FREEVIEW: MULTI-VIEW DRIVING VIDEO GENERATION WITH SPATIOTEMPORAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust autonomous driving requires diverse multi-view observations of common traffic scenarios and long-tail events. Dashcam and online monocular driving videos provide a rich source of these scenarios. Their limited view coverage makes them difficult to use directly in autonomous driving systems that require synchronized multi-camera observations. Expanding these videos into coherent multi-view sequences requires preserving the reference scene across viewpoints and integrating observations from different times. We propose FreeView, a multi-view driving video generation framework with spatiotemporal memory. Our reference-as-view design incorporates the reference video as an additional conditioning view, using shared view-wise self-attention for memory encoding and cross-view attention for memory interaction. This design leverages the strong representation and interaction capabilities of the pretrained generation model. We further retrieve relevant reference observations across time and use relative temporal and spatial encodings to guide attention matching. Experiments demonstrate the effectiveness of our method in preserving visual fidelity and consistency with the reference scene. We also demonstrate generation under flexible camera configurations and applications to real-world dashcam videos.

1 INTRODUCTION

Robust autonomous driving systems require data covering diverse environments and long-tail events. Dashcam and online videos provide abundant real-world observations from a single viewpoint. Monocular video generation models (Gao et al., 2025b; Wan et al., 2025; Gao et al., 2024b) further expand the range of available scenarios. However, their limited view coverage cannot provide the synchronized multi-camera observations required by autonomous driving systems (Hu et al., 2023). Expanding them to specified camera layouts can support driving training and validation (Gao et al., 2024a; Liu et al., 2026; Wang et al., 2026a).

Generating multi-view videos requires preserving the reference scene while completing unobserved regions. Since relevant scene details may appear at different times in the reference video, the model needs to integrate these observations to preserve scene content. Beyond reference fidelity, the generated videos should remain consistent across views and over time. In particular, overlapping views should remain consistent with each other.

Geometry-based methods (Lin et al., 2026; Yang et al., 2026) project reference content into target views and use generative models to fill missing regions or refine the results. However, these methods rely on overlap between reference and target views. Large viewpoint changes in driving scenes often lead to limited overlap, making these methods prone to failure. AnyView (Van Hoorick et al., 2026) generates target videos directly from reference videos and camera conditions without explicit reconstruction. However, low view overlap makes reference details difficult to preserve, while generating views separately makes it harder to maintain cross-view consistency.

Concurrent works Sensor2Sensor (Wang et al., 2026a) and OpenLongTail (Liu et al., 2026) explore generating multi-view sensor observations from monocular videos. Sensor2Sensor generates multi-view images frame by frame. Extending this framewise approach to video is prone to generation drift. OpenLongTail uses pose conditioning and cross-view memory to transform a monocular video into another viewpoint. To generate multiple views, it uses each generated video to condition the next view. This cascade can cause generation drift and deviations from the reference scene.

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.



Figure 1: **Multi-view driving video generation with FreeView.** Given a monocular reference video (left), FreeView generates videos for a specified camera layout. **(a)** Surround-view generation from an internet dashcam video, with camera intrinsics and extrinsics estimated using MapAnything (Keetha et al., 2026). **(b)** Generation from a vehicle front-view video. **(c)** Cross-layout generation with a seven-camera target layout. **(d)** Zero-shot generation from Waymo (Sun et al., 2020) videos, with no Waymo data used during training. **(e)** Generation with a five-camera target layout not used during training. Real videos from the source camera layouts are shown for comparison.

A natural approach is to adopt a multi-view video generation architecture (Gao et al., 2025a; Wen et al., 2024; Ali et al., 2025; Ren et al., 2025a) that generates target views together rather than sequentially. This helps maintain cross-view consistency and reduce generation drift. The challenge is how to **represent reference observations** and **use them to guide generation**. Recent methods jointly process reference and noisy target content through self-attention (Bai et al., 2025; Yu et al., 2025; Wang et al., 2026a). These methods process only a small set of memory frames at each step, either by generating frames autoregressively or by retrieving relevant observations (Wang et al., 2026a; Yu et al., 2025; Shen et al., 2026). In contrast, our task uses a full reference video to generate multiple target videos jointly. In this setting, joint self-attention over the reference video and multiple target videos is computationally expensive. Beyond efficient reference representation, the model also needs to establish correspondences between reference observations and target views. In dynamic driving scenes, ego motion and moving objects can substantially change their spatial and temporal relationships, making these correspondences harder to establish.

To this end, we propose FreeView, a multi-view driving video generation framework with spatiotemporal memory that supports specified camera layouts. We adopt a **reference-as-view** design that incorporates the reference video as an additional conditioning view, using shared view-wise self-attention for memory encoding and cross-view attention for memory interaction. By integrating the reference view into the existing generation workflow, we leverage the strong representation and interaction capabilities of the pretrained model. Reference memory representations evolve across layers and guide target-view generation. To explicitly leverage observations beyond the synchronized

time step, we further retrieve relevant reference memory across time and use **relative temporal and spatial encodings** to guide attention matching. In addition, we encode camera poses to support different camera configurations. As shown in Figure 1, FreeView supports **different view counts and layouts** and expands real dashcam videos into multi-view sequences.

