# TOWARD PHYSICALLY CONSISTENT DRIVING VIDEO WORLD MODELS UNDER CHALLENGING TRAJECTORIES

**Anonymous authors**
Paper under double-blind review




Figure 1: Qualitative comparison of video generation under diverse trajectory conditions, with front-view results shown. Prior methods (e.g., DiST-4D) exhibit artifacts and geometric distortions under physically challenging trajectories, whereas **PhyGenesis** preserves physical consistency and high visual fidelity. Additional videos are provided in https://iclr3.github.io/PhyGenesis/.

## ABSTRACT

Video generation models have shown strong potential as world models for autonomous driving simulation. However, existing approaches are primarily trained on real-world driving datasets, which mostly contain natural and safe driving scenarios. As a result, current models often fail when conditioned on challenging or counterfactual trajectories—such as imperfect trajectories generated by simulators or planning systems—producing videos with severe physical inconsistencies and artifacts. To address this limitation, we propose **PhyGenesis**, a world model designed to generate driving videos with high visual fidelity and strong physical consistency. Our framework consists of two key components: (1) a *physical condition generator* that transforms potentially invalid trajectory inputs into physically plausible conditions, and (2) a *physics-enhanced video generator* that produces



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.

high-fidelity multi-view driving videos under these conditions. To effectively train these components, we construct a large-scale, physics-rich heterogeneous dataset. Specifically, in addition to real-world driving videos, we generate diverse challenging driving scenarios using the CARLA simulator, from which we derive supervision signals that guide the model to learn physically grounded dynamics under extreme conditions. This challenging-trajectory learning strategy enables trajectory correction and promotes physically consistent video generation. Extensive experiments demonstrate that PhyGenesis consistently outperforms state-of-the-art methods, especially on challenging trajectories. Our project page is available at: https://iclr3.github.io/PhyGenesis/.

## 1 INTRODUCTION

Video world models have recently emerged as a central paradigm for autonomous driving research (Bruce et al., 2024; Guo et al., 2025a; NVIDIA, 2025; Gao et al., 2024b; Hu et al., 2022; 2023), offering a scalable alternative to expensive real-world data collection and high-fidelity physical simulators. Recent driving world models (Gao et al., 2025; Zhao et al., 2025; Guo et al., 2025b; Wen et al., 2024; Chen et al., 2025; Zeng et al., 2025a) can synthesize high-fidelity multi-view future scenes while preserving controllability through structured conditions such as vehicle trajectories. These capabilities have enabled a variety of downstream applications, including closed-loop evaluation in simulation (Yang et al., 2025b; Yan et al., 2025), high-risk scenario synthesis (Zhou et al., 2025; Xu et al., 2025), and integration with end-to-end planners for decision making and motion forecasting (Zeng et al., 2025b; Shi et al., 2025; Xia et al., 2025; Li et al., 2025d).

Despite these advances, current driving world models struggle when deployed under *challenging trajectory conditions* produced by trajectory simulators, planning systems, or user interactions. We identify two fundamental limitations of existing approaches. **First, current models lack physical awareness of trajectory feasibility.** Trajectory conditions generated by simulators or planners can be imperfect and may violate fundamental physical constraints, such as vehicles penetrating other agents. However, existing models lack explicit physical reasoning and largely behave as condition-to-pixel translators. When forced to follow such physically inconsistent inputs, they often produce videos with severe rendering artifacts and structural failures. **Second, current models lack physics-consistent generation capability.** Most existing approaches (Gao et al., 2025; Guo et al., 2025b; Wen et al., 2024; Chen et al., 2025) are predominantly trained on real-world driving datasets dominated by safe and nominal behaviors. Consequently, they struggle to generate realistic dynamics in rare scenarios such as collisions or off-road departures—even when the trajectories themselves are physically feasible. Consequently, prior approaches (e.g., DiST-4D) often produce severe artifacts and physically inconsistent videos, as shown in Fig. 1.

In this work, we introduce **PhyGenesis**, a physics-aware driving world model designed to address both limitations. Our key insight is that physically consistent world modeling requires joint handling of trajectory feasibility and physics-consistent video generation. To this end, PhyGenesis introduces a novel module called the *Physical Condition Generator*, which transforms arbitrary trajectory conditions into physically consistent ones by resolving potential physical conflicts. The rectified conditions are then fed into a *Physics-enhanced Video Generator*, which synthesizes high-fidelity and physically consistent multi-view driving videos. To support this learning process, we construct a heterogeneous training dataset that combines real-world driving data with a physically challenging dataset generated using the CARLA simulator (Dosovitskiy et al., 2017). While real-world data provides abundant nominal driving behaviors, the CARLA-generated dataset introduces diverse extreme scenarios such as collisions and off-road departures. These events are uniquely informative, providing dense supervision for learning complex object–environment interactions—priors that are fundamentally scarce in routine real-world driving data. With these designs, PhyGenesis substantially outperforms prior methods, particularly under challenging trajectory conditions.

Our main contributions are summarized as follows:

- We propose **PhyGenesis**, a physics-aware driving world model for high-fidelity and physically consistent autonomous video generation. By explicitly handling both trajectory feasibility and physics-consistent video generation, PhyGenesis is the first framework capa-

ble of synthesizing physically consistent multi-view driving videos even when conditioned on initially physics-violating trajectory inputs.

- We introduce a *Physical Condition Generator* that converts arbitrary trajectory inputs into physically feasible 6-DoF vehicle motions. To enable this capability, we formulate a novel counterfactual trajectory rectification training task that equips the model with intrinsic physical priors for resolving physics-violating trajectories.
- We develop a *Physics-Enhanced Video Generator* and train these models on a heterogeneous physics-rich dataset combining real-world driving logs with physically extreme synthetic scenarios generated with the CARLA simulator. This hybrid training paradigm enables the model to learn complex object–environment interactions and significantly improves video generation under challenging trajectory conditions.

## 2 Related Work

**Nominal Driving World Models.** Driving video generation has advanced rapidly, with most methods using structured spatial priors for controllability. BEVGen (Swerdlow et al., 2024) encodes layouts via BEV maps but discards height information, limiting 3D representation. BEVControl (Yang et al., 2023) addresses this with height lifting. MagicDrive (Gao et al., 2024a) improves 3D-aware generation through geometric constraints and cross-view attention, while MagicDrive-V2 (Gao et al., 2025) adopts DiT for high-resolution synthesis. DiST-4D (Guo et al., 2025a) and WorldSplat (Zhu et al., 2025) use metric depth to lift videos into 4D representations for novel-view synthesis. Genesis (Guo et al., 2025b) and UniScene (Li et al., 2025a) target LiDAR–RGB generation via sequential DiT and occupancy-centric voxels. While these methods achieve high fidelity under routine driving, their reliance on nominal datasets limits robustness to challenging and/or physics-violating trajectory inputs.

**High-risk Driving Video Generation.** Generating high-risk driving scenarios has attracted growing attention. Early efforts such as AVD2 (Li et al., 2025b), DrivingGen (Guo et al., 2024), and Ctrl-Crash (Gosselin et al., 2025) synthesize accident scenarios from single-view dashcam footage; however, their data are single-view and low-quality, making it difficult to transfer trained models to high-fidelity, multi-view simulators. More recent methods, SafeMVDrive (Zhou et al., 2025) and Challenger (Xu et al., 2025), combine trajectory simulators with multi-view video generators for safety-critical videos. Nevertheless, their generators are trained only on nominal data, so quality remains limited and cannot generate physical interactions such as collisions.

In summary, existing driving world models handle either nominal scenarios or high-risk synthesis, but rarely both. **PhyGenesis** bridges this gap through a Physical Condition Generator and a Physics-Enhanced Video Generator trained on both real-world and simulation-derived extreme data.

## 3 Method

### 3.1 Overview of PhyGenesis

The framework of **PhyGenesis** is illustrated in Figure 2. As shown in (a), our framework is trained on a heterogeneous multi-view dataset (Section 3.2), which enables the model to learn both high visual fidelity and physical consistency. Given trajectory inputs, the *Physical Condition Generator* (Section 3.3) first rectifies potentially invalid trajectories into physically consistent 6-DoF vehicle motions. These rectified conditions are then passed to the *Physics-Enhanced Multi-view Video Generator* (PE-MVGen) (Section 3.4), which synthesizes multi-view video sequences. This unified pipeline enables high quality video generation even when the input trajectories violate physical constraints.

