Design of Multimodal Behavior Analysis and Evaluation System for Digital Leadership of University Teachers Integrated with Perceiver IO

Xin Wang, Biyi Jiang
School of Marxism, Northeast Agricultural University, Harbin 150030, Heilongjiang, China
Corresponding author: Biyi Jiang (e-mail: jby202411@163.com)

ABSTRACT University ideological and political education faces difficulties in quantitatively evaluating teachers' digital leadership in complex digital teaching environments. This study proposes a multimodal evaluation framework based on the Perceiver IO architecture, using cross-modal attention and structured queries to model digital leadership behaviors across pedagogical contexts. Multimodal inputs, including speech, visual behavior, instructional content, and interaction logs, are mapped into a shared latent space and decoded into six leadership dimensions. Supervised learning enables structured interpretation of leadership patterns and dimension-specific scoring. Experimental results demonstrate strong model performance, with an F1-score of 0.88 for technology application, 0.80 for value guidance, and an AUC of 0.90 for digital influence. These results indicate that the proposed framework can accurately and robustly assess digital leadership in complex teaching environments. By integrating speech, visual, content, and log-based information through a unified latent representation, the method provides an engineering-oriented approach for multimodal behavior analysis, interpretable teaching evaluation, and cross-modal decision support.

INDEX TERMS multimodal behavior analysis, digital leadership evaluation, perceiver io, cross-modal attention, structured queries

I. INTRODUCTION

CURRENTLY, ideological and political education in universities is undergoing deep integration into the digitalization process, extending into professional disciplines. This trend is particularly evident in engineering, where digital technologies are reshaping design, manufacturing, and supply chain ethics, generating complex teaching scenarios that require integrated value guidance. In this context, teachers must master digital pedagogies while effectively conveying the values and responsibilities associated with technological change, directly influencing value communication and learning effectiveness in ideological and political education. Digital leadership encompasses not only technological application but also the ability to guide ideological cognition, maintain digital classroom order, and integrate teaching resources [1,2]. Without a systematic evaluation mechanism, teachers' leadership roles in digital environments are difficult to reflect, limiting precise development of university teaching teams [3,4].

To address the absence of a unified multimodal modeling and evaluation mechanism for teacher digital leadership, this paper proposes a multimodal evaluation method for ideological and political education classrooms based on the Perceiver IO architecture. Speech recognition and text encoding are used to extract guiding content from teacher speech, while a visual Transformer model captures value-related visual cues from videos and courseware images. Teaching platform logs are analyzed to model behavioral sequences in resource scheduling and technical operations, and facial action encoding vectors represent emotional feedback during instruction. After embedding, all modal features are fused within the Perceiver IO latent space through cross-attention.

II. RELATED WORK

Against the background of higher education digital transformation, teachers' digital leadership has emerged as a critical capability for advancing educational governance modernization and classroom paradigm change. Hamzah et al. synthesized international research and highlighted the central role of digital leadership in educational organizations, emphasizing the coordinated development of technical literacy, strategic capacity, and interpersonal communication in driving educational change [5]. This framework provides guidance for redefining university teachers' roles and supports cross-disciplinary integration of digital leadership concepts [6,

7]. Building on this perspective, Jing et al. investigated digital academic leadership practices of middle and senior managers in Chinese universities through questionnaires and interviews, extracting four core dimensions and revealing the influence of background differences on cognition and implementation, thus forming a localized theoretical foundation [8]. This line of research strengthened the application orientation and cultural adaptability of digital leadership theory in local contexts [9,10]. Jameson et al. reviewed 36 empirical studies published between 1999 and 2022 and found that although research volume has increased, issues such as theoretical fragmentation and methodological singularity persist, leading to a call for standardized maturity assessment frameworks [11]. Overall, existing studies clarify the multidimensional structure of digital leadership but remain limited in quantitative evaluation, micro-level behavior modeling, and measurement tools for frontline teachers [12-14].

