# Risk-Controllable Multi-View Diffusion for Driving Scenario Generation

Anonymous CVPR submission

Paper ID *****

## Abstract

*Generating safety-critical driving scenarios is crucial for evaluating and improving autonomous driving systems, but long-tail risky situations are rarely observed in real-world data and difficult to specify through manual scenario design. Existing generative approaches typically treat risk as an after-the-fact label and struggle to maintain geometric consistency in multi-view driving scenes. We present RiskMV-DPO, a general and systematic pipeline for physically-informed, risk-controllable multi-view scenario generation. By integrating target risk levels with physically-grounded risk modeling, we autonomously synthesize diverse and high-stakes dynamic trajectories that serve as explicit geometric anchors for a diffusion-based video generator. To ensure spatial-temporal coherence and geometric fidelity, we introduce a geometry-appearance alignment module and a region-aware direct preference optimization (RA-DPO) strategy with motion-aware masking to focus learning on localized dynamic regions. Experiments on the nuScenes dataset show that RiskMV-DPO can freely generate a wide spectrum of diverse long-tail scenarios while maintaining state-of-the-art visual quality, improving 3D detection mAP from 18.17 to 30.50 and reducing FID to 15.70. Our work shifts the role of world models from passive environment prediction to proactive, risk-controllable synthesis, providing a scalable toolchain for the safety-oriented development of embodied intelligence.*

## 1. Introduction

Autonomous driving systems are ultimately judged by how safely and reliably they behave in the open world, including rare but safety-critical events [5, 10, 39]. Since real-world logs under-sample these cases, driving scenario generation has become essential for data augmentation, stress testing, and systematic safety evaluation [17, 22, 32]. With production vehicles relying heavily on camera suites, generating temporally coherent, multi-view consistent driving sequences is particularly valuable [19, 41].

Despite rapid progress, generating long-tail risky scenarios remains a significant bottleneck [11]. A common practice is to generate a large pool of driving scenes and then mine rare events, but truly high-risk situations occur with extremely low probability [40]. Another practice is to handcraft scenarios by specifying parameters such as near-collisions or road intrusions [37]. While offering direct control, these "sampling-and-filtering" or manual design approaches inevitably reflect human priors and lack the efficiency to synthesize the vast, non-intuitive space of hazardous interactions. What is missing is a general pipeline to transform "risk" from an after-the-fact label into a proactive, time-resolved control signal that can autonomously guide the synthesis of diverse long-tail hazards.

At the same time, the multi-view, outdoor, dynamic nature of driving scenes introduces a second bottleneck: geometry and robustness [2]. While diffusion models have advanced visual realism, driving applications demand more. They require cross-view metric-level consistency, motion continuity, and stability under occlusion and lighting changes. For consistency, many attempts rely on large-scale question-answer datasets to teach VLMs spatial concepts, but this supervision rarely provides a solid geometric prior that enforces physically grounded alignment across cameras [24, 30]. This motivates mechanisms that inject 3D priors into diffusion training and align the model's capacity with localized, motion-dominant regions that matter most for driving risk.

In short, these challenges are twofold: *making risk controllable* for long-tail scenario generation, and *making multi-view geometry reliable* under outdoor dynamic conditions. Therefore, we propose risk-controllable multi-view diffusion for driving scenario generation, a framework that makes risk a first-class control knob while improving geometric fidelity and localized dynamic realism. Our key idea is to turn risk into a time-resolved conditioning signal, generate trajectories and 3D boxes at a target risk level, and use them as structured controls for multi-view diffusion. The contributions can be summarized as follows:

- We propose a general pipeline for risk-controllable synthesis that reformulates driving risk as a proactive control signal, enabling the autonomous generation of dynamic

trajectories and 3D boxes at target risk levels.
- We introduce a geometry-appearance alignment module that injects compact 3D priors into multi-view diffusion to ensure metric-level cross-view consistency and geometric plausibility.
- We develop multi-view consistent RA-DPO with motion-aware masking to align scene synthesis with localized preferences in dynamic regions, significantly enhancing temporal and motion realism.

## 2. Related Work

This section reviews driving scenario generation, risk-aware modeling and simulation, geometric priors for multi-view diffusion, and preference alignment for generative models.

### 2.1. Driving scene and multi-view video generation

Driving scenario generation has evolved from neural simulators to large-scale driving world models [18]. DriveGAN [14] learns an end-to-end simulator conditioned on ego actions. GAIA-1 [12] models driving as token sequences with video/text/action conditioning. Diffusion-based world models further improve realism and controllability, such as DriveDreamer [31] and DriveDreamer-2 [43].

Multi-view driving video synthesis introduces explicit structure controls to improve cross-view consistency. MagicDrive [7] conditions on camera poses, maps, and 3D boxes with cross-view attention. DrivingDiffusion [16] generates multi-view videos from 3D layouts and enforces cross-view and cross-frame consistency. MagicDrive-V2 [9] scales to longer and higher-resolution synthesis, and recent work explores stronger spatiotemporal-view attention [21]. Despite these advances, long-tail risk is still hard to target explicitly and multi-view geometry remains fragile in dynamic outdoor scenes, motivating our risk-controllable and geometry-aware design.

### 2.2. Risk modeling and risk-aware simulation

Risk modeling often relies on trajectory-based criticality metrics (e.g., TTC variants) [34, 38] and formal safety envelopes such as RSS [25]. Recent work moves toward risk forecasting under interaction and uncertainty. RiskNet [20] combines field-theoretic interactions with multi-modal trajectory prediction for time-evolving risk estimation, and SafeDrive [45] quantifies perceived risk via risk fields to guide an LLM driving agent.

Risk-aware testing uses risk signals to search for rare failures efficiently. Scenic [6] specifies scenarios as distributions with constraints, VerifAI [4] supports simulation-based falsification and parameter synthesis, and Adaptive Stress Testing [15] uses RL to discover failure trajectories. Most toolchains still operationalize risk via thresholds or specification violations rather than enabling generation at a desired, time-resolved risk level, which motivates our risk-as-control formulation.

