A Diffusion Model for Generating Safety-Critical Rural Driving Video Data

A Technical Report Submitted to the Rural Safe Efficient Advanced Transportation Center (R-SEAT) and United States Department of Transportation

FINAL REPORT

Principal Investigator:
Ruwen Qin, Ph.D.
Associate Professor and Graduate Program Director
Department of Civil Engineering
Stony Brook University
2427 Old Computer Science Building
Phone: +1(631) 632-9341
E-mail: ruwen.qin@stonybrook.edu

Research Assistant:
Ke Li, M. Sc.
Ph.D. Student
Department of Civil Engineering
Stony Brook University
1208 Old Computer Science Building
E-mail: ke.li.1@stonybrook.edu

Research Assistant:
Kaidi Liang, M. Sc.
Ph.D. Student
Department of Civil Engineering
Stony Brook University
1208 Old Computer Science Building
E-mail: kaidi.liang@stonybrook.edu

December 2025

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under the grant 69A3552348321 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

METRIC CONVERSION CHART

When You Know	Multiply by	To Find
Length		
inches (in)	25.4	millimeters (mm)
feet (ft)	0.305	meters (m)
yards (yd)	0.914	meters (m)
miles (mi)	1.61	kilometers (km)
Volume		
fluid ounces (fl oz)	29.57	milliliters (mL)
gallons (gal)	3.785	liters (L)
cubic feet (ft^3)	0.028	meters cubed (m^3)
cubic yards (yd^3)	0.765	meters cubed (m^3)
Area		
square inches (in^2)	645.1	millimeters squared (mm^2)
square feet (ft^2)	0.093	meters squared (m^2)
square yards (yd^2)	0.836	meters squared (m^2)
acres	0.405	hectares (ha)
square miles (mi^2)	2.59	kilometers squared (km^2)

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle A Diffusion Model for Generating Safety-Critical Rural Driving Video Data		5. Report Date 12/15/2025
		6. Performing Organization Code 59-0977035
7. Author(s) Ruwen Qin https://orcid.org/0000-0003-2656-8705 Ke Li https://orcid.org/0009-0001-4958-3302 Kaidi Liang https://orcid.org/0009-0001-9129-2744		8. Performing Organization Report No.
9. Performing Organization Name and Address Stony Brook University		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 69A3552348321
12. Sponsoring Agency Name and Address Rural Safe Efficient Advanced Transportation Center 2525 Pottsdamer Street Tallahassee, FL 32310		13. Type of Report and Period Covered Final Report Period Covered: 06/01/2024 – 12/31/2025
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Safety-critical scenarios in rural driving are rare in public datasets, yet rural contexts exhibit elevated fatality rate per vehicle mile traveled (VMT) and present distinct challenges such as sparse infrastructure, and weak nighttime illumination. To address the scarcity of rural safety-critical video data, this project investigates diffusion-based approaches for generating driving videos and controllable rural hazard scenarios. The Historical Motion Prior Diffusion Model (HMPDM) is presented as a diffusion-based driving video prediction approach that incorporates historical motion context to improve temporal and motion consistency in multi-frame generation. In addition, since safety-critical scenario synthesis requires higher controllability than prediction-oriented generation, a rural Sim-to-Real pipeline is developed to script hazardous events in simulation and render them into realistic driving videos using an HD-map-conditioned video foundation diffusion model (Cosmos-Transfer1-7B-Sample-AV) guided by structured prompts. The pipeline is implemented in CARLA and instantiates six rural safety-critical scenario types, with further diversity introduced through adverse weather and illumination variations. Overall, this framework mitigates the lack of realistic and controllable rural safety-critical video data by enabling synthesis of long-tail hazardous events. Moreover, it can support safety evaluation studies that inform the development of more robust and reliable driving assistance and automated driving technologies in rural environments.

17. Key Words	18. Distribution Statement
Safety-critical scenarios generation, Rural areas, Diffusion model, Deep learning, Sim-to-real	No restrictions

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	42	

ACKNOWLEDGEMENTS

This project was sponsored by the Rural Safe Efficient Advanced Transportation Center (R-SEAT) and United States Department of Transportation. The Principal Investigators would like to thank the representatives of the R-SEAT Center for their valuable feedback throughout the project activities.

EXECUTIVE SUMMARY

This project investigated the challenges associated with improving driving safety in rural areas through a series of complementary studies. The findings indicate that the scarcity of safety-critical driving scenario video data, particularly under adverse weather and low-light conditions, remains a key obstacle to current research. To close this gap, this project provides a practical pipeline to synthesize realistic rural safety-critical videos for safety-oriented research by combining diffusion-based video prediction with a Sim-to-Real generation method that conditions a video foundation model on HD maps and structured prompts. Within the broader context of improving driving safety through synthetic data, our work aims to generate controllable and realistic video data for rural safety-critical scenarios, thereby providing sufficient data support for safety-related research. The main contributions are summarized as follows:

First, we developed a Historical Motion Priors Diffusion Model (HMPDM) for future driving video prediction. This spatiotemporal model leverages past frames to learn motion regularities and generates multi-frame predictions with improved temporal coherence and realistic scene dynamics. Beyond prediction accuracy, this study provides methodological evidence that diffusion-based generation can preserve motion continuity and spatiotemporal consistency in driving videos, offering useful insights for building realistic driving video generators.

Besides, based on the existing diffusion-based video foundation model COSMOS, we propose a Sim-to-Real generation pipeline designed to meet the higher controllability requirements of rural safety-critical scenario synthesis. This approach first constructed and scripted representative rural safety-critical events in CARLA, and then transformed into realistic driving videos via HD-map-conditioned generation. In this setting, HD maps serve as an explicit spatial constraint to preserve rural road geometry and infrastructure layout, while structured prompts specify the scenario type and environmental factors such as weather and lighting, enabling systematic and repeatable generation of rare events under diverse conditions.

Together, these two parts form a coherent technical narrative for rural safety-critical data generation. HMPDM establishes that diffusion models can generate temporally consistent driving dynamics by effectively leveraging motion priors, motivating diffusion as a suitable backbone for driving video modeling. Building on this, Sim-to-Real pipeline further translates this motivation into a practical, controllable approach by combining scenarios in simulation platform with generation conditioned on HD maps and prompts, producing diverse rural safety-critical clips that address data scarcity in long-tail hazardous situations.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This section provides the background information and context for the research project on safety-critical scenarios generation in rural areas.

1.1. Safety-critical Scenarios in Rural Areas

Rural roadways pose unique safety challenges that demand dedicated data-driven attention. While only about one-fifth of the U.S. population resides in rural areas, rural roads account for a disproportionately large share of roadway deaths, reported to be around 40-43% in recent summaries (U.S. Department of Transportation, 2025). Moreover, the fatality rate per 100 million Vehicle Miles Traveled (VMT) remains substantially higher in rural settings than in urban settings. For example, National Highway Traffic Safety Administration (2024) reports that the rural fatality rate has been about 1.5 times the urban rate in recent years. Despite this elevated risk, truly safety-critical events such as collisions and near-miss interactions are inherently rare, and real-world driving datasets therefore contain limited coverage of long-tail rural hazards, especially under adverse weather and low-light conditions.

This data scarcity creates a practical barrier to developing and stress-testing data-driven perception and prediction models for automated driving and improved road safety in rural environments, where collecting diverse safety-critical clips is costly and difficult. As a result, there is increasing interest in scalable and controllable synthetic data generation that can expand coverage of rare rural events while maintaining realistic scene structure. Recent work (Siddharth et al, 2025) has begun to explore diffusion-based approaches to improve the realism and utility of synthetic driving data, and diffusion-enhanced datasets have been reported to improve model robustness in underrepresented rural scenarios. Motivated by this trend, our project focuses on generating rural safety-critical driving videos with explicit controllability, using simulation to define scenario intent and HD-map-conditioned foundation diffusion models to synthesize realistic driving clips under diverse environmental conditions.

1.2. Gaps and Project Objectives

Despite the clear safety burden on rural roadways, existing driving datasets and scenario generation methods remain insufficient for studying and modeling rural safety-critical events at scale. Real-world videos of rural safety-critical events are inherently sparse, and the coverage of long-tail hazards is particularly limited under adverse weather and low-light conditions. Meanwhile, many existing scenario generation efforts emphasize urban settings, while rural environments present distinct characteristics such as sparse infrastructure, and irregular road geometry that are not well captured by common benchmarks. Even when simulation can construct hazardous interactions, a persistent gap exists between simulation outputs and realistic dashcam video appearance, limiting the direct usability of simulated data for vision-based models. Motivated by this desire, the project aims to realize objectives as follows:

- Develop a diffusion-based driving video model informed by historical motion priors to improve spatiotemporal consistency and realistic motion evolution in generated future driving sequences
- Build a Sim-to-Real generation pipeline for rural safety-critical scenarios by scripting representative hazardous scenarios in simulation, exporting HD-map representations, and

using an HD-map-conditioned diffusion foundation model with structured prompts to generate realistic driving videos

1.3. Project Relevance to the R-SEAT Themes and USDOT Strategic Plan

As an R-SEAT-funded study, this project develops new methods to address a key rural safety challenge: the lack of safety-critical driving data in rural environments. The project supports USDOT priorities and RD&T strategic goals on safety by enabling scalable generation of rare rural hazardous events. Specifically, we propose diffusion-based approaches for spatiotemporally consistent driving video modeling and a Sim-to-Real pipeline that uses HD maps and structured prompts to generate controllable rural safety-critical driving videos. These outputs can complement existing USDOT and public datasets by expanding coverage of long-tail rural scenarios for future safety-focused research and model evaluation.

1.4 Organization of the Report

This report is organized to guide the readers through the project's major activities. Section 2 presents the proposed diffusion-based driving video prediction model conditioned on historical motion priors. Section 3 then introduces a Sim-to-Real pipeline for generating rural safety-critical scenarios using HD-map-conditioned foundation diffusion models and structured prompts, which can augment existing datasets for future research. In the end, Section 4 summarizes major findings and outputs from this project.

2. HISTORICAL MOTION PRIORS DIFFUSION MODEL FOR VIDEO PREDICTION

To enable safety-critical scenario generation in rural areas, we propose a generative model built on historical motion priors. This model is designed to preserve spatiotemporal consistency in driving scenes by maintaining coherent motion evolution and realistic dynamics across frames. It also establishes a solid foundation for the subsequent Sim-to-Real generation pipeline. Contents of this section will be published as a conference paper in IV Symposium 2026.

2.1 Background

In the context of intelligent vehicles, modern autonomous driving systems need to not only perceive the present scenarios (Li et at., 2025) but also anticipate their evolution over time, which is crucial for path planning, especially in safety-critical scenarios (Teng et al., 2023). However, the complex dynamics and frequent occlusions in real-world driving scenes cause traditional object-centric predictors, which are based on low-dimensional states (e.g., trajectories), to discard essential appearance, structural, and contextual information. Emerging diffusion-based video prediction methods aim to forecast entire future scenes conditioned on past or present visual context. Crucially, scene-level video prediction offers a comprehensive and holistic understanding of the driving environment, capturing both the static background and the motion of dynamic traffic agents. Furthermore, by leveraging historical motion patterns, they naturally handle occlusions and can better represent the dynamic evolution of driving scene into the future.

Despite notable advances in diffusion-based video prediction, several key challenges remain. Many models struggle with temporal consistency, since the visual quality degrades and objects appear distorted within the long-horizon prediction. Another limitation lies in the lack of robust historical motion modeling. Without properly leveraging historical motion priors, predicted agents often exhibit unrealistic trajectories or blurry motion patterns in driving scenarios. Accompanied with it is the concern of computational efficiency and adaptability. Despite promising performance, recent models often require substantial computation, complex training pipelines, and extra multimodal inputs, revealing the need for simpler yet effective solutions for driving video generation.

To bridge the gaps, the project proposes a diffusion-based driving video prediction framework that is aware of historical motion, named Historical Motion Priors-informed Diffusion Model (HMPDM), making the following contributions:

- To implicitly inject historical motion context, Temporal-aware Latent Conditioning (TaLC) is introduced to feed the model with latent representations of past frames as a learnable prior.
- Complementing this, a Motion-aware Pyramid Encoder (MaPE) hierarchically encodes multi-scale motion features from the historical dynamic.
- To mitigate error accumulation in long-term generation, we implement Self-Conditioning (SC), a strategy where the model conditions on its own intermediate predictions.

2.2 Related Work

The literature relevant to this study includes research focused on developing indicators or measures for mobility and accessibility, determining their relationship with topological features, and examining the challenges faced by rural seniors.