Our contributions are summarized as follows:

- We propose FreeView, a multi-view driving video generation framework that supports flexible camera counts and layouts through explicit camera pose conditioning.
- We develop a reference-as-view design that leverages the representation and interaction capabilities of a pretrained generation model for spatiotemporal memory modeling, complemented by cross-time retrieval and relative spatiotemporal encodings.
- We conduct extensive experiments on multiple datasets to demonstrate the effectiveness of FreeView for multi-view driving video generation.

2 Related Work

2.1 Novel View Video Generation

NeRF (Mildenhall et al., 2021) and 3D Gaussian Splatting (Kerbl et al., 2023) support driving scene reconstruction and novel view rendering through explicit scene representations (Yang et al., 2024; Chen et al., 2026; Yan et al., 2024; Chen et al., 2025). Generative methods combine geometric conditions with generation priors to complete unobserved content (Wu et al., 2025; Wang et al., 2025; Guo et al., 2025; Zhu et al., 2026; Chen & Peng, 2026; Ren et al., 2025b; Lin et al., 2026; Yang et al., 2026), or directly generate target videos from reference videos and camera poses (Bai et al., 2025; Van Hoorick et al., 2026). OpenLongTail (Liu et al., 2026) uses geometric and semantic memory to expand monocular driving videos into multi-view data. However, most general novel-view video generation methods produce a single target view and do not directly meet the multi-view requirements of autonomous driving.

2.2 Reference Memory in Video Generation

Some methods jointly process reference and target content through self-attention to implicitly establish their associations (Bai et al., 2025; Wang et al., 2026a; Ren et al., 2025a; Van Hoorick et al., 2026). ReCamMaster concatenates reference video features with noisy target features as input to the generative backbone. Other methods use additional reference branches to construct reference representations and guide generation through cross-attention (Zhou et al., 2026; Liu et al., 2026). Among them, OpenLongTail encodes conditioning views into dense geometric memory and compressed semantic memory. Beyond memory representation, prior work retrieves relevant observations to condition generation (Yu et al., 2025; Xiao et al., 2025; Shen et al., 2026). CaM selects historical frames based on camera field-of-view overlap to reduce the amount of reference content processed. In contrast, our method efficiently encodes reference memory and supplements it with retrieved observations. Relative spatiotemporal encodings guide its interaction with target views.

3 Method

3.1 Problem Definition

Given a monocular reference video and its camera intrinsics and extrinsics, FreeView generates synchronized multi-view videos for a specified camera layout. We denote the reference video by X^{ref} and the target videos by $\{X^{(v)}\}_{v=1}^{V}$, where V is the number of target views.

3.2 Multi-View Video Generation Architecture

To avoid the high computational cost of joint self-attention over multi-view videos (Ali et al., 2025), we adopt an internal pretrained multi-view video generation model. The model operates in the latent space of a video autoencoder. As shown on the left of Figure 2, each DiT block first builds per-view

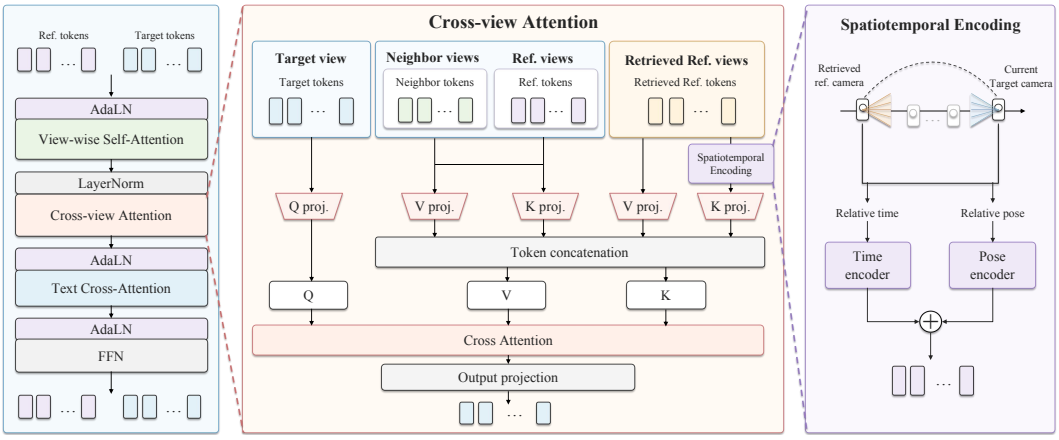

Figure 2: **Overview of FreeView. Left:** Reference and target tokens are processed separately through shared pretrained view-wise self-attention, updating reference memory tokens across DiT layers. **Middle:** Through cross-view attention, target view tokens jointly attend to neighboring target view tokens and reference tokens at the same latent time step, together with reference memory tokens retrieved from other times. All sources share Key and Value projections. **Right:** Relative temporal and spatial embeddings are added to retrieved reference memory tokens before Key projection, while the Value branch uses the original tokens.

representations through shared view-wise self-attention. For efficient interaction across views, these features then attend to neighboring views through cross-view attention. This design suits our need to encode reference videos and interact with multiple target views efficiently. Text cross-attention further provides semantic guidance during denoising. The target latents are decoded into videos.

