

GeoFlow: Efficient Driving Video Generation via Geometry-Aligned Priors

Jiazheng Liu¹, Hang Li¹, Jiawei Zhang¹, Jiahe Li¹, Xiaohan Yu⁵, Shengyin Fan⁶, Jin Zheng^{1,2,4*}, and Xiao Bai^{1,3,4*}

¹ School of Computer Science and Engineering, Beihang University

² State Key Laboratory of Virtual Reality Technology and Systems, Beijing

³ State Key Laboratory of Complex & Critical Software Environment, Beijing

⁴ Jiangxi Research Institute, Beihang University

⁵ Macquarie University

⁶ Tianyi Transportation Technology Co., Ltd, Suzhou 215133, China

Abstract. Generative models like Diffusion Models and Flow Matching have demonstrated remarkable capabilities in synthesizing high-fidelity driving videos, but are severely constrained by high inference latency due to the requirement of extensive sampling steps. We argue that this inefficiency stems from the prevailing reliance on a standard Gaussian source distribution, where consecutive frames are initialized as independent Gaussian noise. This paradigm disregards the rich spatiotemporal correlations inherent in driving videos, compelling the model to regenerate deterministic scene structures existing in previous frames from noise, which is both computationally redundant and prone to geometric inconsistency. To address this problem, we propose GeoFlow, a novel framework designed to achieve efficient driving video generation by harnessing explicit geometric priors. Instead of sampling from standard Gaussian noise, we leverage multi-view geometry and spatially-adaptive noise injection to construct a Geometry-Aligned Prior (GAP) distribution as starting point. This initialization bridges the gap between source distribution and data distribution, yielding a significantly straighter and shorter sampling trajectory. Extensive experiments demonstrate that GeoFlow can achieve remarkable efficiency of both training and inference: merely several hours of fine-tuning on baseline models can significantly boost few-step generation quality, while fully converged training drastically reduces number of inference steps required for state-of-the-art video generation.

Keywords: Autonomous Driving · Efficient Video Generation · Flow Matching

* Corresponding authors: Jin Zheng (jinzheng@buaa.edu.cn) and Xiao Bai (baixiao@buaa.edu.cn).

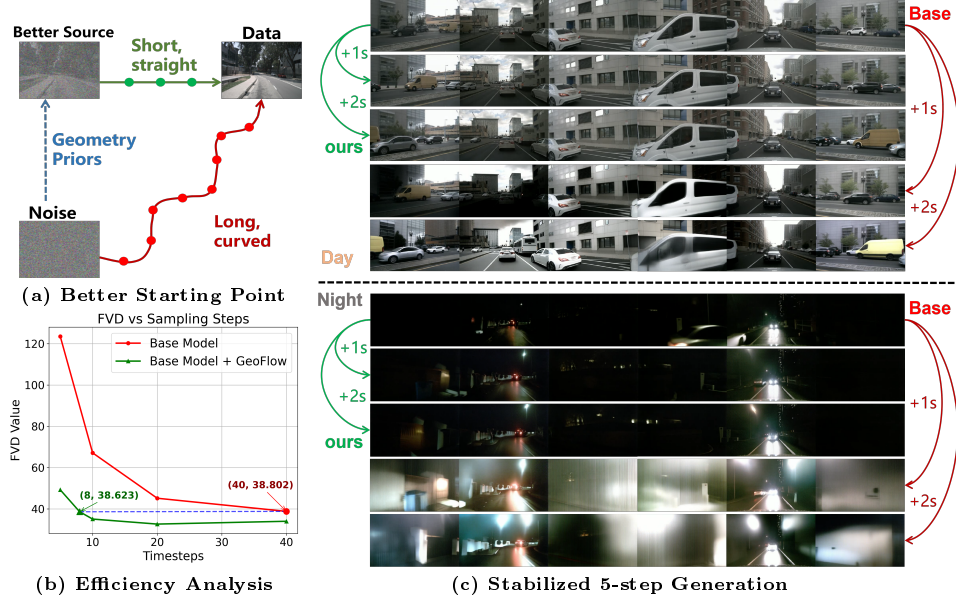


Fig. 1: Efficient driving video generation with structure-aware priors. (a) Instead of sampling from noise, we construct a structure-aware prior distribution as a better starting point. (b) This strategy leads to rapid convergence, achieving SOTA quality in just 8 steps (**Green line**) compared to the baseline (**Red line**). (c) Visual comparisons in 5 steps. Our method (**Green**) significantly improves structural stability and temporal consistency in few-step driving video generation.

1 Introduction

The rapid evolution of autonomous driving (AD) systems is increasingly fueled by the availability of massive, diverse, and high-quality driving data. However, acquiring such data from the real world is limited by high costs and safety risks, particularly for scarce corner cases such as extreme weather conditions and hazardous traffic scenarios. To overcome these limitations, generative models [6, 8–10, 43] have emerged as a promising paradigm for synthesizing high-fidelity driving data, serving as a scalable engine for both data augmentation and closed-loop simulation.

Among existing approaches, diffusion probabilistic models and flow matching frameworks have recently established themselves as the state-of-the-art. These generative models offer superior visual fidelity, diversity, and precise controllability over 3D geometry and scene layout. Representative works, such as MagicDrive [8, 9] and DriveDreamer [43, 58], have demonstrated remarkable capabilities in generating photorealistic, multi-view driving videos conditioned on rigorous control signals. Beyond achieving controllable data generation, these pioneering works demonstrate a potential for establishing generative closed-loop simula-

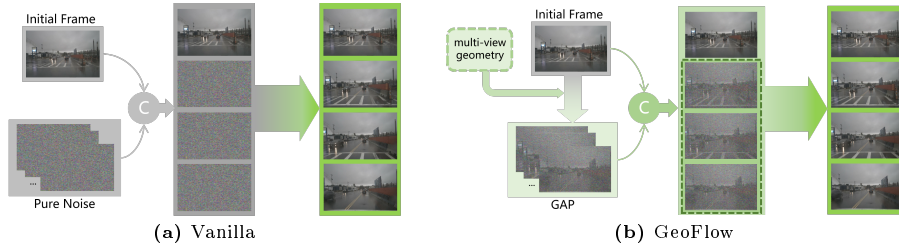


Fig. 2: Comparison between Vanilla Driving Video Generation and **GeoFlow**. The former uses standard Gaussian as source distribution, while GeoFlow leverages multi-view geometry to build an informative source distribution in latent space.

tion environments [45, 47], serving as interactive platforms for the closed-loop evaluation and training of autonomous driving algorithms.

