What is the impact of noise injection for differential privacy on the convergence stability of multimodal LLMs

Assignee Research

June 10, 2026

Abstract

Large language models (LLMs) can handle a wide variety of general tasks with simple prompts, without the need for task-specific training. Multimodal Large Language Models (MLLMs), built upon LLMs, have demonstrated impressive potential in tackling complex tasks involving visual, auditory, and textual data. However, critical issues related to truthfulness, safety, o1-like reasoning, and alignment with human preference remain insufficiently addressed. This gap has spurred the emergence of various alignment algorithms, each targeting different application scenarios and optimization goals. Recent

1 Introduction

This paper examines: Aligning Multimodal LLM with Human Preference: A Survey. Research question: What is the impact of noise injection for differential privacy on the convergence stability of multimodal LLMs during few-shot image-text alignment tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.1/10.

3 Results

12 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.1/10.

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
MME-RealWorld, MMStar, MMBench, MMT-Bench, BLINK, MathVista, SQA3D, MMMU, MVBench, Mantis-Instruct are benchmarks for ev	×	0.03
Object HalBench, VideoHallucer, VALOR-Eval, POPE, HaELM, OpenCHAIR, GAVIE, AMBER, Mementos, MMHal-Bench, VLind-Bench, M-	×	0.02
AdvDiffVLM, RTVLM, VLGuard, MultiTrust, VLLM-safety-bench, MOSSBench, MM-RLHF-SafetyBench are benchmarks for evaluating	×	0.03
Q-Bench, LLVisionQA, LLDescribe, LLaVA-Bench-Wilder, LiveBench, Vibe-Eval are benchmarks for evaluating conversation in	×	0.02
M-RewardBench, VL-RewardBench, Reward-Bench, MJ-Bench, MLLM-as-a-Judge, MM-RLHF-RewardBench are benchmarks for evaluating	×	0.02
Arena-Hard, AlpacaEval-V2, AlignBench, MM-AlignBench are benchmarks for evaluating alignment in multimodal models.	×	0.04
Fact-RLHF is the first multimodal RLHF algorithm, utilizing 10K human-labeled samples for the reward model and 50K hold-	×	0.03
DDPO assigns higher weights to corrected data in its loss function compared to standard DPO, using 1.4K manually refined	×	0.02
FDPO reuses InstructBLIP's existing data.	×	0.02
The creation of alignment datasets involves three core factors: data sources, model responses, and preference annotation	✓	0.20
Most alignment algorithms are designed for specific tasks such as addressing hallucinations, ensuring safety, and improv	×	0.12
This survey is the first to specifically focus on the alignment of MLLMs.	×	0.09

References

- http://arxiv.org/abs/2503.14504v2

- http://arxiv.org/abs/2210.09263v1
- http://arxiv.org/abs/2411.12259v1