For camera control, the backbone uses dynamic Plücker embeddings (He et al., 2024) to describe relative motion between adjacent latent time positions of each camera. We introduce additional static Plücker embeddings to describe each camera relative to a common reference camera. For a camera ray associated with a pixel, the Plücker coordinates are $\mathbf{p} = (\mathbf{o} \times \mathbf{d},\ \mathbf{d}) \in \mathbb{R}^6$, where $\mathbf{o}$ and $\mathbf{d}$ are the ray origin and unit direction in the corresponding coordinate system. Both are determined by camera intrinsics and extrinsics. Static and dynamic embeddings describe the camera layout and relative motion between adjacent times, respectively. Both are mapped to geometric features by a shared pose encoder and used to modulate video features through AdaLN (Peebles & Xie, 2023).

We train the model with a flow matching objective (Lipman et al., 2022). Given clean multi-view latents z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we construct the interpolation $z_\sigma = (1-\sigma)z_0 + \sigma\epsilon$ at noise level $\sigma \in [0, 1]$ and predict the velocity $u^* = \epsilon - z_0$. The training loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0,\epsilon,\sigma,c}\left[\frac{1}{|\Omega|}\sum_{i\in\Omega} w(\sigma_i)\,\|u_\theta(z_\sigma,\sigma,c)_i - (\epsilon - z_0)_i\|_2^2\right], \tag{1}$$

Here, c denotes the text and camera conditions together with available video conditions. The set Ω contains the target latent positions used for supervision. Here, σ_i is the noise level at position i and w weights the training loss.

3.3 Spatiotemporal Memory Modeling

Memory representation. Effective reference conditioning requires both informative memory representations and interactions that connect them to target generation. We adopt a reference-as-view design that incorporates the reference video as an additional conditioning view. Shared view-wise self-attention performs memory encoding, while cross-view attention enables memory interaction with target views. This design integrates memory processing into the existing generation workflow, leveraging the representation and interaction capabilities of the pretrained model without a separate reference encoder. The reference view shares the video VAE encoder $\mathcal{E}$ and patch embedding layer $\mathcal{P}$ with target views. Reference latents remain clean, while target latents are noised at level σ. The initial reference memory tokens and target view tokens are $m_0 = \mathcal{P}(\mathcal{E}(X^{\mathrm{ref}}))$ and $h_0^{(v)} = \mathcal{P}(z_\sigma^{(v)})$,

respectively. Reference memory tokens evolve across DiT blocks as $m_\ell = \mathcal{F}_\ell(m_{\ell-1}; c^{\text{ref}})$, where $\mathcal{F}_\ell$ shares parameters with the target branch and c^{ref} contains camera and text conditions. Each view processes its own tokens independently through view-wise self-attention. The reference branch does not attend to target view tokens to avoid contaminating reference representations with noisy generated content. At each latent time step, a target view jointly attends to the reference view and its neighbors in the multi-view adjacency graph at the same time step through cross-view attention. Retrieved reference memory tokens from other times also participate in this interaction.

Memory retrieval. As the ego vehicle moves forward, a scene observed by the front camera may enter the field of view of a back camera several seconds later. Inspired by CaM and Lyra 2.0 (Yu et al., 2025; Shen et al., 2026), we retrieve reference latent frames using the camera pose of the original video frame corresponding to each latent frame. The reference memory tokens corresponding to the selected latent frames are then used for cross-view interaction. To focus on nearby regions relevant to driving, we sample 3D points within the target camera frustum over a depth range of 1–30 m. We project these points into candidate reference images. We estimate view coverage from the projections that fall within each image. Since reference observations at the same time already participate in cross-view interaction, we follow Lyra 2.0 (Shen et al., 2026) in selecting the candidate with the largest additional coverage. This process only requires camera intrinsics and extrinsics, without depth estimation. All layers share the retrieval indices and read their corresponding reference memory tokens. Retrieval details are provided in Appendix A. We then encode their temporal and spatial relationships with the target query, as shown on the right of Figure 2.

Relative temporal encoding. Ego motion and moving objects cause differences between reference observations and the target scene over time. We encode the relative frame offset to account for these temporal differences during memory interaction. The offset is first mapped to a sinusoidal embedding (Vaswani et al., 2017) and then processed by a layer-specific two-layer MLP.

Relative spatial encoding. Although the pretrained cross-view attention has strong spatial matching capabilities, it mainly operates between neighboring cameras in the predefined adjacency graph. Retrieved observations may come from views not directly connected to the target view in this graph, such as front and back views. Ego motion further changes their relative positions and orientations across time. These large relative pose changes make direct matching more difficult. We therefore encode the relative camera pose to guide memory interaction. We construct Plücker embeddings on the reference pixel grid using the relative pose and reference camera intrinsics. These embeddings are processed by the pretrained pose encoder and a layer-specific output mapping to obtain spatial embeddings aligned with the reference memory tokens.