In multimodal teaching behavior analysis, model design has progressively shifted toward balancing fusion accuracy and computational efficiency. Xu et al. integrated multimodal posture features of the head, hands, and body with Feature Pyramid Network and Convolutional Block Attention Module mechanisms to construct a classroom teacher–student behavior recognition system, enabling automated interaction analysis and quantitative behavioral feedback [15]. This approach effectively captured spatiotemporal behavioral characteristics and demonstrated practical value in classroom scenarios [16,17]. Building on behavior recognition, Tuerhong et al. proposed an Adaptive Multimodal Transformer based on an exchanging mechanism to achieve feature complementarity and hierarchical fusion across modalities, enhancing contextual understanding robustness [18]. Its structural design improved model stability in complex educational environments [19]. To further optimize computational resource allocation, Nagrani et al. introduced an attention bottleneck architecture that restricted cross-modal interaction to a compact latent unit, reducing redundant computation while maintaining performance across audio-visual tasks [20]. In summary, current research has made some progress in fusion strategies and module design, but still lacks in-depth exploration of the structured representation of complex behaviors and the unified modeling mechanism for high-dimensional heterogeneous inputs [21,22].

III. MULTIMODAL BEHAVIOR MODELING METHOD FOR DIGITAL LEADERSHIP

A unified multimodal behavior modeling framework for digital leadership is constructed, covering data collection, preprocessing, feature extraction, fusion, and evaluation. The framework adopts a layered and tightly coupled structure to model teachers' digital leadership behaviors. Multimodal data from audio, video, courseware images, and platform interaction logs are synchronized and aligned before feature extraction. Each modality is encoded using deep learning models to transform heterogeneous inputs into a unified feature space. Cross-attention is then applied to fuse modal features and enhance behavioral representations. The resulting latent representation is used as input for behavioral label prediction, providing high-quality features for evaluation. This framework ensures standardized data processing and strengthens cross-modal dependencies, supporting accurate identification of digital leadership behaviors. Figure 1 presents the overall architecture and relationships among data input, feature extraction, fusion modeling, and evaluation modules.

A. MULTIMODAL DATA COLLECTION AND PREPROCESSING

During multimodal data collection and preprocessing, the system performs high-frequency synchronized acquisition centered on university teachers' classroom behaviors, covering voice tracks, video sequences, courseware images, and platform operation logs. Audio data are recorded in real time, sampled, denoised, and spectrally enhanced, then transcribed by a Transformer-based Automatic Speech Recognition engine to produce time-aligned text sequences. Video data are captured by a fixed-angle high-definition camera and decoded at a constant frame rate, with image quality inspection and redundancy removal applied to ensure valid visual encoding input. Courseware content is obtained by parsing teaching presentation files to extract structured text and graphical information and synchronizing it with live instruction. Platform operation logs are collected through embedded event listeners, recording time series of interface switching, resource access, and interaction events, with each record containing timestamps, durations, and event intervals [23,24]. All modalities are encoded as time-aligned vectors and projected onto a unified temporal grid to support subsequent fusion.

To achieve cross-source alignment, a timestamp-based linear interpolation strategy is applied using the teacher's voice as the primary temporal reference. Image frames, courseware images, and log events are bound to corresponding audio segments through a window-based alignment mechanism with a constrained deviation threshold. Segment pairs with the highest event density consistency across modalities are retained, and synchronized samples are organized into fixed-length sequences for unified vector encoding.

To improve modeling reliability, a multi-channel quality screening mechanism is applied. Speech segments are filtered using transcription confidence and language model perplexity. Video sequences are retained only when the teacher's facial region is continuously detectable across frames, the estimated head pose parameters remain within a bounded angular variation range over time, and the image sharpness measured by gradient-based blur metrics exceeds a predefined threshold. Operational logs are filtered by behavioral frequency and temporal coverage to remove sparse segments that could disrupt sequence consistency.

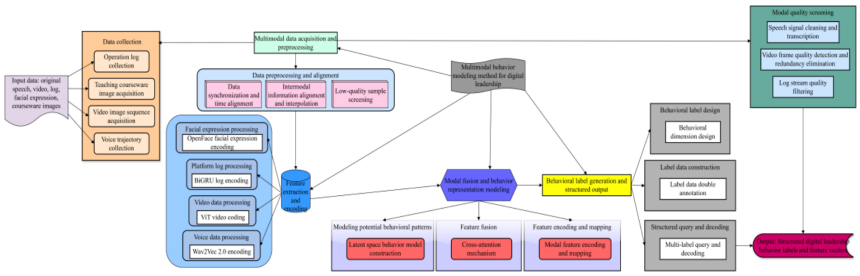

FIGURE 1. Framework of the multimodal behavior modeling method for digital leadership

Each input unit is represented by a vector group:

$$X = \{x_t^{(a)}, x_t^{(v)}, x_t^{(p)}, x_t^{(l)}\}_{t=1}^{T} \tag{1}$$

In Formula (1), $x_t^{(a)}$ represents the voice text vector at time t; $x_t^{(v)}$ represents the video frame image feature vector; $x_t^{(p)}$ represents the courseware image visual feature; $x_t^{(l)}$ represents the interaction log event vector; T is the maximum number of frames with a uniform step size within the sample duration window. To ensure data transmission and storage standardization, the system normalizes all input sample sequences and stores them in a standard Tensor format. A cross-modal index matrix is established for attention path control and behavior label position mapping in the subsequent modeling stage.