However, this high fidelity comes at a steep computational cost. These methods typically rely on iteratively solving Ordinary Differential Equations [24, 26, 37] (ODEs) or Stochastic Differential Equations [15] (SDEs) to map a standard Gaussian distribution to the complex video data distribution (Red curve in Fig. 1a). This process requires extensive sampling steps (often dozens or hundreds) during inference, resulting in significant latency and computational overhead. This inefficiency significantly inflates the cost of data production and severely compromising the interaction frequency of generative simulation pipelines.

We identify that a fundamental source of this inefficiency lies in the ubiquitous assumption that the source distribution is typically *standard Gaussian noise* (Fig. 2a). Under this assumption, the model is forced to reconstruct the entire scene, including road, buildings, and vegetation, from pure noise for every single frame. This formulation ignores the strong spatiotemporal correlations between frames in driving videos, consequently, not only leading to significant computational redundancy (generating existed contents in history frames) but also increasing the difficulty of maintaining temporal consistency, as initiating consecutive frames from independent noise often results in texture flickering and geometric drift (Base in Fig. 1c), especially in few-step generation.

In the context of autonomous driving, we argue that the source distribution assumption should be revisited. Unlike open-domain videos, driving scenarios exhibit remarkably strong *spatiotemporal coherence* and *predictability*. The visual evolution of the scene is strictly governed by multi-view geometry and the ego-vehicle’s motion: static background elements shift continuously according to the camera extrinsics, while dynamic agents follow trajectories in given conditions. This implies that given a reference frame and future condition signals, the content of the subsequent frames is largely predictable rather than random. Motivated by this insight, we propose a novel perspective: *Why not start the generation from a coarse prediction of the future, rather than from pure noise?*

A straightforward solution is to initialize the generation directly from the given reference frame or history frame [25]. However, this cannot explicitly account for the multi-frame dynamics induced by ego-motion and dynamic agents, making it difficult to model highly dynamic driving scenes. We propose to instead leverage metric depth estimation [18, 23, 46] to warp the reference frame to future poses, producing a geometry-aligned coarse prediction. Nevertheless, the warped result alone is still insufficient to serve as a good source distribution: due to depth estimation inaccuracies and the dynamic nature of the environment, it inevitably contains artifacts. To address this, we propose a **Spatially-Adaptive Noise Injection** strategy to construct a robust **Geometry-Aligned Prior Distribution**, where we derive a continuous pixel-wise reliability mask based on geometric consistency and semantic priors to handle artifacts without extra learnable parameters. As shown in Fig. 1, this strategy successfully shortens the transport path from the source distribution and the target distribution, enabling high quality driving video generation within several sampling steps.

Our contributions are summarized as follows:

- We propose **GeoFlow**, an efficient driving video generation framework that generates future frames starting from reference frame information rather than pure noise, fundamentally straightening and shortening the transport path.
- We construct a **Geometry-Aligned Prior** distribution that leverages the inherent geometric consistency of driving videos, where **Spatially-Adaptive Noise Injection** is integrated to effectively mitigate artifacts.
- Extensive experiments demonstrate that GeoFlow can significantly improve the few-step generation quality at a low adapting cost.

2 Related Work

2.1 Video Generation for Autonomous Driving

Data synthesis and closed-loop simulation are of great importance for autonomous driving systems, while traditional reconstruction methods [7, 12, 22, 55] have limitations in fidelity and diversity. Generative models [11, 15, 24, 26] thus have become indispensable to the autonomous driving lifecycle. DriveDreamer [43] pioneered the incorporation of structural constraints, achieving controllable video synthesis conditioned on Bird’s-Eye-View (BEV) maps and 3D bounding boxes. MagicDrive [9] further extended these capabilities to ensure strict multi-view geometric consistency. Subsequent research [8] has continued to push the boundaries of controllability, spatial resolution, and temporal duration, leading to significant leaps in the fidelity and realism of synthesized driving scenarios. Fueled by this progress, the scope of video generation has expanded beyond dataset creation to more tasks such as neural scene reconstruction [32, 58, 59] and closed-loop simulation [45, 47]. However, despite these successes, a fundamental challenge persists: these methods inherently require a large number of denoising steps (sampling

iterations) to obtain high-quality results. This requirement creates a conflict between inference efficiency and generation quality, thereby hindering the efficient deployment of applications like generative closed-loop simulation.

2.2 Efficient Video Generation

To mitigate the high inference latency of diffusion models, existing acceleration strategies primarily focus on designing advanced samplers [27, 37, 61] to optimize the numerical solution of ODE, or employing model distillation [26, 36, 38, 48] to compress the multi-step generative trajectory, including recent advances in the video domain [17, 29, 42, 49, 56]. Despite enabling rapid synthesis, distillation-based methods typically incur expensive training costs. And these approaches retain the standard Gaussian source assumption, essentially ignoring the strong spatiotemporal and geometric correlations inherent in driving scenarios. In contrast, GeoFlow tackles efficiency from the perspective of source distribution construction. By establishing a geometry-aligned prior, our method naturally shortens and straightens the generative trajectory, which is orthogonal to existing frameworks.

