2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control

Anthony Chen*,1,2, Wenzhao Zheng*,3, Yida Wang*,2
Xueyang Zhang[2], Kun Zhan[2], Peng Jia[2], Kurt Keutzer[3], Shanghang Zhang†,1

***Abstract*— Recent advancements in world models have revolutionized dynamic environment simulation, allowing systems to foresee future states and assess potential actions. In autonomous driving, these capabilities help vehicles anticipate the behavior of other road users, perform risk-aware planning, accelerate training in simulation, and adapt to novel scenarios, thereby enhancing safety and reliability. Current approaches exhibit deficiencies in maintaining robust 3D geometric consistency or accumulating artifacts during occlusion handling, both critical for reliable safety assessment in autonomous navigation tasks. To address this, we introduce *GeoDrive*, which explicitly integrates robust 3D geometry conditions into driving world models to enhance spatial understanding and action controllability. Specifically, we first extract a 3D representation from the input frame and then obtain its 2D rendering based on the user-specified ego-car trajectory. To enable dynamic modeling, we propose a *dynamic editing* module during training to enhance the renderings by editing the positions of the vehicles. Extensive experiments demonstrate that our method significantly outperforms existing models in both action accuracy and 3D spatial awareness, leading to more realistic, adaptable, and reliable scene modeling for safer autonomous driving. Additionally, our model can generalize to novel trajectories and offers interactive scene editing capabilities, such as object editing and object trajectory control.**

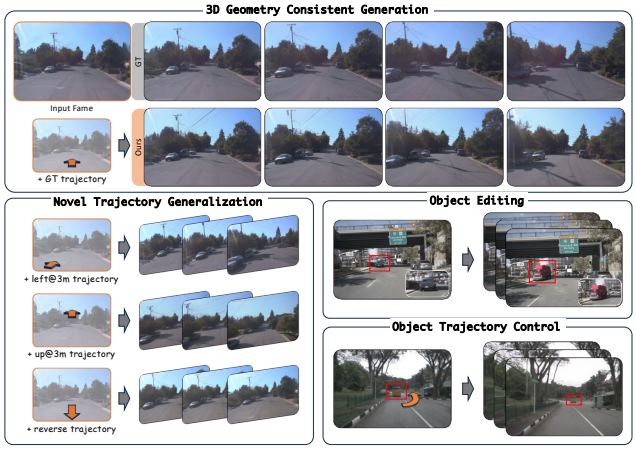

Fig. 1: GeoDrive enables multiple capabilities from a single input frame. **Top**: 3D geometry-consistent generation closely matching ground truth in road geometry and lane markings. **Bottom-left**: Novel trajectory generalization to unseen viewpoints (3m lateral/vertical shifts, reverse), maintaining 3D consistency. **Bottom-right**: Scene editing (vehicle removal/replacement) and non-ego vehicle trajectory control.

I. Introduction

Driving world models simulating 3D dynamic environments enable critical capabilities including trajectory-consistent view synthesis [44], physics-compliant motion prediction [18], and safety-aware scenario reconstruction [44] and generation ([7], [27]). Particularly, generative video models have emerged as effective tools for ego-motion forecasting and dynamic scene reconstruction ([3], [14], [37]). Their ability to synthesize trajectory-faithful visual sequences proves crucial for developing autonomous systems that anticipate environmental interactions while maintaining physical plausibility.

Despite these advancements, most existing methods lack sufficient 3D geometric awareness [39] due to their reliance on 2D space optimization [7]. This shortcoming results in structural incoherence across novel views and physically implausible object interactions ([45], [38]), which is particularly detrimental for safety-critical tasks like collision avoidance in dense traffic. Also, existing methods usually depend on dense annotations (e.g., HD-map sequences and 3D bounding box tracks) for controllability ([38], [23]), which only reproduce prescribed motions without understanding vehicle dynamics. A more flexible approach is to infer dynamic priors from single (or few) images while conditioning on the desired ego-trajectory. However, current methods that fine-tune on numerical camera parameters lack 3D geometry awareness, compromising their action controllability and consistency ([7], [1], [16]). A reliable driving world model should satisfy three criteria: 1) rigid spatio-temporal coherence across static infrastructure and dynamic agents; 2) 3D controllability over ego-vehicle trajectories; and 3) kinematically constrained motion patterns for non-ego agents.

We achieve these demands by first building a 3D structural prior from monocular input, performing projective rendering along user-specified trajectories, and refining the results through video diffusion with specialized geometric conditioning. For dynamic objects, we introduce a dynamic editing module that adjusts vehicle positions under motion constraints. Experiments demonstrate that GeoDrive reduces trajectory-following errors by 42% compared to Vista [7] while improving all video quality metrics (LPIPS, PSNR, SSIM, FID, FVD), using 99.7% less training data. Our model also generalizes to novel view synthesis, surpassing StreetGaussians [44], and offers interactive scene editing and VLA planning assistance.

