OPENLONGTAIL: GENERATIVE SCALING OF LONG-TAIL DRIVING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling robust driving policies is fundamentally bottlenecked by the scarcity of edge cases in curated datasets. While the real world continuously captures these critical events, such long-tail events remain underutilized when collected from heterogeneous sources. Specifically, diverse but valuable in-the-wild long-tail videos lack the full view coverage required for training policy models, often missing multi-view poses or originating solely from monocular dash cameras. This modality gap prevents these ubiquitous observations from being converted into scalable training data for long-tail generalization. We introduce ***OpenLongTail***, an open-source generative data engine for scaling autonomous driving policies under long-tail events. To transform heterogeneous data sources into view-aligned and temporally coherent multi-view assets that are useful for policy learning, we develop a pose-informed extrapolative view synthesis pipeline that generates the missing views. We further incorporate Plücker-ray camera conditioning and lane-weighted supervision to provide geometric guidance for missing-view synthesis while placing greater emphasis on sparse lane structures during training. By synthesizing heterogeneous long-tail data, we observe an effective improvement in closed-loop driving robustness in handling long-tail events. Using extrapolative view-synthesis and pose-recovery metrics, we validate OpenLongTail in terms of visual fidelity and ego-trajectory accuracy.

1 INTRODUCTION

Empowered by large-scale driving datasets (Sun et al., 2020; Ettinger et al., 2021; Caesar et al., 2020; Xu et al., 2025), recent end-to-end driving policies have made substantial progress on common scenarios by learning end-to-end mappings from scene understanding to vehicle control under well-structured training conditions (Wang et al., 2025c; Zhou et al., 2025; 2026; Xu et al., 2024a; Yuan et al., 2025; Jiang et al., 2025b; Hwang et al., 2024; Jiang et al., 2024; Feng et al., 2025; Jiang et al., 2025a). However, real-world autonomy requires these policies to remain reliable not only in common scenarios but also in long-tail situations, where unusual obstacles and atypical environmental conditions can lead to safety-critical failures. Expanding the coverage of such situations through dedicated data collection remains challenging, as rare events involving animals on the road or unusual work zones are difficult to capture at scale with calibrated multi-camera rigs. Instead of relying solely on new data collection, we seek to expand long-tail training data by integrating driving logs already recorded across diverse sources. These include large-scale autonomous-driving repositories such as NVIDIA PhysicalAI Autonomous Vehicles (PAV) (NVIDIA Corporation, 2025), dashcam datasets such as Nexar (Moura et al., 2025), and individual dashcam videos recorded in the wild. However, recording an event does not necessarily provide the camera views required by a target multi-camera driving policy. These observations often follow incompatible annotation formats and provide missing or unreliable metric camera poses. This modality gap prevents common real-world long-tail videos from being directly converted into synchronized multi-view training assets, limiting the continued scaling of learned driving policies toward robust long-tail generalization.

Bridging this gap requires a conversion engine that turns unposed or monocular driving data into synchronized, pose-grounded assets compatible with state-of-the-art models (Zhou et al., 2025; 2026; Gao et al., 2026). Existing approaches fall short because the target output is a target-rig rollout whose side and rear cameras may observe regions absent from the input video. While recent video generation and camera-controlled synthesis methods have made strong progress in producing high-

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

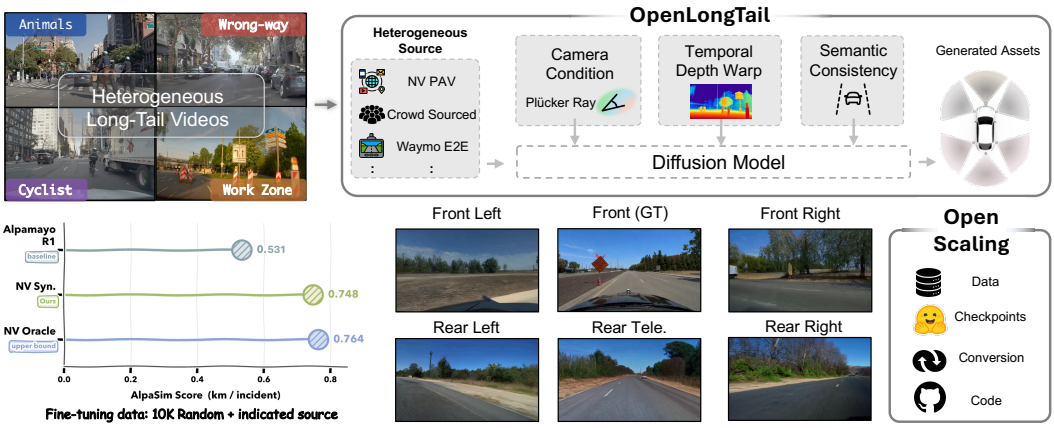

Figure 1: **OpenLongTail** transforms heterogeneous long-tail driving videos from diverse sources into pose-grounded, synchronized multi-view assets under a target camera rig. It conditions a diffusion model on Plücker-ray and geometry reprojection to synthesize extrapolative views beyond the observed front camera. The lower-left plot shows that fine-tuning with our synthesized long-tail data substantially improves closed-loop driving performance and approaches training with ground-truth multi-view data. To support open scaling, we release the generated data, model checkpoints, conversion tools, and code.

quality sequences under controllable poses (Yu et al., 2025; Bai et al., 2025; Ren et al., 2025b; Lin et al., 2026; Van Hoorick et al., 2026; Yu et al., 2024; Wu et al., 2025; Yang et al., 2026), they struggle when extrapolating into unobserved spaces. Specifically, they usually either lack cross-view consistency or rely on conditions like 3D bounding boxes and LiDAR, which are absent in heterogeneous videos (Gao et al., 2023; Wang et al., 2024a; Zhao et al., 2025; Ren et al., 2025a; Hu et al., 2023; Russell et al., 2025; Wang et al., 2024b). Explicit reconstruction with neural rendering (e.g., 3D Gaussian Splatting, 3DGS (Kerbl et al., 2023)) faces the limitation where those methods depend on sufficient view overlap and cross-view photometric consistency, making them brittle under narrow monocular coverage, dynamic objects, and capturing artifacts. 3DGS conversion also depends on accurate camera trajectories, which are hard to acquire from the video sequences in the wild.

