IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3277-3287
Journal homepage: http://iieta.org/journals/ts

Modeling Internal Nonverbal Communication Patterns and Predicting Team Performance via Video-Based Behavioral Representation Learning

Juanjuan Mao[1*], Hanwen Huang[2]

[1] School of Business, Hunan International Economics University, Changsha 410205, China
[2] School of Information Engineering, Hunan Industry Polytechnic, Changsha 410208, China

Corresponding Author Email: sw_mjj@163.com

https://doi.org/10.18280/ts.420619

Received: 22 May 2025
Revised: 9 November 2025
Accepted: 18 November 2025
Available online: 31 December 2025

Keywords:
nonverbal communication, team performance prediction, video representation learning, multimodal multi-granularity CNNs, cross-member attention, latent variable modeling, computational behavioral science

ABSTRACT

A persistent challenge in organizational research is the objective quantification of team collaboration states. Nonverbal communication remains insufficiently captured by existing measurement frameworks, which are limited by single-granularity representations, inadequate modeling of cross-member interactions, and so on. To address these limitations, an end-to-end, multi-task, multimodal, multi-granularity Convolutional Neural Network (CNN) was developed, in which nonverbal communication patterns are represented as learnable latent variables embedded within intermediate network layers. Joint optimization is performed through a primary task of team performance prediction and an auxiliary self-supervised task of nonverbal behavior reconstruction. This design enables the accurate extraction of individual- and team-level skeletal features together with frame- and clip-level Red-Green-Blue (RGB) representations, while modeling interaction dependencies among team members. Experimental results demonstrate that the proposed framework consistently outperforms representative baseline methods, achieving lower prediction errors on a self-constructed ONC dataset and exhibiting robust generalization across public benchmarks. Empirical analyses indicate that higher nonverbal synchrony is positively associated with team performance, that more egalitarian attention distribution benefits creative task outcomes, and that dynamic changes in synchrony during task sprint phases provide stronger predictive power than static descriptors. Ablation studies further confirm that the synergistic integration of multi-granularity fusion, cross-member attention, and latent variable decoding is critical to performance gains. Analyses of SHapley Additive exPlanations (SHAP) values highlight the superior representational power of automatically learned latent variables over traditional handcrafted features. The proposed approach establishes a quantitative paradigm for nonverbal communication analysis, extends the application boundary of video representation learning to multi-agent team interaction scenarios, and offers a practical tool for team collaboration diagnosis and performance forecasting.

1. INTRODUCTION

The core value of team collaboration is derived from implicit interactions among members. Nonverbal communication, as a fundamental component of interpersonal exchange, encompasses behaviors such as bodily synchrony, gaze coordination, and facial expression transmission [1-3], and has been identified as a critical determinant of coordination efficiency, emotional resonance, and ultimate team performance. Foundational research in organizational behavior has long established the importance of such implicit interactions; however, traditional investigations have predominantly relied on questionnaire surveys and manual observation [4, 5]. These approaches are constrained by strong subjectivity, coarse analytical granularity, and an inability to capture dynamic interaction processes in real time [6-8]. As a result, the objective quantification of team nonverbal communication patterns and the establishment of their relationship with performance have remained long-standing unresolved challenges in the field.

From a technological perspective, substantial advances have been achieved in video-based behavioral representation learning within the computer vision community. Models such as Inflated 3D ConvNets (I3D) and Graph Convolutional Networks (GCNs) have enabled increasingly accurate recognition of individual-level actions [9, 10]. Nevertheless, three fundamental bottlenecks persist in team collaboration scenarios. First, the insufficient capture of multi-granularity hierarchical features limits the ability to represent team structures spanning individuals, interactions, and collective dynamics. Second, the absence of explicit modeling of cross-member dynamic interaction dependencies prevents effective characterization of coordinated behavioral coupling among team members. Third, objectives have remained largely confined to action classification, without being tightly coupled to performance prediction tasks central to organizational management research.

Recent interdisciplinary efforts integrating computational

behavioral science with organizational behavior theory [11, 12] have provided a promising pathway for addressing these limitations. The construction of end-to-end quantitative modeling frameworks has been shown to not only mitigate methodological constraints inherent in traditional management research but also to extend the application boundary of video representation learning, enabling technology-driven empowerment of management practice and playing a crucial role in promoting the synergistic development of both fields [13]. Despite these advances, notable limitations remain at both the technical and theoretical levels. From a technical standpoint, existing approaches exhibit modality and granularity designs that are misaligned with the hierarchical structure of teams. Attention mechanisms have predominantly focused on individual actors while neglecting cross-member interaction dependencies. Moreover, the construction of communication patterns has relied heavily on expert-defined rules, which not only introduce subjective bias but also limit generalization across scenarios. The decoupling of communication modeling from performance prediction has further resulted in fragmented optimization pipelines [14-16]. From a theoretical standpoint, quantitative evidence supporting the intrinsic mechanisms linking nonverbal communication to team performance remains limited. The differential effects of communication patterns across task types have not been clearly established, and prior studies have largely emphasized static descriptors while overlooking the dynamic evolution of nonverbal interaction patterns and their temporal relationship with performance outcomes [17-19].

The central objectives of this study are threefold. First, an end-to-end multimodal video-based representation learning model is designed to enable the automatic decoding of latent variables characterizing team nonverbal communication. Second, testable research hypotheses grounded in organizational behavior theory are formulated to precisely quantify the associations between communication patterns and team performance. Third, a team performance prediction framework with both strong generalization capability and interpretability is constructed to provide effective analytical tools for organizational management practice. In alignment with these objectives, three core hypotheses were proposed:

H1: Team nonverbal synchrony is significantly and positively correlated with task performance.

H2: In creative tasks, an egalitarian cross-member attention distribution pattern is more predictive of high performance than a centralized attention pattern.

H3: The rate of increase in communication synchrony during task sprint phases exhibits greater predictive power for performance outcomes than static synchrony measures.

This study introduces four key innovations. First, it proposes an end-to-end framework that combines latent variables with multi-task learning, allowing communication patterns to be learned directly from data, rather than relying on expert-defined features. Second, it develops an interpretable cross-member attention mechanism and multi-level feature fusion strategies to model interpersonal interactions and complementary features. Third, it creates a closed research loop that links theory and methodology—organizational behavior hypotheses shape the model and experiments, and are later tested using computational methods. Fourth, it uses SHAP analysis and pattern pathway mining together to provide practical insights for improving management practices.

The remainder of this study is organized as follows. Section 2 describes the model's architecture and the overall research framework. Section 3 presents the experimental results used to evaluate the model and test the hypotheses. Finally, the key findings are summarized, and suggestions for future research are provided.