B. MODAL FEATURE ENCODING STRATEGY

After multimodal data alignment, all modalities are transformed into a unified vector representation for fusion modeling. Voice modality encoding adopts Wav2Vec 2.0. Raw audio is converted into frame-level acoustic features by extracting MFCC and prosodic descriptors and passing the waveform through the convolutional front end of Wav2Vec 2.0. The resulting representations are concatenated and linearly projected into a shared feature dimension. In parallel, speech signals are transcribed into text and embedded through word segmentation into a high-dimensional semantic space, where contextual dependencies and pragmatic information are modeled. The encoding process integrates convolutional feature extraction with Transformer layers to capture both local acoustic patterns and global semantic relationships. The output is a time-aligned text vector sequence $\{x_t^{(a)}\}_{t=1}^{T}$. Residual connections are used to stabilize feature propagation, enhancing the model's ability to model language rhythm and repetitive structure [25-27].

Visual features from video and courseware images are encoded using the Vision Transformer (ViT) to produce spatial embeddings, while short-term motion embeddings are extracted with a lightweight 3D convolutional network. Facial action unit vectors from OpenFace are appended, and the combined visual representation is linearly projected into the shared feature dimension. The input frame image is first divided into fixed-length patches, then linearly embedded and positionally encoded to form image vector sequences $\{x_t^{(v)}\}$ and $\{x_t^{(p)}\}$. The ViT structure establishes non-local dependencies across the image sequence. A multi-head self-attention module extracts spatial representation features and image semantic information of the teacher's behavior, capturing regional patterns that are highly correlated with teaching behavior. To enhance the contribution of facial expressions in model perception, the system simultaneously applies the OpenFace library to parse the action unit parameters of the face area in the video and construct the action tensor $x_t^{(f)} \in \mathbb{R}^{n_f}$. After splicing it with the ViT output, it forms a joint visual vector $\tilde{x}_t^{(v)} = [x_t^{(v)}; x_t^{(f)}]$, improving the video encoding ability to depict the teacher's non-verbal behavior signals.

The platform interaction log mode is modeled by inputting the event time series into the Bidirectional Gated Recurrent Unit (BiGRU) network. Each discrete event is encoded into a fixed-length attribute vector through categorical embedding and time-interval encoding. Then, the event vectors are input to the BiGRU, and the BiGRU output is linearly projected to the shared dimension. Each log event is encoded by word vectors to generate a discrete representation $x_t^{(l)}$. The BiGRU network propagates event dependencies along the positive and reverse time order, captures the sequential logic and behavior chain closed-loop structure in the operation behavior sequence, and outputs a bidirectional splicing sequence $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. To align the input to a unified dimension, all modal output vectors are linearly transformed and projected to the shared dimension d to construct the behavior time step tensor input set:

$$Z_t = W_a x_t^{(a)} + W_v \tilde{x}_t^{(v)} + W_p x_t^{(p)} + W_l h_t + b \tag{2}$$

In Formula (2), $W_a, W_v, W_p, W_l \in \mathbb{R}^{d \times d_i}$ are the linear mapping matrices of speech, vision, courseware, and log modes, respectively; $b \in \mathbb{R}^d$ is the bias term; d_i is the original dimension of each mode; d is the unified mapping dimension. The linear projections use learnable matrices that map heterogeneous source dimensions into the fixed 512-d latent space so that subsequent attention layers operate on consistent vector dimensions.

After encoding, the modal features are normalized to a unified semantic scale, effectively suppressing the behavioral representation offset caused by structural differences between modalities. To further improve the model's response to key behavioral instantaneous features, the system sets a modality selection gating mechanism at the input:

$$\alpha_t^{(m)} = \sigma(W^{(m)} z_t^{(m)} + b^{(m)}) \tag{3}$$

In Formula (3), $\alpha_t^{(m)}$ represents the activation weight of the m-th modality at step t; $z_t^{(m)}$ is the original modal input; $W^{(m)}$ and $b^{(m)}$ are the gating weights and bias; σ is the sigmoid activation function. The preprocessed modal encoding samples meet the requirements of unified fusion in terms of structural consistency and semantic integrity, ensuring that the model fully captures cross-modal complementarity and temporal dependencies during the behavioral representation learning process.