2.3 Beyond Standard Gaussian Source Distribution

To break free from the standard Gaussian source assumption for specific tasks, Diffusion Bridges [62] extend the standard diffusion framework to learn the transport between two arbitrary probability distributions. These methods have achieved superior performance in specific data-to-data translation tasks, such as image super-resolution [52], in-painting [13, 33], and restoration [40, 51]. Similarly, Flow Matching (FM) [24, 26] is theoretically grounded in regressing vector fields defined by the interpolation between sample pairs, naturally supporting mappings between any two arbitrary distributions. While FM has been successfully applied to data-to-data translation tasks [5, 28, 30, 34, 39, 53], mainstream research in video generation predominantly retains the standard Gaussian source assumption to ensure training stability and generation diversity.

In the context of video synthesis, Video Bi-flow [25] first investigates this flexibility by using the previous frame with noise injection as the starting point of flow matching. Their experiments demonstrate that evolving the flow from the previous frame yields a shorter transport distance and higher sampling efficiency compared to standard methods that treat the previous frame merely as a condition. GeoFlow shares a similar insight of starting from a better source distribution, yet it fundamentally diverges in how this distribution is constructed. We explicitly model the transition between frames using physical priors to construct the source distribution, which makes it geometrically aligned with the target manifold.

2.4 Geometry-Constrained Video Generation

Introducing geometric priors into video generation tasks has recently been an active area of research. Existing works [3, 14, 16, 35, 50, 57] typically integrate

3D geometric information [19–21, 41, 54] as conditioning signals to guide the generation process, achieving precise camera control and enhancing the temporal coherence of synthesized videos. GeoFlow fundamentally differs from these works in both rationale and methodology. The primary objective of these prior works is to enforce controllability and consistency through explicit 3D conditioning, meaning the generative model still traverses the complete trajectory from pure Gaussian noise to the target data manifold during inference. In contrast, GeoFlow is designed from the perspective of generation efficiency. Rather than conditioning, we construct a geometrically aligned and much closer source distribution, which significantly shortens and straightens the transport path, thus improving sampling efficiency.

3 Method

3.1 Problem Formulation

Task Definition. Given a past video sequence (or a single reference frame $\mathbf{I}_{ref}$), the goal is to generate the future video chunk $\mathbf{V} \in \mathbb{R}^{L \times N \times H \times W}$ that strictly follows a set of control signals $\mathcal{C}$. We formulate this as learning a conditional distribution $p(\mathbf{V}|\mathbf{I}_{ref}, \mathcal{C})$.

Control Conditions. Following standard practices in driving world modeling [9, 43], the condition set $\mathcal{C}$ encapsulates the geometric and semantic layout of the driving scene: it includes (i) **ego-motion and camera parameters** $\mathcal{C}_{cam}$, i.e., the camera extrinsics $\mathbf{T} \in SE(3)$ and intrinsics $\mathbf{K}$ for each frame, defining the 3D-to-2D projection rules; (ii) the **3D object layout** $\mathcal{C}_{obj}$, represented as a set of 3D bounding boxes $\{(b_k, c_k)\}$ that describe the position and category of dynamic agents (e.g., vehicles, pedestrians) in the future scene; and (iii) a **road map** $\mathcal{C}_{map}$, such as an HD-Map or BEV layout providing road boundaries and lane information.

Flow Matching Backbone. We employ a latent flow matching model parameterized to estimate the vector field v_t towards the clean data x_1 (latent representation of V). The generative process is governed by the ODE: $dx_t/dt = v_t(x_t, t, \mathcal{C})$. The model $F_\theta(x_t, t, \mathcal{C})$ is trained to minimize the flow matching objective:

$$\mathcal{L} = \mathbb{E}_{t, x_1, x_0, \mathcal{C}} [||F_\theta(x_t, t, \mathcal{C}) - v_t||^2], \quad (1)$$

where $x_t = (1 - t)x_0 + tx_1$ and $v_t = x_1 - x_0$.

3.2 GeoFlow Framework

Standard driving video generation models typically initialize from pure Gaussian noise ($x_0 \sim \mathcal{N}(0, \mathbf{I})$), forcing the model to construct complex driving scenes entirely from randomness. This leads to highly non-linear generation trajectories and requires numerous integration steps to solve, which fundamentally limits the generation efficiency. We address this problem by introducing the **GeoFlow** framework. Its core principle is to replace the standard Gaussian noise with a

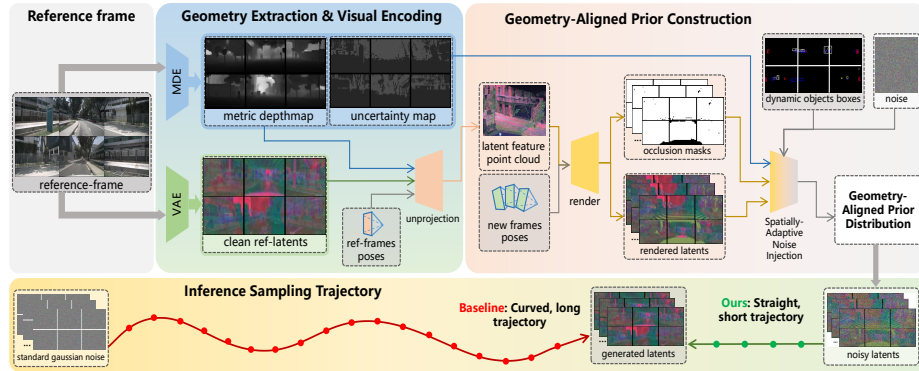


Fig. 3: Schematic of the proposed GeoFlow framework. To bridge the gap between the source and target manifolds, we leverage multi-view geometry to construct a **Geometry-Aligned Prior Distribution**, where an Spatially-Adaptive Noise Injection strategy is introduced to reduce error accumulation due to depth and rendering artifacts. **Bottom Panel:** As visualized in the sampling trajectory comparison, our initialization significantly *shortens and straightens* the flow, thereby enabling high-fidelity synthesis within several inference steps.

Geometry-Aligned Prior Distribution. This makes the source distribution geometrically aligned with the target, thus significantly shortening and straightening the generation path, enabling efficient high-fidelity driving video synthesis with substantially fewer sampling steps (Fig. 3).