2. RESEARCH METHODOLOGY

2.1 Overview of the research framework

An end-to-end, multi-task, multimodal, multi-granularity CNN framework is developed with the objective of integrating multimodal video data, accurately decoding latent nonverbal communication patterns, and achieving efficient team performance prediction. The overall architecture comprises four core components: multi-granularity feature extraction, latent variable decoding of nonverbal communication, dual-task output, and multi-task optimization. The central processing pipeline is structured below. Multimodal video data, consisting of RGB video and skeletal data are provided as inputs. Individual- and team-level skeletal features, together with frame- and clip-level RGB features, are extracted through the proposed network architecture. Latent variables representing core nonverbal communication patterns are subsequently decoded at intermediate layers. Finally, a dual-output layer is employed to simultaneously generate team performance predictions and reconstructed nonverbal behavior representations. To strengthen the intrinsic associations between features and the ultimate objective of performance prediction, a multi-task joint optimization strategy is adopted. The primary task corresponds to team performance prediction, with an associated loss function denoted as L_{main}, while the auxiliary task corresponds to self-supervised reconstruction of nonverbal behaviors, with an associated loss function denoted as L_{aux}. The overall loss function is defined as:

$$L=\lambda L_{main}+(1-\lambda)L_{aux} \tag{1}$$

where, λ is set to 0.7 and determined via cross-validation to balance the relative priorities of the primary and auxiliary tasks. This ensures that the latent variables are capable of capturing essential nonverbal communication information while effectively serving the team performance prediction task.

2.2 Dataset construction and preprocessing

Well-established collaborative task paradigms from organizational behavior research were adopted to ensure the authenticity of the collected nonverbal communication data. The task set encompassed creative decision-making scenarios (the winter survival task), problem-solving scenarios (the moon survival task), and execution-oriented scenarios involving real corporate project collaboration. Data collection was conducted across both laboratory-simulated collaboration settings and authentic corporate meeting room environments. A total of 62 teams were recruited, with team sizes ranging from 3 to 8 members. Team compositions spanned three major industry sectors—technology, education, and finance—yielding a cumulative effective data duration of 128 hours. During data acquisition, multimodal information was

synchronously recorded, including RGB video at a resolution of 1080p and 30 frames per second (fps), three-dimensional skeletal data comprising 25 keypoints extracted via MediaPipe, and auxiliary audio recordings retained as backup signals. To preserve the ecological validity of the dataset, task designs were closely aligned with real-world organizational workflows. Standardized project requirement documents and realistic time constraints were incorporated, and team compositions were structured to reflect typical corporate role distributions, including leaders, executors, and coordinators. The data collection process adhered to a minimal-intervention principle; concealed camera setups were employed to mitigate behavioral distortion and to ensure the naturalistic expression of nonverbal communication.

A multidimensional performance labeling system was constructed based on the input-process-output framework, and a triangulation approach was employed to ensure label reliability. The labeling scheme comprised three core dimensions. The task performance dimension was defined using objective indicators, including task completion rate, decision accuracy, and execution efficiency. The team vitality dimension integrated objective and subjective measures, encompassing the variance in member speaking frequency and the mean level of emotional positivity derived from facial expression recognition. The member satisfaction dimension was obtained through subjective assessments, including post-task peer evaluations of collaboration quality and communication fluency, as well as self-reported satisfaction scores provided by participants. Labels across all dimensions were integrated using a weighted averaging method to produce a composite performance score normalized to the [0,1] interval. Weight assignments were determined by organizational behavior experts according to the relative contribution of each dimension to overall team performance. The annotation process was jointly conducted by three organizational behavior experts and two technical engineers. Inter-annotator reliability was evaluated using Krippendorff's α, with all coefficients exceeding 0.88, thereby satisfying established reliability standards for empirical research.

During skeletal data preprocessing, median filtering was applied for noise reduction, and missing frames were completed via linear interpolation. Individual skeletal sequences and team-level spatial position matrices were subsequently constructed to meet model input requirements. For RGB video preprocessing, frame sampling at 15 fps was first performed to balance data volume and computational efficiency. Face and body regions were detected using the You Only Look Once version 8 (YOLOv8) model, followed by region-of-interest cropping to remove background interference and to focus on core interaction areas. To enhance model generalization and to accommodate auxiliary task training, targeted data augmentation strategies were implemented, including random frame flipping, temporal sequence shuffling, and adjustments to brightness and contrast. All preprocessing procedures were designed to preserve the integrity and authenticity of the original nonverbal communication characteristics.

2.3 Detailed design of the proposed model

2.3.1 Multi-granularity feature extraction module

The primary objective of the multi-granularity feature extraction module is to accommodate the hierarchical structure of team nonverbal communication by capturing core representations of individual-team-level features and frame- and clip-level features from both skeletal and RGB modalities, providing fine-grained representations for subsequent interaction modeling and team performance prediction. For the skeletal branch, a dual-granularity design is adopted. The individual skeletal keypoint subnetwork is constructed based on Spatial-Temporal Graph Temporal Convolutional Networks (S-GTCNs). The input consists of temporal sequences with length T, each containing 25 three-dimensional skeletal keypoints. Through the combined application of temporal convolutions and graph convolutions, dynamic body motion patterns of individual members are extracted. The output is an individual-level feature matrix F_{ind} of dimensionality $T \times d$, where the feature dimensionality d is set to 256 to balance representational capacity and computational efficiency.

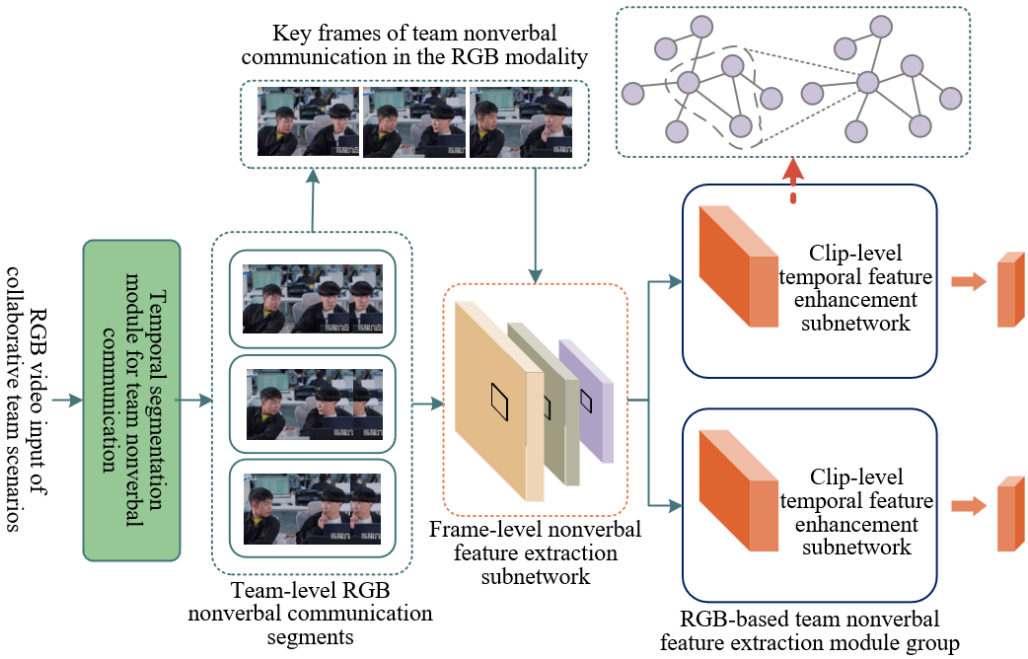

Figure 1. Temporal feature extraction module for team nonverbal communication in the RGB modality

The team interaction skeletal subnetwork is designed to capture interaction-related features among team members. A graph structure is constructed based on pairwise spatial distances between members, in which nodes correspond to individual team members and edge weights are defined as the inverse of inter-member spatial distances, such that closer proximity corresponds to stronger interaction association. Feature aggregation is performed over this graph using a GCN, enabling effective capturing of team-level interaction patterns, including bodily coordination and spatial orientation consistency. The resulting team-level interaction representation is denoted as F_{team}, matching the dimensionality ($T \times d$) of the individual-level features and achieving hierarchical coverage of skeletal representations from individual actions to collective team interactions.