### 2.3. 3D priors for multi-view generation

Multi-view diffusion increasingly incorporates explicit 3D signals to stabilize generation and reduce cross-view drift [27]. In driving, MagicDrive [7] and DrivingDiffusion [16] condition on 3D controls and add cross-view modeling, while DriveScape [35] scales to high-resolution synthesis via multi-view feature fusion with 3D guidance. Even so, models can "understand the language" but still cannot "see through the geometry": existing conditioning often provides surface-level reinforcement and remains brittle under occlusion and fast motion. This motivates our geometry-appearance alignment that injects compact 3D priors into diffusion training.

### 2.4. DPO for generative models

DPO is a practical tool for aligning generative models with preference data. Diffusion-DPO [28] adapts DPO to diffusion models, and DSPO [46] improves objective matching under score-based training. For video, DenseDPO [36] motivates finer temporal preference signals, and LocalDPO [13] introduces region-level mask-guided preference optimization. Existing preference alignment is rarely designed for multi-view driving, where preferences should be consistent across cameras and concentrated on motion-dominant regions, motivating our multi-view motion-aware masking and region-aware DPO with progressive corruption fusion.

## 3. Problem Formulation

We consider the task of generating multi-view driving scenarios at a user-specified risk level. Let $\mathcal{V} = \{1, \dots, V\}$ denote the camera set and $\mathbf{I}_{1:T} = \{I^v_{1:T}\}_{v\in\mathcal{V}}$ denote the observed multi-view history. Given structured context $\mathbf{C}$ (including HD maps $\mathbf{M}$ and text $\mathbf{y}$), the user provides a target risk control $\mathbf{r}^*$, which can be a scalar $r^*$ or a time-resolved profile $\mathbf{r}^*_{1:H}$ over a future horizon $H$.

Unlike passive prediction models, our framework reformulates driving risk as a proactive control signal used for structured conditioning. Let $\mathcal{A}$ denote the set of traffic participants. For each agent $a \in \mathcal{A}$, we define the motion control signal $\mathbf{U}$ as the set of future trajectories and 3D bounding boxes:

$$\mathbf{U} = \left\{ \left( \mathbf{x}^a_{1:H}, \mathbf{b}^a_{1:H} \right) \right\}_{a\in\mathcal{A}}. \tag{1}$$

To achieve risk-controllable synthesis, we decouple the process into two stages. First, a risk control module $g(\cdot)$ generates physically-grounded motion $\mathbf{U}$ conditioned on the target risk:

$$\mathbf{U} = g(\mathbf{I}_{1:T}, \mathbf{M}, \mathbf{y}, \mathbf{r}^*)\,. \tag{2}$$

Second, a multi-view diffusion generator $f(\cdot)$ renders the future frames $\hat{\mathbf{I}}_{T+1:T+H}$ conditioned on the structured motion control:

$$\hat{\mathbf{I}}_{T+1:T+H} = f(\mathbf{I}_{1:T}, \mathbf{M}, \mathbf{y}, \mathbf{U}) \,. \tag{3}$$

Our objective is to generate diverse long-tail scenarios where the induced risk matches $\mathbf{r}^*$ while preserving cross-view consistency and temporal plausibility.

# 4. Risk-Controlled Motion Generation

This section presents how we model driving risk, reveals the risk distribution in typical scenarios, and shows how to control the risk level to generate corresponding trajectories and 3D boxes.

## 4.1. Per-frame Risk Computation

At each time step $t$, we assume the ego vehicle state $(\mathbf{p}_e^t, \mathbf{v}_e^t)$ and a set of surrounding agents $\mathcal{A}_t = \{1, \dots, N_t\}$ with states $(\mathbf{p}_i^t, \mathbf{v}_i^t)$ are available from logs or annotations. We compute a per-agent risk contribution $R_i^t$ and optionally aggregate them into a per-frame scalar risk $r_t$. Our formulation follows the field-inspired, direction-aware risk modeling principle in RiskNet [20], and adapts it to an efficient per-frame computation.

Let the relative displacement and its unit direction be

$$\mathbf{r}_i^t = \mathbf{p}_i^t - \mathbf{p}_e^t, \qquad \hat{\mathbf{r}}_i^t = \frac{\mathbf{r}_i^t}{\|\mathbf{r}_i^t\|_2 + \epsilon}. \tag{4}$$

We use two dot products to indicate whether ego and agent $i$ are moving towards each other:

$$d_{e\to i}^t = (\mathbf{v}_e^t)^\top \mathbf{r}_i^t, \qquad d_{i\to e}^t = (\mathbf{v}_i^t)^\top (-\mathbf{r}_i^t). \tag{5}$$

A larger positive $d_{e\to i}^t$ implies ego is moving towards agent $i$, and a larger positive $d_{i\to e}^t$ implies agent $i$ is moving towards ego.

Moreover, we map the approaching cues to a discrete interaction weight $\omega_i^t$ that reflects asymmetric driving risk:

$$\omega_i^t = \begin{cases} \omega_{\text{bi}}, & d_{e\to i}^t > 0 \ \wedge \ d_{i\to e}^t > 0, \\ \omega_{\text{agent}}, & d_{e\to i}^t \le 0 \ \wedge \ d_{i\to e}^t > 0, \\ \omega_{\text{ego}}, & d_{e\to i}^t > 0 \ \wedge \ d_{i\to e}^t \le 0, \\ \omega_{\text{away}}, & \text{otherwise.} \end{cases} \tag{6}$$

Typically $\omega_{\text{bi}}$ is the largest since both parties are closing, while $\omega_{\text{agent}} > \omega_{\text{ego}}$ reflects that an approaching agent can be riskier when ego may not react in time.

Also, we introduce an agent-type coefficient $\mu_i$ to reflect severity differences across categories (e.g., heavy trucks vs. pedestrians):

$$\mu_i = \text{TypeCoeff}(\text{cls}_i), \tag{7}$$

where $\text{cls}_i$ is the category label.