• This work was supported by the National Natural Science Foundation of China (62476011), and by the Beijing Natural Science Foundation (L252060).
*Equal contribution. † Corresponding Author.
[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University. [2]Li Auto Inc. [3]UC Berkeley

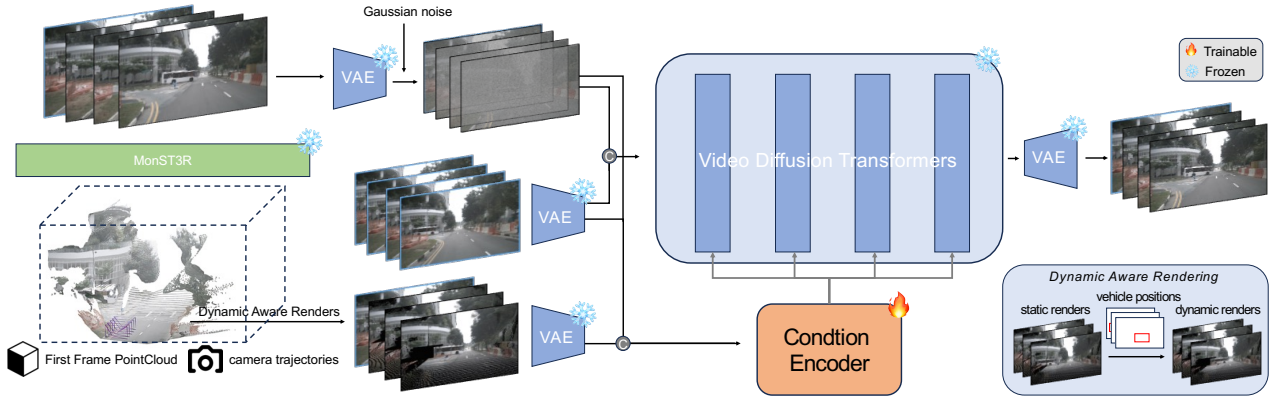

Fig. 2: **Overview of our training pipeline.** We use pretrained dense stereo model to obtain 3D point clouds and camera trajectories. A dynamic video is rendered from the first-frame point cloud using our *dynamic editing* technique. The noisy latent representation and rendered video are encoded via VAE and concatenated as input for our *condition encoder*, modulating the DiT model's features. The DiT then generates photorealistic video that accurately follows the specified action conditions.

II. Related Work

Driving World Models. World models enable agents to anticipate and act in dynamic environments. Existing approaches rely on point clouds ([51], [19], [49]), occupancy grids ([26], [55], [8], [48]), or images ([38], [53], [23]) for scene representation, with image-based models being most promising due to sensor flexibility and data abundance ([17], [7], [33]). However, current systems rely on dense annotations or weakly aligned control vectors. Our approach instead encodes actions as visual conditions naturally aligned with latent representations, yielding stronger, more stable control.

Conditional Video Generation. Diffusion models have progressed from image synthesis to video generation with increasing emphasis on conditional control ([52], [30], [24]), including depth ([6], [42]), trajectory [47], and camera conditioning ([9], [41], [43], [31]). However, these offer only coarse or indirect guidance. In driving video synthesis, DriveDreamer [38] and DrivingDiffusion [23] rely on dense maps and box tracks, while Vista [7] and GAIA ([16], [33]) inject control vectors into latent features, but misalignment with visual space weakens control. Our work instead encodes actions as visual conditions, aligning control with latent dynamics for stable trajectory generation.

III. Preliminary

Diffusion Model is built from two stochastic chains: a *forward* (noising) process q and a *reverse* (denoising) process p_θ [34] . Starting with a clean sample $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$, the forward process incrementally injects Gaussian noise, producing $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where the schedule satisfies $\alpha_t^2 + \sigma_t^2 = 1$. where α_t and σ_t denote the signal and noise coefficients of the diffusion schedule, satisfying $\alpha_t^2 + \sigma_t^2 = 1$. The reverse process seeks to remove this noise using a neural predictor $\boldsymbol{\epsilon}_\theta$, trained by minimizing

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\right\|_2^2\right]. \quad (1)$$

To ease the heavy computation of pixel-space generation, latent diffusion models (LDMs) [28] compress RGB video $\mathbf{x} \in \mathbb{R}^{L\times 3\times H\times W}$ into a lower-dimensional tensor $\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{L\times C\times h\times w}$ via a frozen VAE encoder $\mathcal{E}$, where L is the number of video frames and C is the latent channel dimension. Both forward and reverse chains operate in this latent space, after which a decoder $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$ reconstructs the final video.

We adopt the pretrained CogVideo-I2V [46] as our backbone, as its single-image animation capability aligns with our objective of predicting future scenarios from a single input.

IV. Methodology

Given an initial reference image $I_0 \in \mathbb{R}^{H\times W\times 3}$ and ego-vehicle trajectory $\{C_t\}_{t=1}^{L}$, our framework synthesizes realistic future frames that follow the input trajectory. We leverage 3D geometric information from the reference image to guide world modeling. First, we reconstruct a 3D representation (Sec. IV-A), then render video sequences along user-specified trajectories with dynamic object handling (Sec. IV-B). The rendered video provides geometric guidance for generating spatio-temporally consistent videos that follow the input trajectory (Sec. IV-C). See Figure 2 for illustration.

A. Extracting 3D Representations from Reference Image

To ensure 3D-consistent generation, we extract geometry from the input image I_0 using MonST3R [50], which jointly predicts 3D structure and camera poses. Since MonST3R requires cross-view matching, we duplicate the reference image during inference.

