CROSS-MODAL LEARNING FOR HUMAN MOTION UNDERSTANDING

Draft of March 4, 2026 at 20 : 24

BY

ABHI KAMBOJ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2026

Urbana, Illinois

Doctoral Committee:

Professor Minh N. Do, Chair
Professor David Forsyth
Professor Katie Driggs-Campbell
Professor Alexander Schwing
Professor Romit Roy Choudhury

ABSTRACT

Multimodal representation learning has achieved remarkable success in aligning static modalities like images and text, but its application to time-series sensor data remains challenging due to the temporal nature of the data, modality heterogeneity, and inefficient reasoning architectures. To systematically address these issues, this work introduces novel theoretical frameworks, alignment algorithms, and multi-agent systems designed to capture and transfer temporal dynamics across modalities. We propose (i) a comprehensive taxonomy of cross-modal transfer learning for sensor-based human activity recognition, identifying critical gaps and unifying disparate methods under early, middle, and late transfer stages; (ii) Cross-modal Temporal Alignment (XTA), a method that mitigates temporal collapse by aligning latent trajectories rather than single global vectors, significantly improving zero-shot transfer from video to IMU data; (iii) Pairwise Alignment (PA), an analytic formulation based on singular value decomposition that constructs a shared latent subspace, providing theoretical grounding and closed-form solutions for multimodal alignment; and (iv) Temporal Video Agents (TVA), a multi-agent framework that probes the temporal reasoning of Multimodal Large Language Models (MLLMs), uncovering that simulated tool use outperforms explicit execution for larger models, ultimately enhancing temporal reasoning efficiency and accuracy. Together, these contributions advance the theoretical understanding and practical deployment of temporal cross-modal systems, bridging the gap between high-level multimodal reasoning and low-level sensor intelligence.