To calibrate longitudinal and lateral factors, we define a closing speed along the line of sight

$$s_i^t = \max\Big(0, \ (\mathbf{v}_e^t - \mathbf{v}_i^t)^\top \hat{\mathbf{r}}_i^t\Big), \tag{8}$$

and convert it to a longitudinal amplification factor

$$\alpha_i^t = \exp(\kappa \, s_i^t), \tag{9}$$

where $\kappa$ controls sensitivity. To penalize lateral (non-colliding) motion, we compute the sine term using the relative velocity $\mathbf{v}_{\text{rel}}^t = \mathbf{v}_e^t - \mathbf{v}_i^t$:

$$\sin^2\theta_i^t = \frac{\|\mathbf{v}_{\text{rel}}^t \times \hat{\mathbf{r}}_i^t\|_2^2}{\|\mathbf{v}_{\text{rel}}^t\|_2^2 + \epsilon}, \qquad \beta_i^t = \exp(-\lambda \, \sin^2\theta_i^t), \tag{10}$$

where $\lambda$ controls the lateral attenuation.

Therefore, the final per-agent risk contribution is

$$R_i^t = K \, C \cdot \frac{\omega_i^t \, \mu_i \, \alpha_i^t \, \beta_i^t}{\|\mathbf{r}_i^t\|_2 + \epsilon}, \tag{11}$$

where $K$ and $C$ are calibration constants. We optionally aggregate per-agent risks into a per-frame scalar risk

$$r_t = \sum_{i \in \mathcal{A}_t} R_i^t \quad \text{or} \quad r_t = \max_{i \in \mathcal{A}_t} R_i^t, \tag{12}$$

depending on whether we emphasize cumulative risk or the most critical interaction.

## 4.2. Risk Mining in Typical Scenarios

Driving risk is often not visually obvious. Many safety-critical situations arise from subtle conflicts among the ego vehicle, surrounding traffic participants, and road geometry, where small timing differences, limited sight distance, or tight lateral clearance can quickly increase crash probability. To reveal such "hidden" hazards, we quantify risk per frame and use the resulting signals to analyze when and why risk accumulates in typical driving scenarios.

Large-scale crash statistics further motivate this analysis. NHTSA introduced a pre-crash scenario typology that categorizes hazardous situations into 37 distinct scenarios using the General Estimates System (GES) crash database [23]. A later national analysis reorganized these into nine major groups (rear-end, crossing paths, lane change, road departure, control loss, opposite direction, animal, pedestrian, and pedalcyclist) using 2011–2015 FARS and GES data [26]. Among police-reported crashes where the light vehicle performs the critical action, rear-end and crossing-path conflicts dominate crash exposure (about 34% and 23%), followed by lane-change conflicts (about 13%) and road-departure events (about 11%). Opposite-direction crashes occur less frequently (about 2%) but show a much higher

fatal-to-all crash ratio (about 3.23%), indicating that rare scenarios can be disproportionately severe. These findings suggest that risk-aware generation should target not only frequent crash types but also rare, high-severity scenarios often overlooked in manual scenario design.

We mine latent risks in diverse traffic scenarios by computing a per-frame risk signal and analyzing when risk accumulates over time. Figure 1 presents three representative left-turn cases. In Fig. 1(a), the ego vehicle performs an unprotected left turn at an unsignalized intersection while an oncoming vehicle proceeds straight, creating a lateral collision threat. The risk rises sharply around the mid-turn phase when the ego vehicle becomes laterally exposed to the oncoming stream. In Fig. 1(b), two vehicles execute parallel left turns with small lateral separation, where slight asynchrony in turning trajectories can induce a sideswipe risk that is hard to anticipate from high-level intent alone. In Fig. 1(c), a parking-lot exit appears immediately after a left turn, forming a localized blind spot, and the risk spikes as the ego vehicle approaches the exit region where cross-traffic may emerge. These examples show that non-intuitive risks are concentrated in specific moments and regions, motivating the use of time-resolved risk as a controllable signal for long-tail motion synthesis and scenario generation.

### 4.3. Controllable Motion Generation

We treat risk as a conditioning variable for motion synthesis. Given a target risk $\mathbf{r}^*$ (either a scalar $r^*$ or a profile $\mathbf{r}^*_{1:H}$), our goal is to generate future motion $\mathbf{U} = \{(\mathbf{x}^a_{1:H}, \mathbf{b}^a_{1:H})\}_{a\in\mathcal{A}}$ such that the induced risk matches the target. We build on a GENAD-style conditional motion generator [44] and introduce an explicit risk-conditioning path.

We then encode the target risk with an MLP $h(\cdot)$ and apply feature-wise linear modulation (FiLM) on the BEV (or latent) feature map $\mathbf{F}$:

$$\mathbf{e}_r = h(\mathbf{r}^*), \qquad (\boldsymbol{\gamma}_r, \boldsymbol{\beta}_r) = \psi(\mathbf{e}_r), \qquad \tilde{\mathbf{F}} = \boldsymbol{\gamma}_r \odot \mathbf{F} + \boldsymbol{\beta}_r, \tag{13}$$

where $\psi(\cdot)$ is a linear head and $\odot$ denotes element-wise multiplication. The modulated feature $\tilde{\mathbf{F}}$ is then used to generate multi-modal future trajectories and 3D boxes.

Let the generator produce $M$ motion hypotheses $\{\hat{\mathbf{U}}^{(m)}\}_{m=1}^{M}$. We compute a scalar risk for each mode using the per-frame risk functional in Eqs. (11)–(12), denoted by $\mathcal{R}(\hat{\mathbf{U}}^{(m)})$. We enforce a lower-bound matching and an upper-bound capability constraint:

$$R_{\min} = \min_m \mathcal{R}\left(\hat{\mathbf{U}}^{(m)}\right), \qquad R_{\max} = \max_m \mathcal{R}\left(\hat{\mathbf{U}}^{(m)}\right),$$
$$\mathcal{L}_{\min} = \left|R_{\min} - r^*\right|, \qquad \mathcal{L}_{\max} = \left|R_{\max} - \tau r^*\right|. \tag{14}$$

where $\tau$ is a scaling factor (e.g., $\tau = 2$) that encourages the model to retain the capacity to generate higher-risk modes.