Given RGB frames $\{I_t\}_{t=0}^{T}$, MonST3R outputs per-pixel 3D coordinates $\{O_t\}$ and confidence maps $\{D_t\}$:

$$\{O_t\}_{t=0}^{T}, \{D_t\}_{t=0}^{T} = \text{MonST3R}\big(\{I_t\}_{t=0}^{T}\big), \quad (2)$$

where $\mathbf{O}_t^{i,j} \in \mathbb{R}^3$ denotes the 3D position of pixel (i, j), and $\mathbf{D}_t^{i,j} \in [0, 1]$ indicates reliability. Thresholding $\mathbf{D}_t$ at τ yields the colored point cloud

$$\mathcal{P}_t = \big\{(\mathbf{O}_t^{i,j},\, I_t^{i,j}) \mid \mathbf{D}_t^{i,j} > \tau\big\}. \quad (3)$$

A transformer decoupler separates static from dynamic features using learnable prompt tokens, enabling pose estimation from static correspondences only:

$$\hat{C}_t = \arg\min_{C_t} \sum_{(i,j)\in\mathbf{F}^{\text{static}}} \left\|\pi(C_t\, \mathbf{O}_t^{i,j}) - \mathbf{p}_t^{i,j}\right\|_2^2, \quad (4)$$

where F_{static} is the set of pixel locations classified as static by the decoupler, $\mathbf{p}_t^{i,j}$ are the observed 2D coordinates, and π denotes perspective projection from 3D points to image coordinates. Compared to conventional structure-from-motion [32], this reduces pose error by 38% in dynamic scenes [50]. The resulting $\mathcal{P}_0$ serves as our geometric scaffold.

B. Rendering 3D Videos with Dynamic Editing

To achieve precise input trajectory following, our model renders a video that serves as a visual guide for the generation process. We project the reference point cloud $\mathcal{P}_0$ through each user-provided camera configuration $C_t = (R_t, T_t, f_t)$, denoting rotation, translation, and focal length respectively, using standard projective geometry techniques. Each 3D point $\mathbf{P}_i^w \in \mathcal{P}_0$ undergoes a rigid transformation into the camera coordinate system $\mathbf{P}_i^c = R_t\mathbf{P}_i^w + T_t$, followed by perspective projection yielding image coordinates $\mathbf{p}_i = \left(\frac{f_t\mathbf{P}_i^{cx}}{\mathbf{P}_i^{cz}} + \frac{W}{2}, \frac{f_t\mathbf{P}_i^{cy}}{\mathbf{P}_i^{cz}} + \frac{H}{2}\right)$. We only consider valid projections within a depth range of $\mathbf{P}_i^{cz} \in [0.1, 100.0]$ meters and use z-buffering to handle occlusions, ultimately producing the rendered view $\tilde{I}_t$ for each camera position.

Limitations of Static Rendering. Since we utilize only the first frame point cloud, the rendered scene remains static, creating a significant discrepancy with real-world driving where vehicles are in constant motion.

Dynamic Editing. To address this limitation, we propose *dynamic editing* to produce renderings R with static backgrounds and moving vehicles. Specifically, when users provide a sequence of 2D bounding box information for moving vehicles in the scene, we dynamically adjust their positions to create the illusion of motion in the rendering. This approach not only guides the ego-vehicle's trajectory during the generation process but also directs the movement of other vehicles in the scene. Fig. 3 provides an illustration of this process. Such a design significantly reduces the disparity between static rendering and dynamic real-world scenarios while enabling flexible control over other vehicles—a capability that existing methods like Vista [7] and GAIA [16] fail to achieve.

C. Dual-Branch Control for Spatio-Temporal Consistency

While the point cloud-based rendering accurately preserves geometric relationships between views, it suffers from several visual quality issues. The rendered views often contain substantial occlusions, missing areas due to limited sensor coverage, and reduced visual fidelity compared to real camera images. Addressing these artifacts requires a generative model capable of hallucinating plausible content in occluded regions while respecting the underlying 3D structure. Diffusion models have emerged as the state of the art for high-fidelity visual generation, surpassing both GANs and autoregressive methods in quality and stability. While image diffusion models can refine individual frames, they process each view independently and therefore cannot enforce temporal consistency across a video sequence — leading to flickering, identity drift, and physically implausible motion between frames. Video diffusion models overcome this by jointly modeling spatial and temporal dimensions, enabling the generation of coherent frame sequences that respect both appearance and motion continuity. This makes them naturally suited for our task, where the generated output must simultaneously be photorealistic, temporally smooth, and faithful to the input trajectory. We therefore adapt a latent video diffusion model [4] to refine projected views while preserving 3D structural fidelity through specialized conditioning.

Inspired by VideoPainter [2], we integrate contextual features into a pre-trained diffusion transformer (DiT), with key distinctions: we employ dynamic renderings to capture temporal and contextual nuances for the generation process. Let $\delta_\phi(z_t, t, C)$ represent the feature output at layer i of our modified DiT backbone δ_ϕ, where z_R denotes the dynamic renderings latent via VAE encoder $\mathcal{E}$ and z_t is the noisy latent at timestep t.