2.2.1 Diffusion-based Video Prediction

Previous research primarily relies on LSTM, VAEs, and VRNNs (Villegas et al., 2017; Lotter et al., 2017; Denton and Fergus, 2018; Castrejon et al., 2019; Wu et al., 2021) to learn a probabilistic latent space, model temporal dynamics and capture spatial coherence. However, they meet the limitation on long term video prediction with temporal consistency.

Recently, diffusion-based paradigms have demonstrated remarkable performance in both video generation and prediction tasks. Diffusion model for video prediction is a generative framework that progressively transforms random Gaussian noise into coherent videos through an iterative denoising process ((Ho et al., 2020; Song et al., 2021; Karras et al., 2022)). Most studies build upon Unet denoising architecture, incorporating various conditioning strategies and spatio-temporal consistency techniques (Ho et al., 2022; Hoppe et al., 2022; Voleti et al., 2022; Yang et al., 2023; Mei and Patel, 2023; Ye and Bilodeau, 2024; Pallotta et al., 2025). Ho et al. (2022) introduces a U-Net diffusion framework that jointly models frames in spatial and temporal dimension for video prediction and generation tasks. Voleti et al. (2022) proposes a mask-based conditioning strategy, which randomly hides part of frames to realize video prediction, interpolation, and completion.

2.2.2 Historical Motion Injection

Historical frames naturally provide motion prior as the conditional context for the future video prediction. The most common ways to incorporate historical frames include direct concatenation on frames or channel dimension, or encode then inject to cross attention layer.

Ye and Bilodeau (2023) used a CNN autoencoder to extract appearance features from historical frames and a Fourier feature network to encode their spatio-temporal coordinates. These features are performed as the key, and value in cross attention layer to predict future frames. Similarly, Ye and Bilodeau (2024) leverages difference images from past frames as input to a specialized motion encoder, which disentangles motion and content features. While, Yang, Srivastava and Mandt (2023) compared three different historical motion injection schemes, and illustrated that the direct token concatenation performs best in transformer-based architecture. Integrating both injection methods, Yang et al. (2023) introduced a two-stages training design, where recent local frames are concatenated channel dimension, while longer-range global history is encoded and injected via cross attention mechanism. However, these methods struggle to effectively unify local and global motion priors while maintaining alignment across scales.

2.2.3 Motion-enhanced Multimodal Video Prediction

To enhance time coherence and motion consistency of video prediction, recent approaches incorporated multi-modal data, such as depth, optical flow and contour as complementary information. Both MoVideo (Liang et al., 2024) and FloVD (Jin et al., 2025) leverage optical flow as a motion condition, where MoVideo concatenates it with RGB latent space in channel dimension, while FloVD encodes and additively injects it into the multi stages of U-Net through a dedicated flow encoder. However, ExtDM (Zhang et al., 2024) and LFDM (Ni et al., 2023) designed motion autoencoders to extract the flow-based motion information, which guide the video diffusion process. ExtDM emphasizes the distributional extrapolation of motion cues within its autoencoder, which compresses and reconstructs optical flow and occlusion information to reason about future frames. In contrast, the autoencoder in LFDM is trained to warp a flow sequence and

reconstruct the corresponding image sequence in a latent space, serving as a motion prior for video generation. Syncvp (Pallotta et al., 2025) incorporates conditional depth with RGB video within a synchronous denoising framework. Whereas most of them adopt a two-stage training pipeline, which increases computational cost and complicates cross-modal alignment.

2.3 Methodology

We propose a simple, yet effective, diffusion model for driving video prediction that uses RGB cameras and exploits historical motion priors to enhance the quality of generated data. The overall framework of the proposed method is illustrated in Figure 1. HMPDM aims to generate and predict F future video frames, $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=1}^{F}$, given a set of P past frames, $\boldsymbol{c} = \{\boldsymbol{c}_t\}_{t=1}^{P}$, where $\mathbf{x}_\mathbf{t}$ and $\boldsymbol{c}_t \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ are RGB images. Therefore, the objective is to learn the conditional distribution of future frames $\mathrm{p}(\mathbf{x}|\mathbf{c})$. The proposed framework comprises three main components: temporal-aware latent conditioning, motion-aware pyramid encoder, and self-conditioning strategy.

2.3.1 Temporal-aware Latent Conditioning

HMPDM builds upon the latent diffusion modeling paradigm and follows the Elucidated Diffusion Model framework widely adopted in Stable Video Diffusion. Instead of running diffusion directly on RGB pixels, we first encode past frames and future frames into low dimensional latent representations using a latent autoencoder. This reduces computational cost and allows the diffusion model to focus on learning spatio-temporal dynamics in a compact space.

In temporal aware latent conditioning, we keep the latent representations of past frames clean so the historical context remains stable and reliable. We progressively corrupt the latent representations of future frames by adding Gaussian noise at different noise levels. We then construct a joint input by concatenating the clean past latents with the noisy future latents and feed the joint latent sequence into the UNet. This design implicitly injects temporal context into the UNet and allows its spatio-temporal attention layers to jointly model historical conditioning information and the noisy future targets during denoising.

To differentiate the deterministic nature of the observed domain from the stochasticity of the noisy domain, we define time embeddings for clean past frames and noisy future latents. Past frames use stationary embedding that corresponds to a minimal noise level. Future frames use a noise dependent embedding that changes with the noise scale. A binary temporal mask is used to assign the clean embedding to historical positions and the noise embedding to future positions. This enforces a clear separation between clean history and noisy prediction targets and enhances the diffusion process awareness of motion observed from past frames.

Based on this joint input and the time embedding, the denoiser predicts the clean signal for the future part while preserving the historical part. The diffusion model learns to progressively remove noise from the future latents while using the clean history as a strong condition. This makes the generated future frames more coherent with the observed motion patterns and reduces unrealistic temporal changes.

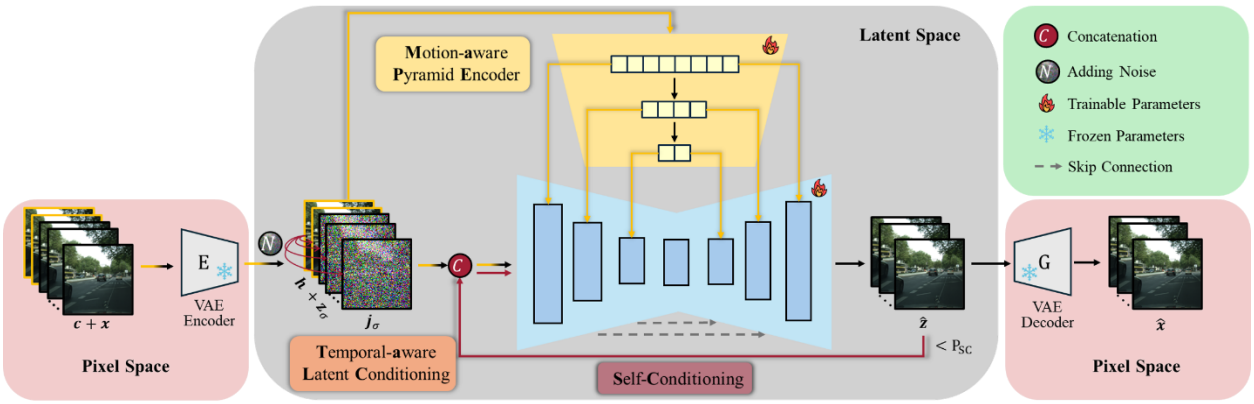


Figure 1. Overview of the proposed HMPDM framework

2.3.2 Motion-aware Pyramid Encoder (MaPE)

Temporal aware latent conditioning provides historical latents to the UNet, but it does not explicitly emphasize motion priors at different spatial resolutions. We therefore introduce the motion aware pyramid encoder to explicitly extract historical motion priors and provide them to the diffusion model in a structured way. The motion aware pyramid encoder encodes the historical latent sequence into multi scale representations. It starts with fine level tokens that preserve local details and then progressively builds coarser level tokens that capture global structure and long-range dependencies. Spatio-temporal attention blocks are applied to model how motion evolves across time, so the encoder can represent both local agent motion and broader scene level changes.

The extracted multi scale motion features are injected into the UNet through cross attention at corresponding stages. The UNet can retrieve the most relevant historical motion cues at a matched scale during denoising. This makes the model more sensitive to motion boundaries, object movements, and scene dynamics. It also improves the alignment between the predicted future frames and the historical motion trend, which enhances temporal stability and visual fidelity.

2.3.3 Self-Conditioning (SC)

Even with strong historical conditioning, long horizon diffusion sampling can accumulate errors and lead to blur and motion drift. HMPDM uses a self-conditioning strategy (Chen et al., 2023; Gupta et al., 2024) to improve stability and refine prediction quality over the denoising trajectory.

During training, the model often produces an intermediate prediction without gradient updates and uses this prediction as an additional condition for the subsequent learning pass. This encourages the denoiser to correct its own errors and to stay consistent across steps. When self-conditioning is not applied, we use a simple condition that repeats the most recent observed frame, which is consistent with common practice in Stable Video Diffusion style training.

During inference, we start from random noisy future latents and iteratively denoise them following the sampling procedure. The prediction from the previous step is appended as self-conditioning input for the next step, so each step benefits from the current estimate of the clean signal. After the denoising process finishes, we decode the predicted future latents back to RGB frames. This strategy improves long horizon consistency and helps preserve motion continuity in the generated video.

2.4 Analysis and Discussions

In this section, we illustrate the implementation details of HMPDM and conduct several experiments to demonstrate the superior of our method.

2.4.1 Implementation Details

The proposed HMPDM was implemented using PyTorch 1.10.0 on a server equipped with an Nvidia L40S featuring 48 GB of memory. The diffusion model is trained for 10^5 steps using AdamW optimizer, with a batch size (B) of 4 and a learning rate 2×10^{-5}.

To evaluate the effectiveness of the proposed framework for driving video prediction, experiments are conducted on Cityscapes (Cordts et al., 2016) and KITTI (Geiger et al., 2013) datasets. Following the standard video prediction protocol, both datasets are resized to 128×128. Specifically, the Cityscapes dataset is partitioned into 2,975 training clips, 500 validation clips, and 1,525 test clips, each containing 30 consecutive frames. The KITTI dataset is divided into 759 training clips and 150 test clips, with each clip containing 9 frames.

Following the protocols established in prior work~\cite{ Voleti et al., 2022; Pallotta et al., 2025; Zhang et al., 2024}, four commonly used metrics are adopted to evaluate the performance of video prediction models: Structural Similarity Index Measure (SSIM) (Wang et al., 2004), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and Fréchet Video Distance (FVD) (Unterthiner et al., 2018). SSIM and PSNR evaluate pixel-wise reconstruction fidelity, while LPIPS measures perceptual similarity in deep feature space. FVD captures the spatio-temporal consistency by quantifying both motion dynamics and temporal coherence. For fair comparison, all metrics are reported on both a randomly sampled subset of 256 clips and the full test set across 10 random future denoising trajectories (#T), from which the best performing result is selected. To compare computational complexity, we additionally report the input modality, and the number of stages in the training pipeline.

2.4.2 Comparison with State-of-the-art Models

To evaluate the effectiveness of the proposed HMPDM on driving video prediction, we benchmark it against SOTA models on Cityscapes and KITTI datasets following the two standard protocols, with results summarized in Table 1. Under a strictly RGB-only input setting, HMPDM demonstrates superior performance on Cityscapes. Specifically, it achieves an FVD of 151.2 on the 256-sample test protocol, representing a 17.8% relative reduction compared to MCVD (Voleti et al., 2022), and 77.0 on the full-test protocol, corresponding to a 28.2% relative reduction to STDiff (Ye et al., 2024). Although HMPDM is limited to a single RGB modality, it attains competitive results even against multimodal models (e.g., R+F, R+D), ranking the second on the 256-sample test set while achieving the best FVD on the full test set. Furthermore, HMPDM retains a one-stage training pipeline and requires fewer optimization steps, emphasizing its efficiency and compact architecture.

In contrast to FVD, HMPDM's performances on SSIM, PSNR, and LPIPS illustrate comparable yet slightly inferior results to the SOTA models. This disparity primarily stems from the inference mechanism. These evaluation metrics are calculated per frame, whereas HMPDM performs a one-time prediction of the entire future segment. Conversely, SOTA methods (Zhang et al., 2024; Voleti et al., 2022; Yang et al., 2023; Pallotta et al., 2025) adopt autoregressive rollouts with

shorter prediction horizons, which attenuate the drift, even in cases where the long-term motion consistency metric FVD is not superior.

