# ViC-MAE Contrastive Alignment for Zero-Shot Video-Text Retrieval Accuracy on MSR-VTT

Assignee Research

June 14, 2026

**Abstract**

We present a simple yet effective end-to-end Video-language Pre-training (VidLP) framework, Masked Contrastive Video-language Pre-training (MAC), for video-text retrieval tasks. Our MAC aims to reduce video representation's spatial and temporal redundancy in the VidLP model by a mask sampling mechanism to improve pre-training efficiency. Comparing conventional temporal sparse sampling, we propose to randomly mask a high ratio of spatial regions and only feed visible regions into the encoder as sparse spatial sampling. Similarly, we adopt the mask sampling technique for text inputs for consisten

## 1 Introduction

This paper examines: Masked Contrastive Pre-Training for Efficient Video-Text Retrieval. Research question: How does ViC-MAE's contrastive alignment mechanism impact zero-shot video-text retrieval accuracy on the MSR-VTT benchmark compared to standard MAE pre-training?.

## 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 7.4/10.

## 3 Results

15 papers retrieved. 16 claims extracted; 12 independently verified. Quality review score: 7.4/10.

## 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. | ✓ | 0.25 |
| Multimodal alignment with masked modeling encourages the model to learn not only cross-modal alignment but also uni-moda | ✓ | 0.22 |
| The proposed method outperforms existing works on several video-text retrieval tasks with fewer FLOPs and faster trainin | ✓ | 0.20 |
| The proposed method achieves competitive results on image-text retrieval tasks, showing flexibility for various tasks. | × | 0.15 |
| Early works on video-text alignment use unimodal pre-trained models, such as video action recognition and image classif | ✓ | 0.25 |
| End-to-end video-language pre-training combining large-scale datasets and pretext tasks has shown great potential. | ✓ | 0.27 |
| Cross-fusion modules such as masked language modeling (MLM), video text matching (VTM), frame order modeling, and masked | ✓ | 0.23 |
| ClipBERT and Frozen propose sparse frame sampling to reduce temporal redundancy, enabling end-to-end training with raw v | ✓ | 0.25 |
| End-to-end VidLP methods still process full-resolution frames for video spatio-temporal information extraction, which is | ✓ | 0.26 |
| Masked modeling is one of the standard pretext tasks for pre-training. | ✓ | 0.19 |
| Masked language modeling (MLM) predicts masked tokens of the input text, showing great generality on various downstream | ✓ | 0.21 |
| Visual input can be processed like language, making masked visual modeling (MVM) possible. | ✓ | 0.20 |
| BEiT and follow-up works utilize dVAE to encode visual patches into discrete semantic tokens, which can be trained in a | ✓ | 0.29 |
| MAE and follow-up works utilize the autoencoder to reconstruct the masked patches. | × | 0.14 |
| 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 WebV4id2M | × | 0.08 |
| The proposed method has 180.7M parameters and 83.3G FLOPs, achieving 38.9% R@1, 63.1% R@5, and 73.9% R@10 on video-text | × | 0.06 |

## References


- http://arxiv.org/abs/2405.12710v3
- http://arxiv.org/abs/2212.00986v2
- http://arxiv.org/abs/2303.12001v3