These renderings are processed through a lightweight condition encoder, which extracts essential background cues without duplicating extensive portions of the backbone architecture. The integration of features from the condition encoder into the frozen DiT is formulated as follows:

$$\delta_\phi(z_t, t, C)_i = \delta_\phi(z_t, t, C)_i + \mathcal{W}\left(\gamma_\phi^{enc}([z_t, z_R], t)_{i//\frac{M}{2}}\right), \tag{5}$$

where γ_ϕ^{enc} denotes the condition encoder processing the concatenated input of noisy latent z_t and renderings latent z_R, with M representing the total number of layers in the DiT backbone. $\mathcal{W}$ is a learnable linear transformation initialized to zero to prevent noise collapse in early training. The extracted features are selectively fused into the frozen DiT in a structured manner, ensuring that only relevant contextual information guides the generation process. The final video sequence is decoded via the frozen VAE decoder $\mathcal{D}$ as $\hat{I}_t = \mathcal{D}(z_t^{(0)})$, where $z_t^{(0)}$ denotes the fully denoised latent after the final reverse diffusion step.

By training condition encoder $g\phi$ alone (6% of total parameters), we maintain the pre-trained model's photorealism and gain precise camera control. Temporal coherence arises naturally from the video transformer's dynamics modeling and the geometric consistency of $\{\tilde{I}_t\}$ features across frames, enabling trajectory-faithful video synthesis.

V. Experiments and Applications

A. Experimental Settings

Training Configuration. We train exclusively on NuScenes [5], processing each clip through MonST3R for 3D reconstruction and camera trajectories, with Dynamic Editing leveraging 2D bounding box annotations. We curate 25,109 video-condition pairs. The base model (CogVideo-5B-I2V [15]) is frozen; the condition encoder trains for 28,000 steps at lr 1×10^{-5} for 4 days.

Inference Configuration. For Trajectory Following experiments (Section V-B), we use the first frame of each video as condition frame, and we estimate trajectory with MonST3R from the video. For fair comparision with baseline methods, we do not include object control as the baseline methods do not accept object bounding box information. For Novel

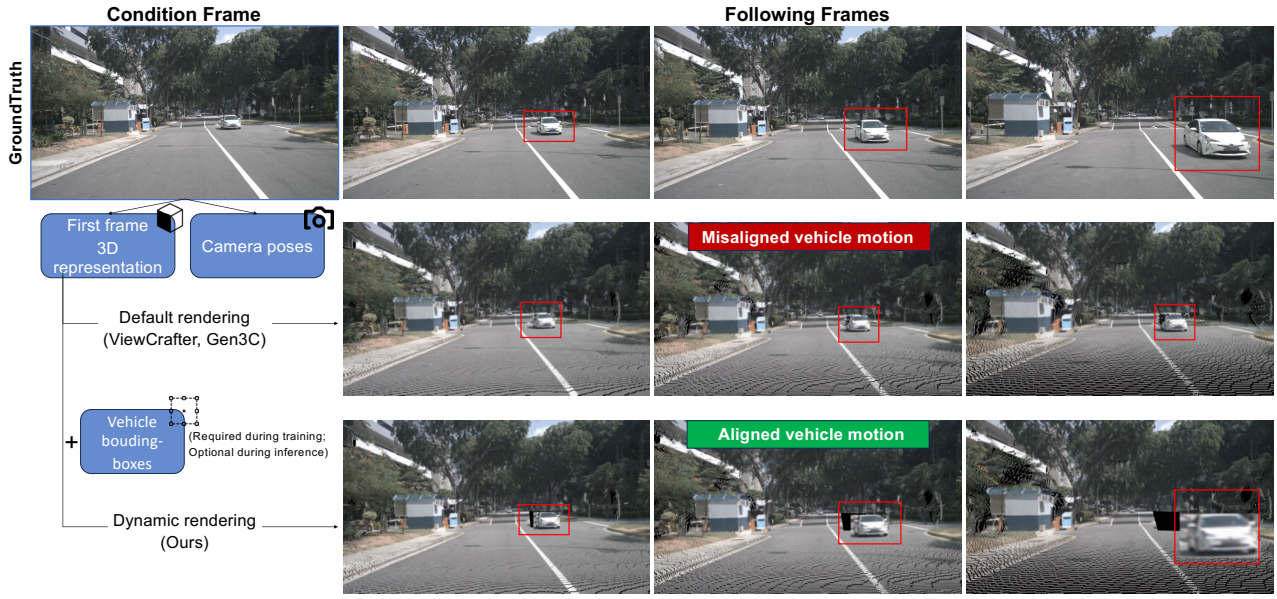

Fig. 3: **Illustration of *dynamic edit* design.** Compared with default rendering, it effectively reduces disparity between static rendering and dynamic real-world scenarios.

TABLE I: **Quantitative results of generation quality and action fidelity on NuScenes [5] validation subset.** We outperform baseline methods on every metric while requiring much less training data.

Method	Data Scale	Prediction Quality					Action Fidelity	
		LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	FVD ↓	$ADE_{\times 10^2}$ ↓	$FDE_{\times 10^2}$ ↓
Vista [7]	1740h	0.351	20.086	0.621	8.35	163.7	2.77	5.28
Terra [1]	1740h	0.455	18.42	0.553	26.73	787.54	5.57	11.8
GeoDrive (Ours)	5h	**0.303**	**21.979**	**0.6535**	**7.17**	**85.22**	**1.62**	**3.1**

View Synthesis experiments (Section V-C), we use the first frame of each video as condition frame, and we estimate the original trajectory with MonST3R from the video. Next, we align the original trajectory with the depth scale from the Lidar point cloud. Then we shift the trajectory left, right and up to obtain novel trajectory input for the experiment.

