Generating the Safety Spectrum: LLM-Guided Counterfactual Diffusion for Autonomous Driving

Abstract— **The development of reliable autonomous driving systems requires comprehensive testing across different safety-critical driving scenarios, but real-world crash data are scarce and near-crash events are difficult to capture systematically. To address these challenges, We present a controllable video generation framework that synthesizes crash, near-crash, and safe driving scenarios from natural language prompts. Our approach adapts existing video generation pipeline through low-rank adaption fine-tuning, introducing end-frame conditioning to guide generation toward specific outcomes and multi-condition classifier-free guidance for independent control over visual anchors and semantic prompts. We employ two complementary prompt strategies and demonstrate that description-based prompts achieve superior performance. Comprehensive evaluation on the Crash-1500 dataset shows that our method achieves 31.086 average CLIPScore, a 8.67% improvement over baseline methods, and attains near 100% category classification accuracy with description-based prompts. The framework generates temporally coherent, physically plausible scenarios suitable for systematic autonomous driving system testing and validation.**

I. Introduction

As autonomous driving systems advance toward widespread deployment, comprehensive safety validation has become critical. Reliable systems require testing across the whole safety spectrum. However, real-world crash data are scarce, and near-crash events are particularly difficult to capture systematically. Traditional physics-based simulators [5], [21] lack photorealistic rendering, while data-driven approaches [8], [15] primarily focus on post-collision scenarios, missing the critical pre-crash moments essential for preventive testing. Moreover, existing methods lack fine-grained control over scenario parameters and struggle to generate realistic scenarios that capture the subtle distinctions between safe, near-crash, and crash situations.

Recent advances in video generation, such as Stable Video Diffusion [2] and FramePack [24], demonstrate impressive capabilities in controllable video synthesis. The integration of large language models (LLMs) with video generation presents unique opportunities for safety-critical scenario generation, as vision-language models [13], [19] and prompt engineering techniques [9], [16] enable encoding complex causal relationships through natural language. However, general-purpose video generation models face significant challenges when adapted to safety-critical scenarios, particularly in capturing subtle dynamics and physical constraints present in automotive crash situations.

To overcome these limitations, we introduce a specialized controllable video generation framework that synthesizes driving scenarios across the safety spectrum guided by natural language prompts. Our approach adapts FramePack [24] through LoRA fine-tuning [12] to specialize in safety-critical scenario generation. Unlike previous video synthesis methods [7], [8], [15] typically condition solely on initial frames, limiting fine-grained trajectory control, we propose dual-frame conditioning: pairing initial frames to establish starting conditions with end frames to constrain target outcomes, ensuring logical progression across crash, near-crash, and safe sates. This dual-anchor conditioning ensures logical scenario progression from start to target state, while our multi-condition guidance mechanism that independently controls visual anchors and semantic prompts. The key contributions of this study are summarized as follows:

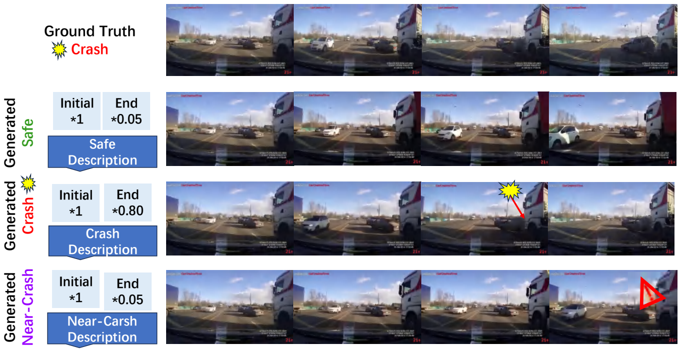


Fig. 1: **Counterfactual video synthesis across the safety spectrum.** Given initial frames and end-state targets, our framework generates temporally coherent scenarios spanning safe driving, near-crash avoidance, and collision events. Natural language descriptions provide semantic control over environmental conditions, agent behaviors, and outcomes.

1) **LoRA-Enhanced Video Generation:** We adapt FramePack through LoRA fine-tuning to enable controllable synthesis of crash, near-crash, and safe driving scenarios, while maintaining the pre-trained capabilities.
2) **Multi-Condition Guidance Mechanism:** We enhance classifier-free guidance to separately manage the initial frames, final frames, and text prompts, allowing precise control of the trajectory across different security categories.
3) **Dual Prompt Strategy:** We compare ground truth temporal labels with LLM-generated descriptions, showing that description-based prompts achieve better semantic alignment and classification performance.