The RGB branch likewise adopts a dual-granularity design to jointly capture static visual attributes and temporal dynamics. The frame-level feature subnetwork is constructed based on the I3D architecture. The input consists of T frames of three-channel RGB images. Through three-dimensional convolution operations, frame-level visual features are extracted, encompassing static nonverbal cues such as facial expressions and body postures. The resulting output is a frame-level feature representation F_{frame} of dimensionality $T \times d$. The clip-level temporal subnetwork is implemented using a three-layer Temporal Convolutional Network (TCN), with the dilation factor set to 2 to expand the receptive field. The frame-level feature sequence is provided as input, and temporal convolutions are applied to capture behavioral evolution across consecutive frames, thereby extracting temporal dynamics of nonverbal communication. The final output is a clip-level temporal feature representation F_{seq} of dimensionality $T \times d$, completing multi-granularity feature extraction in the RGB modality from static frames to dynamic temporal segments. Figure 1 provides an overview of the RGB-based team nonverbal feature extraction module. RGB video streams from team interactions are processed through three steps: dividing into temporal segments, extracting frame-level nonverbal features, and applying clip-level temporal enhancement. This produces multi-level representations of team nonverbal communication in the RGB modality.

2.3.2 Cross-member attention mechanism

The primary function of the cross-member attention mechanism is to quantify the dynamic association strength of nonverbal behaviors among team members, thereby amplifying interaction features relevant to performance while suppressing interference from behaviorally irrelevant individuals. To achieve this objective, preliminary feature fusion is first performed for each member. Specifically, individual skeletal features $F_{ind\_i}{}^t$ and frame-level visual features $F_{frame\_i}{}^t$ are concatenated along the feature dimension, yielding a composite feature vector f_i^t for member i at time step t. The resulting feature dimensionality is two-dimensional, preserving essential individual behavioral information while providing a comprehensive feature basis for computing interaction dependencies among team members.

The computation of inter-member attention weights is formulated using a scaled dot-product attention mechanism. The attention weight α_{ij}^t, representing the influence of member i on member j at time t, is defined as:

$$\alpha_{ij}^t = softmax\left(\frac{f_i^t \cdot W_a \cdot (f_i^t)^T}{\sqrt{2d}} + b_a\right) \tag{2}$$

where, W_a denotes a learnable weight matrix of dimensionality $2d \times 2d$ used to model feature associations between members, b_a represents a bias term, and $(2d)^{1/2}$ serves as a scaling factor to mitigate gradient vanishing issues induced by increasing feature dimensionality. The softmax function normalizes weights to the [0,1] interval, ensuring that their sum equals unity and enabling quantification of relative interaction strengths.

To balance the contributions of individual features and interaction-driven features, residual connections are incorporated into the cross-member feature aggregation process. The equation is as follows:

$$f_{att,i}^t = \sum_{j=1}^{M} \alpha_{ij}^t \cdot f_j^t + \gamma \cdot f_i^t \tag{3}$$

where, M denotes the number of team members, and γ is set to 0.5 as the residual weighting coefficient. This design ensures that critical individual behavioral characteristics are retained while inter-member interaction dependencies are effectively emphasized. Through the dynamic learning of attention weights, the proposed mechanism adaptively focuses on task-critical interaction behaviors, such as gestural coordination during creative discussions and gaze convergence during decision-making phases. As a result, more targeted and interaction-aware representations are produced for subsequent feature fusion.

2.3.3 Multi-granularity feature fusion

The primary objective of multi-granularity feature fusion is to integrate four complementary feature dimensions across skeletal and RGB modalities, thereby generating a comprehensive representation that preserves hierarchical completeness while emphasizing interaction relevance. This integrated representation serves as a high-quality input for subsequent decoding of latent nonverbal communication variables. An adaptive weighted fusion strategy is adopted in place of fixed-weight fusion, enabling the model to automatically learn dynamic feature weights that align with the requirements of team performance prediction. Weight allocation is designed to be positively correlated with each feature's predictive contribution to performance outcomes.

The fusion process is formulated as follows:

$$F_{\text{fusion}}^t = w_1 \cdot F_{\text{ind}}^t + w_2 \cdot F_{\text{team}}^t + w_3 \cdot F_{\text{frame}}^t + w_4 \cdot F_{\text{seq}}^t \tag{4}$$

where, w_1, w_2, w_3, and w_4 denote the dynamic weights associated with individual skeletal features, team interaction features, frame-level visual features, and clip-level temporal features, respectively. Each weight is learned via a sigmoid activation function according to:

$$w_k = \text{sigmoid}\left(W_w \cdot F_k^t + b_w\right) \tag{5}$$

where, W_w represents a learnable weight matrix of dimensionality $d \times d$ that captures the relationship between each feature type and the performance prediction objective, while b_w denotes a bias term. The sigmoid function constrains weight values to the [0,1] interval, allowing each feature's contribution to be quantitatively interpreted while enabling adaptive normalization across feature dimensions.

A key advantage of this fusion strategy lies in its ability to

dynamically adjust feature importance across task types and collaboration phases. For example, during brainstorming stages of creative tasks, higher weights are automatically assigned to clip-level temporal features and cross-member interaction features, whereas during execution-oriented stages, increased emphasis is placed on individual skeletal features and frame-level visual features. Through this adaptive mechanism, the fused representation $F_{fusion}{}^{t}$ is guided to focus on context-specific nonverbal communication cues, effectively leveraging the complementary strengths of multimodal and multi-granularity features. This design establishes a robust foundation for latent variable decoding and subsequent team performance prediction.

2.3.4 Latent variable decoding of nonverbal communication

The latent variable decoding module for nonverbal communication serves as the central intermediate layer connecting multi-granularity fused features to the team performance prediction task. Its primary objective is to automatically distill low-dimensional latent variables that characterize team nonverbal communication patterns from high-dimensional fused representations, thereby avoiding subjective bias introduced by expert-defined constructs and enabling data-driven quantification of communication patterns. The dimensionality of the latent variables directly affects representational capacity and interpretability. Based on extensive cross-validation experiments, the latent dimensionality k is set to 64, a configuration that sufficiently captures essential communication information while mitigating overfitting and preserving interpretability.