To bridge this gap, we introduce **OpenLongTail** (Figure 1), an open-source generative data engine that unifies heterogeneous long-tail driving videos into spatially aligned and temporally coherent multi-view assets for safer policy learning. OpenLongTail utilizes the generalization capability from Geometry Foundation Models (GFMs) to recover a metric-scale ego-trajectory from weakly annotated driving videos, then turns this recovered motion into geometry conditions that guide pose-informed extrapolative view synthesis via a generation engine. The generative diffusion model injects Plücker ray geometry to encode target-camera rays and uses temporal depth warping to propagate visible evidence from the observed front view into the target rig, enabling missing side and rear contexts to be synthesized under a shared camera frame. Ultimately, OpenLongTail transforms the dash-cam driving logs into robust, posed multi-view assets. Through pose-recovery benchmarks, extrapolative view synthesis, and closed-loop policy evaluation, we demonstrate that OpenLongTail significantly improves driving robustness during long-tail events and enhances visual fidelity and ego-trajectory accuracy. OpenLongTail introduces these major contributions:

1. We introduce OpenLongTail, an open-scaling generative data engine that combines metric ego-trajectory recovery and extrapolative view synthesis to convert heterogeneous long-tail videos into pose-grounded multi-view data for VLA policy learning.

2. We propose a pose-informed extrapolative synthesis pipeline that combines front-anchored side-view generation with masked joint rear-view denoising, integrating Plücker ray conditioning and lane supervision to provide geometric guidance and prioritize structure.

3. We evaluate synthesis quality across all five target views using paired-reference metrics and targeted ablations, and demonstrate the downstream utility of front-view-derived assets incorporating synthesized side views through closed-loop evaluation with driving policy.

2 Related Work

Vision-Language-Action Models for Autonomous Driving. Recent vision-language-action (VLA) models unify perception, language-conditioned reasoning, and end-to-end planning (Wang et al., 2025c; Zhou et al., 2025; 2026; Li et al., 2025a; Sima et al., 2024; Shao et al., 2024; Xu et al., 2024b; Wang et al., 2023; Ma et al., 2024), increasingly leveraging reinforcement learning, diffusion, and explainable policies (Jiang et al., 2025b; Hwang et al., 2024; Jiang et al., 2024; Feng et al., 2025; Jiang et al., 2025a; Rowe et al., 2025; Song et al., 2025). Extending multi-camera policies to long-tail events requires compatible training data that meets their geometric and temporal input requirements. Standard benchmarks provide rich sensor logs (Sun et al., 2020; Caesar et al., 2020; Wilson et al., 2023; Mao et al., 2021; Yu et al., 2020; Caesar et al., 2021) and targeted evaluations for long-tail corner cases (Li et al., 2022; Xu et al., 2025; Tian et al., 2024; Xie et al., 2025), but heterogeneous, unposed monocular videos from dashcams or the web lack the synchronized multi-view observations and pose information available in calibrated multi-camera logs. OpenLongTail addresses this data-format mismatch, converting unstructured observations into synchronized, pose-grounded surround-view assets for downstream policy learning.

Controllable Video Generation. Recent world foundation models provide strong generative priors for visual synthesis (Ali et al., 2025; Wan et al., 2025; Yang et al., 2024; Chen et al., 2025). Driving-oriented methods use these for multi-view, long-horizon, and closed-loop simulation (Swerdlow et al., 2024; Yang et al., 2023; Li et al., 2024; Ren et al., 2025a; Gao et al., 2023; Wen et al., 2024; Lu et al., 2024b; Gao et al., 2024; Wang et al., 2025b; Hu et al., 2024; Gao et al., 2025; Fu et al., 2024; Guo et al., 2024; Li et al., 2026b), often conditioned on 3D layouts or occupancy (Yang et al., 2025; Lu et al., 2025). Dream4Drive (Zeng et al., 2026) constructs corner cases through 3D asset insertion into existing multi-view videos, demonstrating downstream perception benefits, whereas we repurpose naturally captured events. Camera-controlled generation synthesizes novel viewpoints from sparse observations (Yu et al., 2025; Bai et al., 2025; Yang et al., 2026; Chen et al., 2024; Lu et al., 2024a; Yao et al., 2024; Guo et al., 2025); extending this task to surround-view rigs requires generating substantial regions unobserved in the source view. More closely related, Sensor2Sensor (Wang et al., 2026) converts monocular driving videos into multi-view camera streams and LiDAR, training on 4DGS-rendered dashcam views paired with real AV logs. Sharing this data-repurposing goal, OpenLongTail couples metric ego-trajectory recovery with pose-conditioned view extrapolation to produce synchronized target-rig assets from naturally occurring long-tail videos, evaluating their training value in closed-loop driving.

3 Methodology

OpenLongTail synthesizes synchronized target-rig videos from monocular front-view input, using camera-ray conditions derived from stabilized metric ego-trajectories and rig calibration as shown in Figure 2. Cross-left/right extrapolation is anchored by geometric reprojection and front-view appearance. The three rear views, which overlap with each other but not the same-time front view, are jointly denoised under front-view conditioning via masked cross-view attention. To prioritize safety-critical road structure in the generated data, we guide rear-view synthesis with road-surface priors and emphasize sparse lane markings through a lane-weighted training objective.

