Semantic Diversity in Pretraining Corpora for Cross-Domain Multimodal Embedding Generalization

Assignee Research

June 11, 2026

Abstract

Sensing human motions through Inertial Measurement Units (IMUs) embedded in personal devices has enabled significant applications in health and wellness. Labeled IMU data is scarce, however, unlabeled or weakly labeled IMU data can be used to model human motions. For video or text modalities, the "pretrain and adapt" approach utilizes large volumes of unlabeled or weakly labeled data to build a strong feature extractor, followed by adaptation to specific tasks using limited labeled data. However, pretraining methods are poorly understood for IMU data, and pipelines are rarely evaluated on out-

1 Introduction

This paper examines: PRIMUS: Pretraining IMU Encoders with Multimodal Self-Supervision. Research question: Does increasing the semantic diversity of pretraining corpora for embedding models improve cross-domain generalization on multimodal structured data tasks more effectively than scaling corpus size?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

9 papers retrieved. 17 claims extracted; 13 independently verified. Quality review score: 7.7/10.

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
The architecture of the IMU encoder consists of both 1D-CNN and GRU layers.	✓	0.20
The IMU encoder has two MLP heads during pretraining: one for multimodal loss and the other for unimodal loss.	×	0.12
After pre-training, only the output of the multimodal head is kept for training downstream tasks.	✓	0.20
The IMU encoder is trained with three objectives: self-supervision loss (LSS), multimodal loss (LMM), and nearest-neighb	✓	0.23
LSS ensures that the IMU encoder remains invariant to noise.	×	0.15
LMM pushes IMU representations towards aligned text and video representations.	✓	0.19
LNN uses the closest examples in representation space as positive pairs.	✓	0.21
The IMU encoder is a Stacked RNN consisting of convolutional, group normalization, and max-pooling layers, topped with a	✓	0.24
PRIMUS achieves a consistent performance improvement of up to 15% in test accuracy compared to state-of-the-art multimod	✓	0.24
The EgoExo4D dataset contains around 250K segments after pre-processing.	×	0.13
The EgoExo4D dataset includes IMU data from head-placed sensors, egocentric videos, and free-form text annotations.	✓	0.23
The test set activities for EgoExo4D include 8 activities: play music, cook, medical test, perform CPR, repair bike, cli	✓	0.27
The test set activities for Ego4D include 10 activities: play music, cook, eat, clean, carpenter, craft, farmer, househo	✓	0.26
The test set activities for REALWORLD include 8 activities: climbing up, climbing down, jumping, lying down, run, walk,	✓	0.19
The EgoExo4D dataset has a sample size of Train: 195K–Test: 53K.	✓	0.15
The Ego4D dataset has a sample size of Train: 555K–Test: 57K.	×	0.15
The REALWORLD dataset has a sample size of Train: 8.3K–Test: 2.6K.	✓	0.15

References

- http://arxiv.org/abs/2506.00969v1
- http://arxiv.org/abs/2411.15127v3
- http://arxiv.org/abs/2411.15497v3