Memory interaction. We extend pretrained cross-view attention to jointly read neighboring target view tokens and reference memory tokens, as shown in the middle of Figure 2. Inspired by Lyra 2.0 (Shen et al., 2026), we add relative temporal and spatial embeddings to the retrieved reference memory tokens only before Key projection. These embeddings guide attention matching, while the Value branch uses the original memory tokens to preserve their content. All sources share Key and Value projections. Keys and Values are separately concatenated for joint cross-attention.

Let t and r denote the target and reference latent time indices. The video VAE downsamples time by a factor of four, so the relative offset in original video frames is $\Delta f = 4(t-r)$. We construct Plücker embeddings Φ for reference pixels using the reference camera intrinsics and express the rays in the target camera coordinate system using the relative pose. The temporal and spatial embeddings are $e_\ell^{\text{time}} = g_\ell(\gamma(\Delta f))$ and $e_\ell^{\text{geo}} = A_\ell(E_{\text{cam}}(\Phi))$. Here, γ denotes sinusoidal encoding and E_{cam} is the shared pretrained pose encoder. The two-layer MLP g_ℓ and output mapping A_ℓ are layer-specific.

Let m_{ret} denote the retrieved reference memory tokens, with layer and time indices omitted. Temporal embeddings are broadcast across spatial tokens, whereas spatial embeddings are aligned with the reference memory tokens at corresponding spatial positions. We add both embeddings before Key projection and use the original memory tokens for Value projection:

$$K_{\text{ret}} = W_K \left(m_{\text{ret}} + e^{\text{time}} + e^{\text{geo}}\right), \qquad V_{\text{ret}} = W_V m_{\text{ret}}. \tag{2}$$

Keys and Values from the three sources are concatenated separately along the token dimension:

$$\begin{aligned} K &= \text{Concat}(K_{\text{nbr}}, K_{\text{ref}}, K_{\text{ret}}), \\ V &= \text{Concat}(V_{\text{nbr}}, V_{\text{ref}}, V_{\text{ret}}). \end{aligned} \tag{3}$$

Table 1: **Quantitative comparison on nuScenes for 41-frame generation.** † Uses Depth Anything 3 (Lin et al., 2025). ‡ Uses SAM 3 (Carion et al., 2026). * Uses the default camera layout without target pose control and is included for reference only. Best results are bold and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
Street Gaussians	12.89	**0.517**	0.655	192.55	1816.41
Difix3D+	12.73	0.463	0.631	126.69	1067.84
GEN3C†	11.57	0.420	0.666	131.90	1552.88
NeoVerse†	9.07	0.252	0.709	121.48	1169.46
Vista4D†‡	12.13	0.396	0.632	65.01	557.79
AnyView	12.14	0.426	0.619	71.64	617.85
OpenLongTail†	12.45	0.441	0.685	146.15	1179.56
Ours	**13.03**	0.450	**0.596**	**50.05**	**372.63**
Cosmos-Drive-Dreams*	12.14	0.415	0.659	79.76	595.84

Table 2: **Quantitative comparison on the internal driving dataset for 41-frame generation.** The telephoto view is excluded. † Uses Depth Anything 3 (Lin et al., 2025). ‡ Uses SAM 3 (Carion et al., 2026). * Uses the default camera layout without target pose control and is included for reference only. Best results are bold and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
GEN3C†	13.32	0.444	0.712	198.90	2908.08
NeoVerse†	9.90	0.292	0.743	145.20	1364.02
Vista4D†‡	13.78	0.505	0.646	65.51	620.19
AnyView	13.73	0.509	0.640	85.42	978.36
OpenLongTail†	13.77	0.508	0.681	142.28	1082.86
Ours	**16.85**	**0.545**	**0.548**	**46.07**	**503.85**
Cosmos-Drive-Dreams*	13.26	0.474	0.665	88.13	914.83

Here, nbr, ref, and ret denote neighboring target view tokens, reference tokens at the target latent time step, and retrieved reference tokens, respectively. All three sources share W_K and W_V within each layer. Target tokens read these Keys and Values through pretrained cross-view attention.

3.4 Training and Inference

We initialize FreeView from the pretrained multi-view video generation model and jointly optimize the backbone and added modules with the flow matching objective. Reference memory tokens are learned through the generation loss on target views. We zero-initialize the output layers of the temporal encoders and spatial mappings so that the added embeddings do not affect the pretrained model at initialization. At inference, we compute target camera poses from the reference camera trajectory and the specified camera layout. We then precompute retrieval indices and jointly denoise the target-view latents to generate multi-view videos.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on nuScenes (Caesar et al., 2020) and an internal autonomous driving dataset. We use nuScenes videos at 12 Hz with six camera views: front, front left, front right, back, back left, and back right. Following prior driving scene generation works (Chen & Peng, 2026; Gao et al., 2025a), we use the official training split of 700 scenes. The internal dataset provides videos at 10 Hz with seven views: front left, front, front right, back right, back, back left, and a forward-facing telephoto view. We train on 300,000 video segments from this dataset.

