SOVEREIGN: What is the impact of imbalanced domain generalization techniques (e.g., SMoES routing) on the robustness of m

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have significantly pushed the frontier of egocentric video question answering (EgocentricQA). However, existing benchmarks and studies are mainly limited to common daily activities such as cooking and cleaning. In contrast, real-world deployment inevitably encounters domain shifts, where target domains differ substantially in both visual style and semantic content. To bridge this gap, we introduce EgoCross, a comprehensive benchmark designed to evaluate the cross-domain generalization of MLLMs in EgocentricQA. EgoCross covers four div

1 Introduction

Analysis of: EgoCross: Benchmarking Multimodal Large Language Models for Cross-Domain Egocentric Video Question Answering. Research goal: What is the impact of imbalanced domain generalization techniques (e.g., SMoES routing) on the robustness of multimodal language models when evaluated on cross-domain visual question answering benchmarks with varying modality distributions?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

14 papers retrieved. 6 claims extracted, 1 verified. Tribunal: 4.2/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
EgoCross is the first cross-domain benchmark for EgocentricQA, covering four distinct domains with ∼1k high-quality QA p	×	0.12
Egocentric-specific MLLMs struggle on EgoCross, with CloseQA accuracy below 55% (random chance: 25%) and OpenQA below 35	×	0.08
A notable performance drop (1.6×) on the same question types from EgoSchema to EgoCross confirms the challenge of cross	×	0.07
The EgoCross benchmark includes four domains: surgery, industry, extreme sports, and animal perspective.	✓	0.17
EgoCross contains approximately 957 QA pairs across 5 datasets with a total duration average of 22.5 seconds per video.	×	0.08
The best-performing MLLM achieves an average accuracy of 44.82% across all domains in the EgoCross benchmark.	×	0.03

References

- https://arxiv.org/abs/2508.10729
- https://www.semanticscholar.org/paper/c9719ea63ff65d897279e006ecb7708a8b7aa192
- https://arxiv.org/abs/2311.00807