Global-local contrastive consistency learning versus transformer attention alignment for DiDeMo retrieval performance

Assignee Research

June 14, 2026

Abstract

Text-video retrieval aims to find the most semantically similar videos with given text queries. However, since videos contain more diverse content than texts, the main semantics expressed by each text-video pair is often partially relevant. The primary methods involve the utilization of language-video attention module to align texts and videos. Though effective, this paradigm inevitably introduces prohibitive computational overhead, resulting in inefficient retrieval. In this paper, we propose a simple yet effective method called Global-Local Contrastive Consistency Learning (GLCCL) to achieve

1 Introduction

This paper examines: Text-Video Retrieval With Global-Local Contrastive Consistency Learning. Research question: How does the global-local contrastive consistency learning method compare to transformer-based attention alignment approaches in terms of retrieval accuracy and computational efficiency on the DiDeMo benchmark under varying video lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

8 papers retrieved. 17 claims extracted; 14 independently verified. Quality review score: 8.1/10.

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
GLCCL achieves 47.6 R@1 and 13.0 MnR on MSR-VTT, surpassing the baseline by +1.5% and +0.2% absolute improvements.	✓	0.24
GLCCL yields +3.4% and +0.7% improvements on R@1 compared with recent methods, i.e., CenterClip and X-Pool, respectively	✓	0.17
The base model is X-CLIP [7].	×	0.11
Experiments are conducted on 4 NVIDIA GeForce RTX 3090 GPUs using PyTorch.	✓	0.18
The text and video encoders are initialized with the parameters of CLIP (ViT-B/32).	✓	0.16
The Adam optimizer with a cosine learning rate schedule is adopted.	✓	0.16
The initial learning rate is set as 1e-7 for CLIP encoders and 1e-4 for others.	✓	0.21
The feature dimension is set as 512.	×	0.14
The model is trained on MSR-VTT, DiDeMo, and VATEX for 5, 20, and 5 epochs, respectively.	✓	0.15
The batch size is 128 for all datasets except DiDeMo (64).	✓	0.17
The word length and frame length are set as 32, 12 in MSR-VTT and VATEX while 64, 64 in DiDeMo.	✓	0.24
During training, the CSC loss weight η is set to 0.1 (in Eq. 19).	×	0.11
All videos are compressed to 3FPS (Frame Per Second) with width 224 or height 224.	✓	0.21
GLCCL outperforms existing methods on most of the evaluation metrics on MSR-VTT, DiDeMo, and VATEX.	✓	0.24
The proposed approach achieves comparable results across three public benchmarks of MSR-VTT [9], DiDeMo [10], and VATEX	✓	0.18
A parameter-free Global-Local Interaction Module (GLIM) is proposed to align text and video semantics with different gra	✓	0.19
An auxiliary Contrastive Score Consistency (CSC) loss is designed to promote consistency learning on positive pairs and	✓	0.28

References

- http://arxiv.org/abs/2405.12710v3
- http://arxiv.org/abs/2008.02531v2
- http://arxiv.org/abs/2605.17959v1