Latent variable decoding is implemented using two fully connected layers. The latent representation at time step t, denoted as Z^t, is computed as:

$$Z^t = \mathrm{ReLU}\left(W_z \cdot F_{\mathrm{fusion}}^t + b_z\right) \tag{6}$$

where, W_z represents a learnable weight matrix of dimensionality $d \times k$, and b_z denotes a k-dimensional bias term. The ReLU activation function introduces nonlinearity to enhance the model's capacity to represent complex communication patterns. The resulting latent variable matrix Z consists of $T \times k$ dimensions that correspond to core nonverbal communication factors automatically learned by the model, such as bodily synchrony, balance of attention allocation, and emotional positivity. These dimensions are not predefined and are entirely induced from data, allowing adaptive capture of task-specific key communication characteristics across diverse team collaboration scenarios.

To promote independence and interpretability across latent dimensions and to prevent information redundancy, an L2 regularization constraint is incorporated during the decoding process. The regularization term is defined as $\lambda_z||Z||_2^2$, where λ_z is set to $1e^{-5}$. By penalizing the L2 norm of the latent variables, this constraint encourages the learning of sparse and relatively independent communication dimensions, thereby clarifying the semantic meaning of each latent factor, supporting subsequent interpretability analyses and further enhancing model generalization capability.

2.3.5 Multi-task output head design

The multi-task output head is designed under a joint optimization framework comprising a primary task and an auxiliary task. The primary task focuses on team performance prediction, while the auxiliary task constrains latent variable quality through self-supervised reconstruction of nonverbal behaviors. The collaborative optimization of these tasks is intended to jointly enhance prediction accuracy and feature representation capability.

The primary task, corresponding to the team performance prediction head, receives two types of inputs: the temporal latent variable features Z and team attribute features A. Team attribute features include the number of members, task type, and role composition. After one-hot encoding and normalization, the attribute feature dimensionality is denoted as m. These attributes are jointly fed with the latent variable features to complement critical non-behavioral information. To capture the dynamic evolution of communication patterns across task phases, a bidirectional Long Short-Term Memory (BiLSTM) layer is employed for temporal modeling. Through the combined operation of forward and backward LSTM units, temporal dependencies are effectively extracted. The resulting temporally fused feature representation is denoted as H of dimensionality $T \times 2k$ and is computed as:

$$H = BiLSTM(Z) \tag{7}$$

To emphasize the differential contribution of critical task phases to overall performance, temporal attention-weighted pooling is applied to H. Temporal attention weights β_t are obtained via softmax normalization:

$$\beta_t = \mathrm{softmax}\left(W_\beta H^t + b_\beta\right) \tag{8}$$

where, W_β denotes a learnable weight matrix of dimensionality $2k \times 1$, and b_β represents a bias term. The features obtained after weighted pooling are passed through a fully connected layer followed by a sigmoid activation function to produce a composite team performance score P in the range [0,1].

$$P = \mathrm{sigmoid}\left(W_p \cdot \left(\sum_{\mathrm{t}=1}^{\mathrm{T}} \beta_{\mathrm{t}} H^t\right) + b_p\right) \tag{9}$$

The loss function for the primary task adopts the mean squared error (MSE) formulation, which is well suited for continuous-valued performance labels. The primary task loss is defined as:

$$L_{\mathrm{main}} = \frac{1}{N}\sum_{i=1}^{N}\left\|P_i - Y_i\right\|^2 \tag{10}$$

where, N denotes the number of samples, and Y_i represents the ground-truth performance label.

The auxiliary task corresponds to a self-supervised nonverbal behavior reconstruction head, whose primary objective is to constrain the latent variable representation Z to preserve essential communication information by reconstructing original nonverbal behavior features. This design enhances the robustness and effectiveness of feature representations. The reconstruction targets include individual skeletal features F_{ind} and frame-level RGB features F_{frame}, as these features directly reflect fundamental nonverbal behavioral cues. The reconstruction network is composed of two transposed convolutional layers followed by a fully connected layer. Transposed convolutions are employed to restore spatial feature structures, while the fully connected layer is used to ensure dimensional alignment. The network

outputs reconstructed features $\hat{\boldsymbol{F}}_{ind}$ and $\hat{\boldsymbol{F}}_{\text{frame}}$. To improve robustness to outliers, the auxiliary task employs an L1 loss function, defined as:

$$L_{\text{aux}} = \frac{1}{N \cdot T}\left(\left\|\mathbf{F}_{ind} - \mathbf{F}_{ind}\right\|_1 + \left\|\mathbf{F}_{\text{frame}} - \mathbf{F}_{\text{frame}}\right\|_1\right) \quad (11)$$

The overall loss function of the model is defined as a weighted sum of the primary task loss and the auxiliary task loss, and is computed as follows:

$$L_{total}=\lambda \cdot L_{main}+(1-\lambda)\cdot L_{aux} \quad (12)$$

where, λ denotes the balancing coefficient between the two tasks. The value is determined via grid search and is set to 0.7. This configuration preserves the central objective of team performance prediction while enabling the auxiliary task to effectively regularize latent variable quality, thereby achieving coordinated optimization of both tasks.

2.4 Model training and optimization details

Model training was conducted using the AdamW optimizer to achieve efficient convergence while incorporating parameter regularization. The weight decay coefficient was set to $1e^{-5}$, effectively mitigating overfitting by suppressing excessive parameter growth. The initial learning rate was configured as $1e^{-4}$, and a cosine annealing scheduling strategy was employed. Under this strategy, the learning rate was decayed to one-tenth of its current value every ten training epochs, thereby preserving exploratory capacity during early training stages while enabling gradual stabilization and convergence in later phases. During training, the batch size was set to 8, balancing GPU memory constraints with gradient estimation stability. The total number of training epochs was fixed at 100, supplemented by an early stopping mechanism. Training was terminated when the mean absolute error (MAE) on the validation set exhibited no improvement for 15 consecutive epochs, thereby preventing redundant iterations and reducing the risk of overfitting. Parameter initialization was performed using the Xavier uniform distribution scheme to ensure consistent variance across layer inputs and outputs, while all bias terms were initialized to zero to provide a stable starting point for training. With respect to regularization, in addition to the L2 constraint applied within the latent variable decoding module, dropout layers with a probability of 0.3 were introduced in fully connected layers and feature fusion modules. By randomly deactivating a subset of neurons during training, model generalization capability was further enhanced, ensuring robust performance across diverse datasets and scenarios. An overview of the proposed framework for team nonverbal communication decoding and performance prediction is illustrated in Figure 2.