Evaluation protocol. We select 20 dynamic video segments from non-training scenes in each dataset. We evaluate 41- and 61-frame generation on nuScenes and 41-frame generation on the internal dataset. All main evaluations use the front video as reference and generate at 480×832. The 20 reference clips each provide five target views, yielding 100 target videos per benchmark. We exclude the input front view and the telephoto view where applicable. We measure fidelity to ground truth using PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) assess image and video distribution quality. On the internal dataset, we use Pi3X (Wang et al., 2026b) to estimate generated and GT trajectories and report absolute trajectory error (ATE), rotation error, and translational and rotational relative pose error (RPE) after Sim(3) alignment. Cross-view consistency is measured by geometric matching success rate (GMSR@10): the percentage of frame and view-pair groups with at least 10 LoFTR (Sun et al., 2021) matches passing confidence and GT-based epipolar geometry checks (Appendix C).

Compared methods. We compare several reconstruction and generation approaches. Street Gaussians (Yan et al., 2024) optimizes a 4D Gaussian representation for each scene. We apply the Difix

Table 3: **Quantitative comparison on nuScenes for 61-frame generation.** † Uses Depth Anything 3 (Lin et al., 2025). ‡ Uses SAM 3 (Carion et al., 2026). Best results are bold and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
Street Gaussians	<u>12.97</u>	**0.517**	0.650	174.59	1982.17
Difix3D+	12.89	0.456	0.621	102.72	975.23
Vista4D†‡	12.26	0.418	0.625	<u>53.76</u>	490.96
AnyView	12.29	0.432	<u>0.610</u>	70.89	<u>465.67</u>
OpenLongTail†	12.10	0.410	0.714	195.91	1643.09
Ours	**13.52**	<u>0.463</u>	**0.590**	**44.51**	**321.97**

Table 4: **Ablation of individual module contributions. Ret.**, **Temp.**, and **Spat.** denote memory retrieval, relative temporal encoding, and relative spatial encoding, respectively. Checkmarks indicate enabled modules. Setting (a) disables all three modules.

Setting	Ret.	Temp.	Spat.	PSNR ↑	SSIM ↑	LPIPS ↓
(a)				12.69	0.340	0.594
(b)	✓			12.57	0.338	**0.589**
(c)	✓	✓		12.84	0.341	**0.589**
(d)	✓		✓	12.86	**0.344**	0.595
(e)	✓	✓	✓	**12.94**	0.337	0.592

neural enhancer from Difix3D+ (Wu et al., 2025) to Street Gaussians renderings. GEN3C (Ren et al., 2025b), NeoVerse (Yang et al., 2026), and Vista4D (Lin et al., 2026) combine feed-forward reconstruction with video generation to refine target-view renderings. Our evaluation pipelines use Depth Anything 3 (Lin et al., 2025) and SAM 3 (Carion et al., 2026) for geometry and segmentation where needed. AnyView (Van Hoorick et al., 2026) and OpenLongTail (Liu et al., 2026) perform pose-conditioned generation, with the latter using reference memory for cascaded view generation. Cosmos-Drive-Dreams (Ren et al., 2025a) jointly generates multiple views. We adapt official implementations as detailed in Appendix B.

Implementation details. Our pretrained backbone has approximately 2B parameters. Training uses 64 NVIDIA H20 GPUs. For nuScenes, we first train for 3 days at a resolution of 240×416 with 29-frame videos and a total batch size of 64. We then train for another 3 days at 480×832 with 61-frame videos and a total batch size of 16. On the internal dataset, we train at 480×832 with a total batch size of 16. We first use 29-frame videos for 7 days and then continue with 61-frame videos for 3 days. We set the learning rate to 1×10^{-4} for the new relative spatiotemporal encoding modules as well as the pose encoder and cross-view attention. The remaining parameters use 1×10^{-5}. FreeView uses 20 sampling steps for inference. Text and pose guidance scales are both 3. The sampling shift is 3.

4.2 Multi-View Video Generation Results

Quantitative comparison. Tables 1 and 2 report 41-frame generation results on nuScenes and the internal dataset. On nuScenes, FreeView achieves better reconstruction accuracy and perceptual quality than the compared methods, while Street Gaussians (Yan et al., 2024) retains higher SSIM. On the internal dataset, FreeView performs best across all reported metrics. These advantages extend to the 61-frame nuScenes evaluation in Table 3. The PSNR margin over Street Gaussians increases from 0.14 dB to 0.55 dB, with FreeView also achieving the lowest LPIPS, FID, and FVD. On the internal dataset, FreeView also yields the lowest trajectory errors and highest GMSR@10 (Table 5).

Qualitative comparison. Figures 3 and 4 compare generated views in a sunny nuScenes scene and a rainy scene from the internal dataset. In the nuScenes example, FreeView preserves recognizable vehicles and building details across target views. The highlighted car is visible in the earlier front reference and the later back-right view, illustrating the value of reference observations from different times. In the rainy example, FreeView retains the road layout and wet-road appearance more faithfully, while several baselines alter the surrounding scene or introduce visible artifacts.