Inference Runtime. We report per-clip inference time on a single NVIDIA H100 GPU. MonST3R reconstruction takes 3s, point cloud rendering 0.5s, and diffusion inference (50 DDIM steps) 500s, totaling approximately 505s per 81-frame clip. The current pipeline is not real-time, but is practical for offline applications such as training data augmentation and safety scenario generation. Diffusion distillation techniques could reduce inference steps from 50 to 4–8, MonST3R reconstruction can be cached per scene, and replacing the backbone with speed-optimized video diffusion models (e.g., LTX-Video [11]) could further reduce generation time toward real-time, offering a viable path for interactive deployment.

B. Trajectory Following

Benchmark and Baselines. We compare GeoDrive to two most-relevant baselines that condition on single image and ego action (Vista [7], Terra [1]), as well as several other driving world models. We adhere to Vista's protocol by computing trajectory from sensor and calibration data that spans the 25-frame clip, as their condition input. We estimate our condition camera poses by running MonST3R on GT video. While we condition on different modalities, the trajectories for all methods are extracted from the same ground-truth video clip, ensuring aligned action conditions. We evaluate all methods on NuScenes validation set. For trajectory control precision evaluation, we sample a subset of 1087 videos with balanced driving trajectories. Visual quality is quantified through PSNR, SSIM [40], LPIPS [20], FID [13], and FVD [36]. While trajectory fidelity metrics employ Average Displacement Error (ADE) and Final Displacement Error (FDE).

TABLE II: **Quantitative results of generation quality on NuScenes validation fullset.**

Models	Data Scale	FID ↓	FVD ↓
DriveGAN [21]	160h	73.4	502.3
DriveDreamer [38]	5h	14.9	340.8
DriveDreamer-2 [53]	5h	25.0	105.1
WoVoGen [25]	5h	27.6	417.7
Drive-WM [29]	5h	15.8	122.7
GenAD [45]	1740h	15.4	184.0
Vista [7]	1740h	6.6	167.7
GEM [12]	4000h	10.5	158.5
GeoDrive (Ours)	5h	**4.1**	**61.6**

Quantitative Results. The quantitative results on nuScenes validation subset are presented in Table I. GeoDrive outperforms baselines in all metrics. Specifically, our trajectory-

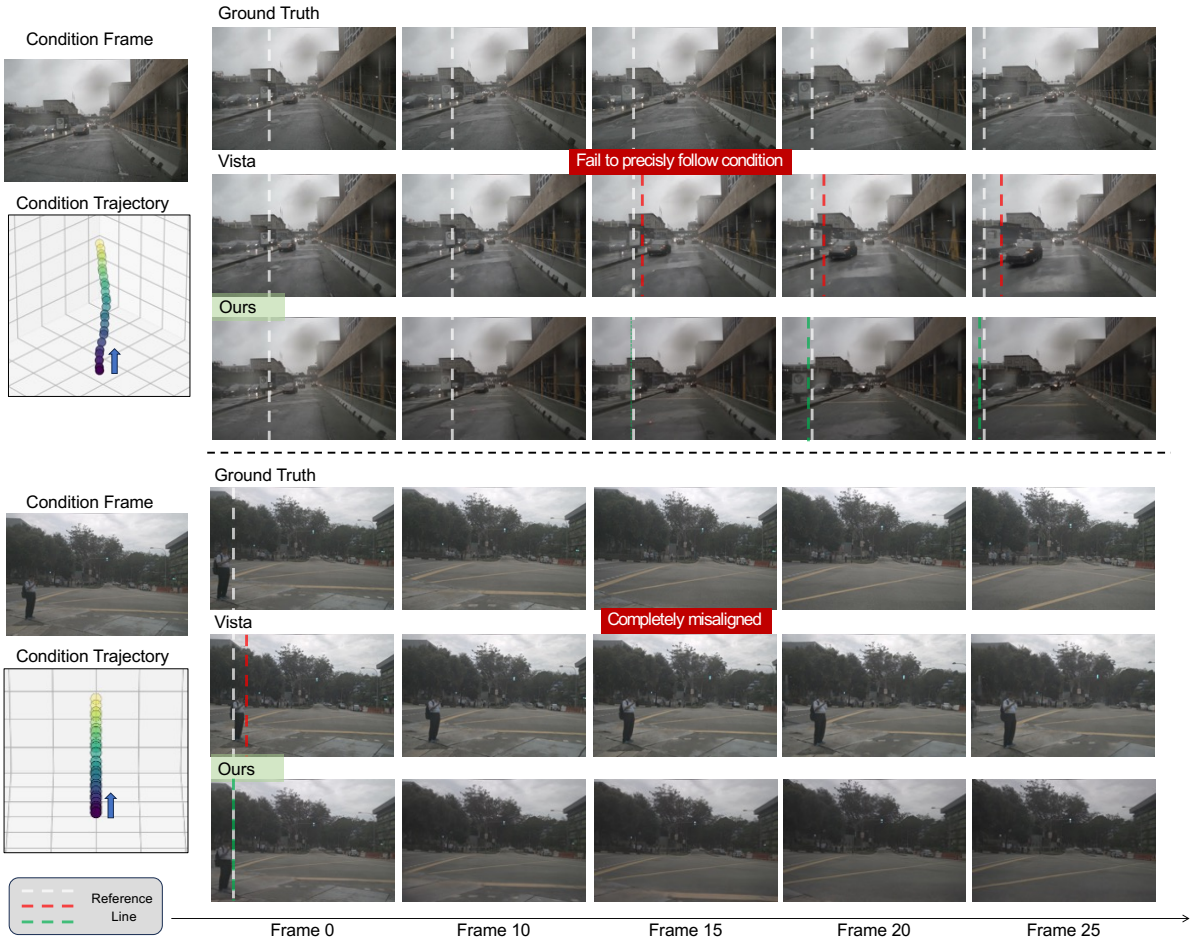

Fig. 4: **Qualitative comparison of action fidelity under the same conditional frame and action control.** Our model precisely follows desired trajectory, while Vista [7] produce misaligned results.