To further assess the generalization of HMPDM in autonomous driving scenarios, we evaluate on KITTI with a short prediction horizon (4 → 5), as shown in Table 2. Given the limited number of predicted frames, FVD is omitted for this dataset and we only report SSIM, PSNR, and LPIPS. The performance of high PSNR and competitive SSIM indicate strong pixel-level alignment and high structural similarity, while the LPIPS margin to the best method suggests remaining room in texture reconstruction. Notably, since HMPDM learns and samples in the VAE latent space at a low dimensional space (128 × 128), high-frequency details may be smoothed, contributing to its modest deficit on LPIPS.

Table 1. Quantitative comparison of video prediction models on Cityscapes datasets. ↓means lower is better, ↑ means higher is better.

Methods	Year	Input	Pipeline	Cityscapes (128×128) 2 → 28			
				FVD↓	SSIM↑	PSNR↑	LPIPS↓
256 Random Samples							
U-ViT	CVPR23	R	1	1045.3	0.362	10.84	0.431
RaMViD	TMLR24	R	1	812.6	0.454	13.14	0.395
VDIM	AAAI23	R	2	724.7	0.539	18.49	0.252
RVD	ARXiv22	R	1	465.0	0.489	17.21	0.226
MCVD	NeurIPS22	R	1	184.8	0.720	22.50	0.121
LFDM	CVPR23	R+F	2	194.9	0.601	20.32	0.157
ExtDM-K4	CVPR24	R+F	2	**121.3**	**0.745**	**22.84**	**0.108**
HMPDM (Ours)	-	R	1	151.2	0.633	21.42	0.142
Full Test Set							
NVPV	CVPR23	R	2	768.0	**0.744**	-	0.183
LGC-VD	ARXiv23	R	2	124.6	0.732	-	**0.069**
STDiff	AAAI24	R	1	107.3	0.658	-	0.136
SyncVP	CVPR25	R+D	2	84.0	0.649	-	0.160
HMPDM (Ours)		R	1	**77.0**	0.626	**21.37**	0.145

Table 2. Quantitative comparison of video prediction models on KITTI datasets. ↓means lower is better, ↑ means higher is better.

Methods	Year	Input	Pipeline	KITTI (128× 128) 4 → 5		
				SSIM↑	PSNR↑	LPIPS↓
Full Test Set						
NVPV	CVPR23	R	2	**0.66**	-	0.279
LGC-VD	ARXiv23	R	2		-	**-**
STDiff	AAAI24	R	1	0.54	-	**0.115**
SyncVP	CVPR25	R+D	2		-	-
HMPDM (Ours)		R	1	0.54	**18.62**	0.149

2.4.3 Effectiveness of Components

Table 3 summarizes the contribution of each component for predicting 28 future frames conditioned on 2 past frames. The architecture and pretrained weights of baseline model are

identical to those of SVD (Blattmann et al., 2023) and then finetuned on the Cityscapes dataset. Compared to this baseline model, adding the TaCL module yields 16.4% lower FVD, 9.2% higher SSIM, 7.4% higher PSNR, and 21% lower LPIPS. These improvements indicate that the TaCL effectively leverages implicit historical context, leading to improved motion dynamics and stronger structural fidelity.

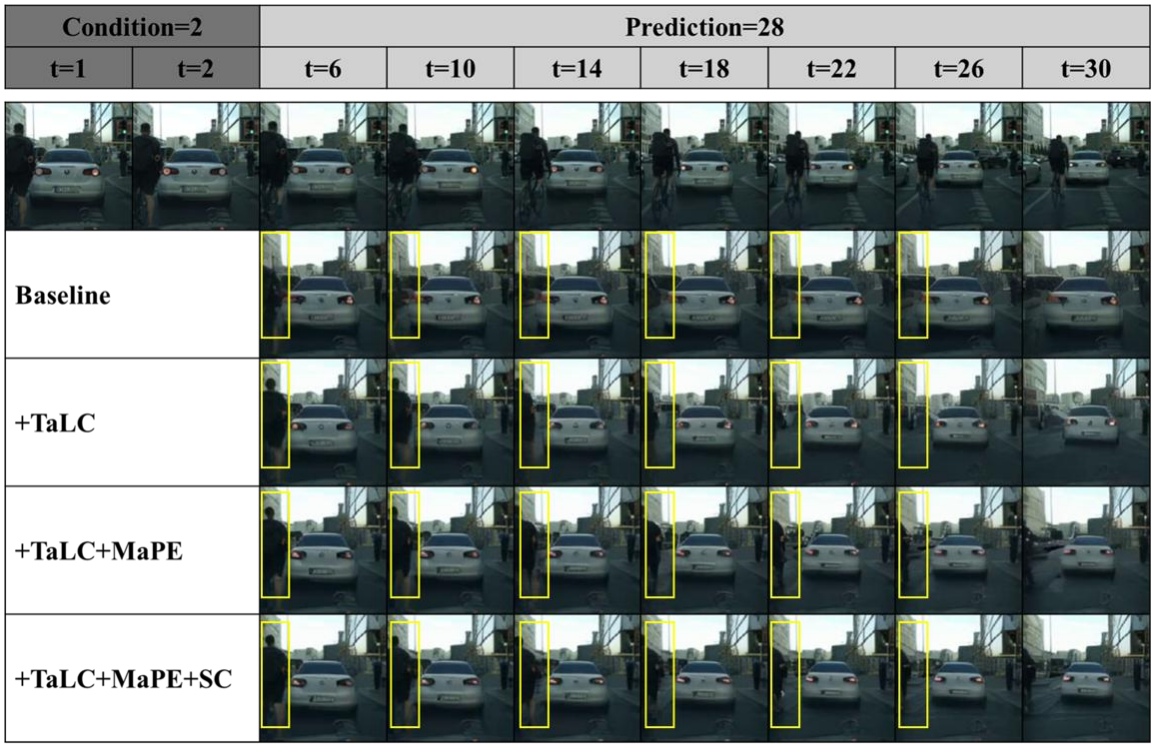
Table 3. Quantitative Ablation Results on Cityscapes. ↓ means lower is better, ↑ means higher is better.

Models	Cityscapes (128×128) 2 → 28			
	FVD↓	**SSIM↑**	**PSNR↑**	**LPIPS↓**
Baseline	236.2	0.574	19.94	0.193
+ TaLC	197.5	0.627	21.42	0.153
+ TaLC + MP	188.7	0.632	**21.52**	0.154
+ TaLC + MaPE	<u>155.7</u>	**0.634**	21.44	0.146
+ TaLC + MaPE + SC	**151.2**	<u>0.633</u>	<u>21.42</u>	**0.142**

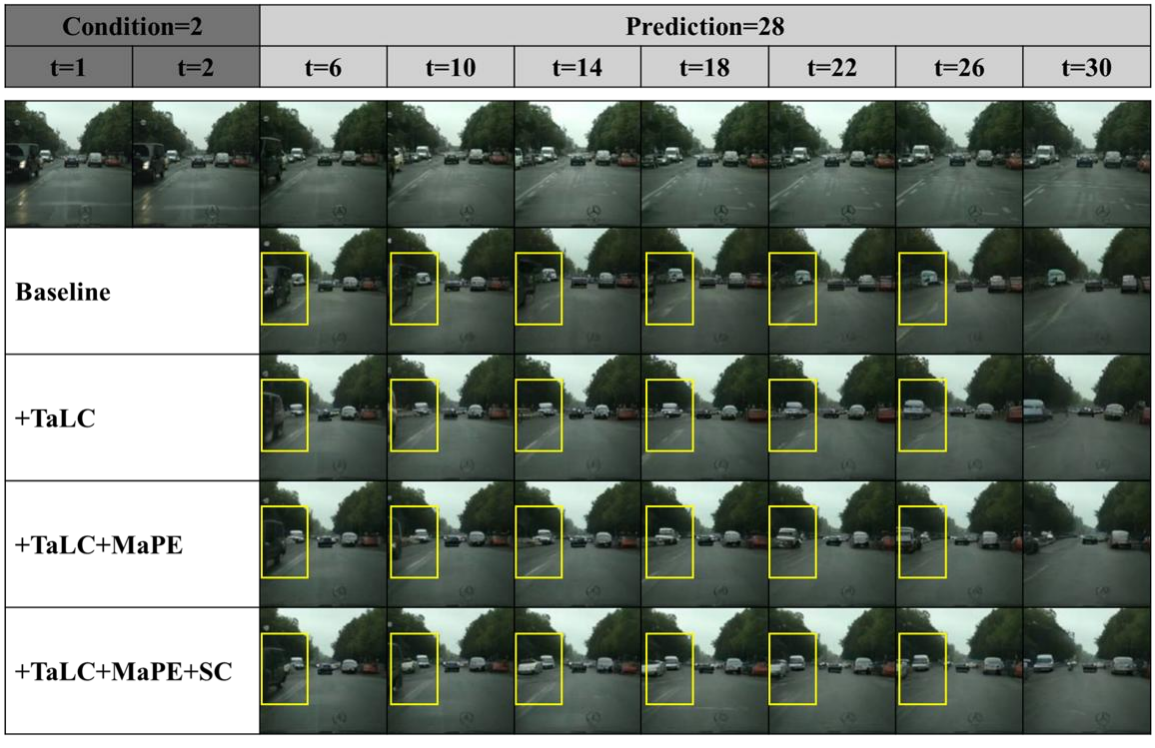
We further evaluate the MaPE module with respect to its pyramid design and patch merging mechanism. First, the variant (+TaLC+MP) improves over the variant (+TaLC) by 4.5% in FVD, highlighting the importance of multi-scale conditioning injected through cross-attention layers for capturing the semantically relevant features. +TaLC+MaPE, which integrates past frames within the spatio-temporal domain through stacked transformers, achieves a 17.5% reduction in FVD compared with the variant (+TaLC+MP), illustrating the effectiveness of the designed transformer blocks.

Finally, adding SC upon the variant (+TaLC+MaPE) constitutes the complete HMPDM framework, which delivers the best FVD and LPIPS and maintains competitive SSIM and PSNR. This strategy mitigates error accumulation and enhances temporal consistency.

The contribution of each designed component is further visualized in Figure 2. Qualitative comparisons demonstrate that the TaLC enhances the spatio-temporal consistency, as evidenced by improved spatial structure. Specifically, in both samples, the front vehicle occupies lager spital area in the frames generated by the baseline model than in the corresponding ground truth frames. Furthermore, the addition of MaPE allows HMPDM to better capture the historical motion priors. Compared with predicted frames generated by (+TaLC) in sample (a), the (+TaLC+MaPE) variant produces clearer motion patterns for the cyclist, as evidenced by more visually coherent leg movements. Integrating the variant (+TaLC+MaPE) with SC, further enables HMPDM to handle occlusions and preserve the fidelity of traffic agents. As shown in sample (b), HMPDM successfully generates the stationary white sedan even though it never appears in the past frames. These improvements enhance the quality of generated traffic videos, thus providing more realistic and reliable information to support vehicles in predicting and reasoning about future scenes.



(a)



(b)

Figure 2. Qualitative Ablation Results on Cityscapes (128×128)

2.4.4 Effectiveness of Conditioning Horizon

Intuitively, providing sufficient historical motion information enhances the dynamic priors required for future frame generation. To investigate the impact of varying conditioning horizons on future driving video prediction, quantitative results on the Cityscapes dataset are reported in Table 4. When increasing the number of historical frames from 2 to 6 while keeping the same prediction length, the FVD decreases from 117.4 to 104.9 (≈10.7%↓). Simultaneously, SSIM improves from 0.679 to 0.694, PSNR increases from 22.92 to 23.37 dB, and LPIPS concurrently decreases from 0.117 to 0.110. These improvements indicate that supplying richer motion context advances both temporal coherence and frame-wise fidelity, with diminishing but still positive returns as the conditioning length grows.

Table 4. Effectiveness of Conditioning Horizon on Video Prediction Performance. ↓ means lower is better, ↑ means higher is better.

Conditioning Horizon	Cityscapes (128×128) 2 → 28			
	FVD↓	**SSIM↑**	**PSNR↑**	**LPIPS↓**
2 → 14	117.4	0.679	22.92	0.117
4 → 14	108.9	0.691	23.24	0.112
6 → 14	104.9	0.694	23.37	0.110

2.5 Summary

This study proposes HMPDM, a diffusion-based video prediction model enhanced by mono-modal historical motion priors and specifically tailored for driving scenarios. TaLC and MaPE modules effectively capture and utilize the historical motion priors from past driving patterns. In addition, SC improves the fidelity of traffic agents and long-term temporal consistency. The proposed HMPDM framework, owing to its efficiency and well-designed historical motion priors, achieves superior performance on the Cityscapes benchmark, outperforming existing methods by 17.8% and 28.2% in FVD under two standard evaluation protocols, respectively.

