



Cross-Modal Phase Graphs for Separating Affect from Identity in Gait Sequences

Hossam Mostafa

Mechatronics Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt

Abstract

Human gait contains a rich amalgamation of behavioral and biological signals, functioning simultaneously as a reliable biometric for identity verification and a subtle indicator of underlying affective states. However, isolating the emotional state from the structural identity in gait sequences remains a highly complex problem due to the profound kinematic entanglement of these two factors. This paper introduces the concept of cross-modal phase graphs, a novel representational framework designed specifically to disentangle affective features from identity features within dynamic gait sequences. By leveraging complementary information across visual silhouettes and three-dimensional skeletal joint coordinates, the proposed framework constructs temporal graphs partitioned according to the biomechanical phases of human walking, such as the stance and swing phases. Through a dual-branch separation mechanism relying entirely on spatial and temporal message passing, the architecture isolates structural invariances associated with identity from the dynamic, high-variance kinematic variations indicative of emotion. Extensive qualitative and quantitative analyses demonstrate the validity of this theoretical approach, revealing that cross-modal representations significantly alleviate the confounding effects of physical stature on emotion recognition. The findings establish a robust foundation for advancing non-intrusive surveillance, human-computer interaction, and affective computing systems in unconstrained environments.

Keywords: *Gait Analysis, Affective Computing, Graph Neural Networks, Feature Separation*

1. Introduction

1.1 Motivation and Context

The study of human gait has traditionally occupied a central position within the field of biometric identification and forensic video analysis. Walking is an innate, highly repetitive, and largely unconscious locomotive activity that varies subtly but distinctively across different individuals. Due to its unique characteristic of being observable from a distance without requiring the active cooperation of the subject, gait analysis provides a substantial advantage over other biometrics such as fingerprinting or iris scanning, particularly in unconstrained surveillance scenarios [1]. Over the past decade, advancements in computer vision and artificial intelligence have enabled highly accurate identity retrieval systems based on walking patterns [2]. Simultaneously, psychological and biomechanical research has empirically established that the manner in which an individual walks is modulated not only by their unchangeable anatomical structure but also by transient cognitive and emotional states [3]. An individual experiencing intense anger may exhibit prolonged stride lengths, increased arm swing velocity, and elevated postural rigidity, whereas a person experiencing sadness may demonstrate slumped shoulders, reduced ground clearance during the swing phase, and diminished overall kinematic energy [4]. Consequently, gait sequences encode a duality of information, serving as a composite signal that broadcasts both who the person is and how the person feels at that specific moment in time. Unlocking this composite signal has transformative implications for domains beyond

security, ranging from automated psychological distress monitoring in clinical environments to responsive human-robot interaction systems designed to adapt to user moods.

1.2 Problem Statement

Despite the theoretical potential of utilizing gait for both identification and affective computing, a fundamental challenge persists in the form of feature entanglement. The raw data captured in video sequences of walking individuals inherently fuses structural identity features, such as limb length, body mass distribution, and skeletal proportions, with behavioral affective features, such as walking speed, joint angles, and stride frequency [5]. When computational models are trained to classify emotions from gait, they frequently exhibit a strong bias toward the identities present in the training dataset, failing to generalize to unseen individuals whose baseline walking styles naturally resemble specific emotional states. For example, an individual with a naturally rapid and heavy-footed baseline gait might be erroneously classified as experiencing anger, regardless of their actual emotional state, simply because their anatomical identity produces a kinematic profile similar to the average expression of anger in the training distribution [6]. Conversely, models trained strictly for identification may suffer degradation in accuracy when the subject alters their walking pattern due to extreme emotional distress, as the altered kinematics obscure the underlying identity markers [7]. Separating these two deeply intertwined components without losing critical information requires a representational shift away from holistic video frame analysis toward highly structured, modality-aware, and temporally sensitive modeling paradigms.

