Pre-Training Dataset Diversity and Robustness in Lightweight Video-Language Adapters

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the correlation between pre-training dataset diversity and robustness scores on YouCook2-P for lightweight video-language adapters. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Prompting Video-Language Foundation Models with Domain-specific Fine-grained Heuristics for Video Question Answering. Research question: What is the correlation between pre-training dataset diversity and robustness scores on YouCook2-P for lightweight video-language adapters?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

13 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
NExT-QA [67] presents complex challenges requiring advanced causal and temporal reasoning. It comprises 5,440 videos ave	×	0.03
MSVD-QA [38] is derived from the MSVD dataset and includes 10s video clips across 1,970 videos. It features around 51K Q	×	0.08
MSRVTT-QA [38] is similar to MSVD-QA but on a larger scale, with 10K videos averaging 15 seconds each and containing 244	×	0.05
SUTD-TrafficQA [68] serves as a critical resource for advancing research in traffic-related question answering, requirin	×	0.03
The method builds on ALPRO [33] as the video-language foundation model, fine-tuned using 4 NVIDIA GeForce 3090 GPUs.	×	0.10
The AdamW optimizer was employed, with a weight decay of 0.001 and a unified learning rate of 5e-5 across datasets.	×	0.02
A linear decay schedule was used to adjust the learning rate dynamically, promoting rapid convergence initially and prec	×	0.03
For raw video processing, videos were rescaled to 224 $\times$ 224, and a random sparse sampling strategy was applied to extrac	×	0.03
The EAPrompter followed a distinct processing strategy, resizing videos to 256 $\times$ 256 before cropping a 224 $\times$ 224 area. A	×	0.03
For the Entity-Action Heuristic Generation, spaCy was used to extract the top 1,000 frequent verbs and nouns as action a	×	0.08
The training was conducted over 10 epochs for NExT-QA and SUTD-TrafficQA, and 15 epochs for MSVD-QA and MSRVTT-QA datase	×	0.02

References

- http://arxiv.org/abs/2410.09380v1
- http://arxiv.org/abs/2103.06561v6
- http://arxiv.org/abs/2407.20962v3