II. Related Work

A. Video Generation Models

The field of video generation has experienced rapid advancement with the introduction of diffusion-based models. Stable Video Diffusion [2] represents a significant milestone in text-to-video synthesis, demonstrating impressive capabilities in generating temporally consistent video content from textual descriptions. Open-Sora [26] has further pushed the boundaries of open-source video generation, providing high-quality synthesis with improved temporal coherence.

Recent controllable video generation frameworks have shown particular promise for domain-specific applications. FramePack [24] introduces sophisticated control mechanisms for video synthesis for fine-grained manipulation of scene dynamics. AnimateDiff [10] proposes motion module adaptation for personalized video generation, while VideoLDM [3] demonstrates effective text-to-video synthesis through latent diffusion approaches. However, these general-purpose models face significant challenges when adapted to safety-critical scenarios, particularly in capturing the subtle dynamics and physical constraints present in automotive crash situations.

B. Crash Scenario Generation

Traditional approaches to crash scenario generation have primarily relied on physics-based simulation environments. CARLA [5] and AirSim [21] provide comprehensive simulation platforms for autonomous driving research, but often struggle with photorealistic rendering and behavioral diversity in critical scenarios. SUMO [20] offers traffic simulation capabilities but lacks the visual fidelity required for vision-based Autonomous driving systems (ADS) testing.

Recent data-driven approaches have emerged to address these limitations. TrafficGen [6] introduces realistic traffic scenario generation using adversarial networks, while ScenarioNet [18] provides large-scale scenario datasets for autonomous driving. Specifically focusing on crash scenarios, Crash-1500 [1] presents a comprehensive dataset of real-world accident videos with detailed annotations. Building upon this foundation, Ctrl-Crash [8] demonstrates controllable crash video generation using textual descriptions and spatial constraints. However, these approaches primarily focus on post-collision scenarios rather than the critical pre-crash moments that are essential for preventive ADS testing.

C. LLM-based Scene Understanding and Generation

Vision-language models have revolutionized scene understanding and generation capabilities. GPT-4V [13] demonstrates remarkable proficiency in visual reasoning and scene description, while LLaVA [19] provides efficient multimodal understanding through instruction tuning. BLIP-2 [17] introduces bootstrapped vision-language pre-training, which can achieve complex image text alignment to complete the task of scene description.

The application of large language models to driving scenarios has shown promising results. DriveLM [22] explores closed-loop autonomous driving with language models, while Talk2Drive [4] demonstrates natural language interaction for driving scene understanding. Recent work on prompt engineering for controlled generation [9], [16] provides valuable insights into achieving precise control over generated content through carefully designed textual instructions.

For crash and safety-critical scenario generation, the integration of LLMs with video generation models presents unique opportunities. The ability to encode complex causal relationships and temporal dynamics through natural language descriptions offers a more intuitive and controllable approach compared to traditional parameter-based methods.

III. Methodology

A. Overview

We adapt FramePack [24] through LoRA fine-tuning to generate three scenario types: crash, near-crash, and safe driving. As illustrated in Figure 2, the pipeline processes initial and end frames with text prompts to synthesize 72-frame videos. Key frames from crash datasets serve as visual anchors, paired with textual descriptions for controllable synthesis. The fine-tuned diffusion model produces temporally coherent, physically plausible sequences across all safety categories (crash, near-crash, or safe) , supporting systematic generation of rare scenarios for ADS testing.

B. Data Preparation

a) Key-Frame Extraction: We extract two critical frames per video, following the Crash-1500 structure [1]: The initial frame (pre-event conditions) and the end frame (outcome state). These serve as visual anchors. We then use optical flow analysis to detect safety-critical transitions. The final extracted sequence is a 25-frame segment centered around the event, standardized to 6 fps and 512×320 resolution.

b) Video Prompt Generation: We utilize two complementary prompt strategies for video description:

- **Method 1:** Uses ground truth temporal labels formatted as structured event markers (e.g., "[0s: Normal driving] [3s: Crash begins] [5s: Impact concludes]").
- **Method 2:** Leverages GPT-5 to generate natural language descriptions covering environmental context, vehicle behaviors, causal factors, and outcome characteristics.