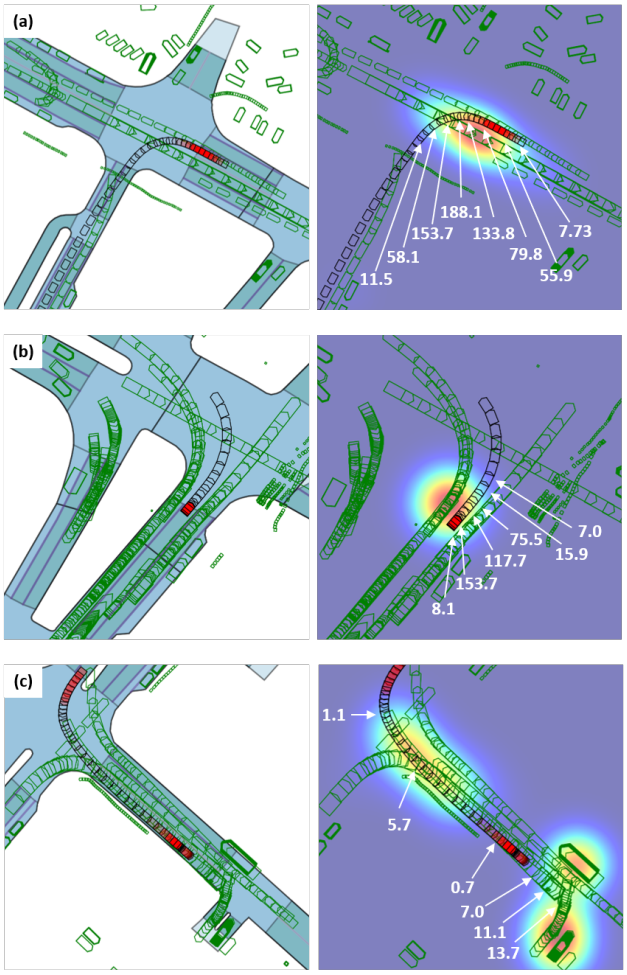


Figure 1. Examples of mined potential risks in typical left-turn scenarios using the proposed per-frame risk quantification. The numbers indicate the risk coefficient of the ego vehicle at that location. (a) Unprotected left turn at an unsignalized intersection with an oncoming straight-moving vehicle. (b) Parallel left turns with tight lateral clearance. (c) Left turn followed by a nearby parking-lot exit that creates a local blind spot.

The final training loss combines the base GENAD objective $\mathcal{L}_{\text{GENAD}}$ with risk guidance:

$$\mathcal{L}_{\text{risk}} = \mathcal{L}_{\text{GENAD}} + \lambda_{\min}\mathcal{L}_{\min} + \lambda_{\max}\mathcal{L}_{\max}. \tag{15}$$

Since $r_t$ is computed per frame, we can localize when risk emerges within a scenario by inspecting peaks or sustained segments of $\{r_t\}_{t=1}^{H}$. This provides interpretable evidence of non-intuitive risk patterns and supports targeted conditioning of motion generation at a desired risk level.

Therefore, our GENAD-based module produces risk-conditioned trajectories and 3D boxes for ego and other agents, which are used as the primary motion controls for the subsequent multi-view diffusion model to generate driving videos consistent with the specified risk level.

# 5. Multi-View Diffusion for Driving Scenarios

Figure 2 illustrates the architecture of the proposed framework, termed *RiskMV-DPO*, which integrates diffusion-based generation with multi-view localized preference alignment for risk-aware driving scenario synthesis. Given risk-conditioned motion controls, including trajectories and 3D bounding boxes, the model generates temporally coherent multi-view driving videos. A multimodal encoder fuses view tokens, motion signals, and scenario context, while geometry-aware conditioning and localized DPO guide the generation toward spatially consistent and dynamically plausible driving scenarios.

## 5.1. Geometry-Appearance Alignment

Multi-view diffusion models can produce visually plausible frames, but they often lack a strong geometric prior and may break metric-level cross-view consistency. We inject compact 3D priors by extracting geometry-aware features from DGGT and its VGGT backbone [3, 29], and aligning them with the diffusion model's appearance features during training.

**Geometric feature extraction.** Given the first-frame multi-view images $I \in \mathbb{R}^{B\times N_c\times C\times H\times W}$, where $B$ is the batch size, $N_c$ is the number of cameras, and $C$ is the number of channels, we extract patch-wise VGGT features

$$F^{\mathrm{VGGT}} = \Phi_{\mathrm{VGGT}}(I), \qquad F^{\mathrm{VGGT}} \in \mathbb{R}^{(BN_c)\times P\times D_{\mathrm{VGGT}}}, \tag{16}$$

where $P$ is the number of patches and $D_{\mathrm{VGGT}}$ is the VGGT feature dimension (e.g., 3072). A learnable projection maps VGGT features to the diffusion latent dimension $D$ (e.g., 1152):

$$F = \mathrm{MLP}_{\mathrm{proj}}(F^{\mathrm{VGGT}}), \qquad F \in \mathbb{R}^{(BN_c)\times P\times D}. \tag{17}$$

**Learnable geometric token compression.** To obtain a fixed number of compact geometric tokens, we introduce learnable queries $Q \in \mathbb{R}^{N_{\mathrm{tok}}\times D}$ (e.g., $N_{\mathrm{tok}}$=16) and compress $F$ by cross-attention:

$$G = \mathrm{Attn}(Q, F, F), \qquad G \in \mathbb{R}^{(BN_c)\times N_{\mathrm{tok}}\times D}. \tag{18}$$

For classifier-free guidance, we apply conditional dropout by replacing $G$ with a learned null token set $G_\varnothing$ using a Bernoulli mask $m_{\mathrm{drop}} \sim \mathrm{Bernoulli}(1-p_{\mathrm{drop}})$:

$$\tilde{G} = m_{\mathrm{drop}}\, G + (1-m_{\mathrm{drop}})\, G_\varnothing. \tag{19}$$

**Appearance feature extraction.** During diffusion training, we extract intermediate token features from the last $K$ backbone layers (e.g., $K$=8). Let $R^{(\ell)} \in \mathbb{R}^{(BN_c)\times N_s\times D}$ be the token features from layer $\ell$, where $N_s$ is the number of spatial tokens. We spatially pool tokens within each selected layer and average across layers:

$$r^{(\ell)} = \frac{1}{N_s}\sum_{i=1}^{N_s} R^{(\ell)}_{:,i,:}, \qquad r = \frac{1}{K}\sum_{\ell} r^{(\ell)}, \tag{20}$$

where $r \in \mathbb{R}^{(BN_c)\times D}$ is the final appearance feature.