Specifically, our system takes as inputs the initial multi-view images $\mathcal{I}_0$, static map $\mathcal{M}$, and future trajectories $\mathcal{T}^{orig}$ for all $N$ agents, *i.e.*, cars. We define $\mathcal{T}^{orig} = \{\mathcal{T}_i^{orig}\}_{i=1}^{N}$, where $\mathcal{T}_i^{orig} = \{\mathcal{T}_{i,t}^{orig}\}_{t=1}^{T}$ and $\mathcal{T}_{i,t}^{orig} = (x_{i,t}, y_{i,t})$ denotes the 2D location of agent $i$ at time $t$. This representation aligns with the standard output format of trajectory simulators and end-to-end autonomous driving planners. Crucially, these trajectories can be physics-violating, e.g., containing overlapping paths that cause object penetration. Given such potentially flawed inputs, our goal is to synthesize a high-

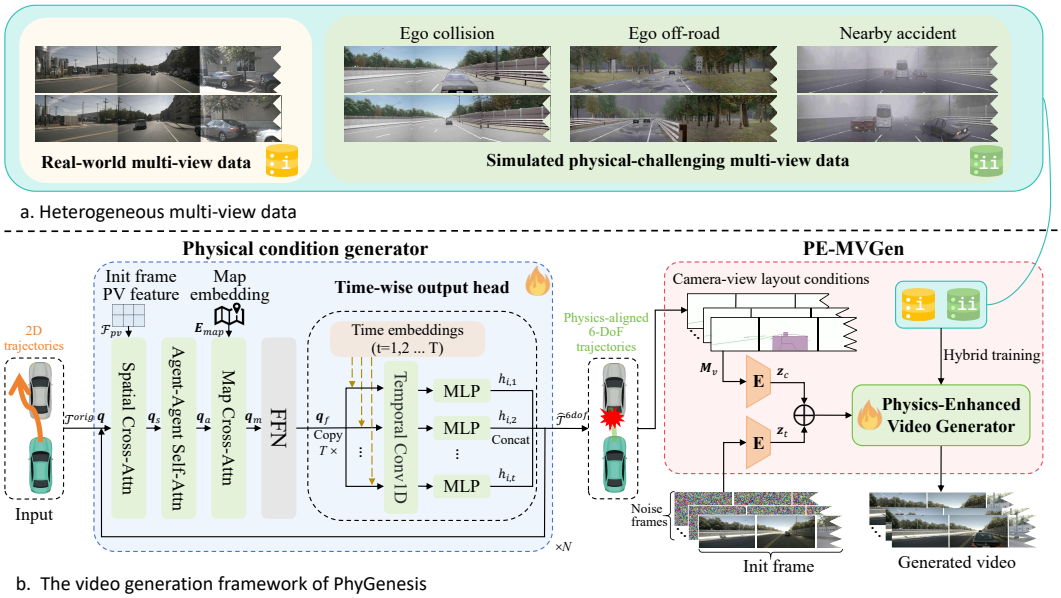


Figure 2: Overview of **PhyGenesis**. (a) Our heterogeneous multi-view dataset consists of both real-world driving data and simulated data that emphasizes physically challenging scenarios, including ego-vehicle collisions and roadway departures, among others. (b) In PhyGenesis, the *physical condition generator* first rectifies arbitrary 2D trajectories—potentially counterfactual or physics-violating—into physically plausible 6-DoF motions. The rectified trajectories are then projected into camera-view layout conditions and fed into a *physics-enhanced video generator*, co-trained on the heterogeneous dataset, to synthesize high-fidelity, physically consistent multi-view videos.

fidelity multi-view video sequence $\mathcal{V}_{1:T}$ that reflects the intended driving behaviors while satisfying real-world physical constraints. Next, we detail each part of our design.

### 3.2 Heterogeneous Multi-view Data

**Real-World Multi-view Data.** Following recent multi-view driving world models (Gao et al., 2025; Zhao et al., 2025), we use nominal real-world driving logs from *nuScenes* (Caesar et al., 2020) to establish a foundational understanding of complex urban environments. However, these data are biased toward safe driving and lack complex physical interactions (eg. collisions and off-road driving). This deficiency causes generative models to produce artifacts and physically inconsistent motion under challenging trajectories.

**Simulated Physically-challenging Multi-view Data.** To equip world models with robust physical understanding and generation ability under challenging interactions, training data containing such events is essential. Since collecting safety-critical real-world data is impractical, modern simulators such as CARLA (Dosovitskiy et al., 2017) provide high-fidelity physics and controllable environments. Prior work such as ReSim (Yang et al., 2025a) incorporates synthetic data into world-model training to mitigate the limited coverage of real-world distributions. However, its data is single-view with only ego-trajectory annotations, making it difficult to train models that control multiple agents. Moreover, its collection is not explicitly focused on physically challenging events, providing limited supervision for physical priors.

To fill this gap, we use CARLA to build a large-scale multi-view synthetic dataset for physically challenging scenarios. Following the Bench2Drive routing setup (Jia et al., 2024), we cover diverse scenes, weather, and traffic events, and curate two subsets: *CARLA Ego*, focused on ego-environment/agent interactions, and *CARLA Adv*, capturing interactions centered on a nearby non-ego agent. We perturb the route and target speed of the ego vehicle (or the Adv agent) to induce collisions, off-road departures, and abrupt maneuvers (Detailed in Appendix H). This results in substantially more aggressive dynamics than *nuScenes*, as shown by the acceleration statistics in Appendix I.

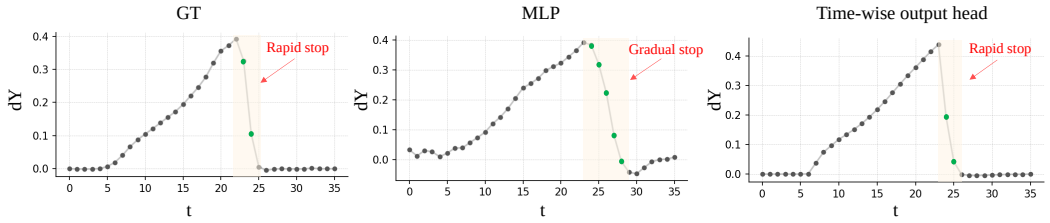


Figure 3: Comparison of MLP and time-wise output head in simulating collision dynamics. The MLP shows a gradual velocity decrease after the collision, while the GT and time-wise head show an instantaneous drop to zero, producing more realistic dynamics.

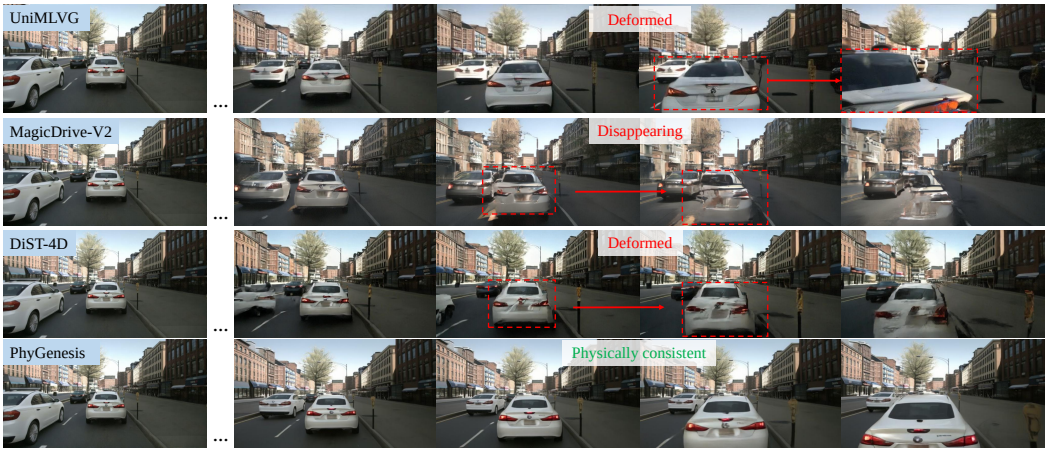


Figure 4: Qualitative comparison with different baselines. Our method maintains the best physical consistency and produces the best visual quality.

We equip the simulation with a sensor suite aligned with the *nuScenes* setup, including 1 LiDAR, 6 surround-view cameras ($900 \times 1600$), 5 radars, and 1 IMU/GNSS unit. To capture physical anomalies, we further add a collision sensor and HD map metadata to record impact and off-road timestamps. Data are collected at 12Hz and organized in the *nuScenes* format for downstream training.

**Heterogeneous Dataset Construction.** In total, we simulate about 31 hours of driving data. The *CARLA-Adv* subset contains 15.5 hours with 760K annotated boxes, while *CARLA-Ego* contains 15.2 hours with 830K boxes. Using collision sensors and map metadata, we design a rule-based filter to localize physical-interaction timestamps and extract 9.7 hours of physically challenging clips. Finally, we combine these clips with 4.6 hours of real-world data to construct our heterogeneous dataset.

