Scaling Masked Spatial-Temporal Tokens in Video-JEPA and MoCo v3 for Video Retrieval

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does scaling the number of masked spatial-temporal tokens affect the performance of Video-JEPA versus MoCo v3 on the VideoRetrieval benchmark. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Factorized Latent Dynamics for Video JEPA: An Empirical Study of Auxiliary Objectives. Research question: How does scaling the number of masked spatial-temporal tokens affect the performance of Video-JEPA versus MoCo v3 on the VideoRetrieval benchmark?.

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 4.2/10.

3 Results

15 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
Motion-Guided Masking improves all reported metrics in the UCF-101 setting (+0.30 pp D-48, +0.14 pp IN-100, +1.38 pp SSv	×	0.10
Kinematic variants degrade D-48 by 2.5–2.9 points while improving IN-100 by 1.5–1.7 points.	×	0.07
FWM-HW-LD achieves +5.92 percentage points on ImageNet-100 and +3.21 percentage points on SSv2 while remaining close to	✓	0.25
LD-JEPA achieves +5.02 pp on SSv2, the largest temporal reasoning gain in the table.	×	0.07
10 of 14 methods lose >5 points on ImageNet-100, and pixel-prediction objectives (AC-JEPA, FAC-JEPA) are particularly we	×	0.09
LD alone boosts SSv2 (+5.02) but hurts ImageNet and Diving-48.	×	0.08
FWM alone boosts ImageNet (+1.88) but hurts SSv2 and Diving-48.	×	0.08
FWM+LD without hard weighting performs poorly on ImageNet (-10.14).	×	0.07
The full FWM-HW-LD combination gives the most balanced result in this ablation.	×	0.08
The +40–45 point improvement confirms kinematic regularization encodes strong temporal structure.	×	0.03
The encoder produces a fixed 768-dimensional embedding that must simultaneously encode (1) what objects are present and	×	0.05
Auxiliary objectives that emphasize temporal structure often coincide with weaker appearance discrimination.	×	0.11

References

- http://arxiv.org/abs/2605.17165v1
- http://arxiv.org/abs/2407.05862v1
- http://arxiv.org/abs/2601.09524v1