4.3 Ablation Studies

We conduct ablation studies on four fixed nuScenes scenes using 41-frame videos at 240×416. All variants are trained with a total batch size of 64 and evaluated on target views excluding the input front camera. Memory representation variants are compared after 1,000 training steps and the remaining variants after 2,000 steps.

Table 5: **Camera motion and cross-view consistency on the internal driving dataset.** Translation uses GT-estimated trajectory units, rotation uses degrees, and GMSR@10 is a percentage. Best results are bold and second-best results are underlined.

Method	ATE ↓	Rot. ↓	RPE_t ↓	RPE_r ↓	GMSR@10 ↑
GEN3C	2.4256	34.6835	1.3073	6.7974	29.33
NeoVerse	1.9337	12.3748	0.9319	1.3461	40.73
Vista4D	<u>0.4970</u>	<u>7.3358</u>	<u>0.2391</u>	<u>0.6267</u>	<u>76.10</u>
AnyView	1.0533	8.0056	0.4435	1.2901	43.35
Ours	**0.3274**	**3.2418**	**0.1738**	**0.4469**	**84.73**

Table 6: **Ablation of spatial encoding. Independent** uses a separate pose encoder for retrieved reference tokens. **Shared** reuses the pretrained encoder. **None** disables relative spatial encoding but retains the original camera pose conditioning.

Pose encoder	PSNR ↑	SSIM ↑	LPIPS ↓
None	12.84	0.341	**0.589**
Independent	12.80	**0.344**	0.592
Shared	**12.94**	0.337	0.592



Figure 3: **Qualitative comparison on nuScenes.** The front-view column shows the input reference, while the other five columns compare generated target views with ground truth. The top inset shows an earlier reference frame at T_0, and the main grid shows frames at T_{24}. Colored boxes highlight differences in scene content.

Component ablation. Table 4 evaluates retrieval and relative spatiotemporal encodings. Retrieval alone improves LPIPS but slightly reduces PSNR and SSIM. Adding either temporal or spatial encoding improves PSNR over retrieval alone. Combining both achieves the highest PSNR. These results suggest that explicit spatiotemporal cues help the model use retrieved observations.

Memory representation. Table 7 compares fixed reference tokens and variants with independent reference self-attention or memory cross-attention. Our shared design achieves the highest PSNR and lowest LPIPS. It performs comparably to independent reference self-attention without introducing separate reference attention parameters. These results support reusing pretrained attention for memory representation and interaction.

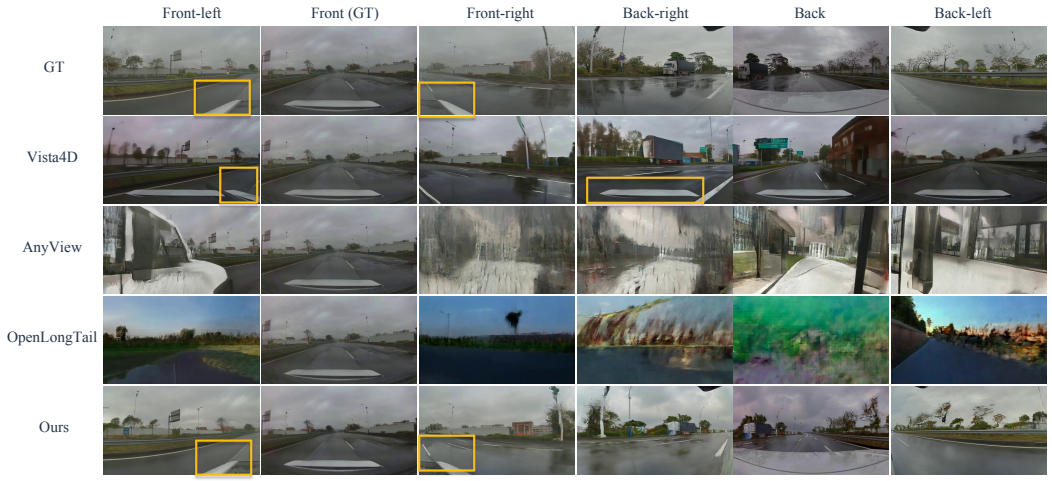

Figure 4: **Qualitative comparison on the internal driving dataset.** We show a rainy driving scene, with the input reference in the front-view column and comparisons between generated target views and ground truth in the other five columns. Colored boxes highlight artifacts and differences in scene content and wet-road appearance.

Table 7: **Ablation of memory representation and interaction. KV Projection** maps fixed reference tokens to Keys and Values through layer-specific projections. **w/o Shared SA** and **w/o Shared CA** use separate reference self-attention and memory cross-attention modules, respectively. **Ours** shares both with the target branch. All variants omit retrieval to isolate representation and interaction.

Reference representation	PSNR ↑	SSIM ↑	LPIPS ↓
KV Projection	11.59	0.285	0.663
w/o Shared SA	12.32	**0.334**	0.601
w/o Shared CA	11.96	0.317	0.618
Ours	**12.34**	0.332	**0.594**

Table 8: **Ablation of temporal encoding designs. Frame position** encodes clip frame indices before self-attention. **Relative time** encodes frame offsets between reference and target frames. **K** and **KV** add temporal embeddings before Key projection only or before both Key and Value projections, respectively. **Top-k** denotes the number of retrieved reference latent frames.