### 3.3 Physical Condition Generator

Since rendering videos from physics-violating 2D layouts often causes artifacts, we introduce a *Physical Condition Generator* as the first stage of our framework. It rectifies possibly physics-violating $T$-frame 2D trajectories $\mathcal{T}^{orig}$ into physically plausible 6-DoF trajectories $\hat{\mathcal{T}}^{\text{6dof}}$. This 6-DoF representation, including $x, y, z$, pitch, yaw, and roll, is crucial because extreme physical interactions often induce vertical and rotational variations that 2D coordinates cannot capture.

**Architecture.** The physical model is illustrated on the left of Figure 2 (b). Given a scenario, input trajectories $\mathcal{T}^{orig}$ are encoded by sine–cosine positional encoding and an MLP into agent tokens $\mathbf{q} \in \mathbb{R}^{N \times D}$, where $N$ and $D$ denote the number of agents and token dimension. We first apply deformable spatial cross-attention to fuse multi-view PV features $\mathcal{F}_{pv}$ based on trajectory coordinates, yielding spatially grounded queries $\mathbf{q}_s$. Then, agent-agent self-attention enables tokens to capture surrounding vehicles' positional and kinematic states, resolving overlap and penetration conflicts.

Table 1: Comparison of video generators under GT trajectory conditions. We report FID, FVD, physical consistency (Phy), and controllability error (CtrlErr); best results are in **bold**.

| Methods | *nuScenes* | | | | *CARLA Ego* | | | | *CARLA ADV* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID ↓ | FVD↓ | PHY ↑ | CtrlErr ↓ | FID ↓ | FVD ↓ | PHY ↑ | CtrlErr ↓ | FID ↓ | FVD ↓ | PHY ↑ | CtrlErr ↓ |
| UniMLVG (Chen et al., 2025) | 16.63 | 113.23 | 0.92 | 0.35 | 34.03 | 240.78 | 0.56 | 1.32 | 30.06 | 264.59 | 0.75 | 0.66 |
| MagicDrive-V2 (Gao et al., 2025) | 13.17 | 88.86 | 0.93 | 0.29 | 32.92 | 181.26 | 0.58 | 1.26 | 33.62 | 214.52 | 0.65 | 0.75 |
| DiST-4D (Guo et al., 2025a) | 10.48 | 45.24 | 0.84 | 0.28 | 19.94 | 133.10 | 0.38 | 1.19 | 16.12 | 105.70 | 0.50 | 0.57 |
| PhyGenesis | **10.20** | **31.14** | **0.97** | **0.25** | **10.98** | **57.02** | **0.69** | **0.85** | **9.07** | **59.44** | **0.83** | **0.37** |

Table 2: Comparison with baselines under 2D trajectory condition: *nuScenes* uses nominal trajectories, while *CARLA Ego* and *CARLA ADV* use physics-violating trajectories. We report FID, FVD, physical consistency (PHY), human preference (Pref.). PhyGenesis achieves the best results across all datasets.

| Methods | nuScenes | | | | CARLA Ego | | | | CARLA ADV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | FVD↓ | PHY↑ | Pref. ↑ | FID↓ | FVD↓ | PHY↑ | Pref. ↑ | FID↓ | FVD↓ | PHY↑ | Pref. ↑ |
| UniMLVG (Chen et al., 2025) | 17.59 | 129.69 | 0.93 | 0.05 | 34.50 | 260.21 | 0.55 | 0.13 | 30.32 | 274.94 | 0.70 | 0.19 |
| MagicDrive-V2 (Gao et al., 2025) | 13.40 | 91.10 | 0.92 | 0.16 | 32.19 | 207.64 | 0.60 | 0.06 | 32.82 | 222.55 | 0.66 | 0.10 |
| DiST-4D (Guo et al., 2025a) | 10.49 | 46.95 | 0.86 | 0.13 | 19.84 | 197.57 | 0.39 | 0.10 | 16.07 | 128.88 | 0.56 | 0.05 |
| PhyGenesis | **10.24** | **40.41** | **0.97** | **0.67** | **11.03** | **72.48** | **0.71** | **0.71** | **9.28** | **77.83** | **0.87** | **0.66** |

Map cross-attention further integrates vectorized map embeddings $\mathbf{E}_{map}$ for off-road awareness. Finally, an FFN refines the aggregated features into queries $\mathbf{q}_f$ for trajectory prediction.

Following the above layers, the refined queries need to be projected into the final trajectories. Since traditional MLPs fail to capture the sudden, high-frequency dynamic impulses indicative of a collision, we design a *Time-Wise Output Head* (detailed in Appendix E) as the final prediction module, which uses step-specific time embeddings and a TCN to capture inter-step dynamic variations, accurately predicting abrupt physical changes at the moment of impact.

**Training Pair Construction.** To enable the Physical Condition Generator to rectify physics-violating trajectories, we construct paired data that maps *physics-violating* inputs to physically feasible targets. We propose a counterfactual trajectory corruption strategy: for each collision clip in our simulated dataset, we keep the original logs before collision, but corrupt post-collision trajectories by extending all agents with pre-collision velocity, synthesizing penetration-style counterfactual conditions. The ground-truth simulation logs, which capture collision dynamics, serve as supervision. To avoid distorting natural driving conditions, we include nominal trajectory-condition pairs from *nuScenes* without corruption, ensuring the model preserves realistic inputs while rectifying invalid ones.

**Optimization.** The Physical Model is optimized using a weighted distance loss between the predicted 6-DoF trajectories and the ground truth:

$$\mathcal{T}_{phy} = \frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{i,t} \left\| \hat{\mathcal{T}}_{i,t}^{6dof} - \mathcal{T}_{i,t}^{gt} \right\|_1 \tag{1}$$

To focus on critical physical moments, we define $W_{i,t}$ with two scalars: an event-window weight $\lambda_{\text{event}}$ that increases the loss within a temporal window around collision/off-road timesteps, and a physical-agent weight $\lambda_{\text{agent}}$ that further amplifies the loss for agents involved in the interaction.

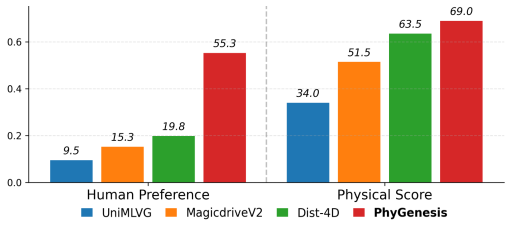


Figure 5: Human preference and Physical score on *nuScenes* under physics-violating trajectories.

Table 3: Performance of the physical condition generator for trajectory rectification. We report the 6-DoF L2 distance to ground truth, comparing w/ and w/o the physical condition generator.

| Method | *nuScenes* | *CARLA Ego* | *CARLA ADV* |
|---|---|---|---|
| w/o Phys. Cond. Generator | 0.21 | 1.78 | 1.05 |
| W/ Phys. Cond. Generator | **0.16** | **0.58** | **0.76** |

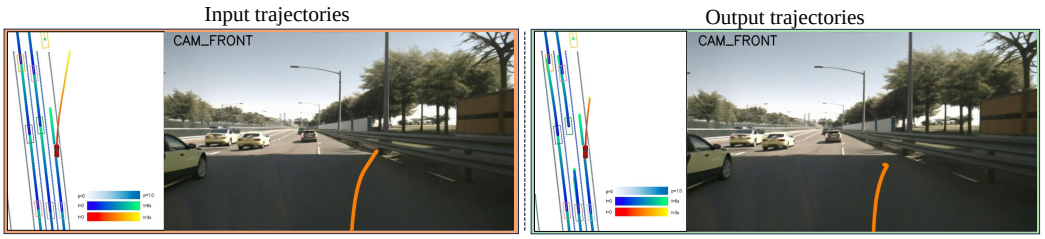


Figure 6: Performance of the physical condition generator in BEV and camera views. The left input trajectory penetrates the guardrail, while our generator rectifies it into a guardrail collision and stop.




Figure 7: Effect of the Physical Condition Generator under physics-violating trajectories: it reduces penetration artifacts between vehicles and the environment.

### 3.4 Physics-Enhanced Multi-view Video Generator

The second stage of our framework is the Physics-Enhanced Multi-View Video Generator (PE-MVGen). Built upon Wan2.1 (Wan et al., 2025), a high-capacity DiT originally conditioned on images and text, we adapt it into a controllable, multi-view generator designed for the autonomous driving.