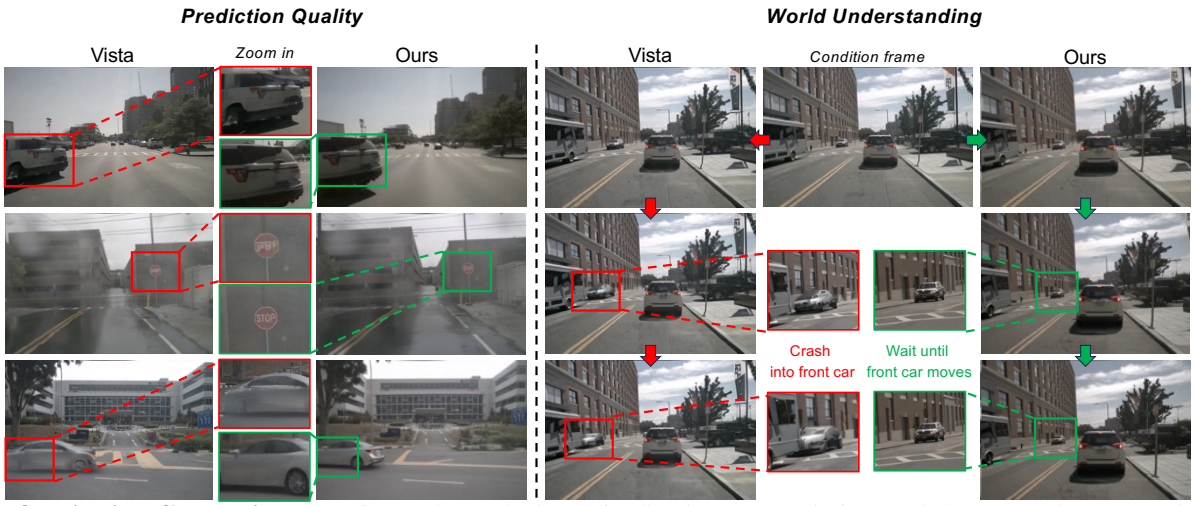

Fig. 5: **Qualitative Comparisons:** Left - Enhanced visual fidelity in our predictions; Right - Superior scene dynamics understanding.

following ability is significantly better than the baselines, yet requiring 99.7% less data, showing the effectiveness of our method. In Table II, GeoDrive outperforms all baselines on FID & FVD results. We note that baselines such as Vista and Terra were designed for broader controllability goals beyond geometry-specific conditioning, which may partially explain the performance gap. However, trajectory-faithful video generation is a core capability that both methods explicitly claim to support, and the evaluation metrics (ADE, FDE) directly measure this shared objective, making the comparison relevant to their stated contributions.

Qualitative Results.