This dual approach enables comparison of structural versus semantic conditioning strategies.

C. Training Pipeline

a) FramePack with LoRA Fine-tuning: We fine-tune FramePack using LoRA (rank $r = 64$), applying it to the model's key components: temporal attention, cross-attention, and spatial convolution blocks. This selective adaptation specializes the model for safety-critical scenario generation. Temporal attention captures progression dynamics, cross-attention enhances text-video alignment, and spatial convolutions learn realistic crash visual details.

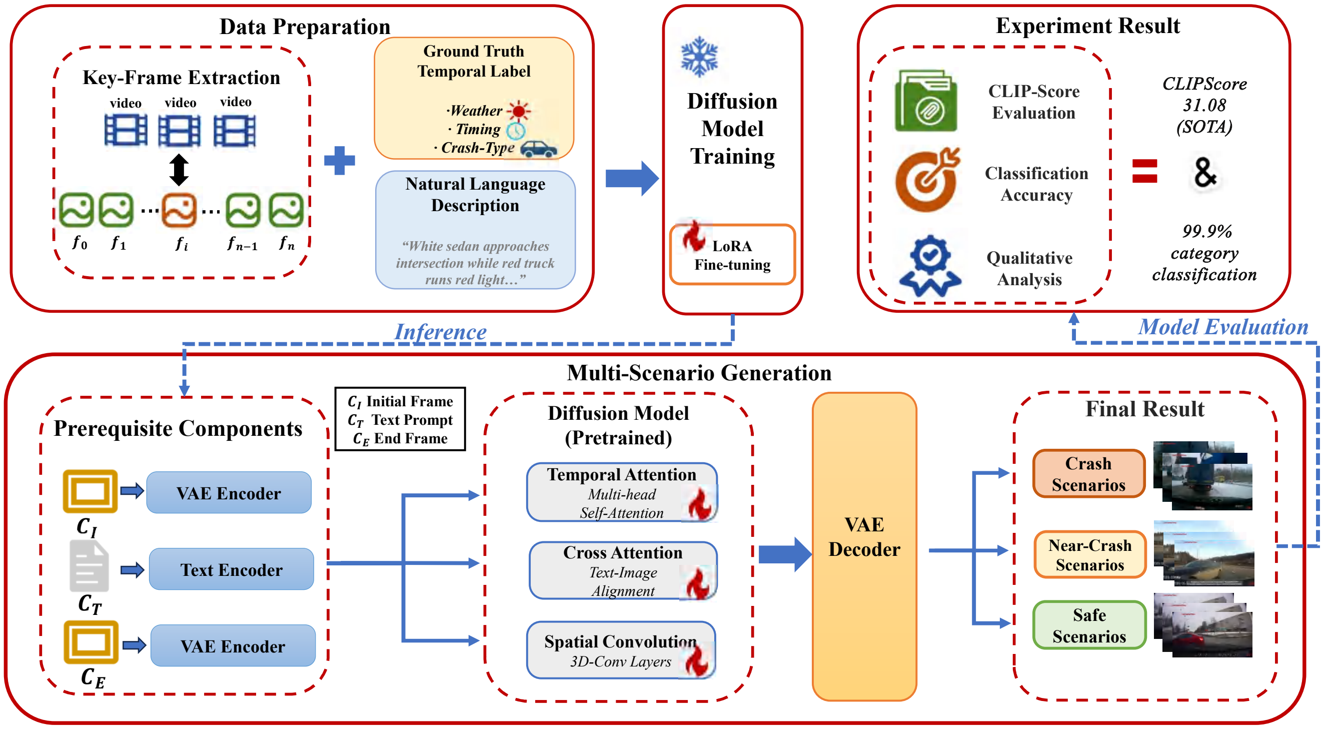


Fig. 2: **Overall Architecture of the Video Generation System.** The diagram illustrates the pipeline for generating crash, near-crash, and safe driving scenarios using the architecture. Structured data, natural language prompts, and fine-tuned large language models drive the scenario generation, while a pretrained diffusion model and VAE-based processing are used for video synthesis. The system outputs videos with 72 frames at a resolution of 512x320.

b) Conditioning Masking Strategy: Following classifier-free guidance, we employ randomized masking during training to prevent over-reliance on any single modality. For Prompt Conditioning, we apply temporal dropout, starting at 20% masking probability for the first 10k steps, then increasing to 40%. For Visual Conditioning, an independent 15% masking is applied to both the initial and end frames. This strategy allows flexible control during inference using guidance coefficients γ_E and γ_T in Equation (2).

