MSR-VTT Cross-Modal Similarity Under Varying Relevant Moments per Query in Multimodal GMR Models

Assignee Research

June 12, 2026

Abstract

Video Moment Retrieval (VMR) aims to localize temporal segments in videos that correspond to a natural language query, but typically assumes only a single matching moment for each query. This assumption does not always hold in real-world scenarios, where queries may correspond to multiple or no moments. Thus, we formulate Generalized Moment Retrieval (GMR), a unified setting that requires retrieving the complete set of relevant moments or predicting an empty set. To enable systematic study of GMR, we introduce Soccer-GMR, a large-scale benchmark built on challenging soccer videos that reflect

1 Introduction

This paper examines: Retrieving Any Relevant Moments: Benchmark and Models for Generalized Moment Retrieval. Research question: What is the effect of varying the number of relevant moments per query on the alignment between textual queries and visual moments in multimodal GMR models, as measured by cross-modal similarity scores on MSR-VTT?.

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 7.2/10.

3 Results

11 papers retrieved. 22 claims extracted; 17 independently verified. Quality review score: 7.2/10.

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
Soccer-GMR is a large-scale Generalized Moment Retrieval (GMR) benchmark comprising 5.5K clips from 139 diverse matches.	✓	0.19
The Soccer-GMR benchmark contains 22.1K query-moment pairs.	×	0.14
The Soccer-GMR benchmark includes naturally occurring in-domain negatives of high semantic similarity.	✓	0.20
The Soccer-GMR benchmark was constructed via a duration-flexible semi-automated pipeline.	✓	0.20
The proposed evaluation protocol includes metrics for null-set rejection, single-moment retrieval, and multi-moment retr	✓	0.18
Conventional Video Moment Retrieval (VMR) measures lack metrics for null-set rejection and multi-moment retrieval.	✓	0.18
The GMR Adapter is a lightweight module compatible with mainstream VMR backbones.	✓	0.19
A GMR-tailored reward was designed for GRPO-based fine-tuning on Multimodal Large Language Models (MLLMs).	✓	0.19
Experiments show that the proposed methods (GMR Adapter and MLLM fine-tuning) outperform existing baselines.	✓	0.18
Early proposal-based VMR methods generate candidate segments via sliding windows or predefined anchors.	✓	0.23
Proposal-free VMR approaches regress boundaries directly from frame-level representations.	✓	0.17
Moment-DETR introduces learnable query slots with Hungarian matching for parallel moment prediction.	✓	0.21
Existing VMR methods lack an explicit mechanism for null-set rejection because their objectives produce no gradient sign	✓	0.22
Prevailing VMR datasets typically provide only a single corresponding moment per query, leaving multi-moment retrieval c	✓	0.24
Recent multimodal large language models applied to temporal grounding default to single-moment outputs.	✓	0.21
Recent multimodal large language models exhibit limited fine-grained temporal 4ocalization ability.	✓	0.20
Fang et al. formalize Open-Set VMR by treating video-irrelevant queries as an out-of-distribution detection problem via	✓	0.28
Discriminative VMR models share a common moment decoding stage that produces query-conditioned cross-modal representatio	×	0.06

References

- http://arxiv.org/abs/2503.14504v2
- http://arxiv.org/abs/2308.12898v2
- http://arxiv.org/abs/2605.02623v1