3 SAFETY-CRITICAL SCENARIOS GENERATION IN RURAL AREAS: FROM SIM TO REAL

Our prior work introduced the HDMP diffusion model for future driving prediction, where historical motion priors are leveraged to improve the temporal coherence and realism of predicted trajectories and video dynamics. This predictive capability provides an important foundation for scenario generation, however, safety-critical scenario generation in rural areas demands substantially higher controllability than prediction-oriented generation. Motivated by this, we build on an existing video foundation model, COSMOS, and condition it explicitly on HD maps as spatial structure. In this section, we present a Sim-to-Real rural safety-critical scenario generation pipeline that uses scripted hazardous events in simulation to define controllable scene intent, and then render them as realistic dashcam-style driving videos via HD-map-conditioned generation. Contents of this section will be submitted to journal 2026.

3.1 Background

Autonomous driving systems must demonstrate reliable behavior not only in routine traffic, but also in rare events that can trigger severe harm. These safety-critical situations, such as sudden cut ins, loss of traction, unexpected obstacles, or delayed perception under poor visibility, occur infrequently in naturalistic driving data, yet they are precisely the corner cases required for validating driving safety. Relying on real world exposure alone is therefore inefficient and potentially unsafe, because the events of interest may require extremely large mileage to observe with statistical confidence. Simulation based testing has long been recognized as a practical path to scale coverage of rare and hazardous events, while reducing public risk and enabling repeatable evaluation (O'Kelly et al., 2018).

Moreover, rural roads as a critical component of the road network, have often been underappreciated and deserve focused attention. Rural roads often feature sparse infrastructure, limited or degraded lane markings, weak illumination at night, and complex geometry such as sharp curves, crests, and narrow shoulders. Traffic composition also differs, with slow moving farm equipment, heavy trucks, wildlife and vulnerable road users appearing under conditions that can strain perception and planning. Most importantly, rural contexts are associated with disproportionately high fatality rate per VMT. Recent U.S. summaries report that the fatality rate per VMT is about 1.5 times higher in rural areas than in urban areas, and that rural roadways account for a large share of roadway deaths despite a smaller share of the population (USDOT, 2025). Independent safety statistics similarly show substantially higher crash death rates per 100 million miles traveled in rural areas than in urban areas, reinforcing that the rural domain represents a critical safety gap rather than a niche corner case (IIHS, 2025).

Despite this, rural safety critical data remains underrepresented in many existing autonomous driving datasets and testing pipelines, which tend to be biased toward dense urban scenes where sensors, maps, and traffic rules are more structured. This imbalance creates a mismatch between where systems are often trained and evaluated and where risk can be most severe. A systematic capability to generate rural safety critical scenarios is therefore valuable for validating autonomous driving techniques, expanding operational design domains, and supporting more equitable safety assurance across geographies.

Recent progress in generative modeling, especially diffusion based generative models, offers a promising approach to address this challenge. Diffusion models learn to synthesize high fidelity samples by gradually denoising from noise toward realistic data, and they have demonstrated strong capability for producing detailed, diverse outputs (Ho, et al., 2020). In a driving context, this paradigm supports generating rare but plausible rural hazards at scale, while maintaining control over factors that matter for safety, such as visibility, road geometry, interacting traffic participants, and the timing of critical events. By combining controlled simulation environments with diffusion driven synthesis, it becomes possible to populate test suites with targeted rural safety critical scenarios that would otherwise be difficult to collect, and reproduce (Zhong et al., 2020).

To generate controllable rural driving videos with diffusion models, strong conditioning signals are often necessary to reduce ambiguity and ensure that the synthesized scene follows the intended safety-critical event. In practice, weak prompts alone rarely guarantee that key decisions such as a left turn versus a right turn, a specific merging behavior, or a particular conflict timing will reliably emerge in the generated sequence. Instead, controllability is typically achieved by conditioning diffusion sampling on structured priors that explicitly encode the driving intent and the road context, such as high-level maneuvers, target trajectories, and map level geometry and topology. These requirements naturally align with simulation, where road topology, intersection design, and trajectories evolution can be precisely specified and systematically varied. Simulation provides a controllable foundation for defining maneuver level constraints and map-based conditions (Pronovost et al., 2023), enabling diffusion models to generate rural safety-critical videos that are not only realistic, but also reproducible for scenario-based evaluation.

3.2 Study Contributions

To address the limitations mentioned above, this study is motivated to develop a novel pipeline that generates safety-critical scenarios from simulation to real-world in rural areas. Contributions of this study are as follows:

- We propose a framework to construct diverse, safety-critical rural driving scenarios in simulation, leveraging configurable traffic assets, multi-agent participants, and map-based scene layouts.
- The Sim-to-Real generation pipeline in rural areas is developed to transfer simulated scenarios into realistic data using generative foundation model, enabling scalable augmentation and repeatable generation under rare and hazardous conditions.

The remainder of the paper is organized as follows. Section 3.3 will describe recent studies relevant to this work. The rural Sim-to-Real pipeline will be presented in Section 3.4. Section 3.5 will further present the implementation details and the experimental results. In the end, Section 3.6 will summarize the findings and suggest future research directions.

3.3 Related Work

Recent related work of this study is concentrating on safety-critical driving scenarios in rural areas, controllable diffusion and foundation models for driving scene generation, and simulation-to-real scenarios generation methods.

3.3.1 Safety-Critical Driving Scenarios in Rural Areas

Several recent studies have addressed safety-critical driving scenarios in rural environments, highlighting both human-driven and autonomous vehicle challenges. Klitzke et al. (2025) analyzed trajectory and video data from rural Germany, revealing recurring risks such as speeding, hazardous overtaking, and lane encroachment by heavy vehicles, which often exacerbated by poor infrastructure and limited protection for vulnerable road users. Complementing this, Sum et al. (2025) conducted a large-scale crash severity analysis in Thailand, identifying factors like head-on collisions, large vehicle involvement, lack of medians, and nighttime conditions as major contributors to severe outcomes on rural roads. Safari et al. (2025) focused on distracted driving crashes on rural curves, finding that motorcycle involvement, older drivers, sharp curves, and higher speed limits significantly increased injury severity. To address overtaking hazards on rural two-lane roads, Vigne et al. Finally, Karacik et al. (2025) introduced a scenario generation pipeline (SCSG) using real crash narratives and language models to simulate high-risk driving conditions within the CARLA (Dosovitskiy et al., 2017) environment, improving the robustness of AV testing. Collectively, these works underscore the unique safety challenges of rural roads and inform both infrastructure-level countermeasures and the development of safer autonomous systems.

3.3.2 Controllable Diffusion and Foundation Models for Driving Scene Generation

Existing research has demonstrated that both diffusion and foundation models can be used to generate realistic and controllable driving scenarios, which are critical for the testing and training of autonomous vehicles. However, two models emphasis on different aspects. Diffusion models can generate realistic scenes that meet specific requirements, meanwhile as core module for driving foundation models. Pronovost et al. (2023) proposed Scenario Diffusion, a model that synthesizes multi-agent traffic scenarios based on HD maps and text prompts. It can produce diverse and physically plausible scenes across different regions. Lu et al. (2023) introduced SceneControl, which learns traffic scene generation from data and applies guided sampling to meet user-defined constraints. This model generates more varied and realistic scenes compared to traditional rule-based methods. Yang et al. (2025) developed DualDiff, a model that separates the generation of foreground objects and background context. By using geometry-aware inputs and refinement strategies, it improves the visual quality and semantic consistency of multi-camera driving videos.

In contrast, driving foundation models are better suited for explicitly defined scene conditions, offering fine-grained controllability over scenario configurations. Ren et al. (2025) presented Cosmos-Drive-Dreams, a large-scale synthetic data generation pipeline built on a pre-trained video foundation model. Their system produces multi-view annotated driving scenes conditioned on HD maps and environmental factors, which significantly improve downstream tasks such as 3D lane detection and object recognition. Zhang et al. (2025) proposed Epona, a world model that links driving video generation with trajectory planning. It predicts future frames and vehicle motions in a sequential manner, allowing long-term simulation and controllable planning. Epona performs well in navigation tasks and supports integrated simulation and decision-making. Overall, these works demonstrate how generative models can enhance the realism and diversity of simulated driving environments and contribute to safer autonomous vehicle development.

3.3.3 Simulation-to-real Scenarios Generation Methods

Recent research has explored diffusion-based methods to bridge the gap between simulation and real-world data for autonomous driving. Zhao et al. (2024) compared GAN-based and diffusion-

based approaches for translating simulator outputs into realistic images, finding that diffusion preserved road geometry and vehicle structure more accurately. Zhou et al. (2024) proposed SimGen, which uses a two-stage diffusion pipeline to convert coarse simulator renderings into high-quality photorealistic images, guided by text and structural inputs. Similarly, Pandey et al. (2025) introduced a real-time Sim2Real Diffusion system that adapts simulator visuals using a vision-language-conditioned diffusion model. Their method enables domain adaptation across simulators with trigger-word conditioning and supports closed-loop driving without retraining. Bu and Yasuda (2024) developed DRIVE, which enhances simulation visuals with diffusion-based editing. Their user study showed improved driver immersion and task performance, demonstrating practical value in human-in-the-loop simulation environments. Extending this line of work, Wang et al. (2025) presented TeraSim-World, an automated data synthesis pipeline that integrates agent behavior simulation and high-fidelity video generation for safety-critical autonomous driving tasks. Starting from any global coordinate, the system retrieves real-world maps and traffic demand, then uses learned behavioral models to simulate normal and adversarial scenarios.

3.4 The Approach

Our study implements a complete simulation to real-world generation pipeline. Specifically, we first construct rural safety-critical scenarios within the CARLA simulation platform. Based on the desired generation objectives, corresponding high-definition maps and structured prompts are then derived and used as conditioning inputs to COSMOS, which generates realistic dashcam-style driving videos. The overview of proposed Sim-to-Real generation pipeline is illustrated in Figure 3.

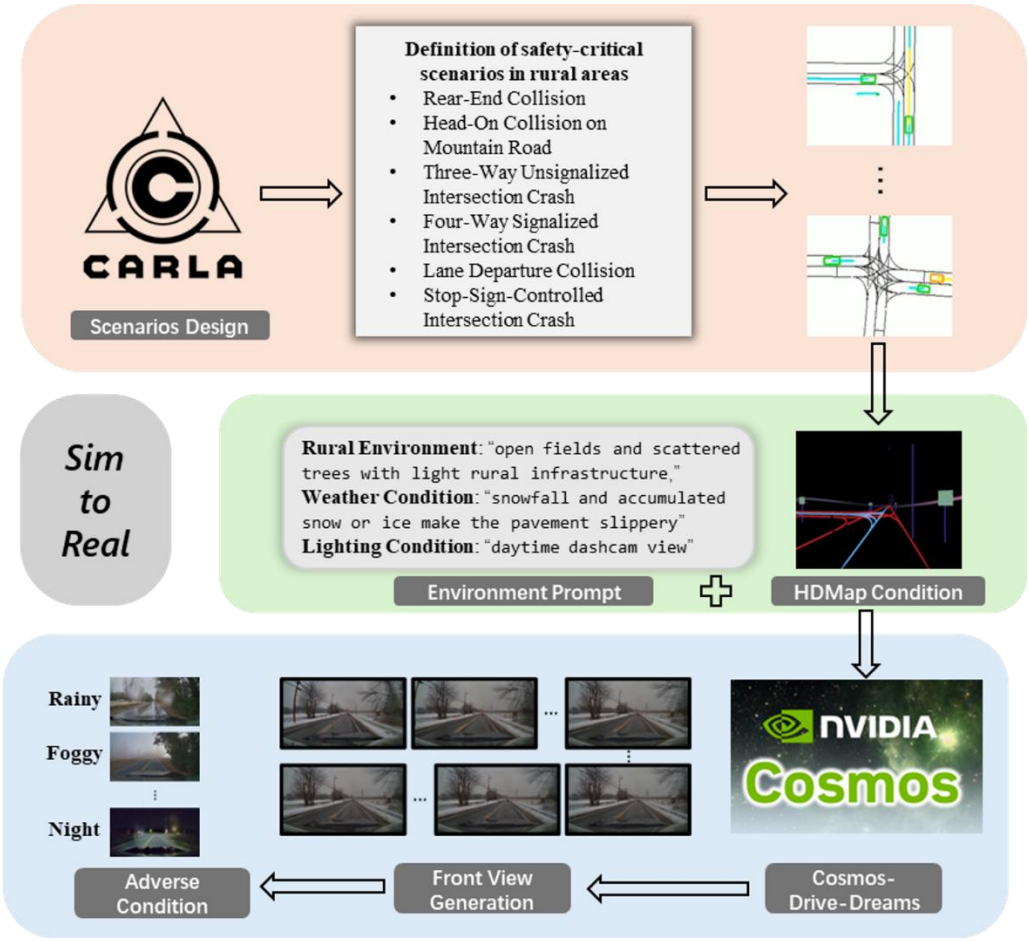


Figure 3. The overview of Sim-to-Real Generation Pipeline in Rural Areas

3.4.1 Simulation-Based Construction of Safety-Critical Scenes in CARLA

In the first stage, we leverage the CARLA simulator to construct a suite of safety-critical driving scenes in a rural context. CARLA provides a flexible platform to model rural road geometries, intersections, and environmental features. We design each scenario's virtual environment to include all relevant road infrastructure: cross walk, lane lines, poles, road boundaries, traffic lights, traffic signs, and wait lines at intersections. These elements are placed in 3D within CARLA to mirror real-world rural layouts and ensure that any downstream model has access to detailed map information. After assembling the scene, we extract a high-definition map (HD map) representation from the simulation. The HD map encodes the precise 3D geometry and location of all the features and serves as a structured blueprint of the scenario. This conversion from CARLA's world to an HD map format is crucial, as it captures the static road context in a form suitable for conditioning generative models later. Consequently, we have a library of simulated rural scenarios, each with an accompanying HD map that delineates the critical road and infrastructure details needed for accurate scene replication.