D. Prerequisite Components

a) Initial Frame (c_I): The initial frame c_I establishes visual grounding (scene appearance, layout, environment) via a pretrained VAE encoder [14]. This anchors the generation to a realistic starting condition, serving as the baseline from which all scenario evolution emerges.

b) End Frame (c_E): The end frame c_E provides crucial target-state conditioning for precise trajectory control. Processed by the same VAE encoder as c_I, c_E guides the diffusion process toward specific outcomes via cross-attention. It constrains temporal dynamics: crash scenarios terminate at collision, near-crash at successful avoidance, and safe scenarios maintain stability.

c) Prompt Conditioning (c_T): Text prompts c_T offer semantic control through structured natural language, integrated via FramePack's cross-attention after processing by a pretrained text encoder. Prompts specify scenario type, environmental parameters, vehicle dynamics, and outcome, supporting both high-level category specification and detailed scenario control.

E. Multi-Condition Classifier-Free Guidance for Rare Scenarios

FramePack is designed to generate three scenario types—Crash, Near-crash, and Safe scenarios by leveraging a multi-condition classifier-free guidance mechanism. This framework provides fine-grained control over the generated scenario progression, from stable initial states to specific target outcomes. The iterative denoising process is formulated as a controlled diffusion process where the noise prediction $\hat{\epsilon}_{\theta,\phi}(\cdot,\tau)$ is a superposition of three concurrent flows in the latent space: the base diffusion trajectory, the end frame steering, and the semantic refinement:

$$\frac{dx_t}{d\tau} = -\hat{\epsilon}_{\theta,\phi}(x_t, c_I, c_E, c_T, \tau), \tag{1}$$

where x_t evolves along the reverse diffusion trajectory parameterized by τ. The guided noise prediction satisfies:

$$\hat{\epsilon}_{\theta,\phi}(\cdot,\tau) = \epsilon_\phi(\cdot, c_I, \tau) + \gamma_E \mathcal{D}_E(\cdot,\tau) + \gamma_T \mathcal{D}_T(\cdot,\tau), \tag{2}$$

The differential operators for End Frame guidance ($\mathcal{D}_E$) and Text Prompt guidance ($\mathcal{D}_T$) are defined as:

$$\mathcal{D}_E := \epsilon_\theta(\cdot, c_I, c_E, \emptyset, \tau) - \epsilon_\phi(\cdot, c_I, \emptyset, \emptyset, \tau), \tag{3}$$

$$\mathcal{D}_T := \epsilon_\theta(\cdot, c_I, c_E, c_T, \tau) - \epsilon_\theta(\cdot, c_I, c_E, \emptyset, \tau). \tag{4}$$

Here, c_I, c_E, and c_T are the initial frame, end frame, and text prompt conditions, respectively, and γ_E, γ_T are the corresponding guidance coefficients. This unified formulation ensures precise control over the complex dynamics of the three target scenario types (see Figure 3 for examples).

IV. Experiments

A. Implementation Details

The training process runs on 4 NVIDIA RTX 4090D GPUs (24GB each) with batch size 4 per GPU, completing 25 epochs in approximately 20 hours. We report averaged results across multiple runs for statistical reliability. Training utilized the Crash-1500 dataset [1], which comprises 1,500 real-world dashcam videos of vehicular crashes. For visual conditioning, we extracted the initial and end frames for each video, as detailed in Section III.

To facilitate a comparative analysis, we generated two versions of each video: one utilizing Method 1 (structured temporal labels combined with frames) and another using Method 2 (natural language descriptions combined with frames). By strategically altering the prompt semantics, we demonstrated controllable synthesis by generating three distinct scenario categories, safe, near-crash, and crash, all derived from the identical initial frames.

B. CLIPScore Evaluation

We evaluate semantic alignment using frame-wise CLIPScore [11]. For each generated video sequence (72 frames), the CLIP similarity (ViT-B/32) is computed and averaged across all frames to obtain the final video-level score. Higher values indicate stronger agreement between the generated visual content and the textual prompt.