**Alignment loss.** We pool geometric tokens into a single vector $g \in \mathbb{R}^{(BN_c)\times D}$, normalize both $g$ and $r$, and align them with cosine similarity:

$$g = \frac{1}{N_{\mathrm{tok}}}\sum_{k=1}^{N_{\mathrm{tok}}} \tilde{G}_{:,k,:}, \qquad \hat{g} = \frac{g}{\|g\|_2}, \quad \hat{r} = \frac{r}{\|r\|_2}, \tag{21}$$

$$\mathcal{L}_{\mathrm{align}} = 1 - \frac{1}{BN_c}\sum_{i=1}^{BN_c} \hat{g}_i^\top \hat{r}_i. \tag{22}$$

This alignment injects compact DGGT/VGGT-derived geometric priors into diffusion training and improves cross-view consistency.

## 5.2. Multi-view Consistent Motion-Aware Masking

In driving videos, most pixels belong to static background, while safety-critical cues are concentrated in dynamic regions such as moving agents and their interactions. Uniform corruption wastes model capacity and weakens learning on motion-dominant areas. We therefore construct a motion-aware mask and enforce multi-view consistency so that corruption focuses on dynamic regions while remaining geometrically aligned across cameras.

**Motion-aware mask.** Let $x \in \mathbb{R}^{B\times N_c\times C\times T\times H\times W}$ denote the video latent (or feature) tensor, where $B$ is batch size, $N_c$ is the number of cameras, $C$ is the channel dimension, and $T$ is the number of frames. We estimate per-frame motion magnitude by temporal differencing and channel averaging:

$$D_{b,c,t} = \left|x_{b,c,:,t+1} - x_{b,c,:,t}\right|,$$
$$M^{\mathrm{mot}}_{b,c,t} = \frac{1}{C}\sum_{j=1}^{C} D_{b,c,t,j}. \tag{23}$$

We normalize $M^{\mathrm{mot}}$ to $[0,1]$ and optionally apply a soft threshold to suppress weak motion.

**Geometric consistency across views.** Motion cues alone do not guarantee cross-view consistency. We additionally render a geometry-consistent mask by sampling 3D points along agent trajectories (or Bézier curves) and projecting them into each camera view using known intrinsics and extrinsics. Let $\mathbf{p} = (x, y)$ be a pixel location

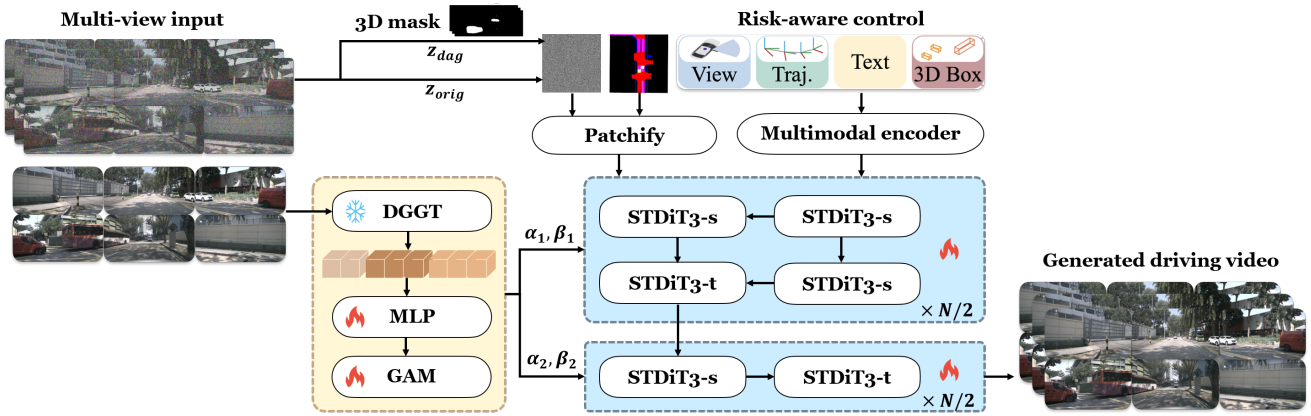


Figure 2. Overview of the proposed *RiskMV-DPO* framework. Given multi-view observations and scenario context, trajectories and 3D bounding boxes generated by the risk control module at a specified risk level are used as structured motion conditions. A multimodal encoder embeds view tokens and motion cues, which are then injected into a diffusion backbone composed of spatial and temporal STDiT3 blocks. Region-aware DPO further aligns the generation toward localized dynamic regions, producing temporally coherent multi-view driving videos consistent with the specified risk-conditioned motion.

and $\mathbf{u}_{bck} = (u^x_{bck}, u^y_{bck})$ be the projected center of the $k$-th point for batch $b$ and camera $c$, with visibility indicator $V_{bck} \in \{0, 1\}$. We rasterize a Gaussian blob:

$$M^{\text{geo}}_{bck}(\mathbf{p}) = V_{bck} \exp\left(-\frac{(x - u^x_{bck})^2}{2\sigma_w^2} - \frac{(y - u^y_{bck})^2}{2\sigma_h^2}\right). \tag{24}$$

We aggregate over sampled points to obtain a per-frame, per-view geometric mask $M^{\text{geo}}_{b,c,t}(\mathbf{p})$.

**Mask fusion.** We fuse motion and geometric cues to form the final corruption mask:

$$M = \text{clip}\big(M^{\text{geo}} \odot M^{\text{mot}},\, 0,\, 1\big), \tag{25}$$

so that corruption focuses on motion-dominant regions while remaining consistent across cameras.