The following sections detail the construction process of this Geometry-Aligned Prior Distribution, which is primarily divided into two subsections: Latent Geometry Extraction (Section 3.3) and Geometry-Aligned Prior Construction (Section 3.4). Finally, we introduce the training and inference pipelines of GeoFlow (Section 3.5).

3.3 Latent Geometry Extraction

Driving scenarios are characterized by strong spatiotemporal coherence. Unlike general open-domain videos, the visual evolution of a driving scene is heavily correlated with the ego-vehicle’s control signals. Specifically, the static scene structures (e.g., roads, buildings, vegetation), which typically occupy the majority of the field of view, shift continuously in accordance with the camera’s ego-motion. This observation implies that a significant portion of the future frame content is not random but can be explicitly derived from historical frames and given conditions.

Motivated by this insight, we propose **explicitly** modeling the geometric relationships between sequential frames to reduce the computational redundancy in video generation. Specifically, we leverage the known future ego-motion to warp the reference frame to expected future locations. This allows the model to

start directly from a **coarse prediction** based on historical context, eliminating the need to redundantly regenerate existing content.

Instead of operating in RGB space, we choose the latent space of the VAE for efficiency. First, we estimate the metric depth $\mathbf{D}_{ref}$ of the reference frame, which also produces the corresponding uncertainty map $\mathbf{M}_{unc}$ to describe the accuracy of depth estimation. And we encode reference frame to get latent representation $\mathbf{Z}_{ref} = \mathcal{E}(\mathbf{I}_{ref})$, then unproject the latent features $\mathbf{Z}_{ref}$ into a 3D feature point cloud $\mathcal{P}_{ref}$ using estimated metric depth and camera parameters, as shown in Fig. 3. Given the relative pose transformation $\mathbf{T}_{rel}$ derived from the ego-vehicle’s control commands or trajectory planning, we transform the point cloud to the target coordinate system:

$$\mathcal{P}_{target} = \mathbf{T}_{rel} \cdot \mathcal{P}_{ref}. \quad (2)$$

We then render $\mathcal{P}_{target}$ using splatting algorithm to obtain the warped latent map $\mathbf{Z}_{warp}$. This process creates a geometrically aligned canvas for the future frame. Detailed computational procedures are provided in the supplementary materials.

Feature Splatting Mechanism. Projecting 3D points to the 2D image plane may result in collisions (multiple points mapping to the same pixel). To address collisions and keep the input shape aligned with base model, we adopt a Z-buffer based feature splatting strategy.

Let $\mathcal{P}_{target} = \{(\mathbf{p}_k, \mathbf{f}_k)\}$ denote the set of transformed 3D points $\mathbf{p}_k$ and their associated feature vectors $\mathbf{f}_k$. For a target pixel coordinate $\mathbf{u} = (u, v)$, we identify the subset of points $\mathcal{N}(\mathbf{u})$ that project onto this pixel neighborhood. The value of the warped latent map $\mathbf{Z}_{warp}(\mathbf{u})$ is determined by the point closest to the camera:

$$k^* = \arg \min_{k \in \mathcal{N}(\mathbf{u})} d_z(\mathbf{p}_k), \quad (3)$$

$$\mathbf{Z}_{warp}(\mathbf{u}) = \mathbf{f}_{k^*}, \quad (4)$$

where $d_z(\cdot)$ denotes the depth value in the camera coordinate system.

3.4 Geometry-Aligned Prior Construction

Directly utilizing the deterministic warped latents $\mathbf{Z}_{warp}$ as the source distribution $\hat{x}_0$ is suboptimal for two main reasons. First, Flow Matching models fundamentally require a stochastic source distribution to ensure training stability and generation diversity. Initializing from a deterministic point restricts the model’s ability to explore the data manifold, potentially leading to severe error accumulation in inference [25]. Second, the geometric prior is inherently imperfect. Due to inevitable inaccuracies in metric depth estimation and rendering artifacts, $\mathbf{Z}_{warp}$ contains structural distortions. Strictly following these flawed features would force the model to hallucinate unrealistic textures to match the geometric errors. Therefore, it is crucial to introduce stochastic perturbations to mitigate artifacts and encourage the generative model to naturally correct and inpaint the warped $\mathbf{Z}_{warp}$.

Limitations of Global Noise. A straightforward solution to introduce stochasticity and mitigate artifacts is to inject global uniform noise (i.e., $\hat{x}_0 = (1 - \alpha)\mathbf{Z}_{warp} + \alpha\epsilon$). However, our experiments suggest that this is a compromise: while it aids convergence in few-step inference, it degrades performance at higher sampling steps compared to the baseline. This is because uniform noise indiscriminately corrupts the artifacts and the high-quality regions of the warped prior, wasting the valuable geometric information we extracted.

Spatially-Adaptive Noise Injection. To resolve this dilemma, we propose a Spatially-Adaptive Noise Injection strategy. Instead of treating all regions equally, we construct a continuous noise injection mask $\mathbf{M}$ to modulate the noise intensity locally. Specifically, low values in the mask identify reliable regions where the extracted geometric prior should be strictly preserved. Conversely, high values signify unreliable areas, such as occlusions or dynamic objects, instructing the model to discard the flawed prior and rely predominantly on standard Gaussian noise for those specific regions.

The mask aggregates three sources of uncertainty, each corresponding to a specific geometric error type and thus providing a controllable fallback mechanism. The first source is geometric invalidity $\mathbf{M}_{occ}$, derived from the Z-buffer during rendering, where regions that are occluded or out-of-view in the reference frame are assigned high noise intensity. The second source $\mathbf{M}_{dyn}$ addresses dynamic objects, as static geometry warping cannot predict object motion; we utilize projected 3D bounding boxes to mask out regions corresponding to dynamic agents, allowing the generative model to redraw their poses. The third source is depth uncertainty $\mathbf{M}_{unc}$, where we utilize the uncertainty map from the depth estimator (normalized to $[0, 1]$) to assign corresponding noise levels to regions with ambiguous geometry, such as sky or transparent surfaces.