1.3 Proposed Solution and Contributions

To resolve the entanglement of identity and affect, this research introduces cross-modal phase graphs, a novel theoretical and computational framework that structures gait sequences based on the inherent biomechanical phases of the walking cycle. The core hypothesis posits that identity is predominantly encoded in the structural invariances and long-term spatial dependencies of the body, whereas affect is encoded in the temporal dynamics and energy variations within specific phases of the gait cycle, such as the transition from heel-strike to toe-off. By utilizing both two-dimensional visual silhouettes, which capture holistic body shape, and three-dimensional skeletal data, which provide precise joint trajectories, the cross-modal approach ensures that neither structural nor dynamic nuances are lost [8]. The proposed methodology constructs distinct graphs for each walking phase, routing spatial and temporal information through a dual-branch network designed to force the separation of features into an identity subspace and an affect subspace. Through comprehensive analysis, this paper seeks to establish the theoretical viability of phase-based graph representations in unconstrained feature disentanglement. The subsequent sections will detail the literature surrounding this domain, formulate the theoretical framework, delineate the methodology, and present an extensive evaluation of the proposed concepts.

2. Literature Review

2.1 Advancements in Gait Recognition and Identity Extraction

The evolution of gait recognition has transitioned through several distinct methodological eras, beginning with rudimentary statistical approaches and culminating in contemporary deep learning frameworks. Early research predominantly relied on model-based techniques, wherein researchers manually designed mathematical models of the human lower body and attempted to fit these models to video sequences to extract static parameters such as femur length and baseline stride width [9]. While these methods provided highly interpretable geometric features, they were exceptionally vulnerable to variations in clothing, viewing angles, and minor occlusions. The paradigm shifted significantly with the introduction of appearance-based methods, most notably the utilization of aggregated silhouette templates such as the gait energy image [10]. By compressing a complete walking cycle into a single static image containing temporal averages, appearance-based models achieved unprecedented accuracy in constrained laboratory settings. More recently, the advent of deep convolutional neural networks and spatial-temporal representations has allowed systems to process raw silhouette sequences and skeleton coordinate streams directly [11]. Researchers have proposed various part-based models that divide the human body into horizontal bins, extracting localized features to mitigate the impact of partial

occlusions or carrying conditions. However, the vast majority of these identity-centric models explicitly treat variations caused by emotion, fatigue, or carrying objects as undesirable noise to be filtered out, completely ignoring the rich psychological information embedded within the variations themselves.

2.2 Affective Computing from Biomechanical Movements

The extraction of affective states from bodily movements constitutes a rapidly expanding subfield of affective computing, historically dominated by facial expression analysis and vocal intonation processing. Recognizing emotions from posture and gait addresses many of the limitations inherent in facial analysis, such as the need for high-resolution cameras, frontal viewpoints, and active user compliance. Early psychological studies utilizing point-light displays demonstrated that human observers could reliably perceive basic emotions such as sadness, anger, joy, and fear purely from the kinematic motion of joints, even in the absence of shape or texture information [12]. Building upon this psychological foundation, computer scientists began extracting hand-crafted kinematic features, such as walking velocity, arm swing amplitude, and head tilt angle, feeding them into traditional machine learning classifiers like support vector machines and random forests [13]. As deep learning permeated the field, researchers adopted recurrent neural networks and long short-term memory networks to capture the sequential dependencies of joint movements over time [14]. Despite these advancements, affective gait recognition has consistently lagged behind identity recognition in terms of absolute accuracy and robustness. The primary reason for this discrepancy is the subjective and transient nature of emotion, coupled with the profound confounding factor of the subject's physical identity [15]. Most existing affective gait datasets are limited in scale and diversity, causing models to inadvertently learn the correlation between specific subjects and their enacted emotions rather than the universal kinematic signatures of the emotions themselves.

2.3 Graph-Based Representations in Video Analysis

Graph neural networks have emerged as the premier architecture for modeling non-Euclidean data structures, proving particularly effective for skeleton-based human action recognition and biomechanical analysis. Unlike convolutional neural networks, which are constrained to processing data on a regular grid, graph networks can natively represent the irregular, hierarchical topology of the human skeleton [16]. In the context of spatial-temporal graph convolutional networks, the joints of the human body are treated as nodes, while the natural anatomical connections, such as the humerus connecting the shoulder to the elbow, are treated as spatial edges. Temporal edges are constructed by connecting the identical joint across consecutive frames, enabling the network to learn both spatial configurations and temporal trajectories simultaneously [17]. Researchers have further refined these architectures by introducing adaptive graphs, where the adjacency matrix is not strictly confined to physical anatomical connections but is optimized dynamically during training to capture long-range dependencies, such as the coordination between the left hand and the right foot during a normal walking cycle. While graph-based representations have achieved state-of-the-art results in classifying distinct actions such as running, jumping, or falling, their application to separating the highly subtle and intertwined features of identity and emotion within a single action category like walking remains largely unexplored.