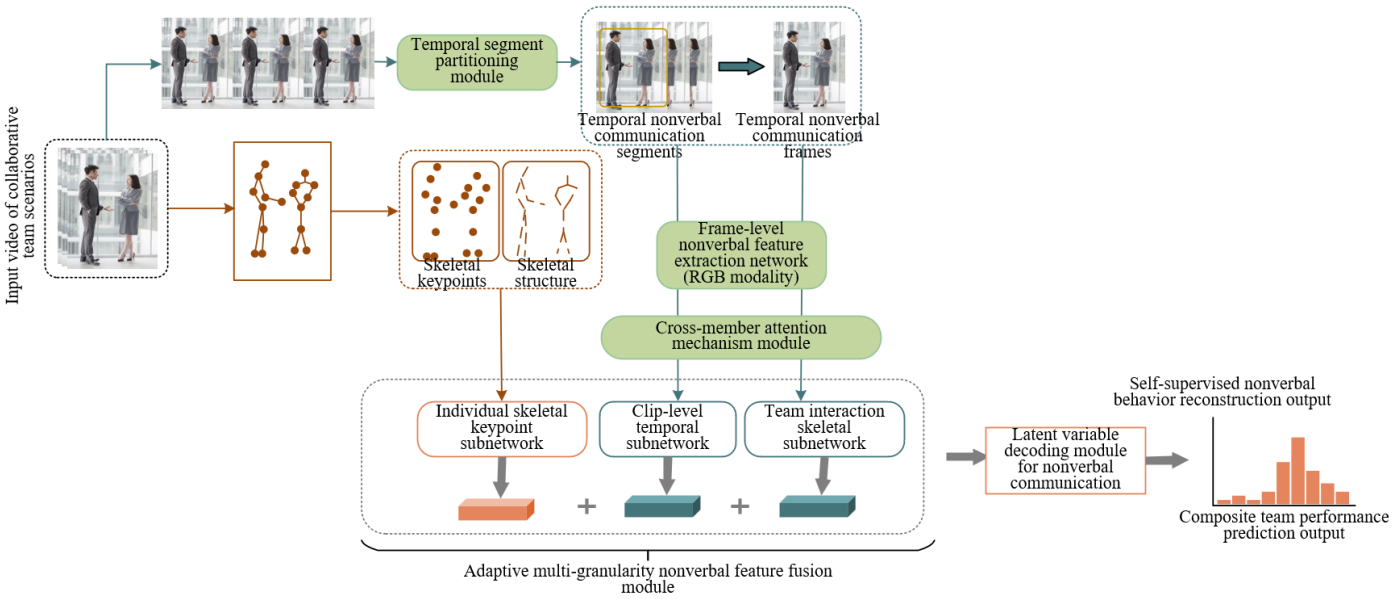

Figure 2. Schematic overview of the proposed framework for team nonverbal communication decoding and performance prediction

3. EXPERIMENTS

3.1 Experimental design and evaluation metrics

The experimental objectives are organized around four dimensions: demonstrating the superior performance of the proposed model in team performance prediction; testing the validity of the three core research hypotheses; determining the necessity of key components, including multi-granularity feature extraction, cross-member attention, and latent variable decoding; and comprehensively evaluating model interpretability, temporal dynamics capture capability, and cross-scenario generalization. Two datasets were employed for experimental validation. The self-constructed ONC dataset comprised 62 team samples and was partitioned into training, validation, and test sets using a 42:10:10 split. The public MPII Group Interaction dataset, consisting of 30 team samples, was used exclusively for generalization evaluation. Dataset partitioning was conducted strictly at the team level to prevent data from the same team appearing across training, validation, and test sets, thereby guaranteeing fair evaluation and robust generalization performance.

Evaluation metrics were designed in accordance with specific validation objectives. Team performance prediction

was formulated as a regression task and was assessed using MAE, root mean squared error (RMSE), and the coefficient of determination (R^2) to comprehensively measure prediction accuracy and goodness of fit. Latent variable interpretability was evaluated by computing correlation coefficients between automatically learned latent dimensions and manually annotated communication dimensions, ensuring that learned representations possess clear behavioral semantics. The three core hypotheses were statistically tested using Pearson correlation coefficients—applied to examine associations between nonverbal synchrony, synchrony growth rates, and performance—and independent-sample t-tests—applied to evaluate performance differences across attention distribution patterns. Model robustness was assessed by analyzing performance variability across different team sizes and task types. Baseline models were selected to comprehensively represent diverse technical paradigms. These include classical dual-stream I3D models from video-based behavioral representation learning, the state-of-the-art temporal Video Swin Transformer, and TeamGCN designed for team interaction modeling; GroupViT and Multi-Agent Transformer (MAT) from multi-agent interaction modeling; and Extreme Gradient Boosting (XGBoost) with handcrafted features and Multilayer Perceptron (MLP) models incorporating team attributes from the performance prediction domain. Comparative evaluations against these baselines were conducted to highlight the advantages of the proposed approach in multimodal fusion, interaction modeling, and end-to-end optimization.

3.2 Results of core hypothesis testing

Figure 3 illustrates the association between team nonverbal synchrony scores and composite performance scores, revealing their distributional characteristics across teams. As shown in the figure, both measures exhibit a consistent upward trend with respect to team ID, and the scatter distribution demonstrates a clear positive relationship. Statistical analysis indicates that the Pearson correlation coefficient between the nonverbal synchrony dimension score and the composite performance score reaches 0.71, suggesting that higher levels of nonverbal behavioral synchrony within teams are associated with superior overall performance outcomes. This finding is consistent with the central tenets of coordination theory, which posit that nonverbal synchrony among team members enhances coordination efficiency and, in turn, improves performance. Accordingly, H1 is supported, providing quantitative evidence that nonverbal synchrony serves as a key predictive factor for team performance.

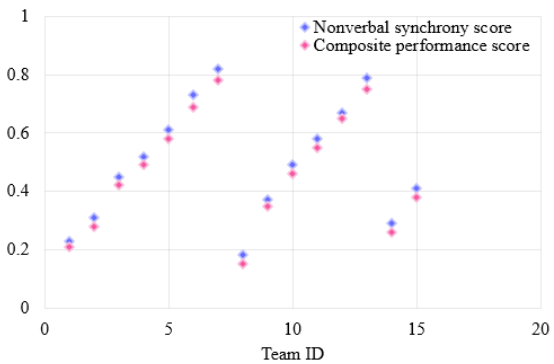

Figure 3. Association between team nonverbal synchrony scores and composite performance scores

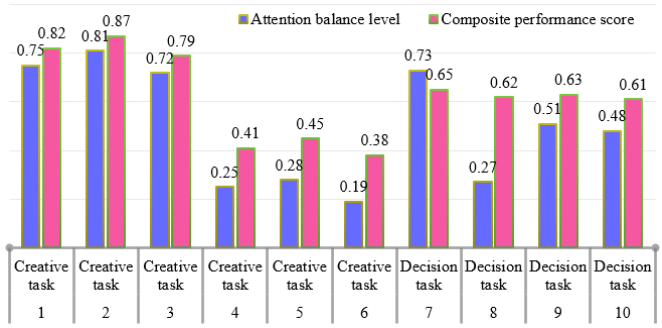

Figure 4. Effects of attention distribution patterns on team performance across different task types

Figure 4 compares team performance differences across attention distribution patterns in creative and decision-making tasks. In creative tasks, teams exhibiting an egalitarian attention distribution (attention balance level > 0.7) achieved a significantly higher mean performance score (0.83) than centralized teams (attention balance level < 0.3), with an independent-sample t-test yielding t = 4.23. In contrast, no substantial performance differences were observed across attention distribution patterns in decision-making tasks, where mean performance fluctuations remained below 0.04. These findings are consistent with expectations derived from team diversity theory. Creative tasks rely on egalitarian interaction and the collision of diverse viewpoints, for which balanced attention allocation facilitates comprehensive information exchange. Decision-making tasks, by contrast, tend to depend more strongly on leadership by core members, thereby attenuating the influence of attention distribution patterns. Accordingly, H2 is supported.