3.1 Metric-Scale Ego-Trajectory Recovery

To construct pose-aligned multi-view assets, we extract a metric-scale camera trajectory from an input video $\mathcal{V}_{1:T}$ using MapAnything Keetha et al. (2025):

$$\hat{\mathcal{T}}_{1:T} = \Phi_{\mathrm{MA}}(\mathcal{V}_{1:T}) = \{\hat{\mathbf{T}}_t\}_{t=1}^{T} \tag{1}$$

where each pose $\hat{\mathbf{T}}_t \in SE(3)$. Because frame-level poses often contain high-frequency fluctuations that destabilize view synthesis, we refine them through a motion-consistent stabilization stage:

$$\tilde{\mathcal{T}}_{1:T} = \Psi_{\mathrm{RTS}}\Big(\Psi_{\mathrm{KF}}\Big(\hat{\mathcal{T}}_{1:T}\Big)\Big). \tag{2}$$

Specifically, a forward Kalman filter (Ψ_{KF}) suppresses local noise, while a backward Rauch-Tung-Striebel smoother (Ψ_{RTS}) uses the complete clip to stabilize the global path. The output trajectory $\tilde{\mathcal{T}}_{1:T}$ preserves metric-scale ego-motion while reducing jitter, yielding smoother pose conditions.

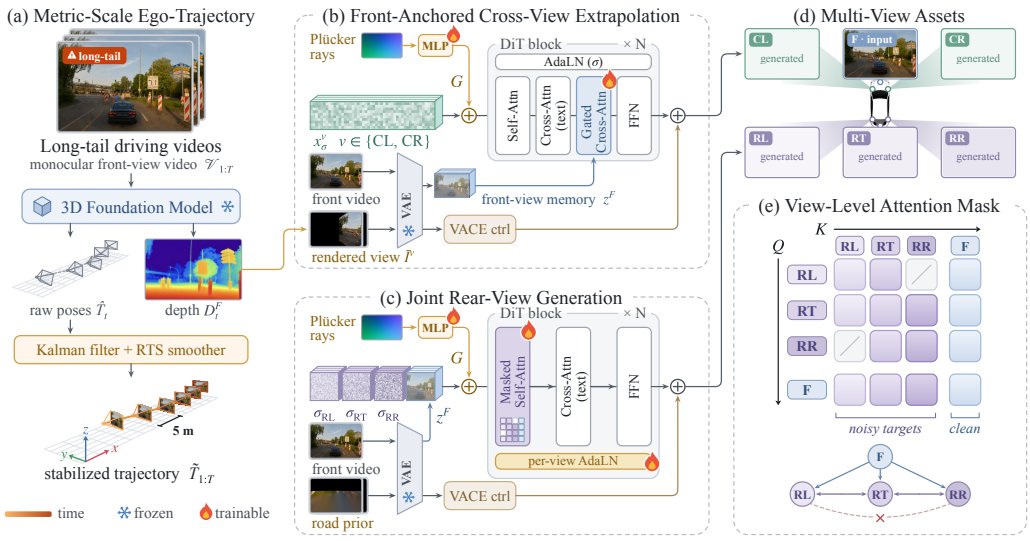

Figure 2: **Overview of OpenLongTail.** Given a long-tail front-view driving video, OpenLongTail recovers a metric-scale ego-trajectory and synthesizes five target views under a target camera rig. Cross-left (CL) and cross-right (CR) views are generated with front-anchored view extrapolation, while three rear views are jointly generated with masked view-level attention and road-surface guidance. Together with the input front view, the outputs form synchronized, posed multi-view assets.

3.2 Front-Anchored Cross-View Extrapolation

We synthesize the cross-left and cross-right videos by combining explicit geometric view transformation with source-conditioned generation. The partial overlap with the front camera provides reprojectable observations, while direct access to the original front video supplies appearance context beyond the rendered input.

Geometry-guided view transformation. Following geometry-conditioned video synthesis (Yu et al., 2025; Lin et al., 2026), we construct target-camera renderings from the front video. Given front frames I_t^F, estimated depths D_t^F, and camera calibration, we lift source pixels into 3D and render them from each target camera $v \in \{\mathrm{CL}, \mathrm{CR}\}$:

$$(\bar{I}_t^v, M_t^v) = \mathcal{R}\big(\mathcal{U}(I_t^F, D_t^F, K_F); K_v, E_v^{-1} E_F\big)\,, \tag{3}$$

where $\mathcal{U}$ denotes depth unprojection, $\mathcal{R}$ performs z-buffered forward splatting, K_v denotes camera intrinsics, and E_v maps camera coordinates to the rig frame. The rendering $\bar{I}_t^v$ places source appearance at target-camera locations under the estimated geometry, while M_t^v marks valid projections. Although incomplete, these renderings provide spatially localized visual conditions for the target viewpoint. We encode the rendered video with the frozen VAE and supply its latents and masks to the VACE control branch (Jiang et al., 2025c), which injects residual features into the denoiser.

Source-conditioned denoising. To retain source information absent from the renderings, we construct dense memory from the clean front-view latent and compact memory through learned-query resampling. Six gated cross-attention modules allow each target latent frame to query both memories within a local temporal neighborhood. Relative-pose and frame-offset embeddings identify the spatial and temporal origins of memory tokens, while noise-dependent gates modulate their residual contributions. Source and target streams remain separate batch items, so backbone self-attention operates within each view. To specify camera geometry, we encode each token's Plücker ray $[d;\, o \times d]$, where o and d denote the camera center and ray direction in a common ego-anchor frame computed using the recovered poses. Their MLP embeddings are added to video tokens together with camera-identity, source/target-role, and ego-motion embeddings. Each side view is sampled separately under visual, geometric, and text conditions, keeping the front latent clean throughout denoising.