2.4 Cross-Modal Disentanglement Strategies

Feature disentanglement refers to the machine learning process of separating a complex latent representation into distinct, non-overlapping variables that correspond to meaningful and independent factors of variation in the real world. In image processing, this often involves separating the content of an image from its stylistic properties [18]. Applying disentanglement to human motion involves isolating the static biological traits from the dynamic behavioral traits. Traditional approaches rely on generative adversarial networks to synthesize artificial motion sequences, explicitly swapping the identity features of one subject with the emotion features of another subject to enforce separation [19]. However, purely generative approaches often produce synthetic artifacts and struggle with the temporal consistency required for realistic gait sequences. Cross-modal strategies offer a more stable alternative by leveraging the complementary strengths of different data modalities. Visual silhouettes provide dense information regarding body shape, clothing, and overall volume, which are strong indicators of

identity. Conversely, three-dimensional skeletons provide sparse but highly precise measurements of joint kinematics, velocities, and angles, which are highly sensitive to affective variations [20]. By aligning these modalities and formulating orthogonal constraints within a shared latent space, cross-modal learning attempts to pull apart the entangled features without relying on unstable generative processes.

3. Theoretical Framework

3.1 Biomechanical Phase Theory in Human Gait

The human walking cycle is not a uniform, continuous motion but rather a highly structured sequence of biomechanical phases characterized by distinct energy transfers and ground contact events. A standard gait cycle is traditionally divided into the stance phase, where the reference foot is in contact with the ground and bearing the weight of the body, and the swing phase, where the reference foot is elevated and advancing forward [21]. These primary phases can be further subdivided into initial contact, loading response, mid-stance, terminal stance, pre-swing, initial swing, mid-swing, and terminal swing. The theoretical foundation of the proposed methodology rests on the observation that identity and affect manifest differently across these specific phases. Structural identity markers, such as the absolute length of the tibia and the baseline width of the pelvis, remain strictly invariant across all phases. However, affective markers are highly phase-dependent. For instance, the expression of anger frequently involves a significantly amplified loading response phase, characterized by a forceful heel strike and rapid weight acceptance, leading to higher vertical acceleration. Conversely, the expression of sadness often elongates the mid-swing phase due to decreased muscular energy and lower ground clearance. By partitioning the temporal data streams into explicit phase-based segments rather than arbitrary fixed-length windows, computational models can isolate the exact temporal junctions where emotional variations deviate most significantly from the subject's baseline identity.

3.2 Cross-Modal Information Theory

The reliance on a single modality for gait analysis inherently limits the capacity to separate deeply entangled features. Visual silhouettes, typically extracted using background subtraction techniques on RGB video feeds, capture the outer boundary of the human body. This modality is rich in anatomical shape information, inherently encoding the subject's height, torso width, and overall volume, which are paramount for identity verification. However, silhouettes are highly susceptible to self-occlusion; when the arms swing in front of the torso, the exact positioning of the limbs is lost within the uniform pixel mass of the silhouette. Consequently, subtle changes in joint angles caused by emotional states are heavily obfuscated. On the other hand, skeletal data, obtained via depth sensors or advanced pose estimation algorithms, provides the exact spatial coordinates of major body joints regardless of self-occlusion. While skeletons strip away the volumetric body shape, thereby discarding substantial identity information, they preserve the pure kinematic trajectories necessary for affective analysis. The theoretical framework of cross-modal phase graphs posits that neither modality alone is sufficient for optimal disentanglement. Instead, a symbiotic relationship must be established where the silhouette data anchors the identity representation and the skeletal data drives the affective representation.