C. FEATURE FUSION AND BEHAVIOR REPRESENTATION CONSTRUCTION

After modal feature encoding, all modality-specific vectors are projected into the latent representation required by the Perceiver IO architecture. Each vector is linearly mapped to a unified latent dimension of 512, with positional encoding applied along the temporal axis. The resulting tensor of size $T \times M \times 512$ is then fed into the Perceiver IO cross-attention encoder for fusion modeling. The encoded voice vector, video visual vector, courseware image vector, and interaction log sequence vector are all linearly projected to the fixed-dimensional space $\mathbb{R}^d$ to ensure structural compatibility and information expression stability. After aligning the encoding results of each modality by time step, a multimodal input tensor $Z \in \mathbb{R}^{T \times M \times d}$ is constructed. Among them, T represents the time step length; M represents the number of modalities; d is the unified feature dimension.

This tensor is input to the Perceiver IO encoder, whose structure adopts a heterogeneous input and homogeneous latent interaction mode, establishes inter-modal dependencies through a cross-attention mechanism, and generates a unified latent representation. The encoding process relies on the Query-Key-Value ternary structure, defines the latent representation matrix $L \in \mathbb{R}^{N \times d}$ as the query matrix, and performs a cross-attention operation with the input tensor Z to construct a cross-modal association mapping:

$$\text{Attention}(L, Z) = \text{softmax}\left(\frac{LW_Q(ZW_K)^T}{\sqrt{d}}\right) ZW_V \tag{4}$$

In Formula (4), $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. This mechanism ensures that the contribution of different modalities to the representation of latent semantics is transmitted to the latent space through a unified channel, effectively alleviating semantic offset and redundant interference between modalities. The model further superimposes multiple layers of cross-attention and feedforward transformation to enhance the ability to recognize behavioral semantics while maintaining information abstraction. To control the influence of different modalities on the latent space, a modality weight normalization strategy is applied. Modality weight gating is added before each layer of attention to form a learnable modality importance score vector $\alpha \in \mathbb{R}^M$, weighting the input tensor:

$$Z'_{t,m} = \alpha_m \cdot Z_{t,m} \tag{5}$$

In Formula (5), α_m represents the global weight of the m-th modality. This parameter is optimized together with the objective function during the training phase to improve the robustness of feature fusion in multimodal imbalance scenarios.

To further compress high-dimensional sequence information and improve semantic abstraction, the latent representation matrix is fed into the projection mapping layer after the last layer of fusion, outputting a unified behavior embedding vector $h \in \mathbb{R}^{d'}$ as the input basis for downstream structured query and behavior decoding. The projection operation is defined as follows:

$$h = \text{LayerNorm}(\text{FFN}(L_{\text{final}})) \tag{6}$$

In Formula (6), L_{final} is the cross-attention output of the last layer; FFN is the feedforward neural network; LayerNorm is the layer normalization function, which is used to stabilize the representation distribution and accelerate convergence.

After feature extraction and encoding, multimodal data are fused through the cross-attention mechanism to enable integrated information modeling and behavior pattern learning. Within the Perceiver IO architecture, features from different modalities interact in a shared latent space to capture cross-modal dependencies. Stacked cross-attention layers progressively refine this interaction and produce a unified latent representation for subsequent behavior evaluation and label decoding. Figure 2 illustrates the cross-attention–based fusion process in Perceiver IO and the semantic alignment of multimodal features in the shared space.