3.3 JOINT REAR-VIEW GENERATION

We jointly generate the rear-left (RL), rear-tele (RT), and rear-right (RR) videos conditioned on the observed front video and camera geometry, allowing the target views to exchange their evolving representations during denoising. Unlike the side cameras, the rear cameras have no same-time field-of-view overlap with the front camera in our rig, precluding the direct same-frame reprojection used in Sec. 3.2. However, the rear cameras observe overlapping regions of the same scene. We therefore model dependencies among the rear targets during generation, instead of synthesizing each view separately from common source conditions.

Joint multi-view denoising. Let $x^v_{\sigma_v}$ denote the noisy latent of rear view $v \in \{\mathrm{RL}, \mathrm{RT}, \mathrm{RR}\}$ at noise level σ_v, and let z^F denote the clean front-view latent. We concatenate the three noisy targets and the clean observation along the token-sequence dimension:

$$H^0 = \mathrm{Concat}_{\mathrm{seq}}\big[\phi(x^{\mathrm{RL}}_{\sigma_{\mathrm{RL}}}), \phi(x^{\mathrm{RT}}_{\sigma_{\mathrm{RT}}}), \phi(x^{\mathrm{RR}}_{\sigma_{\mathrm{RR}}}), \phi(z^F)\big] + G, \tag{4}$$

where ϕ denotes patch embedding and G collects the Plücker-ray, camera-identity, source/target-role, and ego-motion embeddings introduced in Sec. 3.2. In contrast to the separate-batch organization used for side-view synthesis, sequence concatenation allows the backbone self-attention to exchange information across views.

A view-level attention mask specifies the permitted cross-view interactions. Within the rear triplet, rear-tele exchanges information bidirectionally with rear-left and rear-right, while direct attention between rear-left and rear-right is blocked in both directions. For each permitted connection, queries can access the connected view's full spatiotemporal token sequence, not just tokens at the same timestamp. Each lateral rear stream can thus read the evolving rear-tele representation, while rear-tele reads both lateral streams. Across layers, this connectivity also permits indirect information exchange between the lateral rear views through rear-tele.

To distinguish noisy targets from clean context, we apply view-specific noise conditioning $(\sigma_{\mathrm{RL}}, \sigma_{\mathrm{RT}}, \sigma_{\mathrm{RR}}, 0)$, with additional target-pose modulation for the rear streams. The modulation parameters are broadcast over the tokens within each view. Together with the token-level geometry embeddings, this allows the streams to interact while retaining their camera identities and noise states. We fine-tune the backbone self-attention weights to process this masked joint sequence and update all three rear latents together at each sampling step. The front input latent remains clean and is not updated by the sampler, serving as a fixed conditioning source throughout denoising.

Road-surface conditioning. We construct a road-surface prior P^v and confidence map Q^v for each rear view by intersecting target-camera rays with a ground-plane model and assigning appearance from earlier observations or front-view pixels mirrored along the lane direction. The mirrored content serves as an appearance prior instead of a recovered observation. We encode the prior videos with the frozen VAE and supply their latents and confidence maps to the VACE control branch, whose residual features are injected into the joint denoiser to provide spatial guidance.

3.4 TRAINING OBJECTIVE

Both generation stages use latent-space flow matching, with unweighted supervision for side views and lane-weighted supervision for rear views. Under the visual, geometric, and text conditions defined above, the corresponding denoiser predicts $\hat{u}^v$ from a noisy target latent $x^v_{\sigma_v}$. Given a ground-truth target latent z^v, we sample $\epsilon^v \sim \mathcal{N}(0, I)$ and a noise level σ_v to define the noisy latent and target velocity as

$$x^v_{\sigma_v} = (1 - \sigma_v) z^v + \sigma_v \epsilon^v, \qquad u^v = \epsilon^v - z^v. \tag{5}$$

Side-view supervision. Each training step selects either cross-left or cross-right, with balanced frequency across steps. We supervise only the selected target stream using unweighted mean-squared error:

$$\mathcal{L}_{\mathrm{side}} = \frac{1}{BC_z|\Omega|} \sum_{b=1}^{B} \sum_{p \in \Omega} \|\hat{u}_{b,v,p} - u_{b,v,p}\|_2^2, \tag{6}$$

where B is the batch size, Ω is the spatiotemporal VAE latent grid, and $C_z = 16$ is the channel count. The squared norm sums over channels; front-stream outputs do not contribute to the loss.

Table 1: **Extrapolative view evaluation on held-out long-tail clips.** We compare OpenLongTail with novel trajectory video synthesis baselines on held-out clips with paired target-view references, reporting visual fidelity and semantic consistency across all five target views, with FID/FVD computed from pooled target-view feature distributions.

Method	Visual Fidelity			Semantic Consistency				Distribution	
	PSNR↑	SSIM↑	LPIPS↓	compL1↓	mIoU↑	Road IoU↑	Lane IoU↑	FID↓	FVD↓
TrajectoryCrafter (Yu et al., 2025)	10.49	0.4962	0.6535	60.46	0.221	0.517	0.039	114.22	1829.3
Gen3C (Ren et al., 2025b)	13.90	0.5114	0.6282	53.90	0.191	0.322	0.033	63.71	865.1
ReCamMaster (Bai et al., 2025)	13.37	0.4898	0.6421	57.61	0.160	0.277	0.019	60.41	716.5
Vista4D (Lin et al., 2026)	12.92	0.3958	0.6514	51.81	0.199	0.225	0.048	91.67	844.5
AnyView (Van Hoorick et al., 2026)	12.56	0.4643	0.6493	69.27	0.190	0.363	0.024	74.19	765.2
Ours	**14.98**	**0.5198**	**0.5702**	**33.13**	**0.340**	**0.654**	**0.229**	**32.92**	**254.1**

