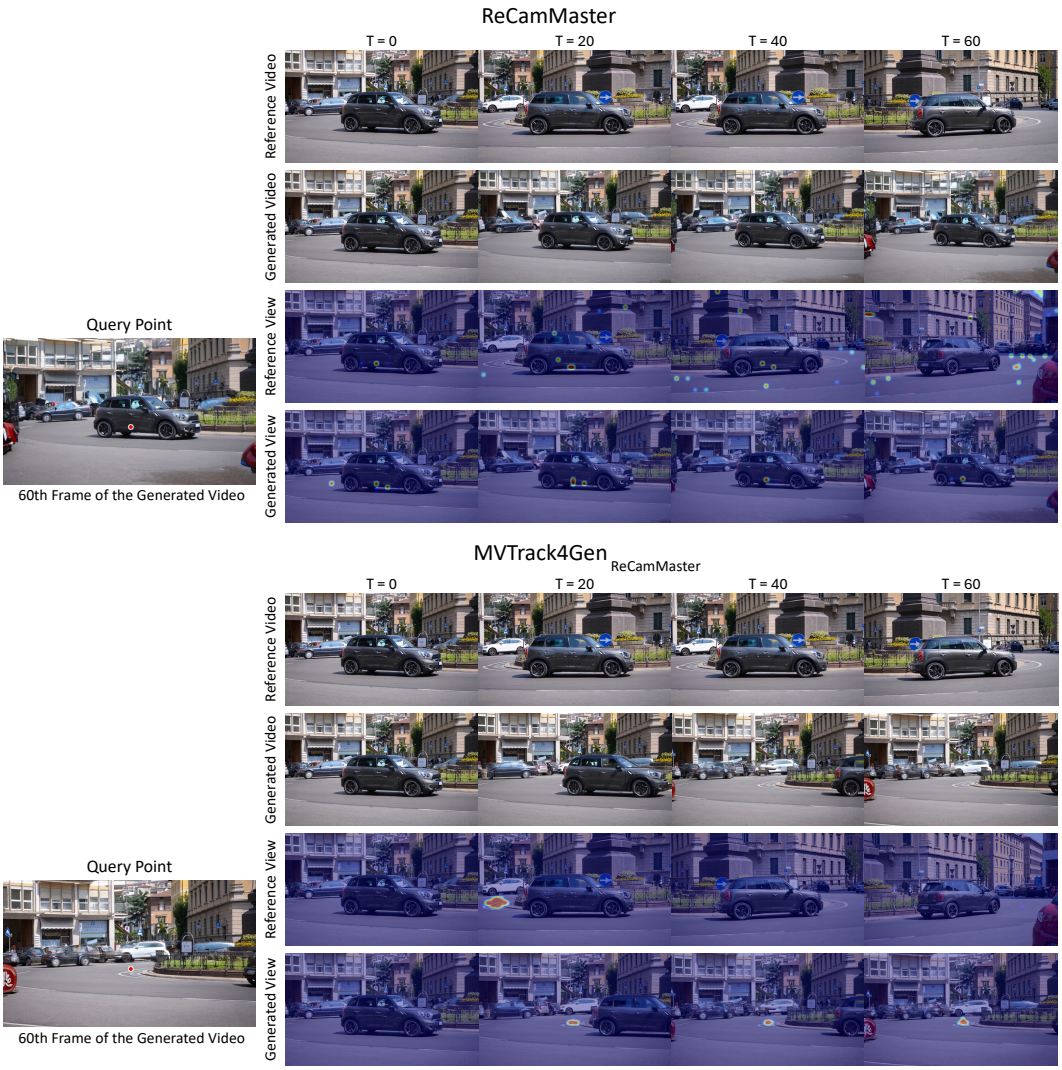


Figure 15: **Correlation between generation quality and attention-map alignment.** For both ReCamMaster (Bai et al., 2025a) (top) and our MVTrack4Gen (bottom), we show, from top to bottom, the reference video, the generated novel-view video, and the corresponding attention maps overlaid on the *reference view* and the *generated view* across frames. Given a query point placed on the 60th frame of the generated video (left), we visualize the attention distribution extracted from the 18th layer, identified as the correspondence-specialized layer in Sec. 3. For ReCamMaster, regions that are incorrectly synthesized in the generated video coincide with attention that is scattered and mislocalized in both the reference and generated views, revealing a breakdown of cross-view correspondence. In contrast, our method produces substantially higher-fidelity generations while keeping the attention concentrated and well-aligned with the true corresponding region of the query point.

