Synthetic Data Evaluation Metrics and Downstream Performance in Multimodal Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do different evaluation metrics (e.g., CLIP similarity vs. human judgment) correlate with the downstream task performance of multimodal models trained on synthetic data, as measured by F1 scores. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Aligning Multimodal LLM with Human Preference: A Survey. Research question: How do different evaluation metrics (e.g., CLIP similarity vs. human judgment) correlate with the downstream task performance of multimodal models trained on synthetic data, as measured by F1 scores on benchmark tasks such as Visual Entailment or SNLI-VE?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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-	×	0.01
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, RewardBench, 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.	×	0.03
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.	×	0.03
DDPO uses 1.4K manually refined samples covering hallucination types such as objects (41.2%), positions (20.3%), numbers	×	0.03
FDPO reuses InstructBLIP's existing data.	×	0.04
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	×	0.11
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/2106.16020v1
- http://arxiv.org/abs/2603.03437v1