Carnegie Mellon University

Penn
University of Pennsylvania

The Ohio State University

Morgan State University

CCAC Community College of Allegheny County

The University of Texas Rio Grande Valley

Humanoid as Human Assessment for the Safety of Vulnerable Road Users

Ding Zhao

ORCID: 0000-0002-9400-8446

Final Report – 8/31/2026

DISCLAIMER

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Humanoid as Human Assessment for the Safety of Vulnerable Road Users	August 31, 2026
	6. Performing Organization Code
7. Author(s)	**8. Performing Organization Report No.**
Miao Li, Chengkai Lai, Ding Zhao	
9. Performing Organization Name and Address	**10. Work Unit No.**
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213	
	11. Contract or Grant No. Federal Grant # 69A3552344811
12. Sponsoring Agency Name and Address	**13. Type of Report and Period Covered** Final Report (July 1, 2025 – July 31, 2026)
US DOT	
	14. Sponsoring Agency Code USDOT

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract

This project builds a pipeline that bridges virtual and physical safety evaluation of autonomous driving under critical scenarios involving vulnerable road users (VRUs). We extend an esmini-based simulator to replay scenarios and control every road user, develop Crash Agent v2 to generate diverse, executable VRU-centric scenarios from NHTSA crash reports and road maps using fine-tuned vision-language and code-generating language models, introduce VRU behavior inpainting to turn trajectory-level scenarios into executable humanoid motions, and align the simulator with humanoid platforms so the same scenarios can be reproduced on a test ground. The result is a foundation for evaluating identical VRU-critical scenarios consistently in both virtual and physical settings.

17. Key Words	**18. Distribution Statement**
AI, humanoid robots, safety, vulnerable road users, evaluation	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161. Enter any other agency mandated distribution statements. Remove NTIS statement if it does not apply.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	8	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Problem

Autonomous driving has been widely deployed on vehicles and is increasingly participating in daily traffic in many regions. To ensure the safety of these systems, in addition to virtual evaluation, physical evaluation is essential because it reduces the sim-to-real gap and captures system performance in a faithful application environment. However, a fundamental limitation of physical evaluation is that critical scenarios are inherently rare and potentially dangerous to reproduce in real-world traffic, making systematic and repeatable evaluation under such conditions challenging. A test ground provides a controlled and safer environment for conducting physical evaluation of critical scenarios. To bridge the gap between the need for such physical evaluation and the scarcity of real-

world critical scenarios, this project aims to leverage the richness and controllability of synthetic critical scenarios and investigate how they can be migrated to a test ground for physical evaluation. In particular, we focus on critical scenarios involving vulnerable road users, which can particularly benefit from the test ground physical evaluation due to the inherent variability and unpredictability of real-world VRU behavior and the safety issue.

This project aims to build a bridge between the virtual evaluation of autonomous driving algorithms and physical evaluation using VRU-centric critical scenarios. We address this goal through a four-stage pipeline, covering scenario representation, scenario generation, behavior migration, and cross-environment evaluation. The first two stages focus on generating VRU-centric critical scenarios in simulation, while the latter two focus on transferring and evaluating these scenarios in a physical test ground. The primary objectives are:

1. Establish a simulation environment that supports both critical scenario replay and the representation of vulnerable road user behaviors. This is built upon the simulator developed in the previous Safety21 project.
2. Develop an agentic system for VRU-centric critical scenario generation based on real-world traffic accidents and infrastructure. The system will support the interpretation of traffic accident reports, understanding of road maps, and automatic generation of scripts describing both the critical scenario layout and vulnerable road user behaviors.
3. Introduce VRU behavior in-painting into the critical scenario generation process. This bridges the gap between the trajectory-level layout descriptions of critical scenarios and the action-level behavior required for physical deployment, enabling complete and executable VRU behaviors to be generated from high-level scenario descriptions.
4. Develop a toolkit for transferring generated VRU behaviors from simulation to humanoid platforms, enabling the generated scenarios to be physically reproduced by humanoids in a controlled test-ground environment. This ultimately enables the same critical scenarios to be reproduced and evaluated consistently in both virtual and physical environments, facilitating direct comparison between the two evaluation settings.

Methodology and Findings

For the first objective, we extend the simulation environment based on esmini. The extended environment supports not only high-throughput parallel replay of trajectories, but also semantic-level representation and execution of VRU actions. In addition to scenario replay, we enable observation collection for all road users, allowing them to be directly controlled through either generated scripts or policies tuned to produce human-like maneuvers and behaviors. This unified control framework also allows any road user to be replaced by or assigned to the system under evaluation, enabling flexible configuration of the critical scenario evaluation subject.