The final noise injection mask is obtained by taking the maximum operation across these components:

$$\mathbf{M} = \max(\mathbf{M}_{occ}, \mathbf{M}_{dyn}, \mathbf{M}_{unc}). \quad (5)$$

Noise Injection. We construct the final source distribution $\hat{x}_0$ via linear interpolation between the warped prior and standard Gaussian noise ϵ , controlled by $\mathbf{M}$:

$$\hat{x}_0 = (1 - \mathbf{M}) \odot \mathbf{Z}_{warp} + \mathbf{M} \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (6)$$

3.5 Training and Inference

Chunk Strategy. Our method relies on the geometric overlap between the reference and target views. To ensure valid reprojection, we adopt a short-horizon frame autoregressive strategy, setting the chunk size to $L = 6$ which contains 1 reference frame and 5 new generated frames.

Optimization Objective. Following the baseline configuration, we maintain the vector field prediction parameterization where the network F_θ estimates the optimal transport velocity v_t . We fine-tune the baseline model to adapt to the proposed Geometry-Aligned Prior Distribution. The intermediate training states

x_t are sampled via linear interpolation along the optimal transport path between our constructed source $\hat{x}_0$ and the ground truth x_1 :

$$x_t = (1 - t)\hat{x}_0 + tx_1, \quad t \in [0, 1]. \quad (7)$$

The training objective is to minimize the prediction error of the vector field:

$$\mathcal{L} = \mathbb{E}_{t, x_1, \hat{x}_0, \mathcal{C}} [||F_\theta(x_t, t, \mathcal{C}) - v_t||^2], \quad (8)$$

where the analytical velocity is defined as $v_t = x_1 - \hat{x}_0$.

Inference. During inference, instead of sampling from a standard Gaussian distribution, we initialize the generative process directly with our Geometry-Aligned Prior Distribution ($x_0 = \hat{x}_0$). From this starting point, we employ standard numerical ODE solvers provided by the baseline architecture to integrate the estimated vector field $F_\theta(x_t, t, \mathcal{C})$ over time.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct our experiments on the NuScenes dataset [2], a large-scale autonomous driving benchmark consisting of 1000 scenes (700 for training, 150 for validation, and 150 for testing). Each scene spans 20 seconds and captures 360-degree views via six cameras. We utilize the training split for fine-tuning baseline models and the validation split for evaluation. The video frames are resized to a resolution of 256×448 for training and inference. **Baselines.** We build our framework upon **OpenDWM** [1] codebase, which reproduces state-of-the-art driving video generation methods and open-source the training, inference, and evaluation codes. We compare our approach primarily against the baseline with Gaussian initialization. Further details, such as a detailed overview of this codebase and the specific selection of the baseline model, are provided in the supplementary materials. To ensure a fair comparison, both the baseline and our model share the exact same model architecture and control conditions, only differ in source distribution.

Metrics. We assess the generation performance from two perspectives: (1) *Video Quality*: We report **FVD** (Fréchet Video Distance) computed on clips of 16 frames to measure the temporal coherence and visual fidelity. (2) *Image Quality*: We report **FID** (Fréchet Inception Distance) to evaluate the visual quality of individual frames. Following UniMLVG [4], we conduct evaluation using 150 scenes from NuScenes validation set, and generate six-view 16-frame videos to calculate those metrics. Except where noted, each video is generated starting from one reference frame.

Implementation Details. We initialize our model from the pre-trained weights of OpenDWM. We fine-tune the pre-trained model on 2 NVIDIA H100 GPUs to adapt to the proposed Geometry-Aligned Prior Distribution. The learning rate is set to $5e-5$. During training, we use MapAnything-v1.0 [18] as metric depth estimator, and we also conduct experiments on other metric depth models during inference.

Table 1: Quantitative comparison on NuScenes validation set. (a) We compare our GeoFlow with state-of-the-art methods. Our **GeoFlow** achieves SOTA quality with significantly less steps. [†] means the metrics are calculated using 3 reference frames. (b) We conduct experiments on base models with different architecture, GeoFlow framework can bring significant improvement on few-step driving video generation.

Method	Steps	FID ↓	FVD ↓
MagicDrive-V2 [8]	30	20.9	94.8
DreamForge [31]	30	14.6	103.6
Drive-WM [44]	50	15.2	122.7
DriveDreamer-2 [60]	-	11.2	55.7
UniMLVG [†] [4]	50	5.8	<u>36.1</u>
OpenDWM [1]	40	<u>6.8</u>	38.8
GeoFlow (Ours)	15	<u>6.8</u>	32.5

(a) Video Quality Comparison

Base Model	5-step FVD
OpenDWM-tvae	204.8
+GeoFlow(-62.5%)	76.7
OpenDWM-vae	121.7
+GeoFlow(-59.6%)	49.2
UniMLVG	72.6
+GeoFlow(-38.6%)	44.6

(b) Generality Analysis of GeoFlow

Table 2: Comparison of FID and FVD metrics across different sampling steps. GeoFlow consistently outperforms the baseline and achieves better quality with 8 steps than base model with 40 steps in FVD.

Method	5-step		8-step		10-step		15-step		20-step		40-step	
	FID	FVD	FID	FVD	FID	FVD	FID	FVD	FID	FVD	FID	FVD
OpenDWM	20.4	123.5	14.7	77.4	12.6	66.1	10.1	52.8	8.7	45.1	6.8	38.8
GeoFlow	11.7	49.2	8.3	38.6	7.5	35.0	6.8	32.5	6.8	32.6	6.9	34.0

4.2 Main Results

Quantitative Comparison. Table 1a compares our method with other SOTA approaches with standard Gaussian initialization. Although existing methods achieve high-fidelity synthesis, they usually require many inference steps. In contrast, GeoFlow synthesizes driving videos with a lower FVD than the baseline (referred to as OpenDWM) using significantly fewer inference steps. This demonstrates that our geometry-aligned prior not only improves the efficiency of the generative process but also naturally enforces better spatiotemporal consistency across frames.

