DPO-Enhanced DONOD Improves Robustness on Adversarial Safety Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the addition of DPO to DONOD affect performance on adversarial safety benchmarks like AdvBench and WildHacks compared to SFT-only models. Predicting the trajectories of surrounding objects is a critical task for self-driving vehicles and many other autonomous systems. Recent works demonstrate that adversarial attacks on trajectory prediction, where small crafted perturbations are introduced to history. 8 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Semi-supervised Semantics-guided Adversarial Training for Trajectory Prediction. Research question: How does the addition of DPO to DONOD affect performance on adversarial safety benchmarks like AdvBench and WildHacks compared to SFT-only models?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

8 papers retrieved. 8 claims extracted; 1 independently verified. Quality review score: 5.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The datasets contain more than 250k real driving scenarios in different cities, such as Miami and Pittsburgh.	×	0.03
For Argoverse 1 and Argoverse 2, each scenario consists of a road graph and multiple agents' trajectories sampled at a f	×	0.03
We choose 20 waypoints as the history trajectory and the models will predict 30 waypoints in the future for Argoverse da	×	0.04
Scenarios in Apolloscape have no maps but 6 waypoints for both history and future trajectories.	×	0.06
Both data-driven and heuristic data augmentation methods offer very limited improvement over the original model.	×	0.04
SSAT method reduces ADE attack error from 5.17 to 3.28 (Table p8 results).	×	0.04
Standard adversarial training reduces ADE attack error from 5.17 to 3.33 (Table p8 results).	×	0.06
The proposed method can significantly improve the system's robust generalization to unseen patterns of attacks.	✓	0.24

References

- http://arxiv.org/abs/2502.11455v1
- http://arxiv.org/abs/2205.14230v2
- http://arxiv.org/abs/2509.09055v1