For the second objective, we implement Crash Agent v2, based on the scenario generation algorithm developed in the previous Safety21 project. In addition to leveraging the spatial reasoning capabilities of vision-language models (VLMs) and the instruction-following capabilities of large language models (LLMs), we introduce two additional modules to improve spatial information extraction and scenario generation quality. These modules address two key challenges: the spatial grounding accuracy of VLMs and the trade-off between generation flexibility and scenario validity. The first challenge can cause VLMs to misinterpret road layouts and the configurations of traffic accidents described in the National Highway Traffic Safety Administration (NHTSA) Crash Investigation Sampling System dataset [1]. The second challenge arises from the trade-off between constraining the generation space to ensure physically and semantically reasonable scenarios and allowing sufficient freedom to generate diverse scenarios. A highly constrained generation space can improve scenario validity but limit diversity, whereas greater generation freedom enables more diverse scenarios but can also result in a higher proportion of distorted or unreasonable scenarios.

To address the first challenge, we fine-tune the vision encoder to support graph-based spatial representation. We notice that the same road layout and traffic scenario can be presented with different image sizes, orientations, offsets, and symbols, resulting in diverse visual representations. Meanwhile, the same scenario can also be described using different natural language expressions. These variations make it difficult for the vision encoder to learn a consistent spatial representation. To mitigate this issue, we define a graph-based representation for road layouts and scenarios as a canonical representation of the same scenario. We fine-tune the vision encoder to extract this graph representation, enabling more stable and accurate spatial information extraction. The fine-tuned vision encoder is then integrated with the LLM backbone to generate the corresponding road network and scenario scripts from the extracted graph.

To address the second challenge, we investigate how to leverage the code generation capability of LLMs to move beyond the predefined scenario templates used in the previous

Safety21 project. This provides greater freedom for scenario generation but introduces a new challenge: the lack of sufficient training data for this task makes conventional imitation learning difficult. To address this, we investigate how to adapt CodeRL [2] to scenario script generation by leveraging our high-throughput simulation environment. Instead of relying on a large amount of labeled data, the LLM generates scenario scripts that are evaluated through simulation, and the resulting feedback is used for online reinforcement learning. We develop an iterative training strategy and reward-shaping method that enable the LLM to progressively improve its ability to generate valid and diverse scenario scripts.

For the third objective, we introduce VRU behavior inpainting to bridge the gap between trajectory-level scenario descriptions and the executable behaviors required for physical deployment. To reproduce VRU-centric critical scenarios in a physical test ground, the generated VRU behaviors need to be translated into executable motions for humanoid platforms. The proposed inpainting module addresses this requirement by first selecting a predefined movement template corresponding to each behavior type and then using a lightweight model to interpolate the motion between adjacent templates. This process generates smooth and continuous trajectories and behaviors, enabling the generated VRU behaviors to be both faithfully replayed in the simulator and transferred to humanoid platforms for physical evaluation.

For the last objective, we evaluate the simulation environment with humanoid platforms to finalize the trajectory and action-space configurations required for physical deployment. This establishes a consistent interface between the virtual and physical environments. It enables scenarios generated by Crash Agent v2 and humanoid trajectories generated by the VRU behavior inpainting module to be transferred from simulation to physical execution on the humanoid platform.

Conclusion

We developed a pipeline for bridging virtual and physical evaluation of VRU-centric critical scenarios. The extended simulation environment provides a unified interface for scenario replay and road user control. Crash Agent v2 enables the generation of diverse and executable critical scenarios from real-world accident reports and road infrastructure. VRU behavior inpainting bridges the gap between trajectory-level scenario descriptions and the executable motions required by humanoid platforms. The simulation and humanoid platforms are aligned through consistent trajectory and action-space configurations, enabling the generated scenarios to be transferred from simulation to

physical testing. Together, these components establish the foundation for evaluating the same VRU-centric critical scenarios consistently across virtual and physical environments.

References

[1] NHTSA crash dataset (https://crashviewer.nhtsa.dot.gov/)

[2] Le, Hung, et al. "Coderl: Mastering code generation through pretrained models and deep reinforcement learning." Advances in Neural Information Processing Systems 35 (2022): 21314-21328. (https://github.com/salesforce/CodeRL)