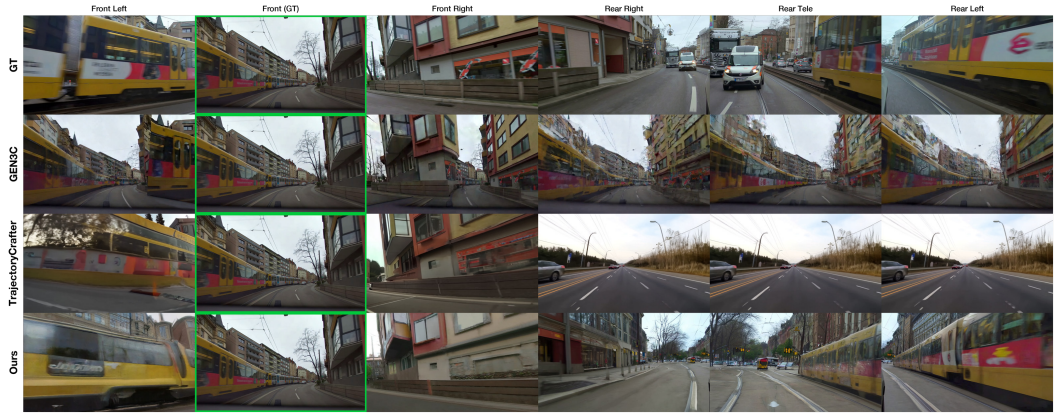

Figure 3: **Qualitative comparison of generated multi-view driving videos.** We compare OpenLongTail with Gen3C and TrajectoryCrafter. Our model produces synchronized target views conditioned on the front camera and target camera geometry, preserving scene layout and semantic consistency across side and rear cameras. Please refer to the suppl. for more qualitative results.

Lane-weighted rear-view supervision. For rear views, we increase the relative supervision assigned to sparse lane markings. We apply Mask2Former Cheng et al. (2022) to one representative RGB frame per latent timestep, decoded from the ground-truth rear latents. We retain the *Lane Marking* class, excluding crosswalks, and spatially max-pool the binary segmentations to the latent grid. A location is marked whenever its pooling region contains a lane-marking pixel.

Let $m_{b,v,p} \in \{0,1\}$ denote the resulting mask and $w_{b,v,p} = 1 + 4m_{b,v,p}$. We optimize

$$\mathcal{L}_{\text{rear}} = \frac{1}{|\mathcal{A}|} \sum_{(b,v)\in\mathcal{A}} \frac{\sum_{p\in\Omega} w_{b,v,p} \, \|\hat{u}_{b,v,p} - u_{b,v,p}\|_2^2}{C_z \sum_{p\in\Omega} w_{b,v,p}}, \tag{7}$$

where $\mathcal{A}$ contains the rear example–view pairs with $\sigma_{b,v} > 0$, excluding views supplied as clean context. Marked locations receive five times the relative weight across all channels. The weighted error is normalized within each view before averaging over supervised views. Lane masks are used exclusively to weight the training loss, not as generation conditions.

4 Evaluation

We evaluate OpenLongTail as a generative data engine for scaling autonomous driving policies with heterogeneous long-tail videos. We first assess the visual fidelity of multi-view videos produced by pose-informed extrapolative view synthesis, then examine the downstream utility of front-view-derived assets in closed-loop long-tail driving. Finally, we evaluate metric-scale ego-trajectory recovery for pose conditioning and ablate the synthesis components.



Figure 4: **Qualitative results on Nexar in-the-wild crowd-sourced dashcams.** We apply OpenLongTail to Nexar videos. Given only the observed monocular front view, our model recovers the ego-trajectory and synthesizes synchronized extrapolative views under the target rig, preserving the overall scene layout and semantic consistency.

4.1 Experimental Setup and Implementation

Data. We curate over 200K driving video clips from approximately 50K scenes, combining NVIDIA PhysicalAI-Autonomous-Vehicles (NVIDIA Corporation, 2025) with PandaSet (Xiao et al., 2021) and nuScenes (Caesar et al., 2020). We fine-tune Wan2.1-VACE-1.3B as our video generation backbone on 32 NVIDIA H200 GPUs for approximately 96 hours. We will release our code and model checkpoints to support open, community-driven expansion of long-tail driving data. **Baselines.** Many driving world models rely on structured scene conditions or focus on future prediction (Yang et al., 2025; Hu et al., 2024), whereas our task is to synthesize missing views of an observed monocular video at the input timestamps. We therefore compare against general camera-controlled novel-view video synthesis methods. Sensor2Sensor is excluded because its code was not publicly available at the time of evaluation. Detailed baseline configurations and evaluation protocols are provided in the supplementary material.

4.2 Pose-informed Extrapolative View Evaluation

We evaluate missing-view synthesis on held-out NVIDIA long-tail videos and on external videos from Waymo E2E (Xu et al., 2025) and Nexar (Moura et al., 2025). Given a monocular front-view video, our pipeline estimates a metric-scale ego-trajectory and combines it with the target camera rig to construct pose conditions. Conditioned on the input video and these estimated poses, OpenLongTail synthesizes the missing target views at the input timestamps. The target rig remains fixed across sources to match a common downstream policy configuration. Please refer to the Appendix for detailed and diverse qualitative results

Using 80 held-out NVIDIA long-tail clips with paired target-view references, we evaluate synthesis quality through complementary metrics; reference construction and implementation details are provided in the supplementary material. As shown in Figure 3 and Table 1, we measure frame-wise *fidelity* against reference frames using PSNR and SSIM (Wang et al., 2004), complemented by LPIPS (Zhang et al., 2018). For *semantic agreement*, we segment generated and reference frames using Mask2Former (Cheng et al., 2022). We measure composition differences irrespective of spatial arrangement using *compL1*, defined as the summed absolute difference in pixel shares across nine semantic groups. Spatial agreement is measured using *mIoU*, averaged over groups occupying more than 0.2% of the pooled reference pixels across the 11 sampled frames of each clip-view pair, with separate IoUs reported for *road* and *lane markings*. We measure *distributional similarity* using FID (Heusel et al., 2017) at the frame level and FVD (Unterthiner et al., 2018) at the video level to compare generated and real target-view feature distributions. We also present qualitative results on Waymo E2E and Nexar using the same generation protocol. Figure 4 compares OpenLongTail with Gen3C Ren et al. (2025b), showing higher visual quality and semantic consistency for our method in the displayed examples.