**Multi-View & Layout Conditioning.** We first encode the input multi-view clips into latents $\mathbf{z} \in \mathbb{R}^{V\times T\times C\times h\times w}$ using a pre-trained 3D VAE (Wan et al., 2025), where $V$ represents the number of views. To enable multi-view modeling without introducing additional parameters (Gao et al., 2025; Wen et al., 2024), we reshape the input into $T \times C \times h \times (V \cdot w)$, concatenating the view dimension into the spatial axis so that the same self-attention can capture cross-view dependencies. Furthermore, to explicitly condition generation on structural layouts, we project the future $T$-frame 3D agent boxes and map polylines onto each camera view using the calibrated intrinsics $\mathbf{K}_v$ and extrinsics $\mathbf{E}_v$. The resulting view-specific control images $\mathbf{M}_v$ are encoded by the VAE encoder to get $\mathbf{z}_c$ , reshaped to match the video latents, concatenated along the channel dimension with the noisy latent input $\mathbf{z}_t$, and fed into the DiT.

**Data-Driven Physical Enhancement.** Current world models fail in physically challenging scenarios because their training distributions lack physical interactions. To solve this, we train PE-MVGen on our heterogeneous dataset, maintaining a balanced 1:1 ratio between nominal real-world logs and simulated physically challenging data. Crucially, we do not use counterfactual trajectories at this stage; the generator is supervised with ground-truth physical trajectories, decoupling physical correction from rendering. PE-MVGen is optimized via Rectified Flow (Esser et al., 2024) and a two-stage curriculum training strategy (detailed formulations are provided in Appendix F).

## 4 Experiment

### 4.1 Experimental Setup

**Datasets:** We use the heterogeneous multi-view dataset introduced in Section 3.2, including *CARLA Ego*, *CARLA ADV*, and *nuScenes*, for training and evaluation. During training, we adopt balanced sampling to keep the simulated-to-real clip ratio close to 1:1. For video generation evaluation, we randomly sample 150 clips per test split and generate videos for all sampled clips. Since the compared

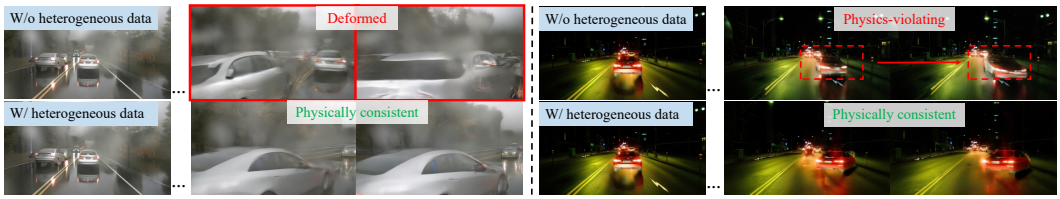


Figure 8: Effect of heterogeneous data: improves visual quality in physically challenging frames.

Table 4: Ablation study. "Mixed Data" indicates co-training our heterogeneous dataset, and "Phy-Model" denotes our Physical Condition generator. The best results are highlighted in **bold**.

| Components | | *nuScenes* | | | | *CARLA Ego* | | | | *CARLA ADV* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mixed Data | Phy-Model | FID ↓ | FVD ↓ | PHY ↑ | Pref. ↑ | FID ↓ | FVD ↓ | PHY ↑ | Pref. ↑ | FID ↓ | FVD ↓ | PHY ↑ | Pref. ↑ |
| ✓ | ✓ | **10.24** | **40.41** | **0.97** | 0.36 | **11.03** | **72.48** | **0.71** | **0.55** | 9.28 | **77.83** | **0.87** | **0.57** |
| ✓ | × | 10.70 | 44.43 | 0.97 | 0.26 | 11.81 | 116.51 | 0.65 | 0.19 | **9.18** | 89.25 | 0.85 | 0.28 |
| × | ✓ | 10.53 | 41.06 | 0.97 | **0.38** | 12.74 | 82.39 | 0.71 | 0.27 | 9.83 | 89.83 | 0.84 | 0.15 |

baselines are primarily trained on *nuScenes*, we translate the CARLA clips into the *nuScenes* visual style using a learned style-transfer diffusion model (details in Appendix D) for fair evaluation.

**Evaluation Metrics:** We evaluate generated videos along three dimensions: visual quality, physical plausibility, and controllability. **(1):** Following prior work, we use **FID** and **FVD** to measure realism. **(2):** For physical plausibility, we adopt WorldModelBench (Li et al., 2025c), a benchmark that can effectively measure the effectiveness of autonomous driving video generation using human preference-aligned VLM-based judges. We report **PHY** as the average of four WorldModelBench metrics: Mass (objects do not deform irregularly), Impenetrability (objects do not unnaturally pass through each other), Frame-wise Quality, and Temporal Quality. In addition, we report human preference (**Pref.**), the normalized score from multi-video choices, with an "uncertain" response split equally among the compared methods. The questionnaire and scoring protocol are provided in Appendix B. **(3):** For controllability, we measure how well generated videos follow input trajectory conditions. Specifically, we compute the Controllability Error (**CtrlErr**) between trajectories extracted from generated videos and ground-truth trajectories, defined as the geometric mean of Rotation Error and Translation Error, following prior work (He et al., 2024; Duan et al., 2025). Camera poses are extracted using ViPE (Huang et al., 2025).

**Baseline:** We compare our method with three baselines: UniMLVG (Chen et al., 2025), MagicDrive-V2 (Gao et al., 2025), and DiST-4D (Guo et al., 2025a). For UniMLVG, MagicDrive-V2, and our framework, the input includes the initial RGB frame, the scene captions, and future layout information. For DiST-4D, it additionally requires the initial-frame depth map, providing extra geometric cues. Implementation details of these baselines are provided in Appendix.

**Implementation Details:** The training of our Physical Condition Generator is conducted on 12Hz data with $T = 36$. Perception-view (PV) features are extracted using a ResNet50 (He et al., 2016) initialized from Sun et al. (2025) and trained with a learning rate of $9 \times 10^{-5}$. The main network is trained with a learning rate of $9 \times 10^{-4}$, and the batch size is set to 256. We set $N = 2$ and, for the trajectory-rectification results in Table 3, use $\lambda_{\text{event}} = 5$ and $\lambda_{\text{agent}} = 10$ to define $W_{i,t}$ (Section 3.3); ablations are provided in Appendix J.

Our video generation model is initialized from the pre-trained WAN2.1 weights and optimized with AdamW. We adopt a two-stage schedule: Stage 1 trains at $224 \times 400$ for 2,850 steps with a learning rate of $5 \times 10^{-5}$ and a global batch size of 480, followed by Stage 2 at $448 \times 800$ for 350 steps with a learning rate of $1 \times 10^{-4}$ and a global batch size of 240. Training is implemented on 48 NVIDIA H20 GPUs. The model generates 33-frame videos at 12 Hz.

### 4.2 Performance of PhyGenesis

In this section, we evaluate **PhyGenesis**. For *nuScenes*, we use the ground-truth $(x, y)$ trajectories as the input trajectory condition $\mathcal{T}^{orig}$. For *CARLA Ego* and *CARLA ADV*, to assess generation under physics-violating trajectory inputs, we construct counterfactual conditions following Section 3.3: we

keep pre-collision trajectories unchanged and extrapolate post-collision motion using each agent's pre-collision velocity. For baselines, we additionally compute yaw angles from $(\Delta x, \Delta y)$ to simulate rotation (Liao et al., 2025). As shown in Table 2, our framework achieves the best physical scores across all datasets and the best visual quality, with the largest gains on the challenging CARLA sets, where physics-violating inputs often cause deformation and penetration in prior methods (Figure 4).

To evaluate video generation under OOD, physically challenging trajectories in real-world scenes, we additionally conduct a *nuScenes* stress test by scaling up ego-vehicle speeds to create physics-violating trajectories while fixing the first frame. Results in Figure 5 and Figure 10 in the Appendix show that **PhyGenesis** remains more physically consistent under these conditions

### 4.3 Performance of Physics-enhanced Multi-view Video Generator

In this section, we evaluate the Physics-Enhanced Video Generator. We condition all methods on *ground-truth* trajectories (i.e., no counterfactual or physics-violating inputs) and report results on *nuScenes*, *CARLA Ego*, and *CARLA Adv* in Table 1. Although the inputs are physically feasible, prior baselines are trained mostly on nominal logs and rarely see physics-rich interactions (e.g., collisions and off-road events), so they still struggle to render these dynamics, resulting in worse visual quality, physical consistency, and controllability. In contrast, our PE-MVGen is co-trained with physics-rich data, better capturing complex interactions and improving generation quality and condition following.

### 4.4 Performance of Physical Condition Generator

In this section, we evaluate Physical Condition Generator. Following Section 4.2, we construct 2D trajectory conditions on *nuScenes*, *CARLA Ego*, and *CARLA Adv*, and report the 6-DoF L2 distance to ground truth before vs. after rectification (on the physical agent and its interaction partner). Table 3 shows consistent error reductions across datasets, and Figure 6 gives a qualitative example of trajectory rectification. Gains can be observed on *nuScenes* since the input provides only 2D $(x, y)$ and the model mainly recovers the missing 4-DoF. For physics-violating trajectories in *CARLA Ego* and *CARLA Adv*, the model achieves substantial error reduction, demonstrating its effectiveness.