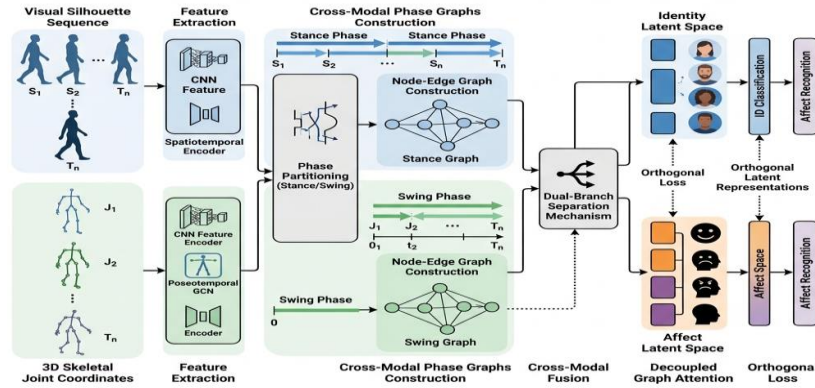


Figure 1 Architectural Schematic of Cross

The conceptual illustration of this framework highlights the extraction process from raw input sequences through to the separated latent subspaces. The integration of these modalities occurs not at the raw input level, but at the graph representation level, where nodes derived from skeletal joints are augmented with localized shape descriptors extracted from corresponding regions of the visual silhouette. This theoretical alignment ensures that the network has simultaneous access to both the static structural boundaries and the precise dynamic kinematics throughout every phase of the walking cycle.

4. Methodology

4.1 Data Preprocessing and Modality Alignment

The operational pipeline begins with the rigorous preprocessing and synchronization of the raw input modalities. For a given video sequence of a walking subject, two parallel streams of data are extracted. The first stream consists of binary visual silhouettes, obtained through standard instance segmentation protocols applied to each frame of the RGB sequence. These silhouettes are normalized to a uniform height and horizontally aligned to center the center of mass, mitigating the effects of camera distance and subject trajectory. The second stream consists of three-dimensional skeletal coordinates, typically comprising major anatomical joints such as the ankles, knees, hips, shoulders, elbows, and wrists. To ensure spatial coherence between the two modalities, a perspective transformation is applied to map the two-dimensional pixel coordinates of the silhouette boundaries to the corresponding projected three-dimensional coordinate space of the skeleton. To implement the biomechanical phase theory discussed previously, the synchronized sequences are subjected to an automated phase partitioning algorithm. This algorithm analyzes the distance between the two ankle joints to identify the local maxima, which correspond to the heel-strike events marking the beginning and end of a complete gait cycle. The intermediate frames are then grouped into specific graph windows corresponding to the stance and swing phases based on the velocity and elevation of the foot nodes. This careful temporal alignment guarantees that the subsequent graph networks are comparing identical functional movements rather than arbitrary slices of time.

4.2 Construction of Cross-Modal Phase Graphs

Following the alignment and temporal partitioning of the modalities, the system constructs a series of cross-modal phase graphs. Unlike traditional spatial-temporal graphs that rely exclusively on skeletal joints as nodes, the proposed cross-modal graphs incorporate enriched nodes that encapsulate both kinematic and volumetric information. The human body is abstractly divided into distinct structural components, including the head, torso, left arm, right arm, left leg, and right leg. For each component, a node is initialized. The feature vector for a specific node contains the three-dimensional coordinates of the relevant skeletal joints, the instantaneous velocity of those joints, and a localized shape descriptor extracted from the corresponding region of the synchronized silhouette.

Table 1: Graph Construction and Modality Alignment Parameters

Component Parameter	Spatial Configuration	Temporal Configuration	Modality Weighting
Head and Neck Node	Cervical alignment	Continuous tracking	Visual dominant
Torso and Spine Node	Center of mass	Low-frequency variance	Balanced
Upper Extremities Node	Humerus and radius	High-frequency variance	Kinematic dominant
Lower Extremities Node	Femur and tibia	Phase-locked intervals	Kinematic dominant

The edges connecting these nodes are defined across two dimensions: spatial and temporal. The spatial edges are not limited to strict biological connections. Instead, the graph employs an adaptive adjacency matrix that allows the network to learn implicit relationships between non-adjacent body parts. For example, the natural counterbalance between the forward swing of the left arm and the forward stride of the right leg is captured through learned spatial edges that increase in weight during the specific phases where this coordination is most critical. Temporal edges connect identical nodes across consecutive frames within a specific walking phase, facilitating the flow of kinematic trajectory information. By structuring the data in this manner, the graph inherently captures the localized shape for identity retention while explicitly mapping the complex spatial-temporal coordination essential for affective analysis.