4.3 Downstream Closed-Loop Policy Performance

Following the view-synthesis evaluation, we examine whether converting monocular long-tail observations into multi-camera training assets improves closed-loop performance beyond nominal-data training. In preliminary experiments, we observed VLA performance degradation when fine-tuning on limited single-source data. We therefore retain 10K randomly sampled nominal driving trajecto-

Table 2: **Closed-loop AlpaSim evaluation in long-tail events.** We report AlpaSim Score (AS) and Collision Rate (CR) across four representative long-tail categories and the overall average. SFT denotes supervised fine-tuning; 10K Rand denotes 10K randomly sampled training trajectories; NV-OOD denotes the NVIDIA ood subset; GT and Syn denote ground-truth and OpenLongTail synthesized multi-view data, respectively. Complex Int. and Uncommon Veh. denote complex intersections and uncommon vehicles. The best and second-best results are shown in **bold** and underlined. AS is reported as the mean with a 95% percentile bootstrap confidence interval.

Model	Additional Training Data					Representative Long-tail Events								Avg. ($N = 106$)	
	10K Rand	NV-OOD		Waymo-E2E		Complex Int.		Cyclists		Uncommon Veh.		Work Zone			
		GT	Syn	GT	Syn	AS↑	CR↓	AS↑	CR↓	AS↑	CR↓	AS↑	CR↓	AS↑	CR↓
Alpamayo R1 (Wang et al., 2025c)						0.457 ± 0.028	69.6%	0.534 ± 0.040	50.0%	0.575 ± 0.027	50.0%	0.683 ± 0.026	50.0%	0.531 ± 0.026	58.5%
Alpamayo 1.5 (Wang et al., 2025c)						0.469 ± 0.030	73.9%	0.347 ± 0.077	100.0%	0.517 ± 0.031	33.3%	0.731 ± 0.022	50.0%	0.501 ± 0.032	60.4%
Alpamayo R1 + SFT (Wang et al., 2025c)	✓					0.659 ± 0.018	26.1%	0.670 ± 0.035	**0.0%**	0.733 ± 0.017	**0.0%**	0.936 ± 0.012	**0.0%**	0.717 ± 0.025	11.3%
	✓	✓				0.743 ± 0.014	**4.3%**	0.688 ± 0.022	**0.0%**	0.758 ± 0.014	**0.0%**	0.938 ± 0.010	**0.0%**	$\mathbf{0.764 \pm 0.019}$	**1.9%**
	✓		✓			$\mathbf{0.747 \pm 0.014}$	6.5%	0.662 ± 0.027	**0.0%**	0.717 ± 0.014	2.8%	0.935 ± 0.012	**0.0%**	0.748 ± 0.021	3.8%
	✓	✓		✓		0.691 ± 0.018	13.0%	$\mathbf{0.691 \pm 0.025}$	**0.0%**	0.735 ± 0.014	**0.0%**	$\mathbf{0.958 \pm 0.007}$	**0.0%**	0.736 ± 0.024	5.7%
	✓		✓		✓	0.699 ± 0.017	13.0%	0.689 ± 0.027	**0.0%**	$\mathbf{0.765 \pm 0.014}$	**0.0%**	0.950 ± 0.009	**0.0%**	0.749 ± 0.023	5.7%

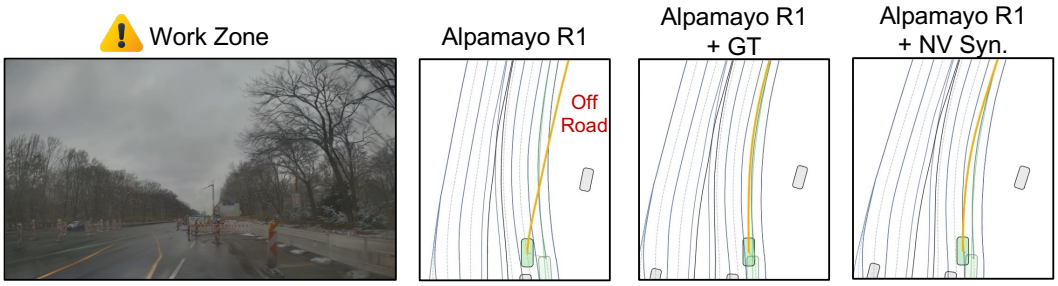

Figure 5: **Qualitative closed-loop evaluation in a work-zone long-tail event.** Left: front-view observation of a rare work-zone scene. Right: BEV rollouts of Alpamayo R1 under different long-tail training data, where NV denotes NVIDIA LT data.

ries as the main training pool and incorporate smaller sets of ground-truth or synthesized long-tail examples as targeted supplements rather than replacements. Through resampling, we match the number of training examples processed and optimization updates across SFT recipes.

We retain the four-camera input interface of Alpamayo-R1 (Wang et al., 2025c) across SFT recipes. For synthesized assets, we retain the original front-view video and obtain front-tele through a deterministic FoV-Z crop with a 30° field of view. OpenLongTail synthesizes CL/CR under the target camera rig to supply the remaining policy inputs. Recovered ego-motion provides both generation conditioning and ego-motion inputs for policy training. The three rear views are not policy inputs; we assess them using paired synthesis metrics and dedicated ablations (Tables 1 and 4(b)).

