Masked Contrastive Pretraining for Efficient Multimodal Video-Text Alignment

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Does the combination of MoCo-style contrastive learning and JEPA masking improve sample efficiency during pretraining for multimodal video-text alignment tasks. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Masked Contrastive Pre-Training for Efficient Video-Text Retrieval. Research question: Does the combination of MoCo-style contrastive learning and JEPA masking improve sample efficiency during pretraining for multimodal video-text alignment tasks?.

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 5.2/10.

3 Results

15 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 5.2/10.

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
Masked contrastive pre-training is a better pretext task than masked prediction for video-text alignment.	✓	0.18
Multimodal alignment with masked modeling encourages the model to learn not only cross-modal alignment but also uni-moda	×	0.15
The proposed method outperforms existing works on several video-text retrieval tasks with fewer FLOPs and faster trainin	×	0.11
The proposed method achieves competitive results on image-text retrieval tasks, showing that MAC is flexible for various	✓	0.17
Early works on video-text alignment use unimodal pre-trained models, such as video action recognition and image classif	×	0.09
End-to-end video-language pre-training combining large-scale datasets and pretext tasks has shown great potential.	×	0.14
Cross-fusion modules such as masked language modeling (MLM), video text matching (VTM), frame order modeling, and masked	×	0.10
ClipBERT and Frozen propose sparse frame sampling to reduce temporal redundancy, enabling end-to-end training with raw v	×	0.14
End-to-end VidLP methods still process full-resolution frames for video spatio-temporal information extraction, which is	×	0.07
The proposed method performs mask sampling on video and text to achieve end-to-end video-text alignment.	✓	0.21
Masked language modeling (MLM) predicts masked tokens of the input text, showing great generality on various downstream	×	0.06
Visual input can be processed like language, making masked visual modeling (MVM) possible.	×	0.13
BEiT and follow-up works utilize dVAE to encode visual patches into discrete semantic tokens, which can be trained in a	×	0.05
MAE and follow-up works utilize the autoencoder to reconstruct the masked patches.	×	0.05
The proposed method achieves 79.3% R@1, 94.7% R@5, and 97.2% R@10 on image-text retrieval tasks using CC3M and WebVid2M	×	0.08
The proposed method achieves 38.9% R@1, 63.1% R@5, and 73.9% R@10 on video-text retrieval tasks with 180.7M parameters a	×	0.08

References

- http://arxiv.org/abs/2206.08262v1
- http://arxiv.org/abs/2210.16870v1
- http://arxiv.org/abs/2212.00986v2