Adversarial Robustness of AdPO in LVLMs Across Perturbation Magnitudes

Assignee Research

June 11, 2026

Abstract

With the rapid advancement of large language models (LLMs), aligning policy models with human preferences has become increasingly critical. Direct Preference Optimization (DPO) has emerged as a promising approach for alignment, acting as an RL-free alternative to Reinforcement Learning from Human Feedback (RLHF). Despite DPO's various advancements and inherent limitations, an in-depth review of these aspects is currently lacking in the literature. In this work, we present a comprehensive review of the challenges and opportunities in DPO, covering theoretical analyses, variants, relevant prefer

1 Introduction

This paper examines: A Comprehensive Survey of Direct Preference Optimization: Datasets, Theories, Variants, and Applications. Research question: How does AdPO's adversarial robustness on LVLMs scale when evaluated against perturbation magnitudes beyond those used in training?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

3 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.7/10.

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
Direct Preference Optimization (DPO) has emerged as a promising approach for aligning policy models with human preferenc	✓	0.36
DPO is an RL-free alternative to Reinforcement Learning from Human Feedback (RLHF).	✓	0.26
An in-depth review of the advancements and inherent limitations of DPO is currently lacking in the literature.	✓	0.23
The paper presents a comprehensive review of the challenges and opportunities in DPO, covering theoretical analyses, var	✓	0.36
Recent studies on DPO are categorized based on key research questions to provide a thorough understanding of DPO's curre	✓	0.30
The paper proposes several future research directions to offer insights on model alignment for the research community.	✓	0.25
An updated collection of relevant papers can be found on https://github.com/Mr-Loevan/DPO-Survey.	✓	0.32

References

- https://doi.org/10.48550/arxiv.2504.17704
- https://doi.org/10.48550/arxiv.2410.15595
- https://doi.org/10.3390/instruments10010008