Training Convergence Speed of Global-Local Contrastive Consistency Versus Attention-Based Alignment for Multimodal Text-Video

Assignee Research

June 12, 2026

Abstract

Adapting large-scale image-text pre-training models, e.g., CLIP, to the video domain represents the current state-of-the-art for text-video retrieval. The primary approaches involve transferring text-video pairs to a common embedding space and leveraging cross-modal interactions on specific entities for semantic alignment. Though effective, these paradigms entail prohibitive computational costs, leading to inefficient retrieval. To address this, we propose a simple yet effective method, Global-Local Semantic Consistent Learning (GLSCL), which capitalizes on latent shared semantics across modal

1 Introduction

This paper examines: Text-Video Retrieval with Global-Local Semantic Consistent Learning. Research question: How does the training convergence speed of global-local contrastive consistency learning compare to attention-based alignment methods for multimodal text-video retrieval on the DiDeMo benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

12 papers retrieved. 11 claims extracted; 9 independently verified. Quality review score: 7.3/10.

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
Adapting large-scale image-text pre-training models like CLIP to the video domain represents the current state-of-the-ar	✓	0.31
Primary approaches for text-video retrieval involve transferring text-video pairs to a common embedding space and levera	✓	0.35
Existing paradigms for text-video retrieval entail prohibitive computational costs leading to inefficient retrieval.	✓	0.24
The proposed method is named Global-Local Semantic Consistent Learning (GLSCL).	✓	0.25
GLSCL introduces a parameter-free global interaction module to explore coarse-grained alignment.	✓	0.25
GLSCL devises a shared local interaction module that employs several learnable queries to capture latent semantic concep	✓	0.35
An Inter-Consistency Loss (ICL) is devised to accomplish concept alignment between the visual query and corresponding te	✓	0.30
An Intra-Diversity Loss (IDL) is developed to repulse the distribution within visual or textual queries to generate more	✓	0.30
Experiments were conducted on five benchmarks: MSR-VTT, MSVD, DiDeMo, LSMDC, and ActivityNet.	✓	0.18
The proposed method achieves performance comparable to State-of-the-Art (SOTA) methods.	×	0.13
The proposed method is nearly 220 times faster than existing methods in terms of retrieval speed.	×	0.14

References

- http://arxiv.org/abs/2405.12710v3
- http://arxiv.org/abs/2605.17959v1
- http://arxiv.org/abs/2005.12419v2