3.4.2 Definition of Safety-Critical Scenarios and Environmental Variations

With the environments in place, we define a set of representative safety-critical scenarios that are especially relevant to rural driving. Each scenario corresponds to a distinct collision or near-collision situation that our pipeline will generate. The six key scenario types are:

- **Rear-End Collision:** A following vehicle fails to brake in time and collides with the vehicle in front on a rural road.
- **Head-On Collision on Mountain Road:** On a mountain road with sharp turns, an ego vehicle encounters an oncoming vehicle, resulting in a frontal collision.
- **Three-Way Unsignalized Intersection Crash:** Two vehicles approach a T-intersection with no traffic signals and near collide due to misjudgment or failure to yield.
- **Four-Way Signalized Intersection Crash:** A near collision at a signalized rural intersection, for example, triggered when a vehicle makes a left turn and enters the wrong lane.
- **Lane Departure Collision:** The ego vehicle veers out of its lane due to slippery conditions in bad weather, causing a near collision with an opposite-direction vehicle.
- **Stop-Sign-Controlled Intersection Crash:** A near crash at a 3-way rural intersection regulated by stop signs, typically stemming from one vehicle failing to stop and colliding with cross traffic.

These scenarios cover a broad range of high-risk events on rural roads, encompassing rear-end and head-on crashes, intersection-related accidents, and incidents influenced by environmental factors. For each scenario type, we script the motions of the ego vehicle and other actors in CARLA to faithfully recreate the precipitating events (e.g., sudden braking, failure to yield, loss of control on a curve).

In addition to the base scenarios, we introduce diverse weather and lighting conditions to each scene to enhance realism and validate the system's robustness. Specifically, for each scenario we vary the conditions such as clear daytime, nighttime darkness, dawn/dusk low-light, heavy rain, fog, snow and other inclement weather. By generating each safety-critical scenario under multiple conditions, the proposed rural Sim-to-Real pipeline can generate driving scenarios including both nominal conditions and challenging situations. This augmentation ensures that the methodology covers safety-critical events in a wide spectrum of real-world conditions, which is essential for training and evaluating models that generalize to the real world.

3.4.3 Simulation-Real Video Generation Using COSMOS

The final stage of our methodology bridges the gap between simulation and real-world appearance by employing COSMOS, a state-of-the-art video foundation model. COSMOS is a generative model designed for Sim-to-Real transformation, which can produce driving videos from structured input representations, allowing our simulated scenarios to be visualized as realistic near-crash videos. We condition COSMOS on two forms of input to achieve controllable and accurate generation of each scenario, including an HD map condition and a textual prompt.

HD Map Conditioning: The HD maps extracted from CARLA are fed into COSMOS as a spatial condition that preserves the exact road layout and infrastructure of the scenario. This ensures that the generated video adheres to the real geometry of the scene. For instance, the curvature of a mountain road, the placement of an intersection's stop line, or the presence of a crosswalk will all manifest correctly in the output frames. COSMOS takes the HD map as an input canvas that constrains where roads and static objects should appear in the generated imagery. The HD map acts as a blueprint, so that lane markings, signs, traffic lights, and other features are rendered in their proper locations and configurations in the photorealistic video. This component is critical for

maintaining consistency between the simulated scenario and the real-world driving output, particularly for evaluating safety-critical details.

Prompt-Based Scene and Event Description: Alongside the HD map, we craft a detailed textual prompt for each scenario to guide COSMOS in generating the correct dynamic content and style. The prompt is written in a descriptive, scene-setting manner. It typically starts with a description of the rural environment. For example, *"a narrow two-lane country road on a foggy morning, with diffuse sunlight and wet pavement"*. Next, the prompt specifies the critical event and participants, such as *"a truck appears around a blind curve ahead while a car is already in the curve, leading to a head-on collision"*. We include details on the vehicles, the cause of the incident (e.g., *"the car skids on the wet road into the opposite lane"*), and any relevant context. This rich description provides COSMOS with high-level semantic guidance on the scenario's dynamics and atmosphere. Essentially, the prompt communicates what is happening, where, and under what conditions, so the model can render appropriate weather effects, lighting, and the unfolding motion of vehicles.

By integrating the HD map and the text prompt, COSMOS generates a coherent video that aligns with both the physical road structure and the narrative of the scenario. The participants and their trajectories are implicitly controlled through this integration: the HD map constrains where vehicles can travel, while the prompt and the model's learned physics prior ensure that the vehicles move and interact as described. In some cases, additional conditional signals (such as depth images or segmentation masks for the vehicles) could be employed to further tighten control over the exact motion paths; however, in our approach the combination of map and textual guidance in COSMOS was sufficient to produce realistic approximations of the intended trajectories and crash outcomes.

Finally, COSMOS outputs a synthetic video for each scenario and condition, typically from a driver's perspective with duration long enough to capture the pre-incident context. The generated videos appear highly realistic, with correct rural scenery details and weather effects, while faithfully reproducing the key safety-critical event. This Sim-to-Real video generation step yields a set of photorealistic clips that correspond to the originally simulated scenarios. These videos can be used for subsequent analysis, such as validating whether an autonomous driving system would correctly perceive and react to the events, under conditions that closely resemble real dashcam footage yet are grounded in known simulation truth. The use of COSMOS in this pipeline thus allows us to scale up the diversity and realism of safety-critical scenarios without the risks and costs of staging real accidents, providing a powerful tool for research in autonomous driving safety.

3.5 Experiments and Implementation Details

We implement the proposed rural Sim-to-Real pipeline in three modules: scenario construction in CARLA, HD map extraction and prompt annotation, and video generation with COSMOS.

3.5.1 Implementation Details

We use CARLA 0.9.16 to create rural road layouts and instantiate six safety-critical scenario types, including rear-end, head-on on mountain roads, three-way unsignalized intersection crash, four-way signalized intersection crash, lane-departure collision, and stop-sign-controlled intersection crash. For each scenario, we place static infrastructure elements, including lane markings, road

boundaries, crosswalks, stop lines, poles, traffic lights, and traffic signs, and script the ego vehicle and other traffic participants to reproduce the triggering behaviors. We record each episode from a front dashcam configuration at a resolution of 1920×1080 and 30 FPS for 121 frames, ensuring that each clip covers the safety-critical context.

For every simulated scene, we export a series of HD maps that encode the geometry of the road and the bounding boxes of participants. Meanwhile, we employ a structured prompt template that specifies (i) rural environment description, (ii) weather and visibility, (iii) lighting, and (iv) the safety-critical event description. For each scenario instance, we generate multiple condition variants by combining weather and lighting settings while keeping the underlying HD map fixed.

For video generation conditioned on the textual prompt and HD map, we use Cosmos-Transfer1-7B-Sample-AV as the video foundation model to synthesize dashcam clips. The HD map is provided as a spatial control signal to preserve road geometry and infrastructure placement, while the prompt guides scene appearance and environment. We generate videos of 120 frames with resolution 1280×704.

3.5.2 Experiments and Results

In our experiments, we generate six specific rural safety-critical scenarios using the proposed Sim-to-Real pipeline. Beyond these base scenario types, we conduct a set of experiments by varying adverse weather and illumination conditions.

Safety-critical scenario types:

(1) *Rear-End Collision*

We first evaluate a rear-end collision scenario on a rural two-lane roadway at nighttime, where illumination is dominated by the ego headlights and roadside lighting is minimal. Detailed HD Maps and generated frames are shown in Figure 4. The scene emphasizes typical rural characteristics including open fields, scattered trees, and sparse built infrastructure, which reduce visual saliency and increase reliance on reflective lane markings. Between 90 and 120 frames, the lead vehicle stops for a red traffic light at a small intersection, while the ego vehicle reacts late and fails to brake sufficiently, resulting in a near rear-end collision on the intersection approach. This case highlights how rural nighttime driving can amplify brake-response risks due to limited ambient lighting and reduced perceptual cues.

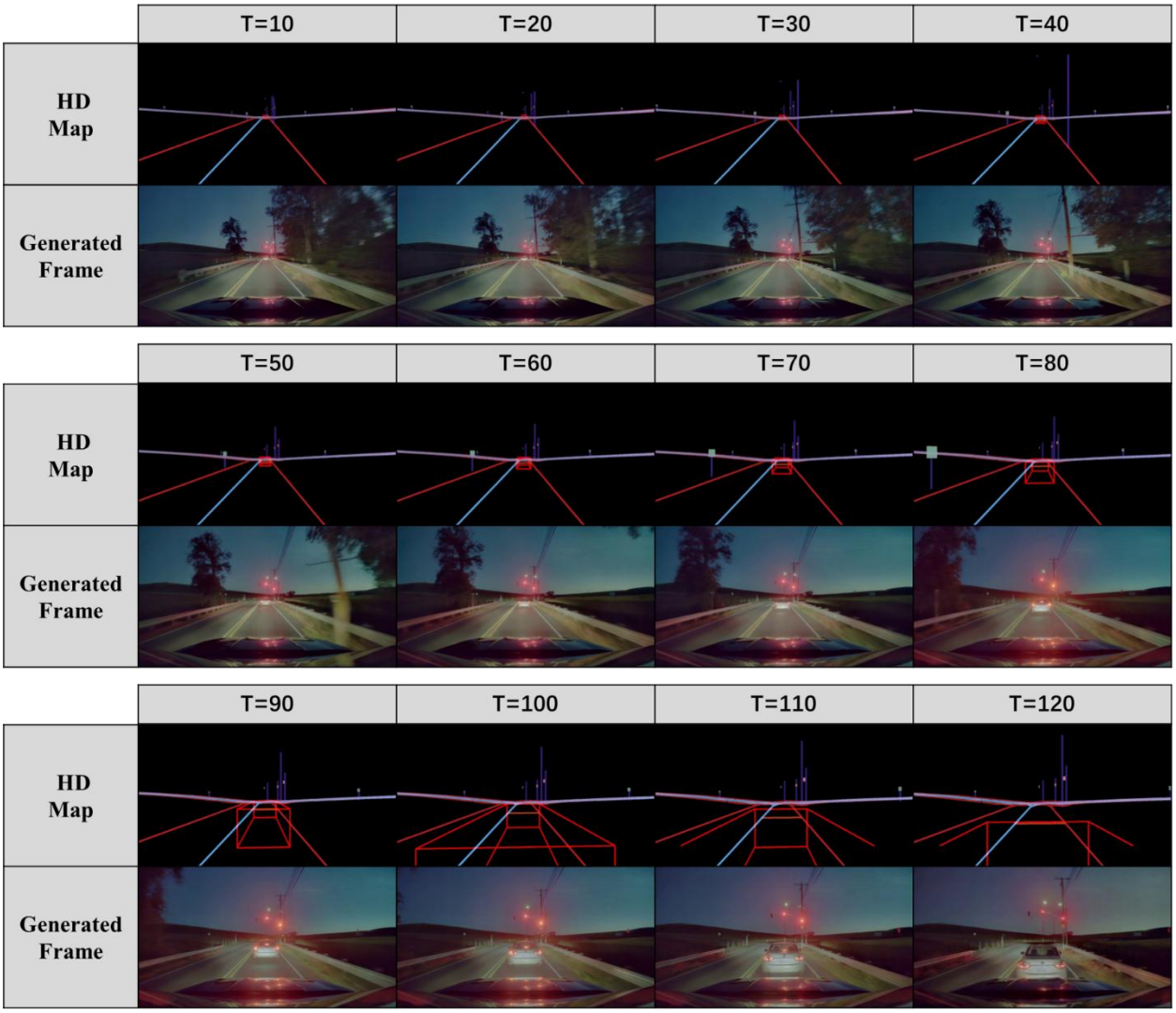


Figure 4. Rear-End near-Collision scenario on a rural road

(2) *Head-On Collision on Mountain Road*

To capture rural mountainous risk factors, we construct a head-on collision on a two-way mountain road featuring a sharp blind curve. Hillsides and dense trees occlude the line of sight, creating extremely limited preview distance and delayed oncoming-vehicle detection. The generated video frames are shown detailed in Figure 5. From frame 50 to 120, the ego vehicle enters the curve at excessive speed, observes the oncoming vehicle only when it emerges around the bend, and the late reaction leads to a frontal collision near the curve apex. This scenario specifically reflects rural mountain-road hazards where geometry-induced visibility constraints and speeding jointly drives severe crash outcomes.