Teachers' digital leadership performance is quantified using multiple behavioral labels, each defined through expert coding and consistency verification. Annotation is conducted by three experts with doctoral training in educational technology and ideological and political education, each with over eight years of research experience. All annotators follow a unified training protocol specifying behavioral boundaries and dimension criteria. Inter-annotator agreement is evaluated using Cohen's Kappa for each dimension, yielding an average value of 0.82 across the six labels, indicating high consistency. Table 1 presents the frequency of annotated labels across technology application, value guidance, organizational coordination, feedback response, resource integration, and digital influence for different teaching experience groups. These statistics represent the dataset-level distribution of annotated behavioral segments and define supervision density during model training. Teachers are grouped by full-time university teaching experience into New Teachers with less than five years, Middle-aged Teachers with six to fifteen years, Senior Teachers with sixteen to twenty years, and Experienced Teachers with more than

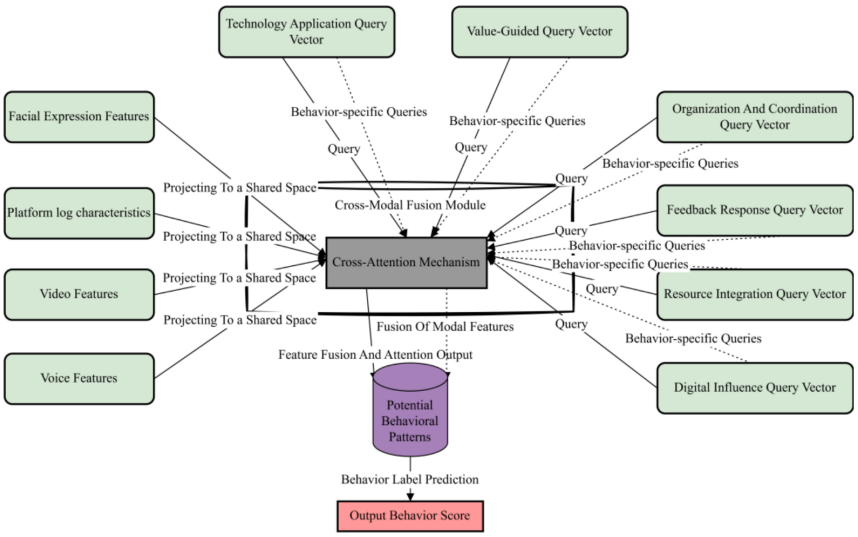

FIGURE 2. Cross-attention mechanism and feature fusion in Perceiver IO

twenty years. Frequency denotes the aggregated count of annotated behavioral segments per dimension within each group, computed over fixed-length segments extracted from continuous teaching sessions.

D. DESIGN OF DIGITAL LEADERSHIP BEHAVIOR LABELING SYSTEM

To support quantifiable modeling and structured evaluation of university teachers' digital leadership, a behavioral labeling system with oversight, distinguishability, and semantic closure is established. The system defines six high-level dimensions covering technology application, value guidance, organizational coordination, feedback response, resource integration, and digital influence. Each dimension is derived from raw behavioral data through expert coding, consistency validation, and semantic constraint control, ensuring functional independence and minimal label overlap. Label semantic definitions are embedded in the system

task space and represented via a structured matrix, mapped to a set of query vectors $\{q^{(i)}\}_{i=1}^{6}$ in the behavioral query space. These label query vectors are constructed offline based on the annotated teacher behavior corpus before model training and remain fixed throughout the optimization process. Each $q^{(i)} \in \mathbb{R}^{d_q}$ represents a semantic abstract vector of a dimension. This semantic vector is extracted from the high-frequency representation subspace of the faculty behavior corpus. Sentence vectors are computed using the Sentence-Bidirectional Encoder Representations from Transformers model, followed by clustering to determine the central representation. Maximum mutual information constraints are used to control similarity overlap between labels.

The construction process of the label vector uses the following optimization objective function to select the cluster center:

$$\arg\min_{\{q^{(i)}\}} \sum_{i=1}^{6} \sum_{x_j \in C_i} \|x_j - q^{(i)}\|^2 + \lambda \sum_{i \neq j} \cos(q^{(i)}, q^{(j)}) \quad (7)$$

This objective function is applied during the offline label construction stage and is not involved in parameter updates during model training. In Formula (7), x_j is the sentence vector of the member in the cluster; C_i is the behavior cluster of the i-th category; λ controls the heterogeneity constraint between labels. The objective is to minimize intra-class variance and maximize inter-class semantic separation. The behavior label vector is projected into the decoding space via a pre-trained semantic mapping function and is invoked by the Perceiver IO output module during the query process.