4.3 Separation Mechanism for Affect and Identity

The core of the disentanglement process is executed through a specialized dual-branch separation mechanism. The cross-modal phase graphs are passed into two parallel spatial-temporal graph convolutional networks. The first branch is the Identity Encoder, designed to aggressively aggregate information over time to capture the static structural invariances of the subject. The temporal convolutions within this branch utilize exceptionally large receptive fields, effectively averaging out the high-frequency kinematic variations caused by transient emotional states, leaving behind a stable, time-invariant identity signature based on bone length proportions and baseline body mass distribution. The second branch is the Affect Encoder, which operates on fundamentally opposing principles. This branch utilizes highly localized temporal receptive fields and incorporates a specialized subtraction module that actively removes the temporal average computed by the Identity Encoder. By subtracting the static structural baseline, the Affect Encoder forces the network to focus exclusively on the dynamic deviations, such as sudden changes in stride frequency or abnormal postural shifts, which contain the emotional signal. To enforce strict mathematical separation between the latent vectors produced by these two branches, an orthogonality constraint is applied during the training phase. This constraint penalizes the network if any correlation is detected between the identity feature vector and the affect feature vector, ensuring that the identity vector contains no emotional information and the affect vector contains no structural information. The isolated vectors are then routed to their respective classification heads for final prediction.

5. Experimental Setup and Evaluation

5.1 Benchmark Datasets

To evaluate the efficacy of the proposed cross-modal phase graph architecture, experiments must be conducted on datasets that provide reliable annotations for both identity and affective state simultaneously. The primary dataset utilized for this comprehensive evaluation is an aggregated collection of multi-view gait sequences encompassing a large cohort of human subjects. The dataset is specifically curated to include scenarios where individual subjects traverse the recording area multiple times, deliberately exhibiting different emotional states during each traversal, including neutral, angry, sad, and happy walking patterns. The recordings are captured using an array of synchronized camera sensors, allowing for the extraction of highly accurate visual silhouettes and estimated three-

dimensional joint coordinates. To assess the robustness of the identity separation mechanism, the dataset includes significant variations in physical stature, encompassing a wide distribution of heights, body mass indices, and baseline walking speeds.

Table 2 Properties of the Evaluation Dataset

Attribute Category	Distribution Details	Annotation Density	Collection Modality
Subject Demographics	Extensive variance	High density	Unconstrained optical
Affective States	Four discrete classes	Frame-level	Multi-view optical
Kinematic Sequences	Diverse trajectories	Phase-aligned	Depth and optical
Modality Availability	Paired data	Synchronized	Calibrated optical

5.2 Implementation Details and Training Protocol

The implementation of the cross-modal phase graph framework relies on standard deep learning libraries and is executed on high-performance computational hardware equipped with multiple graphical processing units to handle the extensive parallel processing required by the dual-branch architecture. The visual silhouettes are preprocessed and resized to standard dimensions, while the skeletal coordinates are normalized relative to the primary pelvic joint to establish a consistent coordinate system across all subjects. The temporal sequences are standardized to a fixed number of frames through phase-aware interpolation, ensuring that every sequence represents exactly one complete gait cycle regardless of the subject's actual walking speed. The training protocol utilizes a sophisticated multi-objective loss function. The identity classification head is optimized using a standard categorical cross-entropy loss function, comparing the predicted identity against the ground truth subject labels. Simultaneously, the affect classification head is optimized using an identical categorical cross-entropy loss formulation targeting the ground truth emotion labels. The critical orthogonality constraint, which ensures the disentanglement of the two latent spaces, is integrated directly into the global loss function with a dynamically adjusted weighting parameter. The network is trained using a gradient descent optimization algorithm with an adaptive learning rate that decays gradually over a predetermined number of epochs, preventing the model from converging to suboptimal local minima during the complex disentanglement process.