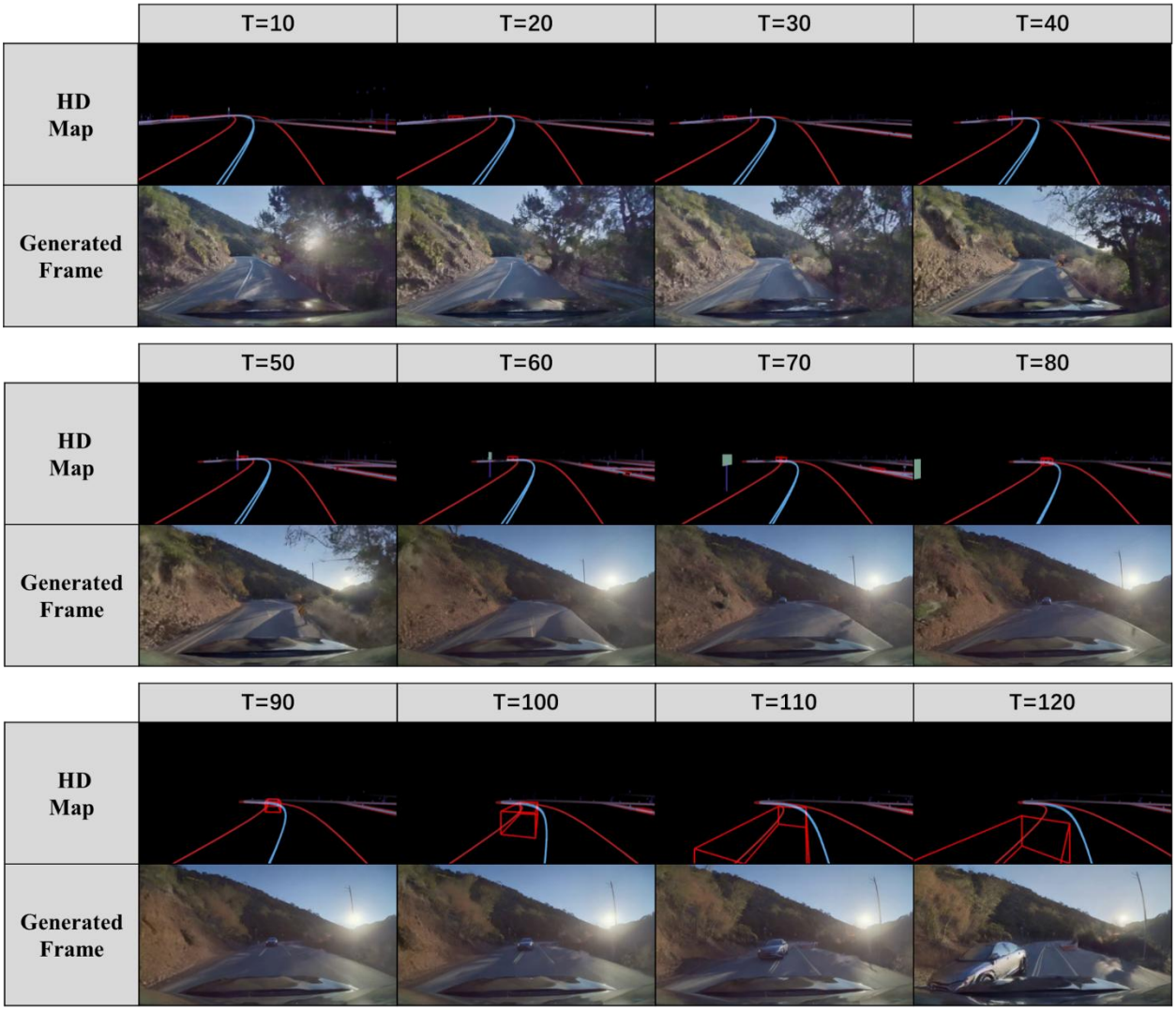


Figure 5. Blind Curve Head-on near Collision Scenario on a Mountain Road

(3) *Three-Way Unsignalized Intersection Crash*

In Figure 6, we further test an unsignalized rural T-junction without stop signs, representing common low-control rural intersections. The environment is characterized by narrow two-lane geometry, open fields, scattered trees, and minimal roadside infrastructure. In the scenario, the ego vehicle proceeds straight at normal speed, while a vehicle from the right-side minor road enters and initiates a left turn to cross the ego path. Due to misjudgment and insufficient braking, a near side-impact collision occurs within the intersection area. This case stresses rural intersection risk driven by sparse traffic control, limited visual cues, and late yielding decisions.

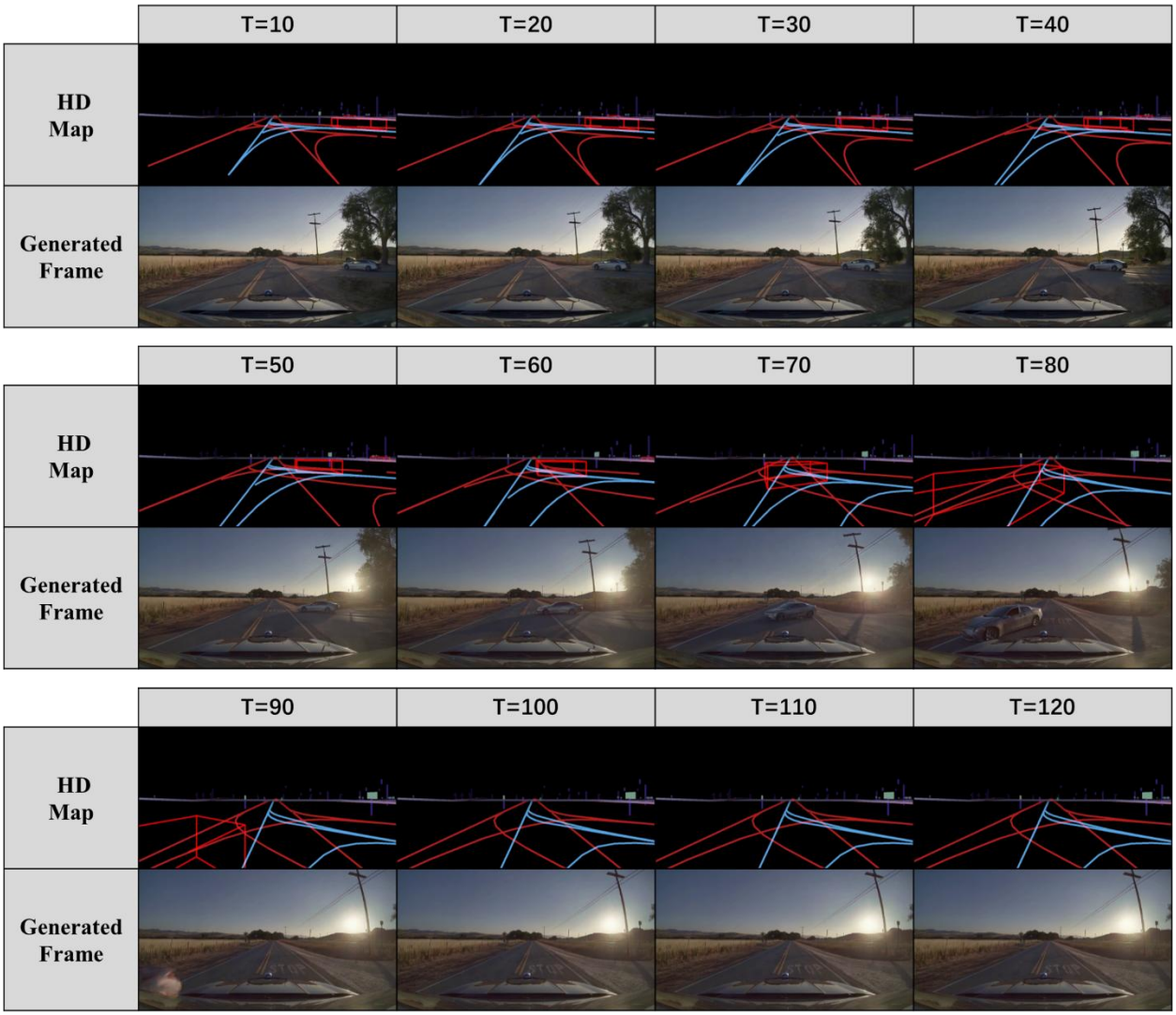


Figure 6. Three-Way Unsignalized Intersection near Collision Scenario

(4) *Four-Way Signalized Intersection Crash*

We consider a rural four-way signalized intersection where the road network is surrounded by open fields and scattered trees with few roadside buildings, reflecting sparse rural development. The video frames are shown in Figure 7. Despite the presence of traffic signals, risky behavior is introduced by having the ego vehicle makes a left turn at frame 40 and enters the intersection in the wrong lane, drifting into opposing lane. Although the synthetic video frames do not strictly adhere to the HD map, likely due to the complexity of the intersection layout, this case provides a useful example for guiding future improvements.

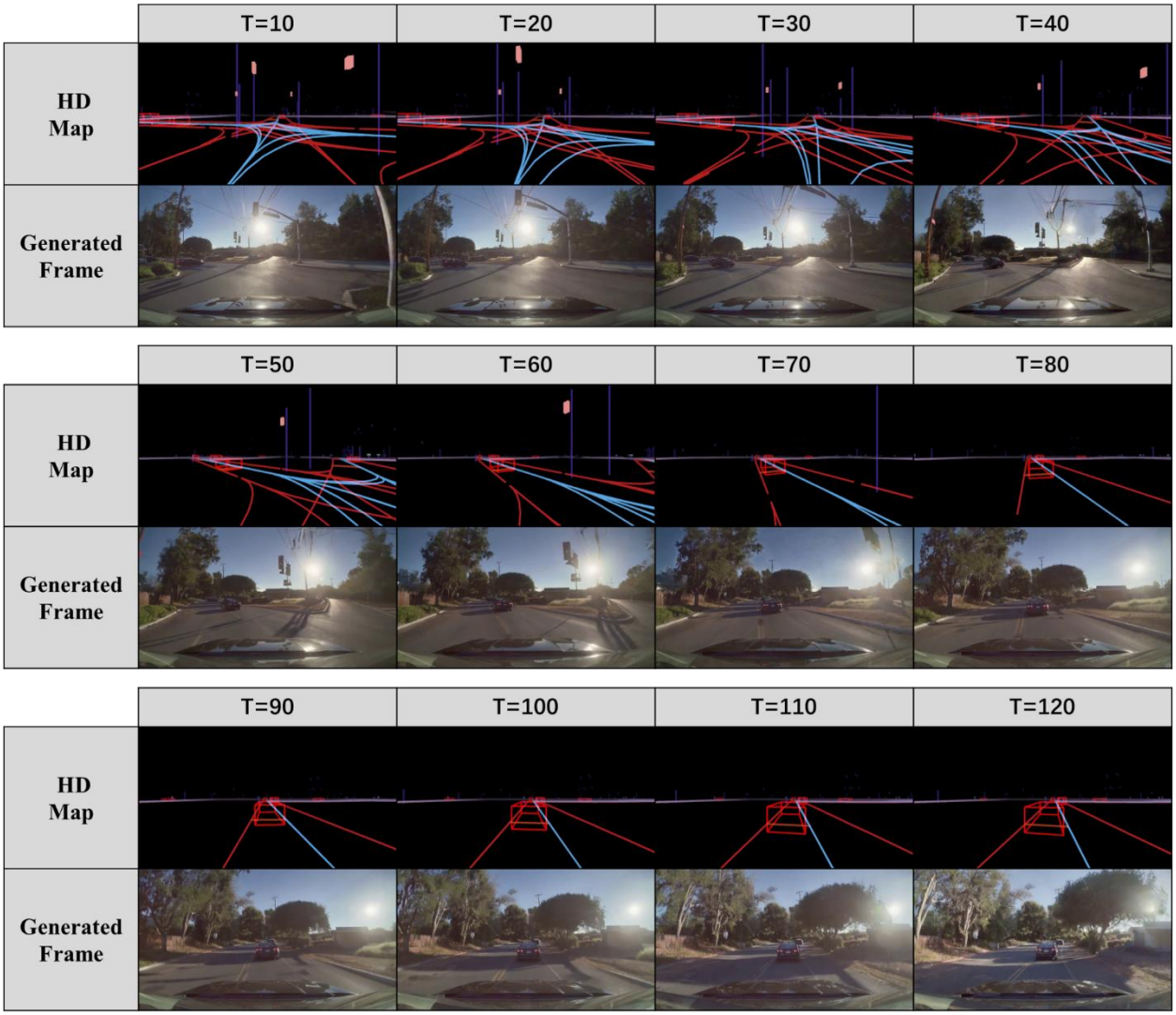


Figure 7. Four-Way Signalized Intersection near Collision

(5) *Lane Departure Collision*

To model rural lane-departure hazards, we generate scenarios on a two-way rural road with open fields, scattered trees, and light rural infrastructure, under adverse conditions that degrade lateral control in Figure 8. Due to snowfall, accumulated snow on the road reduces traction, the ego vehicle gradually drifts across the centerline into the oncoming lane and is unable to recover in time, leading to a near front-to-front collision with an oncoming vehicle. This set of variants highlights rural-specific vulnerability to lane departures: condition-driven loss of traction or perception that can trigger cross-centerline events.