Visual Qualitative Comparison. Qualitatively, our method produces substantially cleaner and more coherent videos in the few-step regime. In particular, with only **8 sampling steps**, GeoFlow already yields visual results that are comparable in FVD to the Baseline using **40 steps**, while the Baseline at few steps often exhibits noticeable blurriness and structural artifacts, as shown in Fig. 1c and Fig. 4. This demonstrates that our geometry-aligned prior provides a much better initialization, enabling the model to reach a high-quality generation manifold with significantly fewer denoising iterations.

Generality Analysis. As demonstrated in Table 1b, our method consistently and significantly improves few-step driving video generation quality across di-

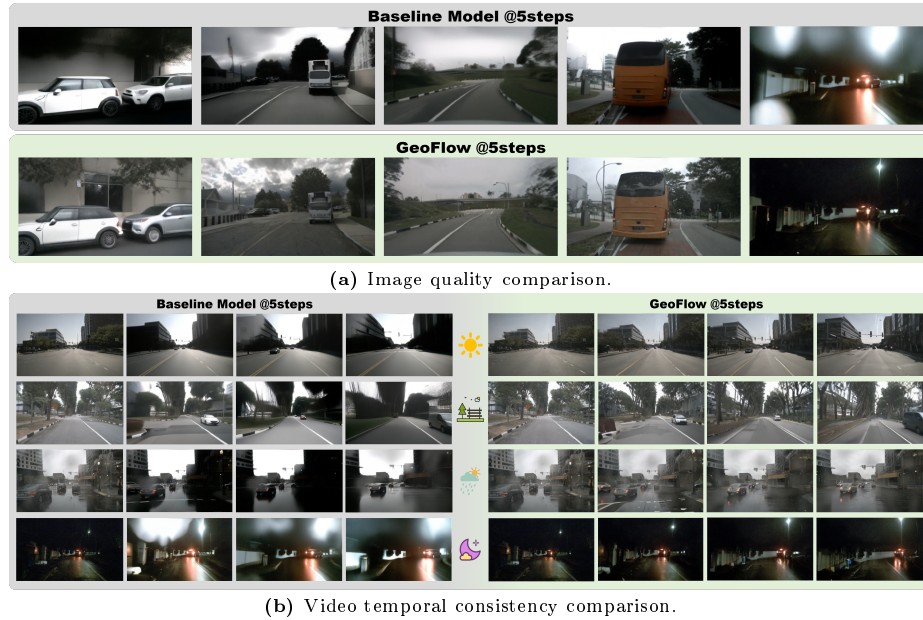


Fig. 4: 5-step generation visual quality comparison between GeoFlow and Baseline model. All these videos are generated with single reference frame.

verse architectures. We validate this by experimenting with various models within the OpenDWM codebase, including those equipped with Temporal (3D) VAE and those with Spatial (2D) VAE. The results show that our approach substantially reduces FVD (as shown in **green** percent number) in the few-step inference setting. Furthermore, when applied to a stronger base model such as UniMLVG, GeoFlow continues to deliver notable improvements. This highlights the **plug-and-play** nature of our approach, offering an efficient path to enhance few-step driving video generators without large computational cost.

4.3 Efficiency Analysis

We provide a comprehensive analysis of the proposed method’s efficiency from both training and inference perspectives.

Fast Training Convergence. A key advantage of GeoFlow is that it simplifies the learning objective from generating from scratch to residual refinement. We argue that the model can efficiently adapt to this new paradigm, because base model has inherently developed such refinement capabilities during its original denoising training phase. To validate this, we monitor the generation quality (FVD) on the validation set at different training milestones. As illustrated in Fig. 5, our model exhibits rapid convergence. The FVD score drops significantly within the first thousands iterations (only hundreds of training steps can achieve significant FVD drop in few-step generation) and saturates around 10,000 iter-

ations. The total adaptive training cost is less than 30 H100 GPU hours. This demonstrates that our framework can be easily adapted to existing base models without requiring expensive distillation or fully retraining.

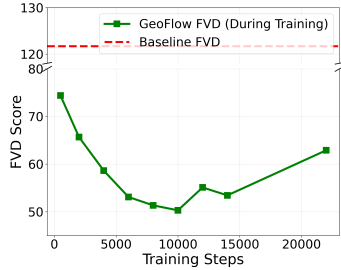


Fig. 5: Training Convergence.

Efficiency-Quality Trade-off. Table 2 illustrates the trade-off between video quality and inference steps for the base model and GeoFlow, demonstrating that the GeoFlow framework substantially enhances the generation quality of the base model. GeoFlow achieves significant acceleration; specifically, our model with only **8 sampling steps** attains an FVD of **38.6**, surpassing the Baseline with **40 steps** (FVD 38.8). This represents a **5× steps reduction** without compromising visual quality. To further visualize the acceleration capability, we plot the FVD scores against the number of sampling steps in Fig. 1b. The Baseline (red curve) exhibits a slow convergence rate, requiring more than 40 steps to reach a saturated performance. In contrast, our method (green curve) starts from a much lower FVD and converges rapidly, achieving high-quality results in around 10 steps.

4.4 Ablation Studies

Effectiveness of Noise Injection. Fig. 6a compares different strategies for constructing the source distribution $\hat{x}_0$. First, **Naïve Warping**, which directly uses warped latents without noise injection, results in poor performance. As shown in Fig. 6b, the deterministic nature of flawed geometric priors causes the model to overfit to rendering artifacts (e.g., stretching textures), leading to severe visual distortions. Also, as illustrated by the red curve in Fig. 6a, the lack of noise injection during training restricts the model’s exploration of the data manifold, leading to error accumulation at higher inference steps and degrading

Inference Efficiency. We further evaluate the inference efficiency. As demonstrated in Sec. 3.5, although our method employs a shorter autoregressive chunk size ($L = 6$) compared to the baseline ($L = 19$) to ensure geometric validity, the total generation time to achieve SOTA quality is still significantly reduced. Specifically, generating a 16-frame clip using our method (8 steps $\times$ 4 passes) is 4.2× faster than the baseline (40 steps $\times$ 1 pass). This confirms that the overhead of geometric warping is marginal compared to the computational savings from reduced denoising iterations. Detailed time costs on one L20 GPU are shown in Table 3.