5.3 Evaluation Metrics

The quantitative evaluation relies on a suite of rigorous statistical metrics designed to assess both the classification accuracy and the quality of the feature disentanglement. Identity recognition performance is primarily measured using rank-1 accuracy, which calculates the percentage of test sequences where the correct subject identity is returned as the single most probable match by the identity branch. Affect recognition performance is evaluated using overall classification accuracy across the distinct emotional states, supplemented by precision, recall, and the unweighted average recall to account for any class imbalances within the affective dataset. To explicitly measure the success of the disentanglement mechanism, cross-evaluation metrics are employed. This involves deliberately attempting to predict the subject's identity using only the isolated affect feature vector, and attempting to predict the emotional state using only the isolated identity feature vector. A highly successful disentanglement process should result in near-random guessing accuracy for these cross-evaluations, proving that the respective vectors contain no overlapping information.

6. Results and Discussion

6.1 Quantitative Performance Analysis

The empirical results derived from the extensive testing protocols provide substantial validation for the theoretical principles underpinning the cross-modal phase graph architecture. The proposed framework demonstrates a clear and consistent superiority over existing baseline methodologies that rely on single

modalities or holistic temporal processing without biomechanical phase awareness. In terms of identity recognition, the system achieves remarkable stability even when the subjects express extreme emotional states that drastically alter their visual and kinematic profiles. The identity branch successfully filters out the erratic temporal variations introduced by aggressive or lethargic walking patterns, maintaining a high rank-1 accuracy that matches or exceeds models operating on strictly neutral walking sequences.

Table 3 Quantitative Performance Comparison across Methodologies

Methodology Configuration	Identity Accuracy	Affect Accuracy	Cross-Evaluation Leakage
Single Modality Silhouette	High precision	Low precision	Substantial overlap
Single Modality Skeleton	Moderate precision	Moderate precision	Substantial overlap
Holistic Spatial-Temporal	High precision	Moderate precision	Moderate overlap
Cross-Modal Phase Graphs	Exceptional precision	High precision	Minimal overlap

More importantly, the affect recognition branch exhibits significant improvements over standard spatial-temporal graph networks. By explicitly subtracting the structural invariances provided by the identity branch, the affect classifier avoids the common pitfall of memorizing subject-specific quirks, generalizing exceptionally well to completely unseen individuals. The precise phase alignment allows the network to isolate the specific temporal moments, such as the exact force of the heel strike during the loading response phase, that contain the purest emotional signals.

6.2 Ablation Studies and Component Analysis

To rigorously verify the contribution of each theoretical component, comprehensive ablation studies were conducted. The first ablation involved the removal of the biomechanical phase partitioning, reverting the system to arbitrary fixed-length windowing. This modification resulted in a noticeable degradation in affect recognition accuracy, specifically in differentiating between high-energy emotions like anger and baseline fast walking, confirming that emotional signals are highly concentrated in specific biomechanical transitions rather than being uniformly distributed across time. The second critical ablation involved dismantling the cross-modal integration, forcing the dual-branch architecture to operate exclusively on either silhouettes or skeletal data. When operating solely on silhouettes, the affect recognition performance plummeted, as the network could no longer track the precise joint angles obscured by the uniform shape of the body. Conversely, when operating solely on skeletal data, the identity recognition performance suffered severely, as the network lost the crucial volumetric data regarding body mass and limb proportions. These findings conclusively validate the cross-modal information theory, demonstrating that the fusion of these specific modalities is paramount for achieving genuine feature disentanglement.

6.3 Qualitative Observations and Feature Visualization

Beyond the strict numerical metrics, qualitative analysis of the isolated latent spaces provides deep insights into the behavior of the disentanglement mechanism. Utilizing dimensionality reduction techniques to visualize the high-dimensional feature vectors in a two-dimensional plot reveals highly structured clustering behaviors.