Figure 5 presents a comparison between the predictive capabilities of static nonverbal synchrony scores and the rate of synchrony increase during task sprint phases. The correlation coefficient between sprint-phase synchrony growth rate and performance reaches 0.76, exceeding the correlation of 0.71 observed for static synchrony scores. Correspondingly, R^2 increases from 0.5041 to 0.5776, representing an improvement of 8.3%. These results indicate that increases in nonverbal synchrony during sprint phases more effectively capture the dynamic optimization of collaborative quality and provide stronger predictive power than static synchrony measures. Accordingly, H3 is supported, highlighting the central importance of dynamic coordination features for team performance prediction.

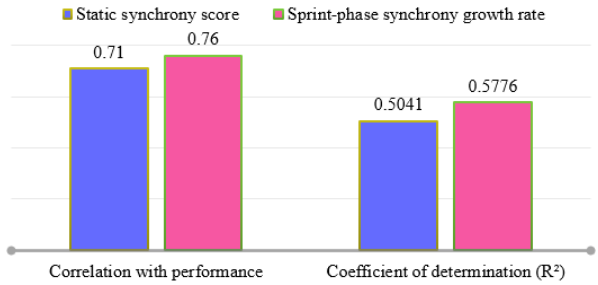

Figure 5. Performance predictive power of static and dynamic nonverbal synchrony features

3.3 Overall performance comparison

Overall performance comparison experiments were conducted on both the self-constructed ONC dataset and the

public MPII Group Interaction dataset to comprehensively evaluate the performance superiority and generalization stability of the proposed model relative to baseline approaches representing diverse technical paradigms. On the ONC dataset, the proposed model achieves an MAE of 0.087, an RMSE of 0.109, and an R^2 of 0.82. Compared with the strongest baseline (MAT), MAE is reduced by 20.3% and RMSE by 18.6%. Statistical testing confirms that these performance differences are highly significant, indicating clear advantages over all comparative models drawn from video-based behavioral representation learning, multi-agent interaction modeling, and traditional performance prediction approaches.

Cross-dataset generalization performance serves as a critical indicator of practical applicability. As reported in Table 1, the proposed model maintains the best overall performance on the MPII Group Interaction dataset, achieving an MAE of 0.095, an RMSE of 0.118, and an R^2 value of 0.79. Relative to MAT, MAE and RMSE are reduced by 14.6% and 12.8%, respectively. Moreover, performance variation between the ONC and MPII datasets remains within 5%, demonstrating stable predictive capability across differing team compositions, task types, and interaction contexts. These results indicate that the proposed approach exhibits substantially stronger generalization ability than baseline models that rely on scenario-specific features. The additional column reporting core model characteristics in Table 1 provides a concise comparison of the examined methods in terms of multimodal fusion, interaction modeling, and end-to-end optimization. This comparison elucidates the technical origins of the observed performance advantages.

Table 1. Comparison of team performance prediction results on the MPII Group Interaction dataset

Model	Core Model Characteristics	MAE	RMSE	R^2	Performance Difference from the Proposed Model (MAE)	Statistical Significance
Two-stream I3D	Dual-modality (RGB + optical flow) and single-granularity frame-level features	0.132	0.165	0.61	0.037	$p < 0.001$
Video Swin Transformer	Single-modality RGB and temporal attention features	0.125	0.158	0.64	0.030	$p < 0.001$
TeamGCN	Single-modality skeleton and team-level graph convolution modeling	0.118	0.149	0.67	0.023	$p < 0.01$
GroupViT	Single-modality RGB and group-level visual Transformer	0.121	0.153	0.66	0.026	$p < 0.01$
Multi-Agent Transformer	Dual-modality and multi-agent interaction attention	0.111	0.135	0.71	0.016	$p < 0.05$
Proposed model	Multimodal, multi-granularity, and latent variables + cross-member attention	0.095	0.118	0.79	-	-

Note: Statistical significance is assessed using independent-sample t-tests comparing the distribution of prediction errors between each baseline model and the proposed model on the test set. Thresholds are defined as follows: $p < 0.05$ indicates statistical significance, $p < 0.01$ indicates high significance, and $p < 0.001$ indicates very high significance.

The performance advantages of the proposed model are attributable to the synergistic effects of three core design elements. First, end-to-end latent variable learning enables the automatic, data-driven induction of communication patterns that are strongly associated with performance, thereby eliminating subjective bias and generalization limitations inherent in expert-defined representations. Second, the cross-member attention mechanism provides precise quantification of dynamic interpersonal interaction dependencies, strengthening the representation of critical collaborative behaviors such as idea exchange and consensus formation. Third, multi-granularity feature fusion is aligned with the hierarchical structure of team nonverbal communication—spanning individuals, interaction, and team—thereby effectively integrating complementary information from fine-grained individual behaviors and global team interactions. Together, these components form a high-performance prediction framework tailored to complex team collaboration scenarios.

3.4 Ablation studies

Ablation studies were conducted by systematically removing or replacing key components of the model, resulting in six variant configurations, aiming to quantitatively assess the contribution of multi-granularity fusion, cross-member attention, latent variable decoding, and the auxiliary task to overall performance. The results are summarized in Table 2. All variants were derived from the original model architecture, with only the target module modified in each case, thereby satisfying the single-variable control principle. Evaluation metrics include MAE, RMSE, and R^2, providing a comprehensive assessment of performance degradation.

The experimental results clearly indicate that the most pronounced performance degradation occurs when the latent variable decoding module is removed, with MAE increasing by 41.4% and R^2 decreasing by 20.7%. This finding demonstrates that latent variables function as the core representational carriers of communication patterns and are essential for enabling data-driven induction of nonverbal communication structures, directly determining the strength of alignment between features and performance objectives. When only a single granularity of features is retained, substantial performance degradation is observed in all cases. Notably, retaining only team-level granularity results in a larger performance loss (MAE +28.7%) than retaining only individual-level granularity (MAE +20.7%), indicating that individual behavioral details constitute the foundational basis of team interactions. A single granularity is therefore insufficient to capture the hierarchical information on team nonverbal communication, underscoring the irreplaceable role of multi-granularity feature fusion. The removal of the cross-member attention mechanism leads to an MAE increase of 13.8%, confirming its effectiveness in capturing inter-member interaction dependencies and its critical role in distinguishing interaction patterns between high- and low-performing teams.