Table I summarizes category-wise performance along with external benchmarks. Both fine-tuned variants outperform the baseline methods across all scenario types. Method 1 (ground-truth temporal labels) yields a notable improvement of roughly 9.8% over Tune-A-Video for near-crash scenarios, indicating that end-frame guidance effectively directs the generative trajectory toward expected outcomes.

Method 2 (rich textual descriptions) achieves the highest overall performance, with an average CLIPScore of 31.08. Relative to Tune-A-Video, this represents an improvement of approximately 12.6%, and it also slightly surpasses ControlVideo. The largest category gain appears in safe scenarios, where Method 2 exceeds Tune-A-Video by around 11.3% and Method 1 by approximately 7.1%. This suggests that detailed semantic descriptions help produce more consistent and stable driving behaviors under non-critical conditions.

Scenario-wise trends further reveal that near-crash sequences consistently achieve the highest CLIPScores (30.30 and 31.18 for Methods 1 and 2). These intermediate-criticality situations appear to benefit significantly from structured prompt conditioning. In crash scenarios, Method 2 maintains an advantage of about 10.4% over Method 1. This suggests that rich, description-based prompts are substantially more effective at capturing the complex, nuanced dynamics of collision events compared to relying solely on sparse temporal labels.

TABLE I: Merged CLIPScore Performance Comparison by Category and Benchmark

Method	**Safe↑**	**Near-Crash↑**	**Crash↑**	**Avg↑**
Tune-A-Video [23]	27.8	27.6	27.3	27.58
ControlVideo [25]	30.9	30.8	30.6	30.79
Temporal Label[a] (Ours)	28.89	30.30	28.18	29.12
Description[b] (Ours)	**30.95**	**31.18**	**31.12**	**31.08**

[a] Method 1: Fine-tuned using Temporal Label prompts.
[b] Method 2: Fine-tuned using natural language Description prompts.

C. Classification Accuracy

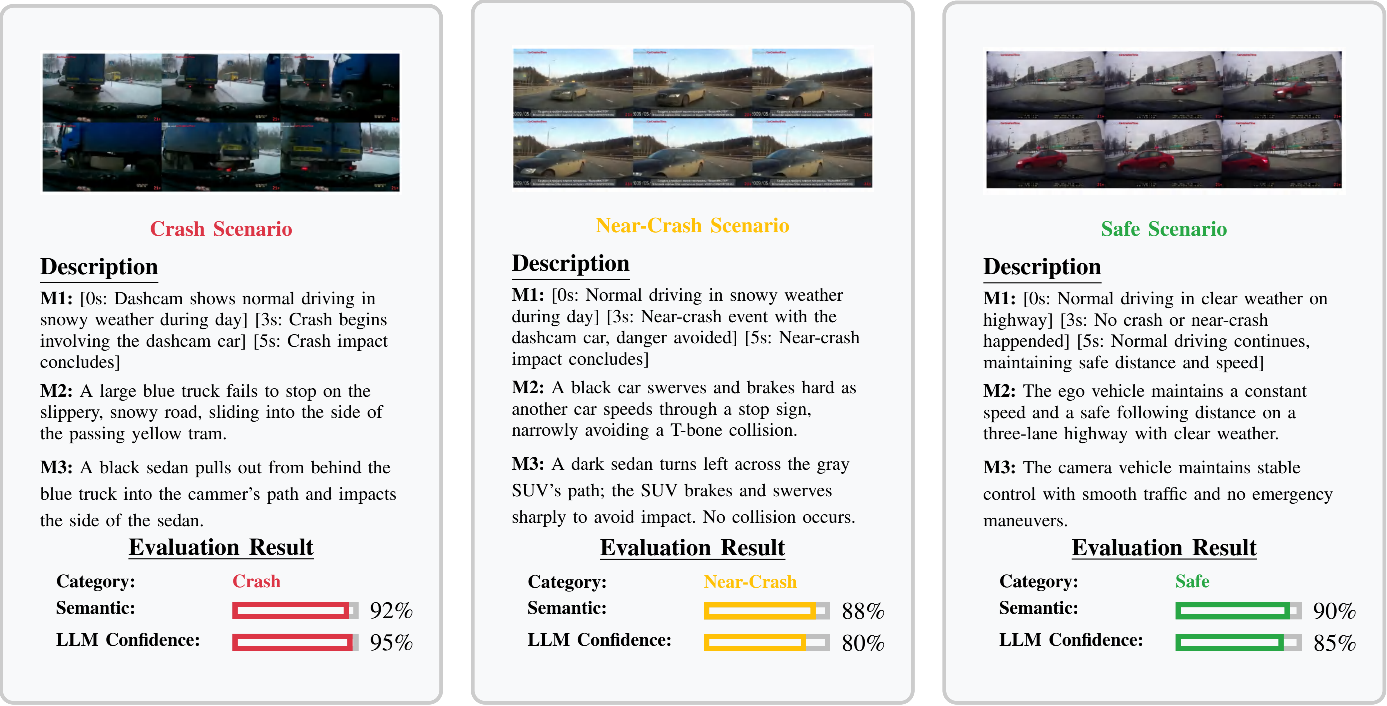
We evaluated the category adherence of the generated videos using Gemini-2.5-flash as an independent, external classifier. The LLM was tasked with classifying scenarios into three categories based purely on visual evidence, without access to the original conditioning prompts. To enhance statistical validity, we increased the sample size by randomly selecting 50 generated videos per category, resulting in a total of 150 samples evaluated for each method.