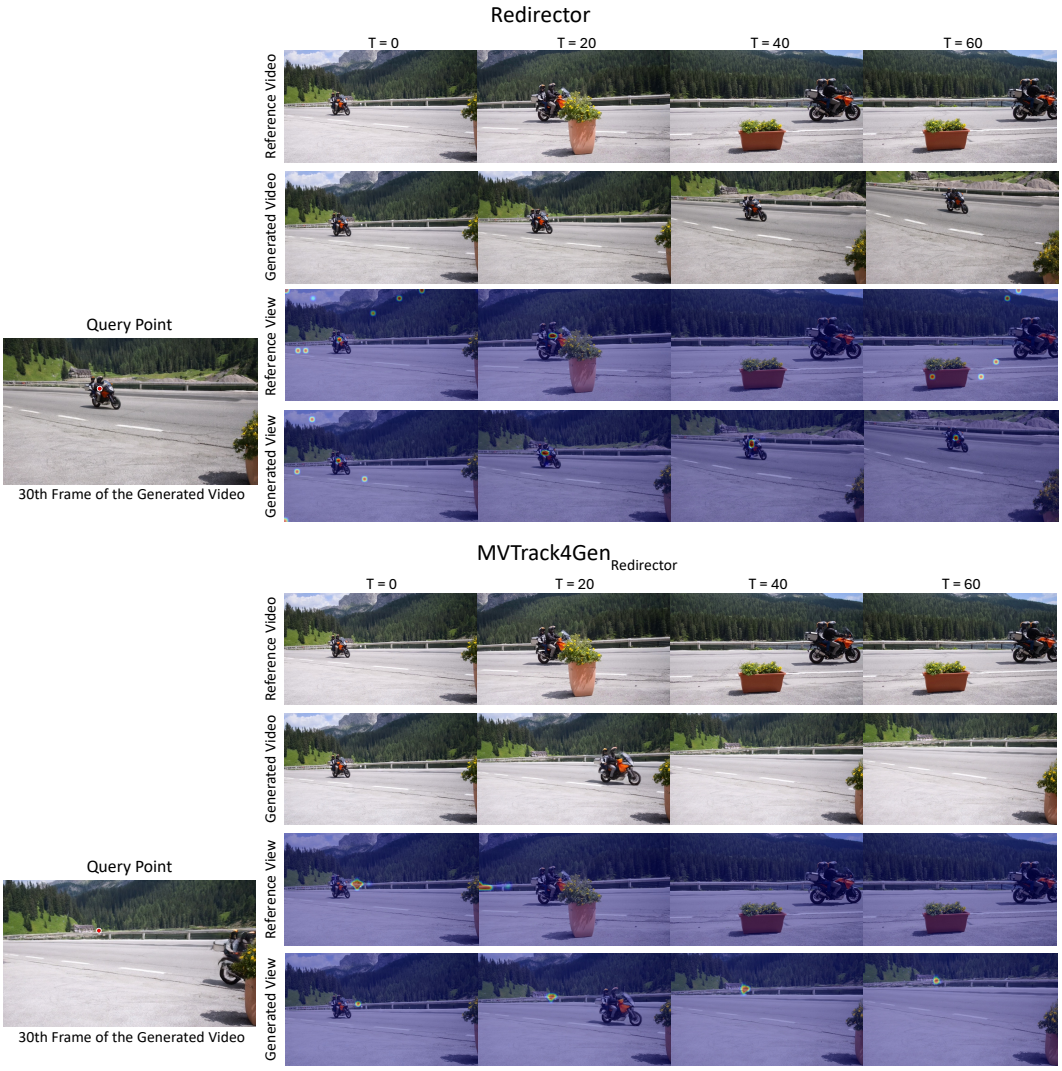


Figure 16: **Correlation between generation quality and attention-map alignment.** For both ReCamMaster (Bai et al., 2025a) (top) and our MVTrack4Gen (bottom), we show, from top to bottom, the reference video, the generated novel-view video, and the corresponding attention maps overlaid on the *reference view* and the *generated view* across frames. Given a query point placed on the 30th frame of the generated video (left), we visualize the attention distribution extracted from the 18th layer, identified as the correspondence-specialized layer in Sec. 3. For Redirector, regions that are incorrectly synthesized in the generated video coincide with attention that is scattered and mislocalized in both the reference and generated views, revealing a breakdown of cross-view correspondence. In contrast, our method produces substantially higher-fidelity generations while keeping the attention concentrated and well-aligned with the true corresponding region of the query point.