In the label data construction stage, to enable the training model to have label-driven discrimination ability, all behavior samples are annotated by two people and undergo consistency verification to form a multi-label supervision matrix $Y \in \{0,1\}^{N \times 6}$, where N is the number of samples. Only samples meeting the inter-annotator agreement threshold are retained in the supervision matrix. Each sample represents a segmented behavioral unit extracted from multimodal teaching data, and all supervision statistics are computed at this level to ensure consistent comparison across teacher groups. Behavioral units may be associated with multiple labels, and supervised training follows a multi-label parallel strategy, where each behavioral unit independently activates multiple dimension-specific supervision signals during optimization. Label distribution statistics, occurrence frequency, and sample coverage are organized to characterize both the structural composition of supervision signals and their empirical distribution across the dataset. Frequency of tag occurrence is defined as the proportion of the total number of validated label assignments contributed by each dimension over the entire annotation corpus, where the denominator is the sum of all dimension-wise label assignments rather

TABLE 1. Distribution of annotated behavioral segments across digital leadership dimensions by teacher group

Teacher Types	Technology Application (Frequency)	Value Guidance (Frequency)	Organizational Coordination (Frequency)	Feedback Response (Frequency)	Resource Integration (Frequency)	Digital Influence (Frequency)
New Teachers less than five years	60	50	40	45	35	30
Middle-aged Teachers six to fifteen years	90	80	70	65	55	50
Senior Teachers sixteen to twenty years	110	100	90	85	75	70
Experienced Teachers more than twenty years	120	110	95	90	80	85

than the number of behavioral units. Under this definition, frequency of tag occurrence reflects the relative contribution of each behavioral dimension within the global supervision signal space. Proportion of sample coverage is defined as the ratio of behavioral units that contain at least one validated annotation of the corresponding dimension to the total number of behavioral units, reflecting the distribution breadth of each dimension across the dataset. All statistics are computed based on 3 420 annotated behavioral samples. Table 2 shows the distribution of various labels.

TABLE 2. Distribution statistics of behavior labels based on label assignment frequency and sample coverage

Label dimension	Number of label query units	Frequency of tag occurrence (%)	Proportion of sample coverage (%)
Technology application	150	21.43	95.5
Value guidance	120	17.14	93
Organizational coordination	100	14.29	91.2
Feedback response	110	15.71	92.4
Resource integration	130	18.57	94.1
Digital influence	90	12.86	89.3

The frequency of tag occurrence is normalized over the total number of label assignments across all behavioral units and represents the relative weight of each behavioral dimension in the multi-label supervision space. To maintain the structural distinction of labels in the attention mechanism, each label embedding vector is enhanced by position shift and semantic perturbation to generate an extended structured query set. The resulting number of label query units determines the resolution of the label attention space and defines how many parallel semantic anchors are allocated to each behavioral dimension during decoding. This structural parameter characterizes the internal richness of the label query space and is independent of dataset-level statistics, while frequency of tag occurrence and proportion of sample coverage describe the empirical distribution of behavioral dimensions across the annotated samples. The perturbation amplitude ϵ is fixed to 0.05 in all experiments, ensuring sufficient semantic diversity while preserving label stability. The enhancement strategy adopts the following perturbation function:

$$\tilde{q}^{(i)} = q^{(i)} + \epsilon \cdot N(0, I) \tag{8}$$

In Formula (8), ϵ is the perturbation amplitude; $N(0, I)$ is the standard normal noise; $\tilde{q}^{(i)}$ is the perturbated label vector. This mechanism improves the model's ability to distinguish between weak label overlap and boundary samples.

E. STRUCTURED QUERY OUTPUT AND EVALUATION GENERATION

The unified behavior embedding is fed into the Perceiver IO output query module, which employs six independent query vectors corresponding to the leadership dimensions of technology application, value guidance, organizational coordination, feedback response, resource integration, and digital influence. These queries are fixed semantic representations defined before training and serve as stable anchors for dimension-specific decoding. Each query interacts with the latent representation through an independent cross-attention path, producing six parallel outputs with one-to-one correspondence between dimensions and output channels. Each dimension is decoded from multimodal indicators: technology application from operational density, value guidance from semantic polarity and gaze fixation, organizational coordination from gesture–speech synchrony, feedback response from reaction interval, resource integration from external call ratio, and digital influence from post-class engagement persistence. The structured query mechanism assigns attention weights to latent features according to query semantics, enabling adaptive semantic activation across heterogeneous modal inputs and enhancing inter-dimension discrimination. The decoding process adopts dot-product attention to activate the corresponding latent behavior representations for each query target. The attention formulation follows a scaled dot-product structure, in which the normalization term depends solely on the latent key dimensionality rather than on any specific key vector. The core calculation of this mechanism is as follows:

$$y_i = \sum_{j=1}^{L} \alpha_{ij} z_j, \quad \alpha_{ij} = \frac{\exp(q_i^\top k_j / \sqrt{d_k})}{\sum_{l=1}^{L} \exp(q_i^\top k_l / \sqrt{d_k})} \tag{9}$$