NVIDIA recordings with paired multi-view ground truth provide a real-data reference for assessing the training value of assets constructed from front-view inputs. Given the limited availability of real-world long-tail events, we conduct a targeted closed-loop comparison on a common set of 53 events in AlpaSim (NVIDIA et al., 2025), with evaluation scenes excluded from policy training. At each step, AlpaSim renders multi-camera observations from the ego state updated by policy actions, testing behavior under covariate shift and error accumulation. For each policy and event, we run two independent rollouts and report average metrics. The values in Table 2 show average AS increasing from 0.531 to 0.717 after nominal-only SFT. Adding NVIDIA synthesized long-tail assets (NV Syn) raises the reported AS to 0.748, compared with 0.764 for ground-truth long-tail assets (NV GT), while the reported collision rate decreases from 11.3% to 3.8%. As shown in Figure 5, the policy fine-tuned with NV Syn remains on the road in the depicted work-zone event, exhibiting behavior similar to the policy trained with NV GT, while the original Alpamayo R1 deviates off-road.

Scaling to external sources. We next extend this asset-construction process to external monocular videos while retaining the same nominal training pool and SFT budgets. Adding assets synthesized from Waymo E2E (Xu et al., 2025) front-view videos (NV + Waymo Syn) leaves the reported average AS nearly unchanged (0.748 to 0.749), while the reported collision rate increases from 3.8% to 5.7%. AS increases for uncommon vehicles (0.717→0.765) and cyclists (0.662→0.689). The work-zone increase (0.935→0.950) contrasts with a decline at complex intersections (0.747→0.699). The

Table 3: **Pose evaluation under metric-scale and Sim(3)-aligned protocols.** Metric-scale evaluation preserves absolute scale; Sim(3) alignment removes global pose and scale before error computation. Light-gray cells denote methods that are not metric-scale.

Method	Metric-scale Evaluation							Sim(3)-aligned Evaluation					
	ATE↓	RPE-r↓	RPE-t↓	RotErr↓	Jerk↓	Acc.Var.↓	LDJ↑	ATE↓	RotAnc↓	RotErr↓	TrErr↓	Jerk↓	Acc.Var.↓
DROID-W (Li et al., 2026a)	—	—	—	—	—	—	—	0.284	0.861	0.217	12.99	4570.6	119.60
MegaSAM (Li et al., 2025b)	—	—	—	—	—	—	—	0.444	0.456	**0.129**	15.57	3704.4	80.69
VGGT (Wang et al., 2025a)	—	—	—	—	—	—	—	0.202	0.393	0.134	13.54	2947.7	61.19
ViPE (Huang et al., 2025)	3.831	0.137	59.47	0.472	929.5	15.51	-9.18	**0.162**	0.472	0.137	**9.86**	1055.5	18.44
MapAnything (Keetha et al., 2025)	**2.212**	0.155	23.52	0.352	4737.4	53.22	-12.65	0.171	0.352	0.155	12.57	4243.9	47.52
Ours	**2.212**	**0.129**	**23.05**	**0.340**	**283.9**	**5.96**	**-7.48**	**0.162**	**0.340**	**0.129**	11.24	**254.5**	**5.33**

Table 4: **Ablation on cross-view extrapolation and joint rear-view generation.** (a) Stage 1: geometric conditioning and pose smoothing. (b) Stage 2: masked self-attention, road-surface conditioning, and lane-weighted supervision.

(a) Stage 1

	Avg. Generation		Closed-loop	
Variant	PSNR↑	LPIPS↓	AlpaSim↑	Coll.↓
w/o geo. condition	10.12	0.80	0.327	51.9%
w/o pose smoothing	14.02	**0.61**	0.738	11.3%
Ours	**14.32**	0.63	**0.748**	**3.8%**

(b) Stage 2: rear views

Variant	PSNR↑	LPIPS↓	Lane IoU↑	FVD↓
w/o masked self-attn	12.51	0.644	0.115	735.7
w/o road-surface conditioning	13.26	0.608	0.103	540.2
w/o lane-weighted loss	14.81	0.607	0.051	551.4
Ours	**15.42**	**0.530**	**0.215**	**260.7**

corresponding ground-truth mixture lowers average AS from 0.764 to 0.736, suggesting that external augmentation does not consistently improve aggregate performance. These mixed effects may reflect distribution differences between the NVIDIA and Waymo subsets (see appendix). OpenLongTail thus broadens the sources usable for multi-camera policy training, while the displayed results suggest category-dependent effects as opposed to uniform gains from external augmentation.

4.4 Ablation Study

Metric-scale pose recovery. We benchmark pose estimators on 218 clips from 109 scenes (Table 3). MapAnything with standard KF/RTS smoothing matches the lowest metric-scale ATE while yielding the lowest jerk and acceleration variance among the evaluated configurations. We therefore use this combination for pose recovery in our pipeline.

Stage 1. Table 4(a) examines geometric conditioning and pose smoothing. Removing geometric conditioning lowers PSNR from 14.32 to 10.12 and AlpaSim Score from 0.748 to 0.327, while the reported collision rate increases from 3.8% to 51.9%. Without pose smoothing, AlpaSim Score decreases to 0.738 and the reported collision rate rises to 11.3%, despite a slightly lower LPIPS (0.61 vs. 0.63). **Stage 2.** In Table 4(b), the full model achieves the best reported results across all four rear-view metrics. Removing masked self-attention lowers PSNR from 15.42 to 12.51 and raises FVD from 260.7 to 735.7. Lane IoU decreases from 0.215 to 0.103 without road-surface conditioning and to 0.051 without lane-weighted supervision. The Stage 1 ablations assess synthesis quality and downstream policy performance, whereas the Stage 2 ablations assess rear-view visual quality and segmentation-based lane agreement.