### 4.5 Ablation Study

In this section, we ablate the two key components of **PhyGenesis**: the *Physical Condition Generator* and the *heterogeneous co-training* with physics-rich CARLA data. Table 4 summarizes the results.

**Effect of the Physical Condition Generator.** Rows (1) and (2) in Table 4 show that the physical condition generator improves both visual quality and physical consistency, with the largest gains on the physically challenging *CARLA Ego* and *CARLA Adv* sets. Figure 7 further illustrates it mitigates penetration and deformation artifacts, yielding more physically consistent interactions among vehicles and nearby structures under physics-violating trajectory conditions.

**Effect of heterogeneous co-training.** Rows (1) and (3) in Table 4 show that physics-rich CARLA data improves performance on challenging data. On *CARLA ADV*, heterogeneous training reduces FVD from 89.83 to 77.83 ($\sim$13.4% relative) and improves human preference. Figure 8 shows that training on nominal *nuScenes* alone often deforms vehicles in challenging interactions, while heterogeneous co-training yields more physically coherent dynamics.

## 5 Conclusion

We present PhyGenesis, a novel driving world model for physically consistent and high-fidelity multi-view video generation. By explicitly handling trajectory feasibility and physics-enhanced video generation, our approach outperforms existing methods in both visual fidelity and physical consistency, particularly under challenging trajectory conditions. Our framework enables more reliable simulation of safety-critical events such as collisions and off-road behaviors. By better aligning planner- or simulator-provided trajectories with physically consistent visual world modeling, PhyGenesis offers a practical building block toward simulation-driven evaluation and safety testing in autonomous driving.

### AI USE STATEMENT

Generative AI was used in the research pipeline and in preparing this manuscript. Our video generator is initialized from Wan2.1 (Section 3.4). We use Qwen2.5-VL to generate video-level descriptions that condition a style-transfer model based on Wan2.1-Fun-V1.1-1.3B-Control, which translates CARLA videos into the *nuScenes* visual domain (Appendix D). The translated videos provide initial-frame conditions and reference videos for FID/FVD evaluation. We use WorldModelBench's VLM-based judges to compute the PHY scores described in Section 4.1. AI-assisted tools were also used to polish the manuscript. The authors are responsible for the final manuscript, including AI-assisted text, claims, and artifacts, and for all reported results.

### ETHICS STATEMENT

This work uses *nuScenes* recordings and CARLA simulations, including collision and off-road scenarios, to study physically consistent video generation. The human evaluation consists of preference judgments over generated videos; the questionnaire, scoring procedure, and response counts are described in Appendix B. Participants in the human evaluation were fairly compensated for their time and effort. Generated videos may still contain physically implausible behavior. Visual quality and physical-consistency scores should therefore not be interpreted as guarantees of safety for a deployed driving system.

### REPRODUCIBILITY STATEMENT

The model formulation and experimental settings are described in Sections 3 and 4.1. Additional details on the style-transfer model, the time-wise output head, the video-generator training procedure, and the simulated scenario construction are provided in Appendices D, E, F, and H, respectively. Appendix B provides the human-evaluation questionnaire, preference-scoring procedure, and response counts. We will make the implementation code and our CARLA dataset available upon acceptance of the paper.

## REFERENCES

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

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.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25453–25463, 2025.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 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. In *International Conference on Learning Representations (ICLR)*, 2024a.

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

Zeyu Gao, Yao Mu, Chen Chen, Jingliang Duan, Ping Luo, Yanfeng Lu, and Shengbo Eben Li. Enhance sample efficiency and robustness of end-to-end urban autonomous driving via semantic masked world model. *IEEE Transactions on Intelligent Transportation Systems*, 25(10):13067–13079, 2024b.

Anthony Gosselin, Ge Ya Luo, Luis Lara, Florian Golemo, Derek Nowrouzezahrai, Liam Paull, Alexia Jolicoeur-Martineau, and Christopher Pal. Ctrl-crash: Controllable diffusion for realistic car crashes. *arXiv preprint arXiv:2506.00227*, 2025.

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, 2025a.

Xiangyu Guo, Zhanqian Wu, Kaixin Xiong, Ziyang Xu, Lijun Zhou, Gangwei Xu, Shaoqing Xu, Haiyang Sun, Bing Wang, Guang Chen, et al. Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency. *arXiv preprint arXiv:2506.07497*, 2025b.

Zipeng Guo, Yuchen Zhou, and Chao Gou. Drivinggen: Efficient safety-critical driving video generation with latent diffusion models. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems*, 35:20703–20716, 2022.

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.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, and Xin Jin. Uniscene: Unified occupancy-centric driving scene generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 11971–11981. Computer Vision Foundation / IEEE, 2025a.

Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huan-ang Gao, Bu Jin, and Hao Zhao. AVD2: accident video diffusion for accident video description. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, pp. 13289–13296. IEEE, 2025b.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025c.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025d.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

NVIDIA. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Chen Shi, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. Drivex: Omni scene modeling for learning generalizable world knowledge in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28599–28609, 2025.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801. IEEE, 2025.

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

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

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.

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.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, et al. Drivelaw: Unifying planning and video generation in a latent driving world. *arXiv preprint arXiv:2512.23421*, 2025.

Zhiyuan Xu, Bohan Li, Huan-ang Gao, Mingju Gao, Yong Chen, Ming Liu, Chenxu Yan, Hang Zhao, Shuo Feng, and Hao Zhao. Challenger: Affordable adversarial driving video generation. *arXiv preprint arXiv:2505.15880*, 2025.

Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27531–27541, 2025.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. *arXiv preprint arXiv:2506.09981*, 2025a.

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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025b.

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. *arXiv preprint arXiv:2510.19195*, 2025a.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

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.

Jiawei Zhou, Linye Lyu, Zhuotao Tian, Cheng Zhuo, and Yu Li. Safemvdrive: Multi-view safety-critical driving video synthesis in the real world domain. *arXiv preprint arXiv:2505.17727*, 2025.

Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wan, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. *arXiv preprint arXiv:2509.23402*, 2025.

## A VIDEOS EXAMPLES

To facilitate a more comprehensive qualitative evaluation, we provide the videos corresponding to several key examples presented in the main paper at https://iclr3.github.io/PhyGenesis/. Compared with static figures, these videos better capture temporal dynamics, visual details, and generation consistency over time.

## B USER STUDY SETTING

In our experiments, participants are shown videos generated by different baselines under the same conditions and are asked to select the one with the highest visual quality. In addition to selecting one of the candidate videos, participants may also choose an "uncertain" option. A selected video is assigned 1 point, while for an "uncertain" response, the score is distributed equally among all compared videos, with each video receiving $1/N$ point, where $N$ is the number of videos in the comparison. The final score is computed as the total points received divided by the total number of comparisons. The questionnaire used in our study is shown in Figure 9.

The user studies reported in Table 2, Figure 5, and Table 4 of the main paper are conducted separately. For each study, we randomly sample generated videos from five scenes in the corresponding evaluation dataset for comparison. Each human preference score is based on 150 responses collected from 30 participants. In total, we collect 1,050 responses across all user studies.

## C QUALITATIVE COMPARISON UNDER OOD AND PHYSICS-VIOLATING CONDITIONS

Fig. 10 shows a qualitative comparison on real-world scenes from the *nuScenes* stress test. All methods use the same first-frame condition, while the future trajectory or layout inputs are constructed as OOD and physics-violating cases. This setting evaluates whether the models can still generate physically plausible scene evolution under abnormal trajectory guidance.

Compared with the baselines, our PE-MVGen better preserves vehicle geometry, spatial relationships, and multi-view consistency. The baselines are more likely to produce distorted objects, inconsistent positions, or unrealistic motion when the input trajectory violates physical constraints. These results show that our physics-enhanced design improves robustness and physical consistency under challenging real-world conditions.

## D STYLE TRANSFER MODEL USED FOR VIDEO GENERATION

Since the compared baselines are primarily trained on *nuScenes*, we introduce a style transfer model to ensure a fair comparison. Our transfer model is built upon Wan2.1-Fun-V1.1-1.3B-Control (Wan et al., 2025) and follows the same rectified-flow video generation framework. Specifically, we adapt the generation model to the style transfer setting by conditioning it on per-frame depth information and detailed textual descriptions to translate CARLA clips into the *nuScenes* visual style. After stylizing the full CARLA videos into the *nuScenes* domain, we can provide all compared video generation methods with *nuScenes*-style initial frames on the CARLA benchmark, and also obtain stylized videos to be the ground truth for computing FID and FVD.