Table 3: Breakdown of 5-step inference latency for a single 6-frame clip.

Stage	Time (s)	Prop.
Geo. Recon.	0.43	2.84%
Feat. Rend.	0.49	3.24%
ODE Solving	13.83	91.65%
Total	15.09	100%

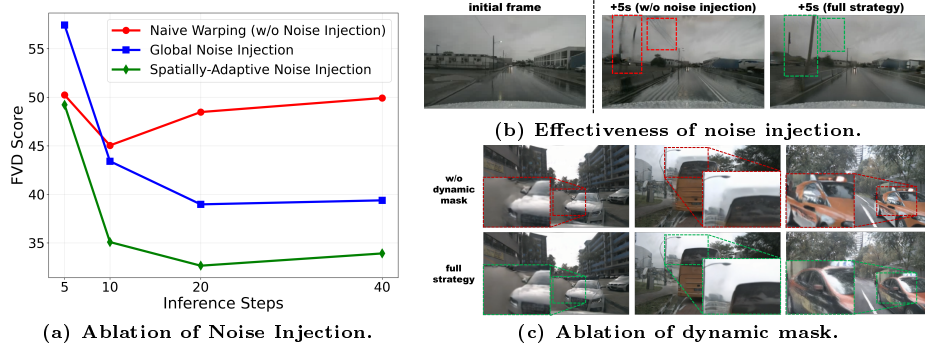


Fig. 6: Qualitative and quantitative ablations. (a) Quantitative comparison of different injection strategies. (b) Visual comparison of generating with and without noise injection. (c) Visual ablation of the dynamic masking strategy.

the generation quality. Second, the addition of **global uniform noise** mitigates these artifacts to some extent. However, as shown in Fig. 6a, its FVD score remains inferior to our method because uniform noise indiscriminately corrupts perfectly aligned static backgrounds (e.g., road markings), counteracting the benefits of geometric warping by forcing the model to reconstruct valid structures from scratch. Finally, our full strategy achieves the best performance by using the proposed reliability mask to selectively inject noise.

Component Analysis of the Noise Injection. We further dissect the contribution of each component in our Noise Injection Mask $\mathbf{M}$ in Table 4. Removing $\mathbf{M}_{dyn}$ leads to significant ghosting artifacts. Since the warping operation projects dynamic objects (e.g., moving cars) to incorrect locations based on their past positions, the model struggles to remove these incorrect features without explicit noise injection (as shown in Fig. 6c). Introducing the dynamic mask significantly reduces artifacts around moving vehicles and improves the visual quality of vehicles in the generated videos. Without $\mathbf{M}_{occ}$, the model receives invalid features (e.g., zero initialization for background rendering) in the out-of-view regions. Masking these areas allows the model to perform standard in-painting.

Robustness to Depth Estimation. We further verify the robustness of our model by substituting the training used Metric Depth Estimation model (MapAnything-1.0) with MapAnything-v1.1 and DepthAnything-3 during inference in a zero-shot manner. As demonstrated in Table 5, the enhanced depth and uncertainty estimation of MapAnything-v1.1 improves video generation performance at low sampling steps. Replacing MapAnything with DepthAnything-3 model can also achieve SOTA performance. Despite this, we show that GeoFlow can boost few-step generation quality even with inaccurate depth or artificially perturbed rendering results, as detailed in the supplementary material.

Table 4: Component analysis of Noise Injection Masks. Inference step is set to 10.

M_{occ}	M_{unc}	M_{dyn}	FID ($\downarrow$)	FVD ($\downarrow$)
-	-	-	8.2	43.6
✓	-	-	7.9	<u>40.0</u>
✓	✓	-	<u>7.6</u>	40.2
✓	✓	✓	7.5	35.0

Table 5: Robustness analysis of depth model selection. The two models are introduced in a zero-shot manner.

Steps	MA-v1.1		DA-3	
	FID($\downarrow$)	FVD($\downarrow$)	FID($\downarrow$)	FVD($\downarrow$)
5	11.1	44.1	13.1	55.1
10	7.9	33.1	8.0	38.2
40	7.8	35.6	7.2	37.3

5 Conclusion

In this work, we present GeoFlow, a plug-and-play framework that achieves efficient driving video generation by bridging the distributional gap between the source and target distributions. By replacing the non-informative standard Gaussian with a Geometry-Aligned Prior Distribution (GAP), we successfully shorten the sampling trajectory, enabling Flow Matching models to generate high-fidelity driving videos in just several sampling steps. Crucially, GeoFlow achieves this efficiency with low adapting cost and no architectural changes, offering a practical and scalable solution for high-throughput data synthesis in autonomous driving.

Acknowledgements

In this work, we are supported by the National Natural Science Foundation of China 62372029, 62276016.