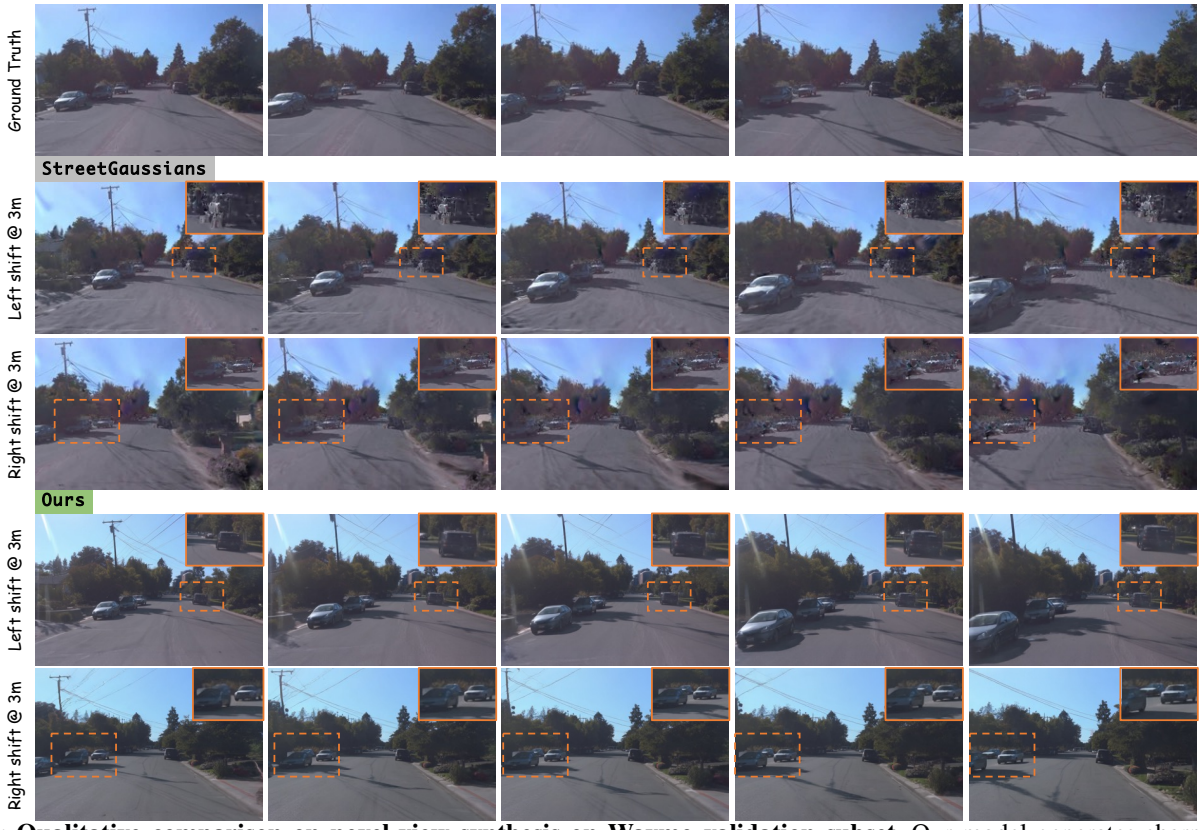

Fig. 6: **Qualitative comparison on novel-view synthesis on Waymo validation subset.** Our model generates sharp results for deviated trajectories in a zero-shot manner, whereas the reconstruction-based method StreetGaussian [44] produces significant artifacts.

As shown in Figures 4 and 5, our method follows specified trajectories more precisely than baselines and produces sharper results with fewer artifacts. Our model also demonstrates stronger scene dynamics understanding, correctly anticipating that a car should wait for the bus ahead, whereas Vista erroneously accelerates forward into a collision.

C. Novel View Synthesis

Benchmark and Baseline.

We compare GeoDrive to StreetGaussians [44] on 5 filtered Waymo validation scenes. Novel trajectories are created by horizontal shifts from the frontal camera trajectory. We report FID and FVD as no ground truth exists for novel views.

Quantitative Results. As demonstrated in Table III, our method achieves lower (and thus better) FID & FVD scores compared to the reconstruction-based baselines. This is due to the difficulty reconstruction methods face in recovering scene structures from the sparsely observed views.

Qualitative Results. Figure 6 illustrates that while StreetGaussians [44] generates projectively correct renderings along given trajectories, it exhibits severe geometric distortions under viewpoint shifts. Such degradations reveal fundamental limitations in 3D scene reconstruction from sparse observational data, particularly the inability to resolve occlusion boundaries and low-textured supervisions. GeoDrive maintains trajectory adherence while preserving photorealistic rendering.

Fig. 7: **Qualitative Results on Scene Editing.** Our approach enables the removal or replacement of vehicles within a scene, allowing for the prediction of seamless future scenarios.

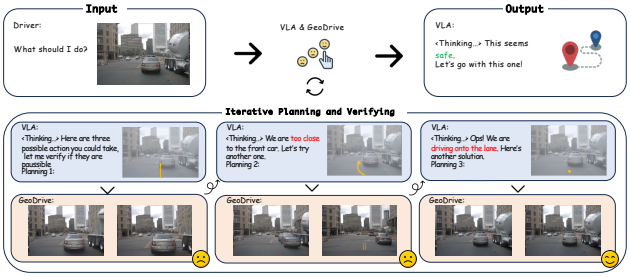

Fig. 8: **Illustration of application to VLA planning.** By simulating each possible planned trajectory, we can assist the VLA model in refining its decisions until it reaches the optimal decision.

D. Applications

Object Trajectory Control. As shown in Figure 1, given extra condition input (i.e., a sequence of 2D bounding boxes) for other vehicles, we can apply *dynamic augmentation* (Sec. IV-B) and thus control the trajectory of vehicles

TABLE III: **Quantitative results for NVS on Waymo validation subset.** GeoDrive is trained solely on the NuScenes dataset, yet it can generalize to Waymo scenes in a zero-shot, feed-forward manner.

Method	Left@3m		Right@3m	
	FID ↓	FVD ↓	FID ↓	FVD ↓
StreetGS [44]	**63.84**	1438.89	69.55	1526.62
Ours	67.13	**1245.23**	**65.67**	**1422.63**

TABLE IV: **Ablation Studies on NuScenes validation subset.** D.E. represents dynamic editing. *w/o* dual-branch means we adjust input channels to adapt renders and finetune backbone.

Method	FID	FVD	$\text{ADE}_{\times 10^2}$
w/o D.E.	7.01	88.68	3.68
w/o dual-branch	8.83	74.76	3.45
GeoDrive	7.17	85.22	1.62

appearing in the condition frame.

Object Editing. GeoDrive enables flexible scene editing by applying off-the-shelf image editing models [22] to condition frames, allowing modification or removal of objects (Fig. 7). This provides valuable control for generating targeted training data, analyzing object influence, and diversifying scenarios from limited data.