When the auxiliary task is removed, MAE increases by 16.1%, indicating that the self-supervised reconstruction task effectively constrains latent variable quality and enhances representation robustness and generalization capability. Furthermore, replacing adaptive fusion with fixed-weight fusion results in a 24.1% increase in MAE, demonstrating that dynamic weight learning enables more precise alignment between feature contributions and performance prediction across varying scenarios, thereby substantially improving fusion effectiveness.

Table 2. Ablation study results of core modules (ONC test set)

Model Variant	MAE	RMSE	R^2	MAE Degradation	R^2 Degradation
Full model	0.087	0.109	0.82	-	-
Variant 1: Cross-member attention removed	0.099	0.126	0.75	13.8%	8.5%
Variant 2: Individual-level granularity only	0.105	0.134	0.72	20.7%	12.2%
Variant 3: Team-level granularity only	0.112	0.143	0.69	28.7%	15.9%
Variant 4: Latent variable decoding removed	0.123	0.158	0.65	41.4%	20.7%
Variant 5: Auxiliary task removed	0.101	0.129	0.74	16.1%	9.8%
Variant 6: Fixed-weight fusion	0.108	0.137	0.71	24.1%	13.4%

Taken together, the high performance of the proposed model is not attributable to the optimization of any single component, but rather to the synergistic integration of multi-granularity feature fusion, cross-member attention, latent variable decoding, and the auxiliary task. Among these components, latent variable decoding serves as the core innovation mechanism, multi-granularity fusion provides the foundation for scenario adaptability, cross-member attention constitutes the key to interaction modeling, and the auxiliary task functions as a performance-enhancing regularizer. Through their mutual reinforcement, these modules collectively establish an efficient framework for modeling team nonverbal communication and predicting team performance in organizational contexts.

3.5 Sensitivity analysis of latent variable dimensionality

Figure 6 illustrates the variation in MAE for team performance prediction on the ONC test set as the dimensionality of latent variables increases from 16 to 256. As shown, MAE decreases monotonically as the latent dimensionality increases from 16 to 64, with MAE values of 0.12 at 16 dimensions, 0.10 at 32 dimensions, and a minimum of 0.087 at 64 dimensions. This trend indicates that increasing dimensionality enhances the model's capacity to represent nonverbal communication patterns, enabling the capture of more fine-grained interaction features. When the latent dimensionality exceeds 64, MAE begins to increase gradually. Excessive dimensionality facilitates the fitting of noise present in the training data, leading to degraded generalization performance.

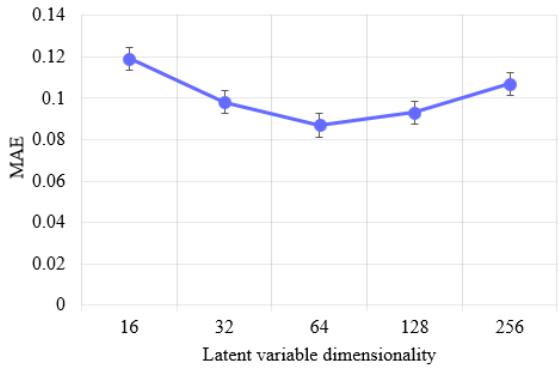

Figure 6. Effect of latent variable dimensionality on MAE for team performance prediction

These results empirically validate the rationality of the selected latent dimensionality. A dimensionality of 64 provides sufficient representational capacity to capture the core patterns of team nonverbal communication while avoiding the overfitting risks associated with higher-dimensional latent spaces. As such, this configuration represents an optimal balance between expressive power and generalization capability for the proposed model.

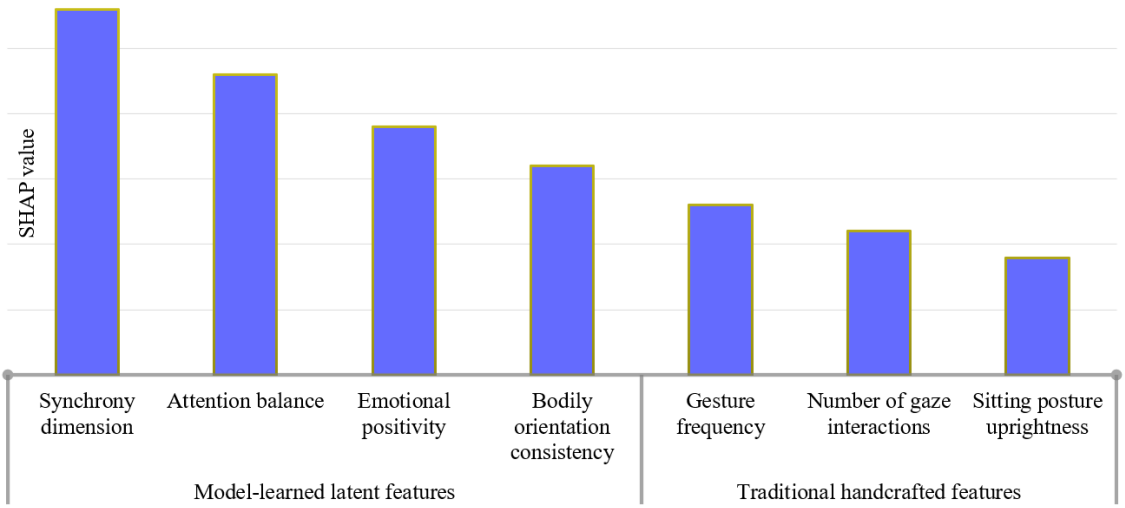

Figure 7. SHAP values of different feature types for team performance prediction

3.6 SHAP-based feature importance analysis

SHAP values were employed to quantify the marginal contributions of individual features to team performance prediction. Figure 7 presents a comparative visualization of the contribution magnitudes between model-learned latent variable features and traditional handcrafted features. As shown, SHAP values associated with latent variable features are substantially higher than those of handcrafted features. Among the latent variables, the synchrony dimension exhibits the highest SHAP value (0.28), followed by attention balance (0.23). Even the latent feature with the lowest contribution—bodily orientation consistency (0.16)—exceeds the mean SHAP value of traditional handcrafted features (0.11).

These results indicate that the automatically induced latent variables of nonverbal communication learned by the proposed model capture core information that is strongly associated with team performance with greater precision. Specifically, the synchrony dimension corresponds to coordination efficiency in collaborative processes, while attention balance reflects interactional equity among team members; both factors constitute critical drivers of team performance. By contrast, traditional handcrafted features focus on isolated behavioral indicators and fail to characterize inter-behavioral relational patterns, resulting in substantially lower contribution levels. This analysis not only clarifies the principal basis underlying the model's performance predictions but also highlights the representational superiority of data-driven latent variable learning over conventional handcrafted feature engineering.