### 5.3. Region-Aware Direct Preference Optimization

Global preference signals can be ambiguous for driving videos and may underweight dynamic regions that dominate risk. We align the model using localized preference pairs constructed on the masked regions, while keeping the unmasked background unchanged.

**Progressive corruption fusion.** Given a noise level $t$ and clean latent $z_0$, we corrupt only the masked region and keep the unmasked region clean:

$$z^{\text{mask}}_t = M \odot z_t + (1 - M) \odot z_0, \tag{26}$$

where $z_t$ is the noised latent at level $t$. To construct localized preference pairs, we use two corruption strengths $t_w < t_l$ on the masked region:

$$\begin{aligned} z^{\text{mask}}_{t_w} &= M \odot z_{t_w} + (1 - M) \odot z_0, \\ z^{\text{mask}}_{t_l} &= M \odot z_{t_l} + (1 - M) \odot z_0. \end{aligned} \tag{27}$$

$t_w$ yields an easier (preferred) target on the masked region, while $t_l$ yields a harder (dispreferred) one, providing localized supervision without altering global context.

**Masked flow-matching proxy.** We maintain an EMA reference model $v_{\text{ref}}$. Using the negative masked flow-matching residual as a proxy for log-probability, we define

$$\begin{aligned} \log p_\theta(y \mid x) &\approx -\,\text{FM}_\theta(y), \\ \delta_\theta &= v_\theta(z^{\text{mask}}_t, t, \mathbf{c}) - (\epsilon - z_0), \\ \text{FM}_\theta(y) &= \frac{1}{\|M\|_1} \big\|M \odot \delta_\theta\big\|_2^2, \end{aligned} \tag{28}$$

where $\mathbf{c}$ denotes the conditioning (history, map, optional text, and motion tokens), and $\|M\|_1$ normalizes by masked area.

**Region-aware DPO objective.** We apply DPO on the localized pair $(y_w, y_l)$:

$$\begin{aligned} \mathcal{L}_{\text{RA-DPO}} &= -\,w(t)\, \log \sigma\Big(\beta\big[\Delta_\theta(y_w) - \Delta_\theta(y_l)\big]\Big), \\ \Delta_\theta(y) &= \log p_\theta(y \mid x) - \log p_{\text{ref}}(y \mid x), \end{aligned} \tag{29}$$

where $\sigma(\cdot)$ is the sigmoid, $\beta$ is the DPO temperature, and $w(t)$ is a noise-adaptive weight. This objective aligns generation with localized preferences in dynamic regions.

| Method | FID↓ | FVD↓ | mAP↑ |
|---|---|---|---|
| DriveDreamer-2 | 25.00 | 105.10 | – |
| Panacea | 16.96 | 139.00 | – |
| MagicDriveV2 | 20.91 | 94.84 | 18.17 |
| **RiskMV-DPO (ours)** | **15.70** | **87.65** | **30.50** |

Table 1. Comparison with representative driving video generators on nuScenes.

**Supervised fine-tuning loss.** We apply supervised training on the masked region using the flow-matching residual.

$$\mathcal{L}_{\mathrm{SFT}} = \frac{1}{\|M\|_1} \left\| M \odot \delta_\theta \right\|_2^2, \qquad \delta_\theta = v_\theta(z_t^{\mathrm{mask}}, t, \mathbf{c}) - (\epsilon - z_0). \tag{30}$$

**Total objective.** The complete training objective is

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{SFT}}\, \mathcal{L}_{\mathrm{SFT}} + \lambda_{\mathrm{RA}}\, \mathcal{L}_{\text{RA-DPO}} + \lambda_{\mathrm{align}}\, \mathcal{L}_{\mathrm{align}}. \tag{31}$$

# 6. Experiments

## 6.1. Experimental Setup

We conduct experiments on the nuScenes dataset [1] using the MagicDrive-STDiT3 backbone. We use 700 scenes for training and 150 scenes for evaluation. All videos are processed at $448 \times 840$ resolution per camera, and each clip contains $T$=16 frames. We compare RiskMV-DPO with MagicDriveV2 [8], DriveDreamer-2 [42], and Panacea [33], and also report ablations of our components.

We report eight metrics that cover visual fidelity, temporal quality, and multi-view geometry consistency. FID measures overall visual quality, and FVD measures temporal coherence. We evaluate 3D realism using mAP from a pretrained 3D detector applied to generated videos. To assess multi-view consistency, we report MV-SSIM computed on overlapping regions between adjacent cameras. For geometry accuracy, we report Depth AbsRel using VGGT-predicted depth against reference depth. We further report Foreground FID (Fg-FID) on cropped foreground objects and DINO-FID in DINO feature space for semantic fidelity.

We initialize from the MagicDriveV2 checkpoint [8] and train for 20,000 steps with AdamW. We use a learning rate of $2 \times 10^{-5}$ and batch size 2 per GPU on 8 GPUs. We adopt BF16 training and an EMA model with decay $\gamma$=0.9999. The trainable parameters include the VGGT geometry adapter ($\sim$45M), the GAM module ($\sim$12M), and the motion-mask generator ($<$ 1M), totaling $\sim$58M trainable parameters on top of a $\sim$2.5B diffusion backbone. All experiments are run on NVIDIA H800-90GB GPUs.

## 6.2. Main Results

After extensive experiments, we find that the trajectories and 3D bounding boxes generated under a specified risk control lead to synthesized scenarios whose measured risk closely matches the target value. Due to space limitations, we omit the detailed analysis here.

Table 1 reports quantitative results on nuScenes. RiskMV-DPO achieves consistent improvements over MagicDriveV2. It reduces FID from 20.91 to 15.70 and improves FVD from 94.84 to 87.65, indicating better visual fidelity and temporal coherence. It also substantially improves 3D detection mAP from 18.17 to 30.50, suggesting that the generated videos contain more realistic and geometrically consistent 3D cues. Compared with other recent methods, RiskMV-DPO improves over DriveDreamer-2 (FID 25.00 $\rightarrow$ 15.70) and Panacea (FID 16.96 $\rightarrow$ 15.70).

