Rationale-Augmented Preference Alignment Enhances Robustness in Embodied Multimodal Agents

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of rationale-augmented preference alignment on the robustness of embodied agents against adversarial perturbations in multimodal task execution. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Automated Driving Planner Robustness against Adversarial Influence. Research question: What is the impact of rationale-augmented preference alignment on the robustness of embodied agents against adversarial perturbations in multimodal task execution?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

11 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
PARAPET's simulation results reasonably approximate evaluation of a protection on a real vehicle against a real adversar	×	0.04
All simulations are performed using CARLA's Scenario Runner with a single camera placed on the front of the vehicle.	×	0.02
The priors to inform the protection were derived from comparing the detections of Faster-RCNN and Yolo3 on observations	×	0.02
For the simulation experiments, we used a bootstrapping set with 100 trajectories to initialize BO search.	×	0.02
Figure 2 plots protection scores against trajectories, demonstrating that the objective function distinguishes between t	×	0.02
When the disturbances are large, the sensor-fusion protection detects the disturbance more often as indicated by more sc	×	0.01
Both protections receive similar scores for a subset of disturbances.	×	0.03
After 150 iterations, BO with an expected improvement acquisition function fails to find a disturbance that induces inva	×	0.01
Each simulation took about one hour of computation, totaling over 10 days for both experiments.	×	0.01
Real world validation experiments were conducted using a vehicle outfitted with a Blackfly S GigE Machine Vision Camera,	×	0.03
Simulations were parameterized to match the state sequence for the test environment to search for effective disturbances	×	0.03
Work in AV security is broader in scope and typically includes other more traditional aspects of security shared with cy	×	0.03
Work in ML adversarial robustness has brought awareness of the susceptibility of ML algorithms to a wide range of attack	×	0.07
System evaluations need context of how ML algorithms are used and protected in specific applications.	×	0.06
AV planners do not implement protections inspired by algorithm defenses against adversaries.	×	0.06
The effectiveness of protections can only be adequately evaluated when the assumptions about threats and valid behavior	×	0.04

References

- http://arxiv.org/abs/2205.14697v1
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2504.14650v1