4. CONCLUSION

This study addressed the longstanding challenge in organizational management of objectively quantifying team collaboration states by integrating computational behavioral science with organizational behavior theory. An end-to-end, multi-task, multimodal, multi-granularity CNN framework was proposed, through which nonverbal communication patterns were automatically induced and team performance was accurately predicted via the coordinated design of multi-granularity feature extraction, cross-member attention mechanisms, and latent variable decoding. The results demonstrated that the proposed model achieved significantly higher performance prediction accuracy than existing baseline methods on both self-constructed and public datasets, with all three core hypotheses quantitatively validated. Ablation experiments and SHAP-based analyses further confirmed the necessity of the key architectural components and the representational superiority of the learned latent variables. The contributions of this study are manifested across three dimensions. At the technical level, the application boundary of video-based behavioral representation learning is extended to multi-agent team interaction scenarios. At the theoretical level, objective quantitative evidence is provided for coordination theory and team diversity theory within organizational behavior research. At the practical level, an implementable technological solution is offered for team collaboration diagnosis and performance prediction in organizational settings.

Despite these advances, several limitations remain. The dataset coverage is primarily concentrated in technology-oriented industries, and emerging work settings such as remote collaboration have not yet been incorporated. Audio modality information has not been integrated, potentially omitting complementary nonverbal cues such as prosody and vocal tone. In addition, the proposed framework has not yet been deployed as an operational management tool, and validation in real corporate environments is still lacking. Future research may be advanced along three directions. First, the multimodal fusion framework may be extended to incorporate audio signals and physiological data in order to enrich the dimensionality of communication features. Second, more advanced temporal modeling strategies may be introduced to enable real-time tracking and prediction of team performance. Third, lightweight enterprise collaboration analysis tools may be developed and deployed in real organizational contexts to validate practical value, while further expanding dataset diversity across scenarios and industries to enhance generalizability and applied relevance.

FUNDINGS

This paper was supported by 2024 Annual Hunan Province Teaching and Educational Reform Research Project (Grant No.: HNJG-202401000014) and 2023 Hunan Provincial Ordinary Higher Education Institutions Teaching Reform Key Project (Grant No.: HNJG-20231233).

REFERENCES

[1] Valencia, S., Steidl, M., Rivera, M.L., Bennett, C.L., Bigham, J.P., Admoni, H. (2023). Nonverbal communication through expressive objects. Communications of the ACM, 67(1): 123-131. https://doi.org/10.1145/3610939

[2] Kudesia, R.S., Elfenbein, H.A. (2013). Nonverbal communication in the workplace. In Nonverbal Communication, pp. 805-832. https://doi.org/10.13140/RG.2.1.2270.6325

[3] Lahiani, R. (2025). Beyond words: Translating nonverbal communication in Arabic poetry a cultural turn approach. Cogent Arts & Humanities, 12(1): 2526941. https://doi.org/10.1080/23311983.2025.2526941

[4] Kazmi, S.W., Ahmed, W., Zahid, T. (2025). Understanding organizational changing behavior for enhancing sustainability performance. RAUSP Management Journal, 60(1): 297-314. https://doi.org/10.1108/RAUSP-04-2023-0064

[5] de Banderali, Z.S., Malavé, J., Alvarado, J.M., Dakduk, S. (2025). Unethical pro-organizational behavior in Hispanic American contexts: The role of organizational and interpersonal factors. Cogent Business & Management, 12(1): 2500679. https://doi.org/10.1080/23311975.2025.2500679

[6] Ahmad, R., Nejati, M., Farr-Wharton, B., Bentley, T. (2024). Impact of leadership on unethical pro-organizational behavior: A systematic literature review and future research directions. Journal of Leadership & Organizational Studies, 31(3): 338-367. https://doi.org/10.1177/15480518241265399

[7] Hustyi, K.M., Hays, T.N. (2024). Organizational behavior management approaches to advancing compassionate care in research and practice. Behavior

Analysis in Practice. https://doi.org/10.1007/s40617-024-00927-z
[8] Zhao, M., Qu, S., Tian, G., Mi, Y., Yan, R. (2024). Research on the moral slippery slope risk of unethical pro-organizational behavior and its mechanism: A moderated mediation model. Current Psychology, 43(19): 17131-17145. https://doi.org/10.1007/s12144-024-05677-3
[9] Shoaib, M., Husnain, G. (2026). Deep learning-based spatiotemporal action recognition in football using I3D and TSN with pose estimation. Biomedical Signal Processing and Control, 111: 108356. https://doi.org/10.1016/j.bspc.2025.108356
[10] Babu, S., Wadhwani, A.K. (2024). Epilepsy diagnosis using directed acyclic graph SVM technique in EEG signals. Traitement du Signal, 41(6): 3163-3172. https://doi.org/10.18280/ts.410632
[11] Jiang, J., Sorensen, A.D. (2026). Experimental and computational study on mechanical behavior of precast members with grouted splice sleeve connectors under sequential compressive and lateral loads. Journal of Structural Engineering, 152(1): 04025239. https://doi.org/10.1061/JSENDH.STENG-14901
[12] Jankelova, N., Dabić, M., Halaszovich, T. (2025). Organizational climate: Citizenship behavior, idiosyncratic deals and leader-member exchange. Management Decision, 63(13): 600-630. https://doi.org/10.1108/MD-01-2025-0012
[13] Albdareen, R., Aljawarneh, N.M., Al-Jedaiah, M.N., Alomari, K.A.K., Alrousan, A. (2025). Organizational factors affecting employee silence behavior: Does psychological empowerment matter? Cogent Business & Management, 12(1): 2512819. https://doi.org/10.1080/23311975.2025.2512819
[14] Sun, H., Ma, Y. (2025). MAViT: A lightweight hybrid model with mutual attention mechanism for driver behavior recognition. Engineering Applications of Artificial Intelligence, 143: 109921. https://doi.org/10.1016/j.engappai.2024.109921
[15] Liu, H., Feng, Z., Guo, Q. (2025). Multimodal cross-attention mechanism-based algorithm for elderly behavior monitoring and recognition. Chinese Journal of Electronics, 34(1): 309-321. https://doi.org/10.23919/cje.2023.00.263
[16] Jayamohan, M., Yuvaraj, S. (2025). Iv3-MGRUA: A novel human action recognition features extraction using Inception v3 and video behaviour prediction using modified gated recurrent units with attention mechanism model. Signal, Image and Video Processing, 19(2): 134. https://doi.org/10.1007/s11760-024-03726-9
[17] Na, C., Zhang, X. (2025). When misunderstanding meets artificial intelligence: The critical role of trust in human-AI and human-human team communication and performance. Frontiers in Psychology, 16: 1637339. https://doi.org/10.3389/fpsyg.2025.1637339
[18] Eldadi, O., Fitoussi, S.J., Tenenbaum, G. (2025). Verbal communication, coordinated effort, and performance in esports teams: An expert-nonexpert paradigm study. Journal of Sport and Exercise Psychology, 47(5): 320-332. https://doi.org/10.1123/jsep.2024-0343
[19] Al Maalouf, N.J., Sarkis, N. (2025). The influence of virtual team-specific factors and communication factors on team performance. The Learning Organization, 32(6): 993-1010. https://doi.org/10.1108/TLO-05-2024-0153