Temporal encoding	Top-k	PSNR ↑	SSIM ↑	LPIPS ↓
None	1	12.57	0.338	**0.589**
Frame position	1	12.72	0.339	0.592
Relative time (KV)	1	12.58	0.336	0.596
Relative time (K)	1	**12.84**	0.341	**0.589**
Relative time (K)	2	12.62	**0.344**	0.595

Temporal encoding. Table 8 compares frame-position encoding before self-attention with relative frame-offset encoding for retrieved references. Adding relative temporal embeddings only before Key projection outperforms both frame-position encoding and adding them before both Key and Value projections. Retrieving two reference latent frames slightly improves SSIM but reduces PSNR and worsens LPIPS. We thus use Key-only relative encoding and retrieve one latent frame.

Spatial encoding. Table 6 compares shared and independent pose encoders. The shared encoder achieves higher PSNR without introducing a separate pose encoder. We therefore reuse the pre-trained pose encoder with layer-specific output mappings.

5 Conclusion

We proposed FreeView, a multi-view driving video generation framework supporting flexible camera counts and layouts. Its reference-as-view design integrates memory encoding and interaction into the pretrained generation model. Cross-time retrieval and relative spatiotemporal encodings enable the model to directly incorporate asynchronous reference observations. Experiments on multiple datasets demonstrate improved scene fidelity and video quality.

AI USE STATEMENT

We used generative AI tools to assist with translation, manuscript editing, literature search and reference checking. The authors reviewed the AI-assisted content and take responsibility for the final manuscript, including its claims, citations, and interpretations.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Shijie Chen and Peixi Peng. Freegen: Feed-forward reconstruction–generation co-training for free-viewpoint driving scene synthesis. In *European Conference on Computer Vision*, pp. 682–699. Springer, 2026.

Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *International Journal of Computer Vision*, 134(3):83, 2026.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo De Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. In *International Conference on Learning Representations*, volume 2025, pp. 85508–85527, 2025.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, volume 2024, pp. 22841–22860, 2024a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28135–28144. IEEE, 2025a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025b.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27231–27241. IEEE, 2025.

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

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

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

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

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.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. Vista4d: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32671–32682, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Lulin Liu, Nuo Chen, Yan Wang, Bangya Liu, Wenyan Cong, Hezhen Hu, Boris Ivanovic, Hao Wang, Ziyao Zeng, Xinyu Gong, et al. Openlongtail: Generative scaling of long-tail driving data. *arXiv preprint arXiv:2607.09655*, 2026.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025b.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Bo Sun, Yijing Bai, Vincent Casser, Songyou Peng, Zehao Zhu, Meng-Li Shih, Xander Masotto, Shih-Yang Su, Kanaad Parvate, et al. Sensor2sensor: Cross-embodiment sensor conversion for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32093–32102, 2026a.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. In *International Conference on Learning Representations*, volume 2025, pp. 58640–58662, 2025.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

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

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26024–26035. IEEE, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49632–49652, 2025.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pp. 156–173. Springer, 2024.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In *International Conference on Learning Representations*, volume 2024, pp. 16739–16766, 2024.

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.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. Uniworld-view: Large-baseline view synthesis via video diffusion models. *arXiv preprint arXiv:2608.04701*, 2026.

Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 68258–68277, 2026.

A Memory Retrieval Details

For target view v at latent time t, we sample 4,096 points uniformly by volume within its camera frustum over an axial depth range of 1–30 m. We project these points into the reference image at the target time and each candidate reference image using camera intrinsics and extrinsics, without scene depth estimation.

Let $\mathcal{S}_{v,t}$ denote the sampled points. We set $a_r(x) = 1$ when point x has positive depth in the reference camera at time r and projects inside its image, and zero otherwise. The additional coverage is

$$s_{v,t}(r) = \frac{1}{|\mathcal{S}_{v,t}|} \sum_{x \in \mathcal{S}_{v,t}} a_r(x)\big(1 - a_t(x)\big), \qquad r^* = \arg\max_{r \in \mathcal{R}_t} s_{v,t}(r), \tag{4}$$

where $\mathcal{R}_t$ is the set of available historical reference candidates within the current clip and $a_t(x)$ denotes coverage by the reference camera at the target time. We retrieve at most one reference latent frame with an additional coverage score of at least 1%. If no candidate qualifies, we retain the original interaction.

The score estimates geometric field-of-view coverage without explicitly accounting for surface occlusion or moving objects. All DiT layers use the same time index and read reference memory tokens from their corresponding layers.

