Auxiliary Factorized Objectives in Video-JEPA for Few-Shot Fine-Grained Video Benchmarks

Assignee Research

June 12, 2026

Abstract

Joint-Embedding Predictive Architectures (JEPA) are a promising framework for self-supervised video representation learning, yet the behavior of auxiliary objectives in small-scale Video-JEPA training is not well characterized. We report a small-scale empirical study of 18 auxiliary objective variants for Video-JEPA across two pretraining regimes: single-dataset (UCF-101) and mixed-dataset (UCF-101 + Something-Something V2 + ImageNet-100). We evaluate frozen representations on three complementary benchmarks: Diving-48 (fine-grained motion), SomethingSomething V2 (temporal reasoning), and Image

1 Introduction

This paper examines: Factorized Latent Dynamics for Video JEPA: An Empirical Study of Auxiliary Objectives. Research question: Do auxiliary factorized objectives in Video-JEPA improve few-shot learning performance on fine-grained video benchmarks relative to non-factorized baselines?.

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 7.5/10.

3 Results

12 papers retrieved. 17 claims extracted; 13 independently verified. Quality review score: 7.5/10.

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 Diving-48 accuracy by 0.30 percentage points in the UCF-101 pretraining setting.	✓	0.18
Motion-Guided Masking improves ImageNet-100 accuracy by 0.14 percentage points in the UCF-101 pretraining setting.	✓	0.17
Motion-Guided Masking improves SSv2 accuracy by 1.38 percentage points in the UCF-101 pretraining setting.	×	0.15
Kinematic variants degrade Diving-48 accuracy by 2.5 to 2.9 points in the UCF-101 pretraining setting.	✓	0.15
Kinematic variants improve ImageNet-100 accuracy by 1.5 to 1.7 points in the UCF-101 pretraining setting.	×	0.13
FWM-HW-LD achieves a +5.92 percentage point gain on ImageNet-100 in mixed-dataset pretraining.	✓	0.19
FWM-HW-LD achieves a +3.21 percentage point gain on SSv2 in mixed-dataset pretraining.	✓	0.16
FWM-HW-LD results in a -0.30 percentage point change on Diving-48 relative to the reference baseline in mixed-dataset pr	✓	0.19
LD-JEPA achieves a +5.02 percentage point gain on SSv2 in mixed-dataset pretraining.	×	0.14
10 out of 14 tested methods lose more than 5 points on ImageNet-100 in the mixed-dataset setup.	✓	0.20
AC-JEPA and FAC-JEPA objectives result in performance losses of 13 to 16 percentage points on ImageNet-100.	✓	0.17
In the ablation study, LD alone boosts SSv2 by +5.02 points but hurts ImageNet and Diving-48.	✓	0.23
In the ablation study, FWM alone boosts ImageNet by +1.88 points but hurts SSv2 and Diving-48.	✓	0.23
FWM+LD without hard weighting performs poorly on ImageNet with a -10.14 point change.	✓	0.24
Synthetic Motion Discrimination shows a +40 to +45 point improvement with kinematic regularization.	✓	0.16
The encoder produces a fixed 768-dimensional embedding.	✓	0.45
Auxiliary objectives emphasizing temporal structure often coincide with weaker appearance discrimination in the experime	×	0.08

References

- http://arxiv.org/abs/2311.10873v2
- http://arxiv.org/abs/2007.06837v6
- http://arxiv.org/abs/2605.17165v1