Table II presents the classification results. Temporal Label Prompt demonstrates a robust average accuracy of approximately 96.7%. The results reveal minor systematic confusion: Safe scenarios achieve perfect recall but experience a drop in precision due to a limited number of false positives. Conversely, the critical categories (Near-crash and Crash) maintain perfect precision but exhibit a slight degradation in recall. This pattern suggests a potential conservative classification bias, where the classifier tends to downgrade borderline or visually ambiguous intermediate stages.

TABLE II: Category Classification Performance Comparison

Prompt	**Category**	**Prec↑**	**Rec↑**	**F1↑**
Temporal Label	Safe	0.909	1.000	0.952
	Near-Crash	1.000	0.950	0.974
	Crash	1.000	0.950	0.974
	Average	*0.970*	*0.967*	*0.967*
Description	Safe	1.000	1.000	1.000
	Near-Crash	1.000	1.000	1.000
	Crash	1.000	1.000	1.000
	Average	*1.000*	*1.000*	*1.000*

Description Prompt achieved perfect classification ($\mathbf{100\%}$ accuracy) across all 150 tested samples with zero false positives or negatives. The perfect scores are attributed to two factors: the inherent high-contrast nature of the classification scheme, and the capacity of the rich semantic descriptions to enforce the synthesis of videos with clear visual features. This evidence suggests that semantic descriptions, unlike sparse temporal labels, effectively enforce strong, clear visual distinctions between safety levels, thereby preventing the generation of ambiguous intermediate states.



Crash Scenario

Description

M1: [0s: Dashcam shows normal driving in snowy weather during day] [3s: Crash begins involving the dashcam car] [5s: Crash impact concludes]

M2: A large blue truck fails to stop on the slippery, snowy road, sliding into the side of the passing yellow tram.

M3: A black sedan pulls out from behind the blue truck into the cammer's path and impacts the side of the sedan.

Evaluation Result

Category:	Crash	
Semantic:	🟥	92%
LLM Confidence:	🟥	95%

Near-Crash Scenario

Description

M1: [0s: Normal driving in snowy weather during day] [3s: Near-crash event with the dashcam car, danger avoided] [5s: Near-crash impact concludes]

M2: A black car swerves and brakes hard as another car speeds through a stop sign, narrowly avoiding a T-bone collision.

M3: A dark sedan turns left across the gray SUV's path; the SUV brakes and swerves sharply to avoid impact. No collision occurs.

Evaluation Result

Category:	Near-Crash	
Semantic:	🟨	88%
LLM Confidence:	🟨	80%

Safe Scenario

Description

M1: [0s: Normal driving in clear weather on highway] [3s: No crash or near-crash happended] [5s: Normal driving continues, maintaining safe distance and speed]

M2: The ego vehicle maintains a constant speed and a safe following distance on a three-lane highway with clear weather.

M3: The camera vehicle maintains stable control with smooth traffic and no emergency maneuvers.