The architecture of the transfer model is shown in Figure 11. Specifically, we use Depth Anything V2 (DAV2) (Yang et al., 2024) to extract per-frame depth maps from the input video, and, similar to the main model, concatenate the resulting depth latents with noisy video latents as generation conditions, without using the initial frame as a conditioning input. In addition, we use Qwen2.5-VL (Team, 2025) to produce a detailed description for the entire video, as the semantic content of the video is already fixed. The prompt used for generating video-level captions is shown in Figure 12.

Importantly, the transfer model is trained entirely on *nuScenes* data, rather than on heterogeneous mixed-domain data. This design encourages the model to learn the appearance characteristics of the *nuScenes* domain more faithfully. Since the initial frame is not used as a conditioning input, the model's generation inherently reflects the *nuScenes* style rather than the specific appearance of the

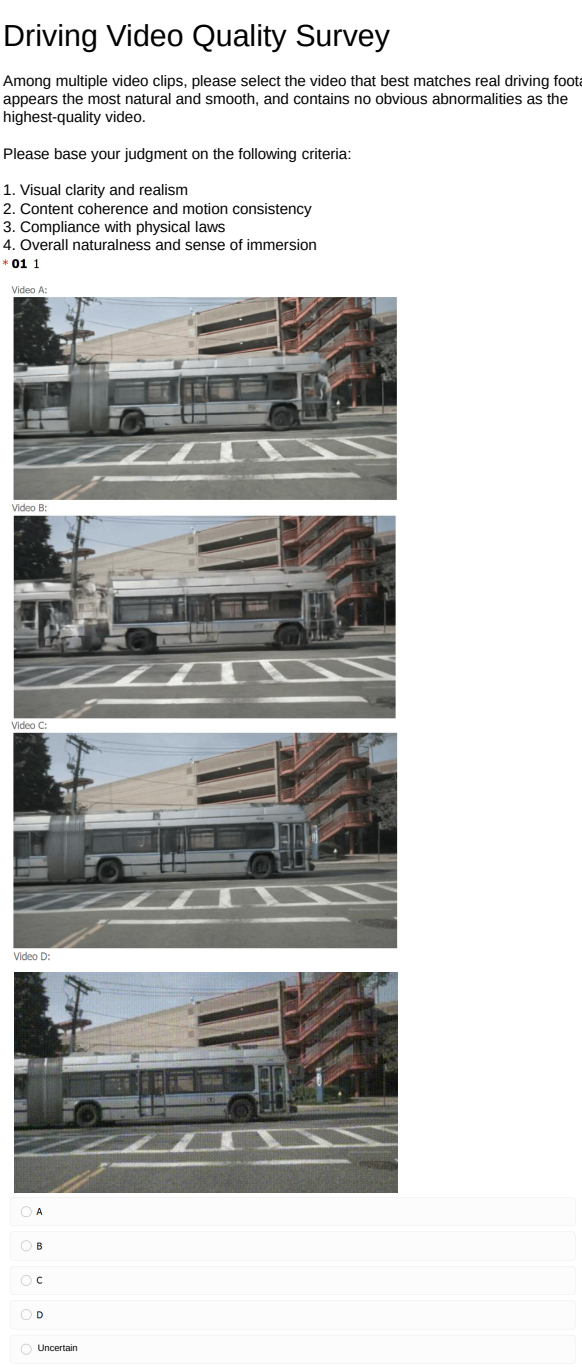


Figure 9: The questionnaire used to evaluate the visual quality of videos generated by different baselines in the user study.

input video. During inference, the trained transfer model is applied to CARLA videos to translate them into the *nuScenes* style.

**Training Objective.** Following the rectified-flow formulation used in the main paper, let $\mathbf{z}_1$ denote the clean target video latent, let $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a random noise latent, and let $t \in [0, 1]$ be the sampled timestep. The noisy latent is constructed as

$$\mathbf{z}_t = t\mathbf{z}_1 + (1 - t)\mathbf{z}_0. \tag{2}$$

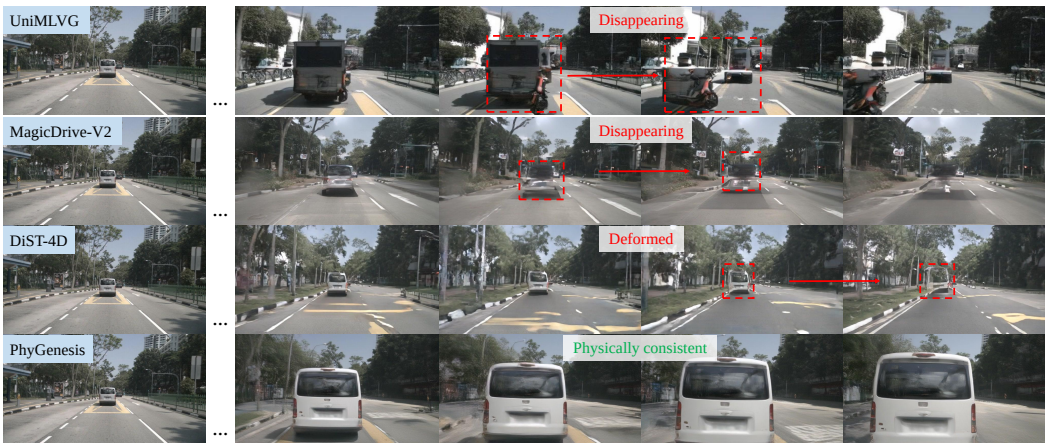


Figure 10: Qualitative comparison on the *nuScenes* stress test under OOD and physics-violating trajectory inputs, with the first-frame condition fixed. Our method shows better physical consistency than the baselines.

The transfer model is trained with the same flow-matching objective, conditioned on the video-level caption $\mathbf{c}_{text}$, and the per-frame depth condition $\mathbf{c}_{depth}$:

$$\mathcal{L}_{\text{transfer}} = \mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,t} \left\| u_\theta(\mathbf{z}_t, t, \mathbf{c}_{text}, \mathbf{c}_{depth}) - (\mathbf{z}_1 - \mathbf{z}_0) \right\|_2^2 . \tag{3}$$

## E Time-Wise Output Head

In this section, we provide additional details regarding the design of our Time-Wise Output Head. As illustrated in Figure 13, traditional MLP-based regression heads tend to produce sluggish and smoothed trajectory predictions, failing to capture the sudden, high-frequency dynamic impulses that occur during collisions. In contrast, our proposed time-wise formulation, combined with step-specific time embeddings, explicitly models the temporal inter-step variations.

Concretely, for the $i$-th refined agent token $\mathbf{q}_f[i]$, we expand it across the $T$ future steps and concatenate it with a step-specific learnable temporal embedding $\mathbf{E}_{time}(t)$. The concatenated feature is then processed by a Temporal Convolutional Network (TCN) to capture local inter-step dynamic variations, before being projected by an MLP to output the exact 6-DoF state:

$$\mathbf{H}_i = \text{TCN}_{1D}\left([\text{Proj}\left(\mathbf{q}_f[i] \parallel \mathbf{E}_{time}(1)\right), \ldots, \text{Proj}\left(\mathbf{q}_f[i] \parallel \mathbf{E}_{time}(T)\right)]\right) \tag{4}$$

$$\hat{\mathcal{T}}_{i,t}^{6dof} = \text{MLP}\left(\mathbf{H}_i[t]\right) \in \mathbb{R}^6 \tag{5}$$

This allows it to accurately predict the instantaneous velocity drop at the moment of physical impact, leading to much more realistic physical dynamics compared to standard regression methods.