B Experimental Details

Baseline implementations. In our evaluation, GEN3C (Ren et al., 2025b), NeoVerse (Yang et al., 2026), Vista4D (Lin et al., 2026), and OpenLongTail (Liu et al., 2026) use Depth Anything 3 (Lin et al., 2025) for depth estimation. For NeoVerse, we cascade along two paths: front $\rightarrow$ front left $\rightarrow$ back left $\rightarrow$ back, and front $\rightarrow$ front right $\rightarrow$ back right. In our experiments, directly transitioning between widely separated views without intermediate cameras often causes generation failures. We therefore insert virtual intermediate cameras to limit each rotation step to at most 25°. At every step, including intermediate cameras, we generate a video and re-estimate its depth with Depth Anything 3 before projecting to the next camera. Vista4D additionally uses SAM 3 (Carion et al., 2026) for segmentation. For Street Gaussians (Yan et al., 2024), we follow the settings in the official repository. The baseline labeled Difix3D+ (Wu et al., 2025) applies the Difix neural enhancer to Street Gaussians renderings. The evaluated Cosmos-Drive-Dreams (Ren et al., 2025a) model does not support target camera pose control and uses its default camera layout.

Method comparison. Table 9 compares the video generation pipelines in our evaluation. The auxiliary estimators are Depth Anything 3 (Lin et al., 2025) and SAM 3 (Carion et al., 2026). FreeView jointly generates target views with camera pose control, without auxiliary depth or segmentation estimation. Street Gaussians and its Difix3D+ enhancement are reconstruction and rendering pipelines and are described separately above.

Table 9: **Comparison of video generation pipelines used in our evaluation.** Single-view methods generate one target view at a time, while multi-view methods jointly generate target views. Pose control refers to specifying target camera poses. Auxiliary information denotes estimated depth or segmentation used during generation, excluding evaluation tools. Depth is estimated using Depth Anything 3 (Lin et al., 2025), and segmentation using SAM 3 (Carion et al., 2026).

Method	Generation	Pose control	Auxiliary information
GEN3C	Single-view	Yes	Depth
NeoVerse	Single-view	Yes	Depth
Vista4D	Single-view	Yes	Depth + segmentation
AnyView	Single-view	Yes	None
OpenLongTail	Single-view	Yes	Depth
Cosmos-Drive-Dreams	Multi-view	No	None
Ours	Multi-view	Yes	None

Evaluation scenes. The nuScenes evaluation uses the following 20 scenes: scene-0012, scene-0013, scene-0015, scene-0038, scene-0093, scene-0102, scene-0105, scene-0106, scene-0108, scene-0109, scene-0110, scene-0268, scene-0329, scene-0520, scene-0522, scene-0524, scene-0552, scene-0559, scene-0560, and scene-0562. We use the first 41 or 61 frames of each scene for the corresponding benchmark, with the front-view video as reference. FID and FVD pool features from all target cameras rather than averaging individual camera scores.

Ablation settings. The ablation subset consists of the first clip of scene-0012, scene-0013, scene-0015, and scene-0038. Each sample uses a full front-view reference without target first-frame conditioning.

C Camera Motion and Cross-View Evaluation

Camera motion. We evaluate 100 generated target videos from 20 internal driving scenes, with 41 frames and five target views per scene, excluding the input front view and the telephoto view. We use Pi3X (Wang et al., 2026b) to estimate trajectories independently from generated videos and their corresponding GT videos. Each generated trajectory is aligned to the GT-estimated trajectory using Sim(3). We report absolute trajectory error (ATE), rotation error, and relative pose error (RPE) in translation and rotation, all as root mean squared errors. Rotation trajectories are expressed relative to their respective first frames, and RPE uses consecutive frames. Translation errors are measured in the coordinate units of the GT-estimated trajectories, not meters. This evaluation measures estimated camera motion consistency rather than absolute scale accuracy or consistency of the multi-camera rig.

Cross-view matching. At each time step, we apply LoFTR (Sun et al., 2021) to full images without ROI cropping for four view pairs: front and front left, front and front right, back right and back, and back and back left. The first two pairs combine the real input front view with a generated view; the remaining two compare generated views. We retain matches with confidence greater than 0.6 and Sampson distance below 10 pixels, using fundamental matrices computed from GT camera intrinsics and extrinsics. Let n_i denote the number of retained matches for group i, where a group is one view pair at one time step. We define

$$\mathrm{GMSR@}k = \frac{100\%}{N} \sum_{i=1}^{N} \mathbf{1}[n_i \geq k], \qquad N = 20 \times 41 \times 4 = 3280. \tag{5}$$

Groups with no matches or fewer than k retained matches count as failures and remain in the denominator. Thus, GMSR measures the fraction of successful groups, not the fraction of correct matches. Table 10 reports results for $k \in \{5, 10, 15, 20\}$. FreeView achieves the highest success rate among the evaluated methods at every threshold.

Table 10: **Geometric matching success rates at different match-count thresholds.** All entries are percentages. The full-image matching protocol is fixed; only the required number of geometrically valid matches changes.

Method	GMSR@5 ↑	GMSR@10 ↑	GMSR@15 ↑	GMSR@20 ↑
GEN3C	36.92	29.33	25.43	22.71
NeoVerse	46.86	40.73	36.65	33.08
Vista4D	80.76	76.10	72.93	70.55
AnyView	50.37	43.35	38.96	35.49
Ours	**91.28**	**84.73**	**78.75**	**72.59**