Evaluation Result

Category:	Safe	
Semantic:	🟩	90%
LLM Confidence:	🟩	85%

Fig. 3: Tri-method assessment for crash, near-crash, and safe scenarios, based on three methods: m1: temporal labels, m2: expert descriptions, and m3: automated semantic analysis. Each panel integrates the synthesized frame, descriptive narratives, and evaluation metrics with color-coded severity indicators.

D. *Qualitative Analysis*

We perform a comprehensive manual evaluation in addition to automated metrics, utilizing three distinct assessment methods. These methods are temporal labels (as objective time markers), expert human descriptions (serving as semantic references), and analysis generated by the Gemini 2.5 Flash LLM. This evaluation focuses on assessing the resulting scenarios' temporal accuracy, semantic consistency (agreement between the LLM and experts), and the calibration of the LLM's classification confidence.

The results in Figure 3 consolidate the imagery, prompts, and manual assessments for the three representative scenarios. A key finding is the high semantic consistency (averaging 90%), indicating the generated videos effectively support multiple valid interpretations while retaining core event fidelity. The LLM confidence shows appropriate variation: Crash (0.95) signifies obvious evidence of a collision, Near-Crash (0.80) represents the inherent ambiguity of boundaries, and Safe (0.85) indicates cautious certainty amid the challenge of proving a negative situation.

All assessment methods agree on the defining elements: Crash scenarios feature clear causal factors and temporal progression; Near-Crash scenarios involve emergency evasive maneuvers; and Safe scenarios emphasize sustained, normal driving patterns without risk indicators. Our framework effectively produces scenarios that are quantitatively accurate, qualitatively interpretable (demonstrating 90% semantic consistency), and temporally precise (achieving 100% alignment with ground truth). These features fulfill the demanding criteria needed for thorough testing of Autonomous Driving Systems (ADS).

V. Conclusions

We introduce a flexible video generation framework designed to create safety-critical driving scenarios, such as crashes, near-crash, and safe situations. It employs LoRA-adapted FramePack with end frame conditioning and a multi-condition classifier-free guidance, allowing for accurate trajectory control while ensuring temporal coherence and physical realism.

Our evaluation shows strong results: description-based prompts achieve 31.086 CLIPScore and nearly 100% category accuracy, demonstrating the value of rich semantic conditioning. The framework generates diverse, controllable scenarios from identical initial frames, enabling systematic stress-testing of autonomous driving systems across varying risk levels.

Future plans involve enhancing the framework to simulate intricate multi-agent interactions along with dynamic weather changes. An important aspect is the integration of this framework with a robust World Model to forecast emergent, stochastic behaviors and facilitate real-time closed-loop testing with ADS decision-making systems. This integration with the World Model is crucial for creating genuinely adversarial and comprehensive safety scenarios.

References

[1] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of*

the 28th ACM International Conference on Multimedia, pages 2682–2690, 2020.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[4] Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, Lingxi Li, Yaobin Chen, Jitesh Panchal, and Ziran Wang. Personalized autonomous driving with large language models: Field experiments. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 20–27. IEEE, 2024.

[5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[6] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. *arXiv preprint arXiv:2210.06609*, 2022.

[7] Yuan Gao, Mattia Piccinini, Korbinian Moller, Amr Alanwar, and Johannes Betz. From words to collisions: Llm-guided evaluation and adversarial generation of safety-critical driving scenarios. *arXiv preprint arXiv:2502.02145*, 2025.

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

[9] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huanang Gao, Bu Jin, and Hao Zhao. Avd2: Accident video diffusion for accident video description. *arXiv preprint arXiv:2502.14801*, 2025.

[16] Haochen Li, Jonathan Leung, and Zhiqi Shen. Towards goal-oriented prompt engineering for large language models: A survey. *arXiv preprint arXiv:2401.14043*, 2024.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[18] Quanyi Li, Zhenghao Mark Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in neural information processing systems*, 36:3894–3920, 2023.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[20] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. Ieee, 2018.

[21] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics: Results of the 11th international conference*, pages 621–635. Springer, 2017.

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

[23] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023.

[24] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[25] Y Zhang, Y Wei, D Jiang, X Zhang, W Zuo, and Q Tian. Controlvideo: Training-free controllable text-to-video generation. arxiv 2023. *arXiv preprint arXiv:2305.13077*.

[26] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.