## 6.3. Ablation Study

Table 2 evaluates the contribution of the key components in our framework. Starting from the baseline (A), introducing Random Mask DPO (B) slightly improves temporal quality and geometric consistency, but leaves FID unchanged. Replacing random masks with motion-aware masks (C) yields a clearer gain, improving FID (20.91 $\rightarrow$ 19.42) and MV-SSIM (0.812 $\rightarrow$ 0.825), indicating that focusing training on dynamic regions provides more informative localized preference signals. Using a 3D multi-view mask (D) further strengthens cross-view consistency, improving MV-SSIM to 0.841 and reducing Depth AbsRel to 0.228.

Next, adding VGGT-GAM without alignment (E) substantially improves geometric accuracy, reducing Depth AbsRel from 0.228 to 0.205, which validates the benefit of injecting geometry priors. Finally, enabling the geometry-appearance alignment loss (F) achieves the best overall performance, with FID 15.70, MV-SSIM 0.856, and Depth AbsRel 0.204.

## 6.4. Qualitative Visualization Across Risk Levels

Figure 3 shows qualitative examples generated by RiskMV-DPO at different target risk levels. Due to space limitations, we show three representative views (front-left, front, and front-right) and visualize a 7-frame clip by sampling one frame every 5 frames. As shown in Fig. 3(a), the higher-quantile scenario depicts the ego vehicle going straight through an unsignalized intersection while encountering a fast left-turning heavy truck at very close distance under partial occlusion, resulting in an imminent collision risk. In the same scene, the medium-quantile example in Fig. 3(b) shows the truck turning more slowly and keeping a larger clearance to the ego vehicle, which reduces the near-collision risk. The lower-quantile case in Fig. 3(c) corresponds to a common straight-driving pattern with no immediate conflict, leading to a low estimated risk.

We note that *high*, *medium*, and *low* risk do not have a strict boundary in our setting. Instead, the reported quantiles indicate the relative position of the induced risk coef-

| Configuration | FID↓ | FVD↓ | MV-SSIM↑ | Depth AbsRel↓ | Fg-FID↓ | DINO-FID↓ |
|---|---|---|---|---|---|---|
| (A) Baseline | 20.91 | 94.84 | 0.812 | 0.250 | 35.0 | 28.4 |
| (B) + Random Mask DPO | 20.91 | 93.01 | 0.818 | 0.243 | 33.6 | 27.9 |
| (C) + Motion-Aware Mask | 19.42 | 91.67 | 0.825 | 0.236 | 32.1 | 27.4 |
| (D) + 3D MV Mask | 19.33 | 89.51 | 0.841 | 0.228 | 31.0 | 26.8 |
| (E) + VGGT-GAM (no Align) | 16.79 | 88.70 | 0.848 | 0.205 | 30.2 | 26.3 |
| **(F) RiskMV-DPO (ours)** | **15.70** | **87.65** | **0.856** | **0.204** | **29.5** | **25.8** |

Table 2. Ablation study on nuScenes.

(a) 0.95 quantile (higher risk) (b) 0.8 quantile (medium risk) (c) 0.2 quantile (lower risk)

Figure 3. Qualitative examples generated by RiskMV-DPO at different target risk quantiles. The quantiles indicate the relative position of the induced risk coefficient among all generated scenarios. (a) Higher risk: an unsignalized intersection with a fast left-turning heavy truck at very close distance under partial occlusion. (b) Medium risk: the same interaction with slower turning and larger clearance, reducing near-collision risk. (c) Lower risk: a common straight-driving scenario with no immediate conflict.

ficient among all generated scenarios, where 0.95, 0.8, and 0.2 correspond to higher, medium, and lower portions of the risk distribution, respectively, which is apparent on the provided examples. Code and additional qualitative results (including videos) are available at https://github.com/venshow-w/RiskMV-DPO.

## 7. Conclusion and Limitations

In this paper, we present RiskMV-DPO, a general and systematic pipeline for physically-informed, risk-controllable multi-view driving scenario generation. It transforms driving risk from an after-the-fact label into an actionable control signal that guides scene generation. By coupling risk-conditioned motion synthesis with diffusion-based multi-view video generation, the framework produces temporally coherent driving scenes consistent with specified risk levels. Experimental results on the nuScenes dataset demonstrate that RiskMV-DPO achieves state-of-the-art performance, notably improving 3D detection mAP to 30.50 and reducing FID to 15.70. More broadly, our work highlights the potential of integrating risk-aware modeling with generative world simulation for safety-oriented development of autonomous driving systems.

Despite these results, several limitations remain. While the proposed risk modeling framework captures a range of typical traffic interactions, many potential risk patterns in complex real-world environments remain underexplored. Besides, the overall performance depends on the quality of trajectories and 3D bounding boxes produced by the risk control module. Improving the robustness and generality of risk-conditioned motion generation is an important direction for future research.

# References

[1] 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 (CVPR)*, 2020. 7

[2] Xuan Cai, Xuesong Bai, Zhiyong Cui, Danmu Xie, Daocheng Fu, Haiyang Yu, and Yilong Ren. Text2scenario: Text-driven scenario generation for autonomous driving test. *Automotive Innovation*, pages 1–26, 2026. 1

[3] Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Hongyang Li, Ya-Qin Zhang, and Hao Zhao. Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. *arXiv preprint arXiv:2512.03004*, 2025. 5

[4] Tommaso Dreossi, Daniel J Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A Seshia. Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *International Conference on Computer Aided Verification*, pages 432–442. Springer, 2019. 2

[5] Yang Fei, Peng Shi, Yang Liu, and Liang Wang. Critical roles of control engineering in the development of intelligent and connected vehicles. *Journal of Intelligent and Connected Vehicles*, 7(2):79–85, 2024. 1

[6] Daniel J Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation*, pages 63–78, 2019. 2

[7] 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. 2

[8] Ruiyuan Gao, Kai Chen, et al. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024. 7

[9] 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*, pages 28135–28144, 2025. 2