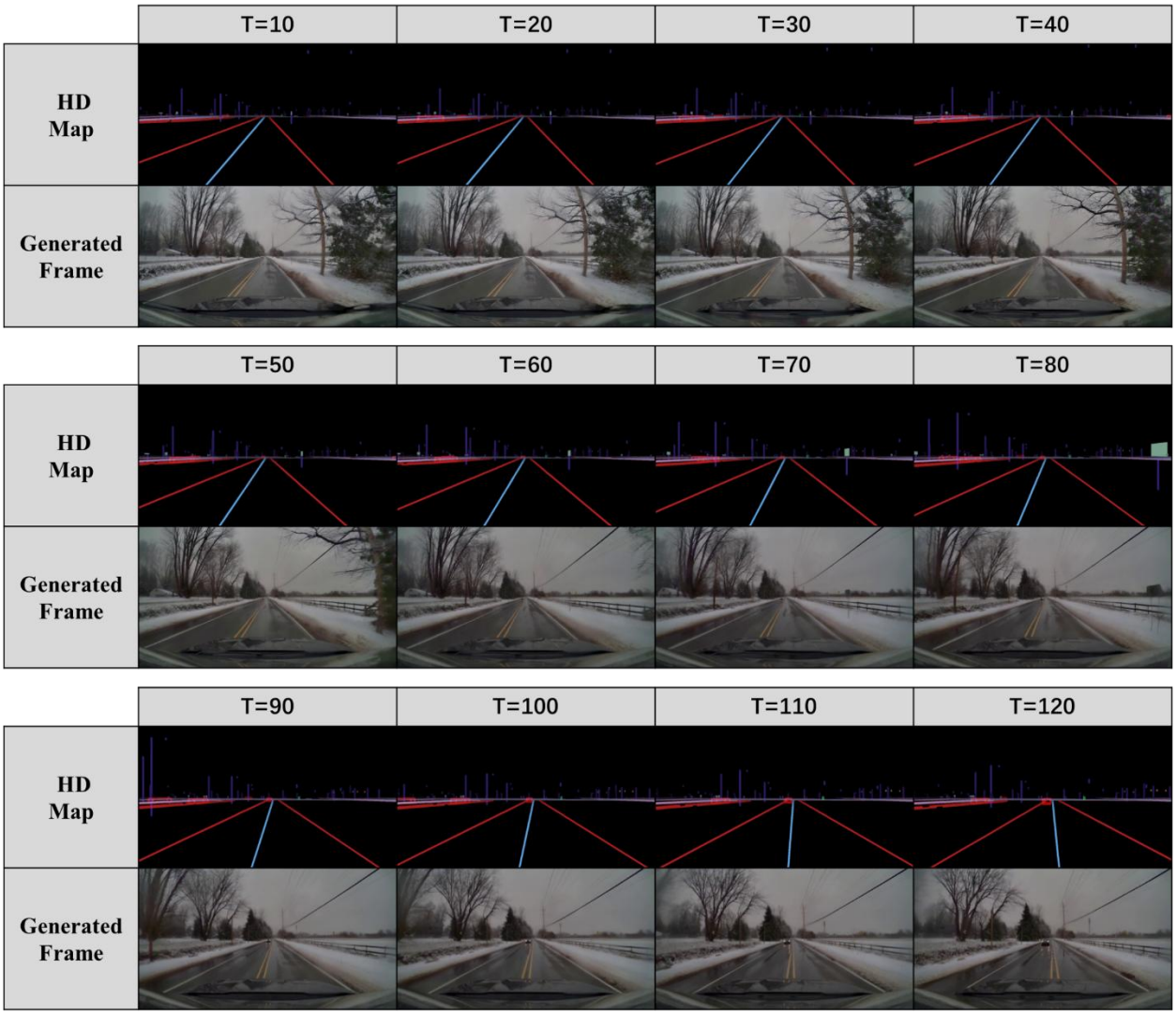


Figure 8. Lane Departure near Collision on a Two-lane Rural Road

(6) *Stop-Sign-Controlled Intersection Crash*

Finally, we examine a stop-sign-controlled rural three-way intersection at night in Figure 9, where the roadway is quiet and street lighting is scarce, leaving headlights as the primary illumination source. The rural context can paradoxically increase risk when drivers underestimate conflict probability. In this scenario, a vehicle from the left continues straight through the junction, while the ego vehicle fails to perceive the stop sign and the cross-traffic vehicle in the dim lighting, proceeds into a left turn without fully stopping, and causes a near side-impact collision in the intersection. This case emphasizes the rural safety challenge that traffic control devices alone may be insufficient under low-light conditions.

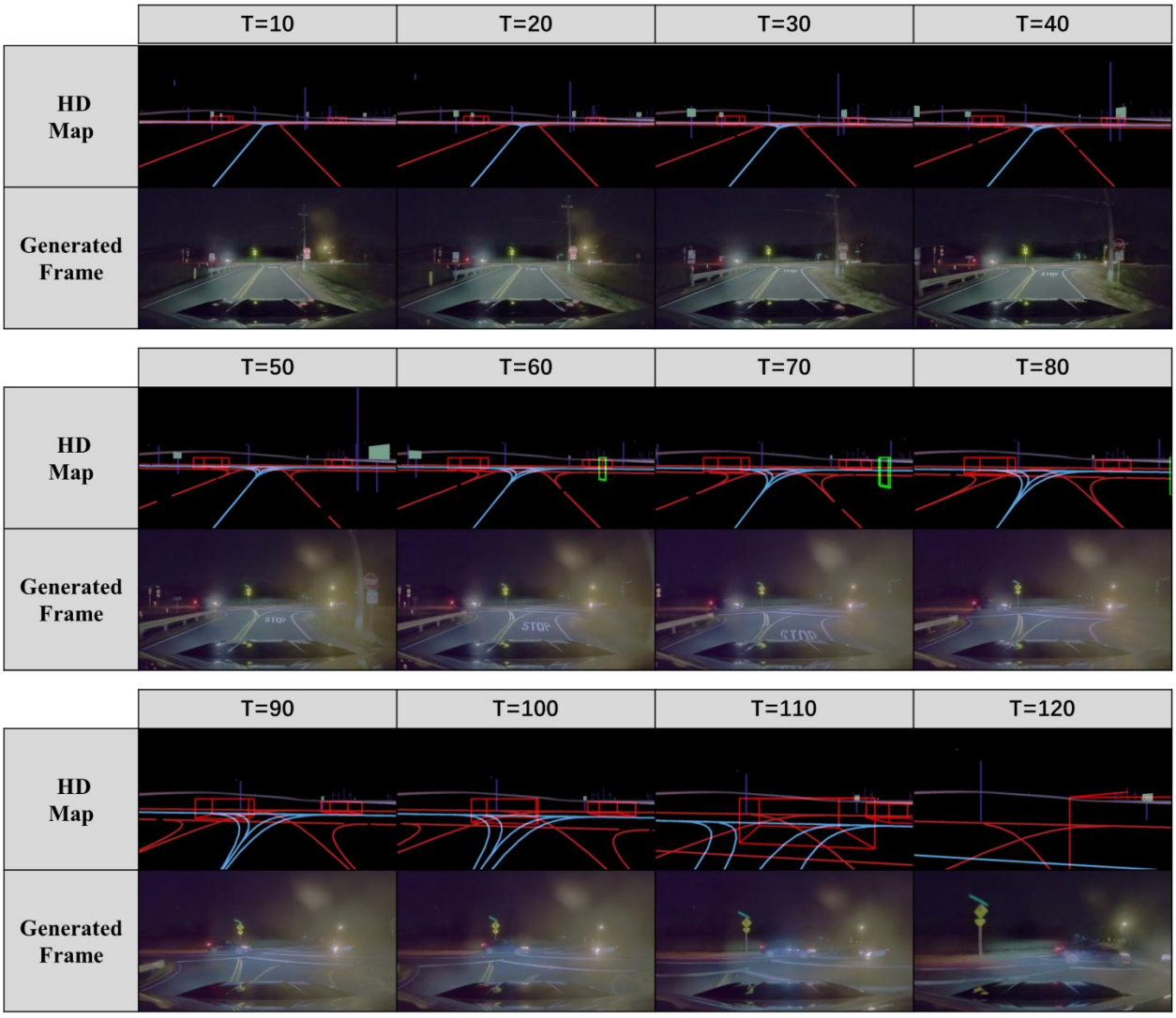


Figure 9. Three-way Stop-Sign-Controlled Intersection near Collision

Adverse Weather and Lighting Conditions:

Figure 10 illustrates the controllability of our HD map conditioned generation under diverse environments. Using the same rural scene layout and the same HD map sequence, the pipeline can synthesize dashcam videos with different appearances, including snowy, rainy, foggy, and nighttime conditions, while preserving the underlying road geometry and scene structure. This capability substantially expands the diversity of rural driving data and directly mitigates the scarcity of real-world data under adverse weather and low light, which are often rare yet highly safety relevant. As a result, the generated clips provide complementary training and evaluation samples for testing perception and prediction models in rural settings, especially for long-tail hazardous conditions where data collection is difficult and costly.

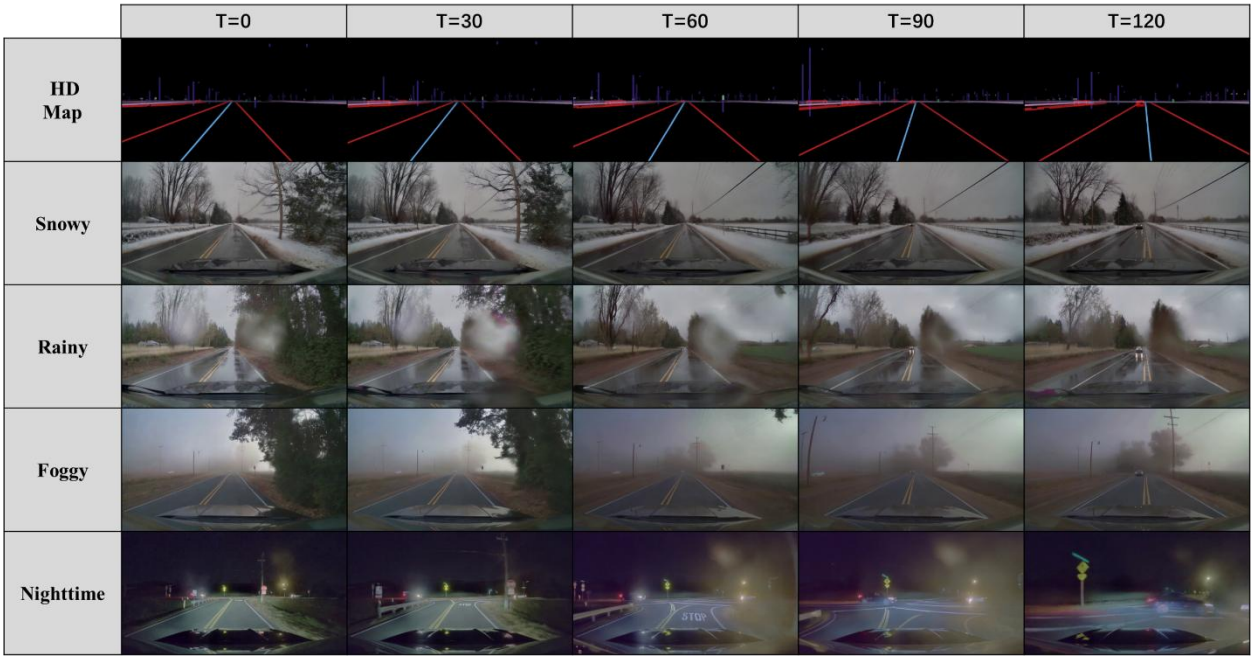


Figure 10. Adverse Weather and Lighting Conditions Generation

3.6 Summary

This study developed a rural Sim-to-Real safety-critical scenario generation pipeline that transforms controllable simulation scenes into real-world driving videos. The pipeline consists of three stages. First, we construct representative rural safety-critical events in CARLA, explicitly modeling rural road geometry and sparse infrastructure, and scripting multi-agent interactions to reproduce near-collision dynamics. Then, we extract HD map representations and pair each scene with a structured textual prompt that describes the environment, weather, lighting, and event type. At last, we employ COSMOS as a video foundation model conditioned on both the HD map and the prompt to generate real-world driving clips that preserve road layout while rendering realistic appearance and event semantics. By design, the pipeline enables the transformation of safety-critical rural scenarios from simulation to realistic driving video generation.

