Multi-Loss Ensemble Effects on Video Representation Robustness Across Foundation Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the impact of varying the number of auxiliary losses combined by MELTR on the robustness of video representation learning, measured by generalization performance on out-of-distribution. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MELTR: Meta Loss Transformer for Learning to Fine-tune Video Foundation Models. Research question: What is the impact of varying the number of auxiliary losses combined by MELTR on the robustness of video representation learning, measured by generalization performance on out-of-distribution datasets like UCF-101 or Something-Something V2?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

9 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
MELTR is applied to UniVL, Violet, and All-in-one video foundation models.	✓	0.19
MELTR is evaluated on four downstream tasks: text-to-video retrieval, video question answering, video captioning, and mu	✓	0.21
Five benchmark datasets are used: YouCook2, MSRVTT, TGIF-QA, MSVD-QA, and CMU-MOSI.	×	0.02
For UniVL, five auxiliary loss functions (LJoint, LAlign, LCMLM, LCMFM, and LDecoder) are used.	×	0.08
Three advanced auxiliary loss functions (LM-Joint, LM-Align, and LM-Decoder) are additionally adopted.	×	0.07
For text-to-video retrieval and video captioning, LAlign and LDecoder are used.	×	0.11
UniVL + MELTR achieves R@1 of 33.7, R@5 of 63.1, R@10 of 74.8, and MedR of 3 on YouCook2.	×	0.02
UniVL + MELTR achieves R@1 of 28.5, R@5 of 55.5, R@10 of 67.6, and MedR of 4 on MSRVTT-7k.	×	0.02
UniVL + MELTR achieves R@1 of 31.1, R@5 of 55.7, R@10 of 68.3, and MedR of 4 on MSRVTT-9k.	×	0.02
Violet + MELTR achieves R@1 of 33.6, R@5 of 63.7, R@10 of 77.8, and MedR of 3 on MSRVTT-7k.	×	0.02
Violet + MELTR achieves Action of 95.4, Transition of 97.5, Frame of 63.4, and MSVD-QA of 51.7 on TGIF-QA.	×	0.02
UniVL + MELTR achieves BLEU-3 of 17.35, BLEU-4 of 11.98, METEOR of 18.19, ROUGE-L of 41.28, and CIDEr of 138 on video ca	×	0.03
UniVL + MELTR achieves BLEU-3 of 24.12, BLEU-4 of 17.92, METEOR of 22.56, ROUGE-L of 47.04, and CIDEr of 190 on video ca	×	0.03

References

- `http://arxiv.org/abs/2605.17165v1`

- http://arxiv.org/abs/2304.06427v2
- http://arxiv.org/abs/2303.13009v1