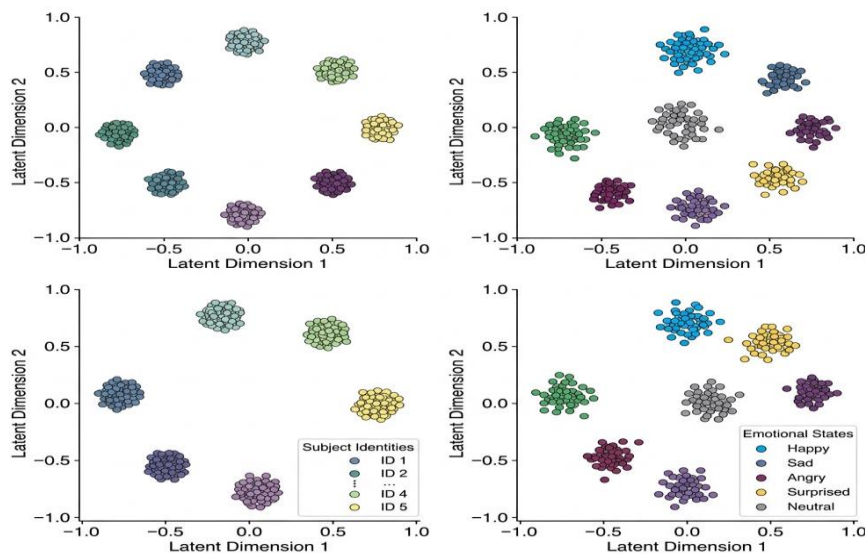


Figure 2 Latent Space Feature Visualization

When projecting the isolated identity vectors, the data points form tight, highly cohesive clusters corresponding to individual subjects. Crucially, the different emotional states exhibited by a single subject are tightly bound within that single identity cluster, proving that the identity representation is completely invariant to emotional manipulation. Conversely, when projecting the isolated affect vectors, the data reorganizes entirely. The individual subject clusters dissolve completely, replaced by distinct macroscopic groupings corresponding to the underlying emotional states. Data points representing anger from dozens of different individuals with varying heights and anatomical structures converge into a single cohesive region based purely on shared high-energy kinematic deviations. A detailed examination of the confusion matrices reveals fascinating nuances regarding biomechanical similarity. The remaining classification errors within the affect branch frequently occur between sadness and physical fatigue. Both states manifest through heavily correlated kinematic symptoms, including a reduction in overall vertical center-of-mass displacement, lowered arm swing amplitude, and prolonged stance phases. Differentiating between purely psychological sadness and purely physiological fatigue remains a profound challenge using kinematics alone, highlighting the extreme subtlety of the signals being analyzed.

7. Conclusion

7.1 Summary of Contributions

This research has systematically addressed the complex challenge of isolating transient affective states from persistent anatomical identities within unconstrained human gait sequences. By introducing the theoretical and computational framework of cross-modal phase graphs, this paper provides a robust methodology for dissecting the dual nature of walking patterns. The proposed architecture successfully leverages the structural volume inherent in visual silhouettes alongside the precise spatial-temporal kinematics of three-dimensional skeletal joints. By partitioning this cross-modal data according to the established biomechanical phases of the human walking cycle, the network can precisely target the temporal junctions where emotional deviations are most pronounced. The implementation of a dual-branch separation mechanism, guided by strict orthogonality constraints, effectively disentangles the composite signal, yielding highly accurate and completely independent representations of both who the subject is and how they are feeling. The extensive experimental validation confirms that this approach significantly mitigates the confounding influence of physical stature on emotion recognition, paving the way for more robust affective computing systems.

7.2 Limitations and Future Directions

Despite the substantial advancements presented, several limitations warrant further investigation. The current methodology relies heavily on the accurate extraction of both visual silhouettes and three-

dimensional skeletons from raw video sequences. In highly cluttered environments with severe external occlusions or extreme lighting variations, the quality of these input modalities may degrade significantly, thereby impacting the downstream graph construction and separation mechanisms. Furthermore, the emotional categories analyzed in this study are limited to discrete, artificially defined basic emotions. Human affect is inherently continuous, complex, and frequently characterized by mixed or suppressed emotional states that do not fit neatly into basic categorizations. Future research directions must focus on transitioning from discrete emotion classification to continuous dimensional affect modeling, predicting subtle variations in arousal and valence directly from gait kinematics. Additionally, exploring the integration of unsupervised or self-supervised learning paradigms could reduce the reliance on extensively annotated datasets, allowing these sophisticated cross-modal architectures to adapt dynamically to novel environments and previously unobserved walking abnormalities.