Despite its effectiveness, the current pipeline has several limitations that motivate future work:

- The safety-critical scenarios are constructed manually in simulation, which is time-consuming, and the scripted interactions in CARLA may not fully reflect real-world physical and behavioral driving patterns. A promising direction is to learn scenario distributions directly from real driving data and to develop data-driven generation strategies. For example, combining diffusion models with reinforcement learning to efficiently produce diverse, physically plausible safety-critical events with controllable risk levels.
- While we demonstrate the feasibility of Sim-to-Real generation, the pipeline has not yet been converted into a standardized, released dataset with consistent annotations and metadata. Future work will focus on using this pipeline to build a curated rural safety-critical video dataset, including scenario labels, environment conditions, and map-based descriptions, to better support reproducible training and benchmarking.
- The generated scenarios have not yet been integrated into downstream evaluation of autonomous driving systems. In the next stage, these synthetic rural safety-critical clips can

be used for open-loop testing of perception and prediction modules and extended to closed-loop simulation-based evaluation to stress-test planning and control policies, providing a practical tool for assessing robustness under long-tail rural hazards.

4 CONCLUSIONS

Rural driving presents a unique and urgent safety problem, as fatal crashes are disproportionately frequent on rural roads and drivers face elevated risk under conditions such as sparse infrastructure and limited illumination. At the same time, truly safety-critical rural driving events are less frequently observed and more sparsely captured than their urban counterparts in real-world driving datasets, making large-scale collection challenging. which motivates the adoption of controllable synthetic data generation. Accordingly, we conduct the corresponding studies in this project, with the main contributions summarized as follows:

First, we proposed the Historical Motion Priors Diffusion Model (HMPDM), and it demonstrates that diffusion models equipped with historical motion priors can generate temporally coherent future driving sequences and preserve realistic scene dynamics across frames, which validates diffusion as an effective approach for modeling spatiotemporal evolution in driving videos. This component focuses on improving spatiotemporal consistency in generated driving videos and provides methodological evidence that diffusion-based generation can maintain coherent motion and appearance over time, which is a desirable property for safety-oriented driving video synthesis.

Furthermore, we developed a practical Sim-to-Real pipeline that constructs representative rural safety-critical events in CARLA, exports HD map representations and structured prompts, and uses COSMOS conditioned on both the HD map and the prompt to generate realistic driving clips while preserving road layout and rendering event semantics. Using this pipeline, this study evaluates six types of rural scenarios, including rear-end, mountain-road head-on, three-way unsignalized intersection, four-way signalized intersection, lane-departure under slippery conditions, and stop-sign-controlled intersection conflicts, and further augments each scenario with adverse weather and lighting variations such as snow, rain, fog, and nighttime. Based on the experimental results, the generated clips expand rural safety-critical data coverage in settings where real-world recordings are scarce, and they provide complementary samples for testing perception and prediction models in rural environments, especially under long-tail adverse conditions where data collection is difficult and costly.

Overall, this project advances rural transportation safety by providing practical methods to generate and expand safety-critical driving video data where real-world coverage is limited. Our work helps address a core data gap that constrains safety-focused research and technology development. These methods can directly support USDOT and the broader transportation community by augmenting existing public datasets with long-tail rural hazards, enabling more systematic training, benchmarking, and evaluation of data-driven perception and prediction models. In the longer term, the proposed pipeline offers a scalable foundation for constructing standardized rural safety-critical scenario datasets and for supporting safety evaluation studies that inform the development of more robust and reliable driving assistance and automated driving technologies in rural environments.

REFERENCES

Arun, S., Panchangmath, T., Celamkoti, S., Wong, V., Duong, C., Sharma, V., O'Brien, S., and Zhu, K. (2025). Enhancing Rural Autonomous Driving Performance with Diffusion-Augmented Synthetic Datasets. In *UrbanAI: Harnessing Artificial Intelligence for Smart Cities*.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., & Rombach, R. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *CoRR*, abs/2311.15127.

Bu, F. and Yasuda, H. (2025). Boosting Visual Fidelity in Driving Simulations through Diffusion Models. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp.1–7.

Castrejon, L., Ballas, N., and Courville, A. (2019). Improved Conditional VRNNs for Video Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, T., Zhang, R. and Hinton, G. (2023). Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. *The Eleventh International Conference on Learning Representations*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223.

Denton, E., and Fergus, R. (2018). Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pp. 1174–1183.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16. PMLR.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11), p.1231–1237.

Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Li, F.F., Essa, I., Jiang, L., and Lezama, J. (2024). Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 33, p.6840–6851.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. (2022). Video Diffusion Models. In *Advances in Neural Information Processing Systems*, pp. 8633–8646.

Hoppe, T., Mehrjou, A., Bauer, S., Nielsen, D., and Dittadi, A. (2022). Diffusion Models for Video Prediction and Infilling. *Transactions on Machine Learning Research*.

Insurance Institute for Highway Safety, 2025. Fatality Facts 2023: Urban/rural comparison. Available at: https://www.iihs.org/research-areas/fatality-statistics/detail/urban-rural-comparison [Accessed: 05 December 2025].

Jin, W., Dai, Q., Luo, C., Baek, S.H., and Cho, S. 2025. FloVD: Optical Flow Meets Video Diffusion Model for Enhanced Camera-Controlled Video Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2040-2049.

Karacik, N. D., Xu, Y., Li, X., Hu, Y., & Liu, Y. (2025). SCSG: Real-World Report–Augmented Safety-Critical Scenario Generation for Autonomous Vehicles. In *Embodied and Safe-Assured Robotic Systems.*

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, pp. 26565–26577.

Klitzke, L., Leschik, C., Lüdtke, R., and Gimm, K. (2025). Investigation on road traffic safety in rural areas using trajectory data: case studies at two measurement sites. *Traffic Safety Research*, 9, p.e000120

Li, H., Yang, Z., Qian, Z., Zhao, G., Huang, Y., Yu, J., Zhou, H., & Liu, L. (2025). DualDiff: Dual-branch Diffusion Model for Autonomous Driving with Semantic Fusion. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Li, K., Zhang, C., Ding, Y., Hu, X., and Qin, R. (2025). Multi-label Scene Classification for Autonomous Vehicles: Acquiring and Accumulating Knowledge from Diverse Datasets. *arXiv e-prints*, p.arXiv–2506.

Liang, J., Fan, Y., Zhang, K., Timofte, R., Van Gool, L. and Ranjan, R., 2024, September. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pp. 56-74.

Lotter, W., Kreiman, G., and Cox, D. (2017). Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In *International Conference on Learning Representations*.

Lu, J., Wong, K., Zhang, C., Suo, S., and Urtasun, R. (2024). SceneControl: Diffusion for Controllable Traffic Scene Generation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16908-16914.

Mei, K., and Patel, V. (2023). VIDM: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9117–9125.

Ni, H., Shi, C., Li, K., Huang, S., and Min, M. 2023. Conditional Image-to-Video Generation With Latent Flow Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18444-18455.

O' Kelly, M., Sinha, A., Namkoong, H., Tedrake, R., and Duchi, J. (2018). Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation. In *Advances in Neural Information Processing Systems*.

Pallotta, E., Azar, S., Li, S., Zatsarynna, O., and Gall, J. (2025). SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13787-13797.

Pronovost, E., Ganesina, M., Hendy, N., Wang, Z., Morales, A., Wang, K., and Roy, N. (2023). Scenario Diffusion: Controllable Driving Scenario Generation With Diffusion. In *Advances in Neural Information Processing Systems*, pp. 68873–68894.

Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J. Z., Chen, R., Kim, S. W., Gao, J., Leal-Taixe, L., Chen, M., Fidler, S., & Ling, H. (2025). Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models.

Safari, M., Effati, M. and Arabani, M. (2025). Analyzing the severity of distracted driving crashes on horizontal curves of rural highways: A hybrid approach of boosting-based ensemble machine learning and discrete choice methods. *Transportation Research Interdisciplinary Perspectives*, 32, p.101542.

Samak, C., Samak, T., Li, B., and Krovi, V. (2026). Sim2Real Diffusion: Leveraging Foundation Vision Language Models for Adaptive Automated Driving. *IEEE Robotics and Automation Letters*, 11(1), p.177-184.

Song, J., Meng, C., & Ermon, S. (2021). Denoising Diffusion Implicit Models. *International Conference on Learning Representations.*

Sum, S., Se, C., Champahom, T., Jomnonkwao, S., Sinha, S., & Ratanavaraha, V. (2025). A random forest and SHAP-based analysis of motorcycle crash severity in Thailand: Urban-rural and day-night perspectives. *Transportation Engineering*, 21, 100369.

Teng, S., Hu, X., Deng, P., Li, B., Li, Y., Ai, Y., Yang, D., Li, L., Xuanyuan, Z., Zhu, F., and Chen, L. (2023). Motion Planning for Autonomous Driving: The State of the Art and Future Perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6), p.3692-3711.

U.S. Department of Transportation. (2025). Rural Roadway Safety. Available at: https://www.transportation.gov/rural/safety [Accessed: 05 December 2025].

United States Department of Transportation, National Highway Traffic Safety Administration, National Center for Statistics and Analysis. (2024). Traffic Safety Facts 2021 Data: Rural/Urban Comparison of Motor Vehicle Traffic Fatalities. (Report No. DOT HS 813 488). Washington, DC: National Highway Traffic Safety Administration.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. *arXiv preprint* arXiv:1812.01717.

Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. (2017). Decomposing Motion and Content for Natural Video Sequence Prediction. In *International Conference on Learning Representations*.

Voleti, V., Jolicoeur-Martineau, A., and Pal, C. (2022). MCVD - Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. In *Advances in Neural Information Processing Systems*, pp. 23371–23385.

Wang, J., Sun, H., Yan, X., Feng, S., Gao, J., & Liu, H. X. (2025). TeraSim-World: Worldwide Safety-Critical Data Synthesis for End-to-End Autonomous Driving.

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), pp. 600–612.

Wu, B., Nair, S., Martin-Martin, R., Fei-Fei, L., and Finn, C. (2021). Greedy Hierarchical Variational Autoencoders for Large-Scale Video Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2318-2328.

Yang, R., Srivastava, P., and Mandt, S. (2023). Diffusion probabilistic modeling for video generation. *Entropy*, 25(10), p.1469.

Yang, S., Zhang, L., Liu, Y., Jiang, Z., and He, Y. (2023). Video diffusion models with local-global context guidance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Ye, X., and Bilodeau, G.A. (2023). A unified model for continuous conditional video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3604–3613.

Ye, X., and Bilodeau, G.A. (2024). STDiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6666–6674.

Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.-X., Cao, X., & Yin, W. (2025). Epona: Autoregressive Diffusion World Model for Autonomous Driving.

Zhang, R., Isola, P., Efros, A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z., Hu, J., Cheng, W., Paudel, D., and Yang, J. (2024). ExtDM: Distribution Extrapolation Diffusion Model for Video Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19310-19320.

Zhao, H., Wang, Y., Bashford-Rogers, T., Donzella, V., and Debattista, K. (2024). Exploring Generative AI for Sim2Real in Driving Data Synthesis. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 3071-3077.

Zhong, Z., Tang, Y., Zhou, Y., Neves, V.D.O., Liu, Y. and Ray, B. (2021). A survey on scenario-based testing for automated driving systems in high-fidelity simulation. *arXiv preprint* arXiv:2112.00964.

Zhou, Y., Simon, M., Peng, Z., Mo, S., Zhu, H., Guo, M., and Zhou, B. (2024). SimGen: Simulator-conditioned Driving Scene Generation. In *Advances in Neural Information Processing Systems*, pp. 48838–48874.