Assistance for VLA planning. GeoDrive enhances the decision-making of VLA (Visual Language Action) models by providing an interactive simulation environment for evaluating driving actions ([56], [35], [54], [10]). As illustrated in Figure 8, it integrates real-time visual input with predictive modeling, allowing the VLA system to simulate outcomes of planned trajectories. This helps foresee hazards, such as proximity to other vehicles or lane deviations, and assess action safety. Consequently, the VLA model refines its decisions by selecting actions predicted to be safe and effective, which ensures driving decisions are contextually appropriate and prioritize safety.

E. Ablation Studies

Impact of Dynamic Editing. To quantify the impact of our dynamic editing strategy, we report the performance of our model trained with and without this component on key evaluation metrics as shown in Table IV.

Impact of Dual-branch Architecture. We further investigate the effectiveness of adopting a dual-branch architecture compared with single-branch architecture. We evaluate the performance of different architectures on key evaluation metrics. The results are shown in Table IV.

VI. Conclusions and Limitations

We presented GeoDrive, a video diffusion world model for autonomous driving that enhances action controllability and spatial accuracy via explicit 3D geometric conditioning. Our method reconstructs 3D scenes, renders along desired trajectories, and refines the output with video diffusion, achieving superior performance over existing models in visual realism and action adherence. Compared to traditional simulators (e.g., CARLA), which offer multi-sensor support and full physical determinism, GeoDrive provides complementary strengths: it preserves the visual distribution of real data, requires no manual 3D assets, and enables counterfactual generation from recorded drives. The two approaches are best viewed as complementary rather than competing. Our approach has several limitations. Performance depends on MonST3R's depth and pose accuracy. The dynamic editing module relies on external 2D bounding box annotations, reducing autonomy, and is purely geometric without modeling appearance changes under motion. These could be addressed by integrating learned motion forecasting or lightweight detection models. Future work will also explore text conditions and VLA understanding to further improve realism and consistency.

References

[1] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. Act-bench: Towards action controllable world models for autonomous driving, 2024.

[2] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control, 2025.

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *CVPR*, 2023.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2023.

[7] 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. *Proc. Adv. Neural Inf. Process. Syst.*, 2024.

[8] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024.

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *Proc. Int. Conf. Learn. Represent.*, 2024.

[10] Ziang Guo, Konstantin Gubernatorov, Selamawit Asfaw, Zakhar Yagudin, and Dzmitry Tsetserukou. Vdt-auto: End-to-end autonomous driving with vlm-guided diffusion transformers, 2025.

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. 2024.

[12] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. *arXiv preprint arXiv:2412.11198*, 2024.

[13] 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.

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video Diffusion Models. *arXiv preprint arXiv:2204.03458*, 2022.

[15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[16] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *Technical Report arXiv:2309.17080*, 2023.

[17] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

[18] Bencheng Huang, Shaoyu Liu, Tianheng Chen, Xinggang Shen, Zeming Zhu, Zhe Wang, et al. Vad: Vectorized scene representation for autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023.

[19] Zanming Huang, Jimuyang Zhang, and Eshed Ohn-Bar. Neural volumetric world models for autonomous driving. In *Proc. Eur. Conf. Comput. Vis.*, pages 195–213. Springer, 2025.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. Eur. Conf. Comput. Vis.*, 2016.

[21] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a Controllable High-Quality Neural Simulation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[22] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[23] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.

[24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023.

[25] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Proc. Eur. Conf. Comput. Vis.*, pages 329–345. Springer, 2025.

[26] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 21486–21495, 2024.

[27] Jiageng Mao, Boyi Li, Boris Ivanovic, Yuxiao Chen, Yan Wang, Yurong You, Chaowei Xiao, Danfei Xu, Marco Pavone, and Yue Wang. Dreamdrive: Generative 4d scene modeling from street view images. *arXiv preprint arXiv:2501.00601*, 2024.

[28] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023.

[29] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 15522–15533, 2024.

[30] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conf. Artif. Intell.*, 2024.

[31] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[32] Jérôme Revaud, Vincent Leroy, Philippe Weinzaepfel, Boris Chidlovskii, and Gabriela Csurka. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023.

[33] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. Learn. Represent.*, 2021.

[35] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models, 2024.

[36] Thomas Unterthiner, Sjoerd Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. 2018.

[37] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope Text-to-Video Technical Report. *arXiv preprint arXiv:2308.06571*, 2023.

[38] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *Proc. Eur. Conf. Comput. Vis.*, 2024.

[39] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Self-supervised latent space optimization with nebula variational coding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(3):1397–1411, March 2024.

[40] 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.

[41] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In Andres Burbano, Denis Zorin, and Wojciech Jarosz, editors, *SIGGRAPH*, 2024.

[42] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Trans. Vis. Comput. Graph.*, 2024.

[43] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

[44] 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 *ECCV*, 2024.

[45] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized Predictive Model for Autonomous Driving. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

[48] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Advancing 3d occupancy prediction in lidar-free environments. *IEEE Transactions on Image Processing*, 2025.

[49] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. In *AAAI Conf. Artif. Intell.*, volume 38, pages 7178–7186, 2024.

[50] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024.

[51] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. In *Proc. Int. Conf. Learn. Represent.*, 2024.

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2023.

[53] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation. *arXiv preprint arXiv:2403.06845*, 2024.

[54] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2drivex: A generalized mllm framework for scene-to-drive learning, 2025.

[55] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *Proc. Eur. Conf. Comput. Vis.*, pages 55–72. Springer, 2025.

[56] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model, 2025.