## F Training Details of the Physics-Enhanced Multi-View Video Generator

**Training Objective.** Following Wan 2.1, PE-MVGen is optimized with the Rectified Flow formulation (Esser et al., 2024). Let $\mathbf{z}_1$ denote the clean video latent encoded from the target multi-view video, and let $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a Gaussian noise latent with the same dimensionality. For a timestep $t \in [0, 1]$ sampled from a logit-normal distribution, the intermediate latent state is obtained by linear interpolation between the noise latent and the clean latent:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1 - t)\mathbf{z}_0. \tag{6}$$

Under this formulation, the target velocity field is defined as the constant displacement from the noise latent to the clean latent:

$$\mathbf{v}_t = \mathbf{z}_1 - \mathbf{z}_0. \tag{7}$$

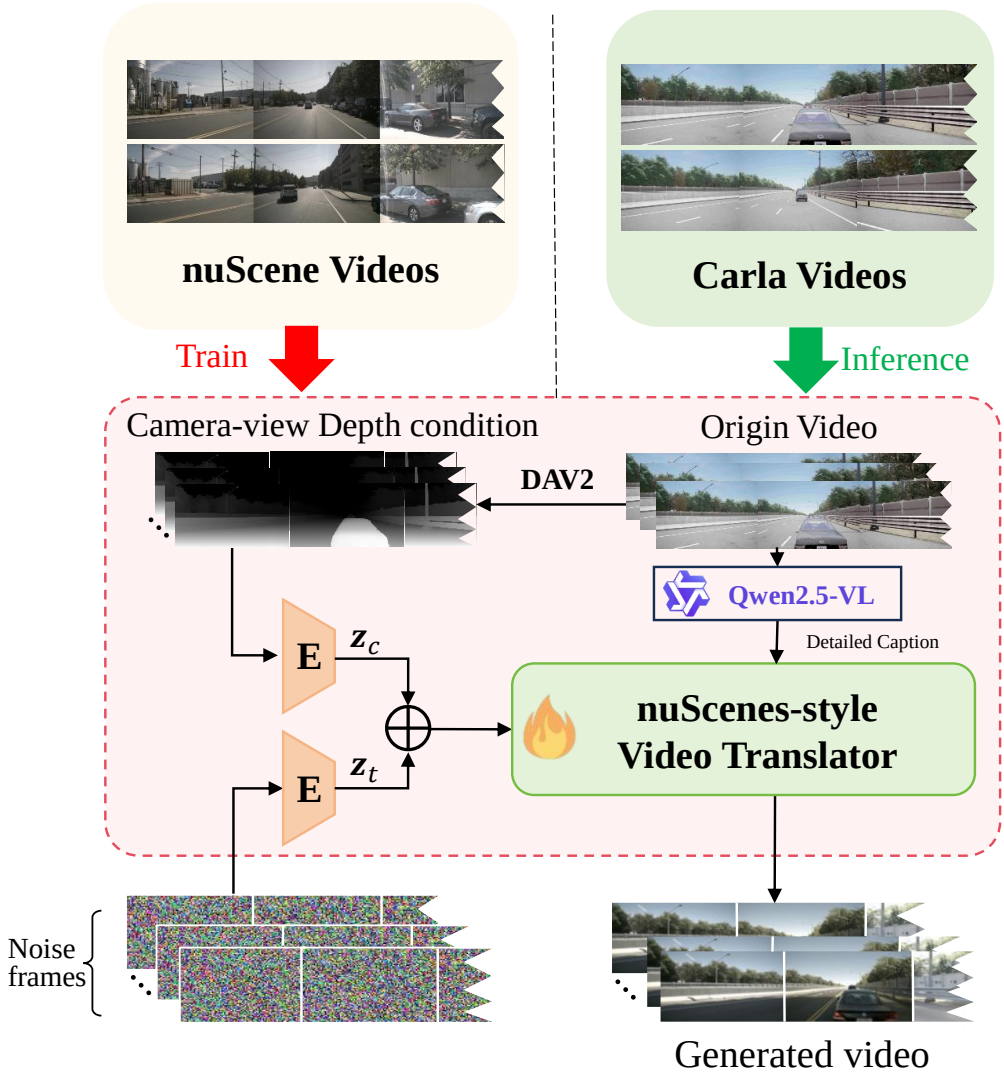


Figure 11: Architecture of the style transfer model used to translate CARLA videos into the *nuScenes* visual style for fair comparison.

The DiT backbone, parameterized by $\theta$, is trained to predict this velocity field from the noisy latent $\mathbf{z}_t$ and the corresponding conditioning information. Specifically, the model predicts

$$u_\theta(\mathbf{z}_t, t, \mathbf{c}_{init}, \mathbf{c}_{text}, \mathbf{c}_{layout}), \tag{8}$$

where $\mathbf{c}_{init}$ denotes the latent representation of the initial context frame, $\mathbf{c}_{text}$ represents the textual scene description, and $\mathbf{c}_{layout}$ encodes the future multi-view layout images. These three conditions provide complementary guidance: the initial frame constrains the appearance and scene context, the text caption provides high-level semantic information, and the future layout images introduce explicit spatial and geometric priors for subsequent multi-view video generation.

The model is optimized using the standard flow-matching mean squared error objective:

$$\mathcal{L}_{FM} = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t} \left\| u_\theta(\mathbf{z}_t, t, \mathbf{c}_{init}, \mathbf{c}_{text}, \mathbf{c}_{layout}) - \mathbf{v}_t \right\|_2^2 . \tag{9}$$

This objective encourages the model to learn a velocity field that transports samples from the noise distribution toward the target video latent distribution. Compared with directly predicting the clean

You are a professional video analyst. The following video shows a driving scenario from a vehicle's perspective. Please describe the scene in detail, covering the following aspects:
1. The approximate time of the video (e.g., daytime or night).
2. Weather conditions during the video.
3. Surrounding environment characteristics (e.g., urban, highway, rural, buildings, trees, road markings, traffic signs).
4. The ego-vehicle's movement and behavior throughout the video, including speed changes (accelerating, decelerating, constant speed), steering actions (going straight, turning left/right), lane changes, and overtaking other vehicles.
5. Any notable objects, obstacles, or vehicles around the ego-vehicle. Please provide all this information in a single continuous paragraph without any lists, line breaks, or formatting.

Figure 12: Prompt used in the VLM for generating detailed video-level captions.

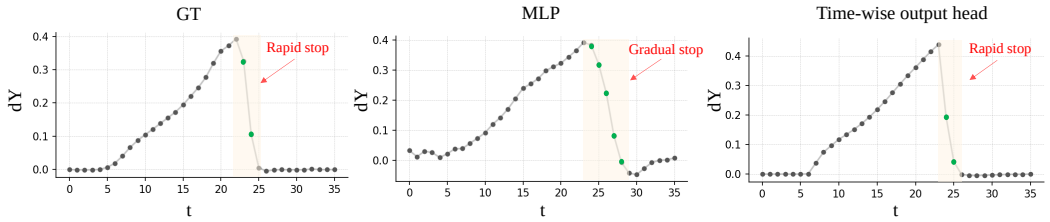


Figure 13: Comparison of MLP and time-wise output head in simulating collision dynamics. The MLP shows a gradual velocity decrease after the collision, while the GT and time-wise head show an instantaneous drop to zero, producing more realistic dynamics.

latent, velocity prediction provides a unified learning target across different noise levels and is consistent with the rectified-flow sampling process used during generation.

**Curriculum Co-Training Strategy.** To improve training efficiency while preserving generation quality, we adopt a two-stage curriculum co-training strategy. In the first stage, PE-MVGen is trained at a lower spatial resolution of $224 \times 400$. This stage is designed to efficiently learn the core correspondence between the conditioning signals and the target multi-view video latents. In particular, the lower-resolution setting allows the model to focus on coarse multi-view consistency, temporal dynamics, and the mapping from physically challenging layout conditions to plausible video structures, while reducing the computational cost of early-stage optimization.

After the model has learned these low-resolution structural priors, we conduct a second-stage fine-tuning at a higher spatial resolution of $448 \times 800$. This stage aims to recover fine-grained visual details and improve the perceptual quality of the generated videos. The high-resolution fine-tuning further refines object boundaries, local textures, and view-dependent appearance, while retaining the multi-view geometric and layout-aware generation ability learned in the first stage.

Overall, this curriculum separates the learning process into two complementary phases: the low-resolution stage emphasizes efficient acquisition of global structure and physically grounded layout-to-video correspondence, whereas the high-resolution stage improves visual fidelity and detailed rendering. This design reduces the difficulty of directly optimizing the full-resolution model from scratch and provides a practical training strategy for physics-enhanced multi-view video generation.

## G Implementation Details of Different Baselines

In this paper, we compare our method with three baselines, all evaluated using their official pretrained weights. For all methods, we generate videos consisting of 33 frames. Our method, MagicDriveV2 (Gao et al., 2025), and DiST-4D (Guo et al., 2025a) generate the entire video in a single pass. In contrast, UniMLVG (Chen et al., 2025) can generate at most 19 frames at a time. Therefore, we follow its official autoregressive inference code to produce a 33-frame video in two stages: the first pass generates frames 1–14, and the second pass generates frames 14–33, using the intermediate frame as the conditioning frame for the second stage.

Our method generates videos at a resolution of $448 \times 800$. MagicDriveV2 and DiST-4D generate videos at $424 \times 800$, while UniMLVG generates videos at $256 \times 448$, since it is trained only at this resolution.

Our method, MagicDriveV2, and UniMLVG all use the initial RGB frame as the generation condition. DiST-4D additionally uses the depth map of the initial frame as an extra condition, which we obtain using its official preprocessing pipeline. For the CARLA stylized domain, we apply a one-epoch adaptation to the Physical Condition Generator on stylized initial frames from the CARLA training set, while keeping the Physics-enhanced Multi-view Video Generator unchanged.