5 Conclusion and Limitations

Conclusion. We present OpenLongTail, an open-scaling framework that transforms heterogeneous, unposed monocular driving videos into synchronized, pose-grounded multi-view assets. Evaluations demonstrate strong visual fidelity and semantic consistency, alongside accurate ego-trajectory recovery. Training with these assets improves closed-loop policy performance on evaluated long-tail events in AlpaSim NVIDIA et al. (2025). Category-dependent gains from external sources further suggest that effective generative scaling depends not only on data *quantity*, but also on source-target *distribution* alignment. **Limitations.** Our framework remains limited by diffusion inference cost and rig-specific camera setup. More fundamentally, front-view evidence cannot uniquely determine unseen rear-view content: visually plausible generations may not reflect the actual scene, and their correctness is difficult to verify without corresponding rear-view ground truth.

AI USE STATEMENT

We used generative AI tools to support manuscript preparation and revision, including draft refinement, language and presentation editing, figure refinement, and code editing and debugging. These tools also assisted with discussions of methodological choices, feedback on experimental design, and the interpretation and presentation of experimental results. The authors reviewed all AI-assisted material and revised it as necessary, retaining final judgment over the research methods, analyses, and conclusions.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9710–9719, 2021.

Bowen Feng, Zhiting Mei, Julian Ost, Filippo Ghilotti, Baiang Li, Roger Girgis, Anirudha Majumdar, and Felix Heide. Verdi: Vlm-embedded reasoning for autonomous driving. *arXiv preprint arXiv:2505.15925*, 2025.

Yongjie Fu, Anmol Jain, Xu Chen, Zhaobin Mo, and Xuan Di. Drivegenvlm: Real-world video generation for vision language model based autonomous driving. In *2024 IEEE International automated vehicle validation conference (IAVVC)*, pp. 1–6. IEEE, 2024.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

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

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

Tian Gao, Celine Tan, Catherine Glossop, Timothy Gao, Jiankai Sun, Kyle Stachowicz, Shirley Wu, Oier Mees, Dorsa Sadigh, Sergey Levine, et al. Steervla: Steering vision-language-action models in long-tail driving scenarios. *arXiv preprint arXiv:2602.08440*, 2026.

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

Xi Guo, Chenjing Ding, Haoxuan Dou, Xin Zhang, Weixuan Tang, and Wei Wu. Infinitydrive: Breaking time limits in driving world models. *arXiv preprint arXiv:2412.01522*, 2024.

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

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. *arXiv preprint arXiv:2309.17080*, 2023.

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

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang, Yunda Dong, et al. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025a.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025b.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025c.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

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

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022.

Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. Droid-slam in the wild. *arXiv preprint arXiv:2603.19076*, 2026a.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pp. 469–485. Springer, 2024.

Yaoru Li, Federico Landi, Marco Godi, Xin Jin, Ruiju Fu, Yufei Ma, Muyang Sun, Heyu Si, and Qi Guo. Far-drive: Frame-autoregressive video generation in closed-loop autonomous driving. *arXiv preprint arXiv:2603.14938*, 2026b.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025a.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025b.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hannan Lu, Xiaohe Wu, Shudong Wang, Xiameng Qin, Xinyu Zhang, Junyu Han, Wangmeng Zuo, and Ji Tao. Seeing beyond views: Multi-view driving scene video generation with holistic attention. *arXiv preprint arXiv:2412.03520*, 2024a.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European conference on computer vision*, pp. 329–345. Springer, 2024b.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27272–27283, 2025.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pp. 403–420. Springer, 2024.

Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

Daniel Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2608–2616, 2025.

NVIDIA, Yulong Cao, Riccardo de Lutio, Sanja Fidler, Guillermo Garcia Cobo, Zan Gojcic, Maximilian Igl, Boris Ivanovic, Peter Karkus, Janick Martinez Esturo, Marco Pavone, Aaron Smith, Ellie Tanimura, Michal Tyszkiewicz, Michael Watson, Qi Wu, and Le Zhang. Alpasim: A modular, lightweight, and data-driven research simulator for autonomous driving, October 2025. URL `https://github.com/NVlabs/alpasim`.

NVIDIA Corporation. PhysicalAI-Autonomous-Vehicles. `https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles`, 2025.

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

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

Luke Rowe, Rodrigue de Schaetzen, Roger Girgis, Christopher Pal, and Liam Paull. Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. *arXiv preprint arXiv:2506.11234*, 2025.

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. *arXiv preprint arXiv:2503.20523*, 2025.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15120–15130, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Nan Song, Bozhou Zhang, Xiatian Zhu, Jiankang Deng, and Li Zhang. Lmad: Integrated end-to-end vision-language model for explainable autonomous driving. *arXiv preprint arXiv:2508.12404*, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4):3578–3585, 2024.

Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv preprint arXiv:2407.00959*, 2024.

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

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

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

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

Xiaodong Wang, Zhirong Wu, and Peixi Peng. Longdwm: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546*, 2025b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025c.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 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, 2024b.

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

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

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pp. 3095–3101. IEEE, 2021.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6585–6597, 2025.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025.

Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024a.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024b.

Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023.

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

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. *arXiv preprint arXiv:2512.12751*, 2025.

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.

Yining Yao, Xi Guo, Chenjing Ding, and Wei Wu. Mygo: Consistent and controllable multi-view driving video generation with camera control. *arXiv preprint arXiv:2409.06189*, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 100–111, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.

Kai Zeng, Zhanqian Wu, Kaixin Xiong, Xiaobao Wei, Xiangyu Guo, Zhenxin Zhu, Kalok Ho, Lijun Zhou, Bohan Zeng, Ming Lu, et al. Rethinking driving world model as synthetic data generator for perception tasks. In *International Conference on Learning Representations*, volume 2026, pp. 135507–135534, 2026.

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

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13782–13790, 2026.

Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025.