[10] Xu He, Ji Li, Chuan Hu, Mingming Liu, and Hongming Xu. Dual-discriminator generative adversarial network with long-tail feature capture for extreme scenarios in human–machine shared driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 1

[11] Yixu He, Hongyi Lin, Lan Yang, and Yang Liu. Generative models for the evolution of transportation systems. *Journal of Traffic and Transportation Engineering (English Edition)*, 2025. 1

[12] 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. 2

[13] Zitong Huang, Kaidong Zhang, Yukang Ding, Chao Gao, Rui Ding, Ying Chen, and Wangmeng Zuo. Mind the generative details: Direct localized detail preference optimization for video diffusion models. *arXiv preprint arXiv:2601.04068*, 2026. 2

[14] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5820–5829, 2021. 2

[15] Mark Koren, Saud Alsaif, Ritchie Lee, and Mykel J Kochenderfer. Adaptive stress testing for autonomous vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2018. 2

[16] 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*, pages 469–485. Springer, 2024. 2

[17] Hongyi Lin, Yang Liu, Shen Li, and Xiaobo Qu. How generative adversarial networks promote the development of intelligent transportation systems: A survey. *IEEE/CAA journal of automatica sinica*, 10(9):1781–1796, 2023. 1

[18] Hongyi Lin, Yang Liu, Liang Wang, and Xiaobo Qu. Big data-driven advancements and future directions in vehicle perception technologies: From autonomous driving to modular buses. *IEEE Transactions on Big Data*, 11(3):1568–1587, 2025. 2

[19] Hongyi Lin, Yang Liu, Liang Wang, and Xiaobo Qu. A high-precision calibration and evaluation method based on binocular cameras and lidar for intelligent vehicles. *IEEE Transactions on Vehicular Technology*, 74(5):7404–7415, 2025. 1

[20] Qichao Liu, Heye Huang, Shiyue Zhao, Lei Shi, Soyoung Ahn, and Xiaopeng Li. Risknet: interaction-aware risk forecasting for autonomous driving in long-tail scenarios. *Transportation Research Part E: Logistics and Transportation Review*, 205:104478, 2026. 2, 3

[21] 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*, 2024. 2

[22] Yuewen Mei, Tong Nie, Jian Sun, and Ye Tian. Llm-attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 1

[23] Wassim G. Najm, David L. Smith, and Mikio Yanagisawa. Pre-crash scenario typology for crash avoidance research. Technical Report DOT HS 810 767, National Highway Traffic Safety Administration, 2007. Based on 2004 NASS GES crash database. 3

[24] Xiaobo Qu, Hongyi Lin, and Yang Liu. Envisioning the future of transportation: Inspiration of chatgpt and large models, 2023. 1

[25] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*, 2017. 2

[26] Elizabeth D. Swanson, Frank Foderaro, Mikio Yanagisawa, Wassim G. Najm, and Philip Azeredo. Statistics of light-vehicle pre-crash scenarios based on 2011–2015 national crash data. Technical Report DOT HS 812 745, National Highway Traffic Safety Administration, 2019. Scenario groups and statistics based on 2011–2015 FARS and NASS GES. 3

[27] Xuewei Tang, Mengmeng Yang, Tuopu Wen, Peijin Jia, Le Cui, Mingshan Luo, Kehua Sheng, Bo Zhang, Kun Jiang, and Diange Yang. Priorfusion: Unified integration of priors for robust road perception in autonomous driving. *Communications in Transportation Research*, 5:100229, 2025. 2

[28] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2

[29] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 5

[30] Jie Wang, Guang Li, Zhijian Huang, Chenxu Dang, Hangjun Ye, Yahong Han, and Long Chen. Vggdrive: Empowering vision-language models with cross-view geometric grounding for autonomous driving. *arXiv preprint arXiv:2602.20794*, 2026. 1

[31] 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*, pages 55–72. Springer, 2024. 2

[32] Yong Wang, Daifeng Zhang, Yanqiang Li, Liguo Shuai, Zhicheng Tang, and Yuxiang Hou. Safety-critical scenario test for intelligent vehicles via hybrid participation of natural and adversarial agents. *Journal of Intelligent and Connected Vehicles*, 8(3):9210066–1, 2025. 1

[33] Yu Wen et al. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[34] Lukas Westhofen, Christian Neurohr, Tjark Koopmann, Martin Butz, Barbara Schütt, Fabian Utesch, Birte Neurohr, Christian Gutenkunst, and Eckard Böde. Criticality metrics for automated driving: A review and suitability analysis of the state of the art: L. westhofen et al. *Archives of Computational Methods in Engineering*, 30(1):1–35, 2023. 2

[35] Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, and Chenjing Ding. Drivescape: High-resolution driving video generation by multi-view feature fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17187–17196, 2025. 2

[36] Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025. 2

[37] Chejian Xu, Aleksandr Petiushko, Ding Zhao, and Bo Li. Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8797–8805, 2025. 1

[38] Mingxing Xu, Hongyi Lin, and Yang Liu. A deep learning approach for vehicle velocity prediction considering the influence factors of multiple lanes. *Electronic Research Archive*, 31(1):401, 2023. 2

[39] Cong Zhang, Bangyang Wei, Yang Liu, and Samuel Labi. World model-based long-tail and scenario-specific generation for autonomous driving. *Journal of Intelligent and Connected Vehicles*, 2026. 1

[40] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15459–15469, 2024. 1

[41] Meng Zhang, Jiatong Xu, Ying Gao, Dandan Shen, and Zhigang Xu. Can combined virtual-real testing speed up autonomous vehicle testing? findings from aeb field experiments. *Communications in Transportation Research*, 5: 100216, 2025. 1

[42] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 7

[43] 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*, pages 10412–10420, 2025. 2

[44] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024. 4

[45] Zhiyuan Zhou, Heye Huang, Boqi Li, Shiyue Zhao, Yao Mu, and Jianqiang Wang. Safedrive: Knowledge-and data-driven risk-sensitive decision-making for autonomous vehicles with large language models. *Accident Analysis & Prevention*, 224:108299, 2026. 2

[46] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. International Conference on Learning Representations (ICLR 2025), 2025. 2