## H Physical-Challenging Scenario Construction in CARLA Ego and CARLA Adv

We construct physically challenging scenarios in CARLA (Dosovitskiy et al., 2017) based on the Bench2Drive (Jia et al., 2024) routing setup. Each rollout starts from a valid predefined route while preserving the original map layout, traffic participants, weather, and background behaviors. On top of this standard setup, we perturb the behavior of one designated vehicle to induce physically challenging events. The perturbation is parameterized by two variables: a lateral route offset and a target speed. Specifically, the target speed is sampled from predefined values between 0 and 30 m/s, with uniform random perturbations within the intervals around each value, and the lateral offset is sampled from predefined values between -200 and 200 meters, with uniform random perturbations within the intervals around each value. At scene initialization, we further choose one of three perturbation modes with equal probability: (i) zero lateral offset with randomized target speed, (ii) fixed target speed of 10 m/s with randomized lateral offset, and (iii) both target speed and lateral offset randomized. This design yields a broad spectrum of behaviors ranging from mild deviations to aggressive maneuvers.

For *CARLA Ego*, the perturbed vehicle is the ego vehicle itself. The ego vehicle first follows the default autopilot for a warm-up period of 24 steps. Since our simulator runs at 12 Hz, this corresponds to 2 seconds. After the warm-up, we replace the nominal route with a smoothed and supersampled route, and then apply a cosine-smoothed lateral offset to the future route points so that the vehicle gradually transitions from its current route to the perturbed one. The ego vehicle is then controlled to track this modified route under the sampled target speed. For *CARLA Adv*, we instead perturb a nearby non-ego vehicle. We first spawn an adversarial vehicle near the ego vehicle by sampling a valid waypoint within a local neighborhood of the ego route, including forward, backward, and adjacent-lane candidates within approximately 15 meters. The adversarial vehicle also follows the default controller for the same 24-step warm-up period. Afterward, we transform the ego route into the adversarial vehicle's local initialization frame and apply the same perturbation mechanism. As a result, the physically challenging behavior is centered on a nearby non-ego agent rather than the ego vehicle.

During rollout, we monitor collisions and off-road departures for the perturbed vehicle. Collisions are detected using CARLA collision sensors, and the collided object category is recorded, including vehicles, pedestrians, parked vehicles, traffic objects, and static objects. Off-road events are detected by querying the local lane type and checking whether the vehicle enters sidewalk or shoulder regions. Once an event is triggered, we continue collecting data for 48 additional frames, corresponding to 4 seconds at 12 Hz, and then terminate the rollout. We also terminate the rollout if no event is observed within 120 post-warm-up steps, i.e., 10 seconds. In this way, each collected sequence contains the pre-event context, the event itself, and its short-term consequence. The resulting dataset therefore consists of route-grounded but physically challenging trajectories. *CARLA Ego* emphasizes failures

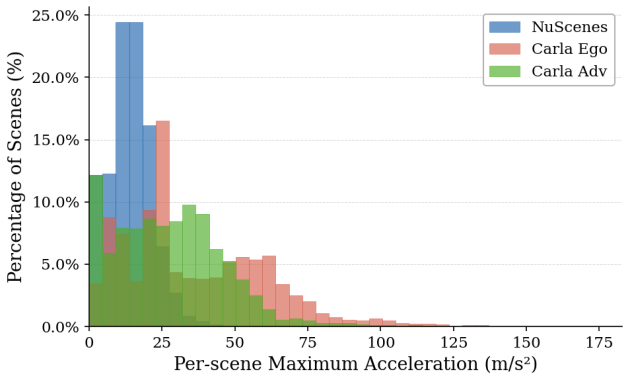


Figure 14: Distributions of maximum ego-vehicle acceleration for *nuScenes*, *CARLA Ego*, and *CARLA Adv*. The simulated CARLA datasets show a clear shift toward higher accelerations, indicating more aggressive dynamics and physically challenging events compared with the predominantly nominal driving behaviors in *nuScenes*.

and abrupt maneuvers induced on the ego vehicle, while *CARLA Adv* emphasizes interactions caused by a nearby perturbed non-ego vehicle.

## I Acceleration Statistics of Different Datasets

To further analyze the dynamic differences between the real-world and simulated datasets, we compare the distributions of maximum ego-vehicle acceleration in *nuScenes*, *CARLA Ego*, and *CARLA Adv*. For each video clip, we compute the maximum acceleration magnitude of the ego vehicle over the future trajectory. As shown in Figure 14, *nuScenes* is mainly concentrated in a lower-acceleration range, reflecting its predominantly nominal driving behaviors. In contrast, both *CARLA Ego* and *CARLA Adv* show clear shifts toward higher accelerations, indicating more aggressive dynamics and more frequent physically challenging events. This supports our motivation for introducing simulation-derived data to complement real-world driving logs and enhance the model's physical understanding.

## J Weighting Design and Ablation Study of $\lambda_{\text{event}}$ and $\lambda_{\text{agent}}$

In the Physical Condition Generator optimization, $W_{i,t}$ is defined using two scalar hyperparameters to prioritize physically important samples. The temporal weight $\lambda_{\text{event}}$ reweights the loss along the time dimension by assigning larger weights to timesteps near critical events such as collision or off-road. For each event timestep $t_e$, we define a forward window $[s_e, e_e]$, where $s_e = \max(0, t_e - 1)$ and $e_e = \min(T - 1, t_e + 10)$, with $T$ denoting the prediction horizon. Within this window, the temporal weight is initialized at $\lambda_{\text{event}}$ and exponentially decays to 1:

$$w_e(t) = \lambda_{\text{event}} \exp\left(\frac{\log(1/\lambda_{\text{event}})}{e_e - s_e}(t - s_e)\right), \qquad t \in [s_e, e_e].$$

If multiple event windows overlap, we take the maximum weight at each timestep:

$$W_t^{\text{event}} = \max_e w_e(t),$$

and set $W_t^{\text{event}} = 1$ for timesteps outside all event windows. This design makes the supervision strongest near the onset of a physical event while preserving standard supervision on the remaining horizon.

The agent weight $\lambda_{\text{agent}}$ further reweights the loss along the agent dimension by assigning a larger multiplicative factor to the agents that are most relevant to the physical event. The selection is determined based on the event annotations in the dataset. When the event corresponds to a collision with a static object, such as a fence or a lamp post, only the ego vehicle (or the Adv vehicle) is

emphasized. When the event corresponds to a collision with a dynamic participant, such as another vehicle or a pedestrian, we additionally emphasize the agent that is nearest to the ego vehicle (or the Adv vehicle) at the collision timestep. In this way, the optimization focuses on the participants that are most directly related to the event, which helps better model the trajectory behavior around the collision moment.

The default configuration, $\lambda_{\text{event}} = 5$ and $\lambda_{\text{agent}} = 10$, is used for the trajectory-rectification results in Table 3. Each ablation varies one weight while fixing the other at its default value.

The ablation study presented in Tables 5 and 6 indicates that different values of $\lambda_{\text{event}}$ and $\lambda_{\text{agent}}$ have limited effect on performance. This suggests that the model is largely robust to these hyperparameters while the weighting mechanism still provides focused supervision around critical events and agents.

Table 5: Ablation study on $\lambda_{\text{event}}$ with $\lambda_{\text{agent}} = 10$.

| **Configuration** | *nuScenes* | *CARLA Ego* | *CARLA ADV* |
|---|---|---|---|
| baseline | 0.21 | 1.78 | 1.05 |
| $\lambda_{\text{event}} = 1$ | 0.17 | 0.56 | 0.69 |
| $\lambda_{\text{event}} = 5$ (default) | 0.16 | 0.58 | 0.76 |
| $\lambda_{\text{event}} = 10$ | 0.16 | 0.57 | 0.77 |

Table 6: Ablation study on $\lambda_{\text{agent}}$ with $\lambda_{\text{event}} = 5$.

| **Configuration** | *nuScenes* | *CARLA Ego* | *CARLA ADV* |
|---|---|---|---|
| baseline | 0.21 | 1.78 | 1.05 |
| $\lambda_{\text{agent}} = 1$ | 0.13 | 0.53 | 0.70 |
| $\lambda_{\text{agent}} = 10$ (default) | 0.16 | 0.58 | 0.76 |
| $\lambda_{\text{agent}} = 20$ | 0.16 | 0.62 | 0.90 |

## K LIMITATION

The current version of PhyGenesis mainly focuses on vehicle-centric physical interactions, such as vehicle collisions and roadway departures. While this setting already covers a wide range of challenging driving dynamics, interactions involving pedestrians and other vulnerable road users are not yet systematically emphasized in our data construction. Extending the framework to richer vehicle–pedestrian interactions is a promising direction for future work.