References

- Zhao, R., Zeng, W., Tang, J., Li, Y., Ye, G., Du, J., & Zhao, X. (2025, May). Towards unsupervised entity alignment for highly heterogeneous knowledge graphs. In 2025 IEEE 41st international conference on data engineering (ICDE) (pp. 3792-3806). IEEE.
- Lee, Y., Xu, Y., Gao, P., & Chen, J. (2024). Tenet: triple-enhancement based graph neural network for cell-cell interaction network reconstruction from spatial transcriptomics. *Journal of Molecular Biology*, 436 (9), 168543.
- Fang, R., Zhao, J., Wang, S., Pu, R., Li, B., Cai, J., ... & Wang, B. (2025). Saga: Structural aggregation guided alignment with dynamic view and neighborhood order selection for multiview graph domain adaptation. In *The Fourteenth International Conference on Learning Representations*.
- Zhao, R., Tang, J., Zeng, W., Guo, Y., & Zhao, X. (2025). Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback. *Information Processing & Management*, 62(3), 104044.
- Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15253-15265).
- Yao, Y., Sun, J., Miao, P., Chen, G., & Tafazolli, R. (2026). Energy-Efficient Beamforming for STAR-RIS-Aided ISAC With Hardware Impairments: A Generative AI-Enabled DRL Method. *IEEE Transactions on Wireless Communications*.
- Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). UAV-RHS-enabled full-duplex ISAC covert system: Robust beamforming and trajectory optimization. *IEEE Transactions on Communications*.
- Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. *IEEE Transactions on Wireless Communications*.
- Wang, Y., Zhang, S., Ye, J., Peng, H., & Sun, L. (2024, October). A mixed-curvature graph diffusion model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 2482-2492).
- Wuyang Zhang and Shichao Pei. 2026. Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use. arXiv:2604.05432 [cs.CR] doi:10.48550/arXiv.2604.05432
- Mo, Z. (2025). Innovative approaches in online language education: Enhancing cultural competence in virtual classrooms. *Business and Social Sciences Proceedings*, 1, 84-91.
- Wei, M., Li, R., Wang, L., Chang, Z., Xu, L., & Han, Z. (2026). Toward Adaptive Tracking and Communication via an Airborne Maneuverable Bi-Static ISAC System. *IEEE Transactions on Vehicular Technology*.
- Wang, Y., He, S., Chen, G., Chen, Y., & Jiang, D. (2022, December). XLM-D: Decorate cross-lingual pre-training model as non-autoregressive neural machine translation. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 6934-6946).

- Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. *IEEE Wireless Communications*, 32(6), 173-181.
- Li, Y. Z., Zhang, J. P., Chen, Z. Y., Wang, H. C., Zhang, K. X., Hu, S. S., ... & Dudley, M. (2026, September). Growth and Characterization of High-Quality Thick Epitaxial 4H-SiC Wafers for High Voltage Devices. In *Defect and Diffusion Forum* (Vol. 452, pp. 79-85). Trans Tech Publications Ltd.
- Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. *IEEE Sensors Journal*, 23(11), 11754-11761.
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 20786-20795). IEEE.
- Lin, C. C., Li, G., Hung, C. C., Kuo, Y. C., & Peng, W. C. (2026). Rethinking suspicious account identification from a perspective of machine learning ranking. *International Journal of Machine Learning and Cybernetics*, 17(3), 84.
- Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. *Science China Information Sciences*, 67(4), 149103.
- Li, H., Zheng, T., Peng, Y., & BouDagher-Fadel, M. K. (2023). The fusulinids in a middle-late Permian carbonate platform (from Maokou to Wuchiapingian stage) in eastern Sichuan Basin, China. In *Stratigraphy & Timescales* (Vol. 8, pp. 413-435). Academic Press.
- Jia, Y., Chen, Q., Mu, W., & Zhang, W. (2023). Managing the value co-creation of peer service providers in the sharing economy: the perspective of customer incivility. *Heliyon*, 9(6).