References

1. Opendwm: Open driving world models (2025), <https://github.com/SenseTime-FVG/OpenDWM>
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. Chen, A., Zheng, W., Wang, Y., Zhang, X., Zhan, K., Jia, P., Keutzer, K., Zhang, S.: Geodrive: 3d geometry-informed driving world model with precise action control. arXiv preprint arXiv:2505.22421 (2025)
4. Chen, R., Wu, Z., Liu, Y., Guo, Y., Ni, J., Xia, H., Xia, S.: Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. arXiv preprint arXiv:2412.04842 (2024)
5. Cohen, E., Achituve, I., Diamant, I., Netzer, A., Habi, H.V.: Efficient image restoration via latent consistency flow matching. arXiv preprint arXiv:2502.03500 (2025)

6. Deng, T., Chen, X., Chen, Y., Chen, Q., Xu, Y., Yang, L., Xu, L., Zhang, Y., Zhang, B., Huang, W., et al.: Gaussiandwm: 3d gaussian driving world model for unified scene understanding and multi-modal generation. arXiv preprint arXiv:2512.23180 (2025)
7. Gao, H., Chen, S., Jiang, B., Liao, B., Shi, Y., Guo, X., Pu, Y., Li, X., Liu, W., Zhang, Q., et al.: Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. *Advances in Neural Information Processing Systems* **38**, 32551–32576 (2026)
8. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 28135–28144 (2025)
9. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
10. Ge, J., Liu, Z., Fan, L., Jiang, Y., Su, J., Li, Y., Zhang, Z., Chen, S.: Unraveling the effects of synthetic data on end-to-end autonomous driving. arXiv preprint arXiv:2503.18108 (2025)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)
12. Gu, M., Zhang, J., Li, J., Yu, X., Luo, H., Zheng, J., Bai, X.: Sparsesurf: Sparse-view 3d gaussian splatting for surface reconstruction. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 40, pp. 4311–4319 (2026)
13. Han, Z., Zhang, B., Zhang, L., Feng, S., Lin, K., Liang, G., Ye, Y., et al.: Asyncdsb: Schedule-asynchronous diffusion schrödinger bridge for image inpainting. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 3374–3382 (2025)
14. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
16. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
17. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
18. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
20. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20775–20785 (2024)
21. Li, J., Zhang, J., Yu, X., Bai, X., Zheng, J., Ning, X., Gu, L.: Dngaussian++: Improving sparse-view gaussian radiance fields with depth normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2026)

22. Li, J., Zhang, J., Zhang, Y., Bai, X., Zheng, J., Yu, X., Gu, L.: Geosvr: Taming sparse voxels for geometrically accurate surface reconstruction. *Advances in Neural Information Processing Systems* **38**, 108809–108837 (2026)
23. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)
25. Liu, C., Ritschel, T.: Generative video bi-flow. *arXiv preprint arXiv:2503.06364* (2025)
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems* **35**, 5775–5787 (2022)
28. Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R.: Srflow: Learning the super-resolution space with normalizing flow. In: *European conference on computer vision*. pp. 715–732. Springer (2020)
29. Mao, X., Jiang, Z., Wang, F.Y., Zhang, J., Chen, H., Chi, M., Wang, Y., Luo, W.: Osv: One step is enough for high-quality image to video generation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 12585–12594 (2025)
30. Martin, S., Gagneux, A., Hagemann, P., Steidl, G.: Pnp-flow: Plug-and-play image restoration with flow matching. *arXiv preprint arXiv:2410.02423* (2024)
31. Mei, J., Hu, T., Yang, X., Wen, L., Yang, Y., Wei, T., Ma, Y., Dou, M., Shi, B., Liu, Y.: Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003* (2024)
32. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 1559–1569 (2025)
33. Peng, J., Li, M., Wang, H.: Stabilizing holistic semantics in diffusion bridge for image inpainting. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. pp. 1756–1764 (2025)
34. Qin, H., Luo, W., Wang, L., Zheng, D., Chen, J., Yang, M., Li, B., Hu, W.: Reversing flow for image restoration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 7545–7558 (2025)
35. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6121–6132 (2025)
36. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512* (2022)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020)
38. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
39. Wang, C., Zhu, Y., Yuan, C.: Diverse image inpainting with normalizing flow. In: *European conference on computer vision*. pp. 53–69. Springer (2022)
40. Wang, H., Zhang, J., Chen, H., Guo, H., Wang, D., Ma, J., Du, B.: Residual diffusion bridge model for image restoration. *arXiv preprint arXiv:2510.23116* (2025)

41. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5294–5306 (2025)
42. Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., Sang, N.: Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109* (2023)
43. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: *European conference on computer vision*. pp. 55–72. Springer (2024)
44. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14749–14759 (2024)
45. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 27531–27541 (2025)
46. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10371–10381 (2024)
47. Yang, X., Wen, L., Wei, T., Ma, Y., Mei, J., Li, X., Lei, W., Fu, D., Cai, P., Dou, M., et al.: Drivearena: A closed-loop generative simulation platform for autonomous driving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 26933–26943 (2025)
48. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6613–6623 (2024)
49. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 22963–22974 (2025)
50. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024)
51. Yue, C., Peng, Z., Ma, J., Zhang, D.: Enhanced control for diffusion bridge in image restoration. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2025)
52. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems* **36**, 13294–13307 (2023)
53. Yun, J.H., Kim, S.B., Lee, S.W.: Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2025)
54. Zhang, J., Li, J., Yu, X., Huang, L., Gu, L., Zheng, J., Bai, X.: Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In: *European conference on computer vision*. pp. 335–352. Springer (2024)
55. Zhang, J., Zhang, Y., Tosi, F., Gu, M., Li, J., Yu, X., Zheng, J., Bai, X., Poggi, M.: Eve3d: Elevating vision models for enhanced 3d surface reconstruction via gaussian splatting. *Advances in Neural Information Processing Systems* **38**, 67934–67960 (2026)

56. Zhang, Z., Li, Y., Wu, Y., Kag, A., Skorokhodov, I., Menapace, W., Siarohin, A., Cao, J., Metaxas, D., Tulyakov, S., et al.: Sf-v: Single forward video generation model. *Advances in Neural Information Processing Systems* **37**, 103599–103618 (2024)
57. Zhang, Z., Chen, D., Liao, J.: I2v3d: Controllable image-to-video generation with 3d guidance. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 13360–13371 (2025)
58. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., et al.: Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 12015–12026 (2025)
59. Zhao, G., Wang, X., Ni, C., Zhu, Z., Qin, W., Huang, G., Wang, X.: Recondreamer++: Harmonizing generative and reconstructive models for driving scene representation. *arXiv preprint arXiv:2503.18438* (2025)
60. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 10412–10420 (2025)
61. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems* **36**, 49842–49869 (2023)
62. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948* (2023)