In Formula (9), y_i represents the decoding output corresponding to the i-th behavior label, q_i is the fixed query vector associated with the i-th semantic dimension, k_j and z_j denote the key and value vectors of the j-th latent position, and L is the length of the latent representation sequence. d_k denotes the dimensionality of the key vectors,

and the inverse square root of d_k is applied as a scalar normalization factor to stabilize the dot-product attention scores in the high-dimensional latent space.

The structured output module uses a multi-label classification head with Sigmoid activation to normalize each behavior score. The output is a six-dimensional real-valued tensor, where each element represents the normalized activation strength of the corresponding structured query over the latent behavior representation. The specific calculation method is as follows:

$$\hat{y} = \sigma(W_o \cdot y + b_o) \tag{10}$$

In Formula (10), $y \in \mathbb{R}^{6\times d}$ is the decoding output representation of the six labels; W_o and b_o are the output weights and bias parameters; σ is the element-by-element Sigmoid function; $\hat{y}$ is the normalized multi-label score result. Label frequency participates in the training process by shaping supervision distribution and gradient magnitude across dimensions.

All label outputs share consistent dimensionality, supporting dimension-wise statistical analysis in the evaluation module. Each query maintains a fixed correspondence to its target dimension during training to preserve semantic independence and prevent feature overlap. Although all labels decode from the same latent representation, independent query vectors guide separate decoding paths, enhancing robustness and accuracy under complex behavior differentiation. This mechanism suppresses modality bias and enforces selective latent activation, ensuring semantic independence across dimensions.

IV. SYSTEM TESTING AND EXPERIMENTAL CONFIGURATION

A. TEACHING DATASET CONSTRUCTION

This study collects classroom teaching processes of ideological and political education courses from multiple universities to construct a multimodal dataset comprising voice, video, courseware images, text, and interaction logs. Data are gathered from six universities, covering twelve undergraduate courses and eighty-seven teachers with complete teaching sessions. Based on experience-based criteria, the dataset includes twenty-one New Teachers, twenty-six Middle-aged Teachers, nineteen Senior Teachers, and twenty-one Experienced Teachers. Each teacher contributes multiple segmented behavioral samples extracted from continuous teaching sessions, forming the annotated supervision data used for modeling and evaluation, with grouping criteria consistent with Table 1. During collection, voice data are recorded in real time using teaching recording devices, video data are captured by fixed-angle high-definition cameras to record facial movements and posture, courseware images are parsed from electronic materials to synchronize visual content with teaching progress, and platform interaction logs record time-stamped operations and resource calls. The integration of these heterogeneous sources provides a comprehensive representation of teachers' digital leadership behaviors in classroom teaching. Table 3 shows the statistical information of each type of data collected, including the frequency, average duration, and total amount of data for each type. This data serves as the foundation for model training and supports multimodal fusion and behavioral analysis tasks.

TABLE 3. Data collection statistics

Data Type	Collection frequency (times/hour)	Total data volume (GB)	Average duration (minutes)
Voice Data	10	15.6	60
Video Data	15	25.2	50
Courseware Content	5	10.3	30
Interaction Log	20	8	20

B. EXPERIMENTAL ENVIRONMENT AND MODEL CONFIGURATION

The experiment is conducted on an NVIDIA A100 server using the PyTorch framework. To evaluate generalization performance, a teacher-level five-fold cross-validation protocol is adopted. All teachers are first partitioned into five mutually exclusive subsets of equal size. In each fold, one subset is held out as the test set, while the remaining four subsets are merged to form the development set. The development set is further divided into training and validation subsets according to a fixed 70 to 15 ratio at the teacher level, ensuring that no teacher appears in more than one subset within the same fold and that all modalities associated with a teacher remain within the same partition. The model is trained for 120 epochs with a fixed learning rate of 0.001, a batch size of 32, and an input vector dimension of 512 for all modalities and latent representations. Standard hyperparameters are used, and weights are optimized via backpropagation. This protocol preserves the standard structure of five-fold cross-validation while introducing an internal validation split for model selection, and all reported test results are obtained exclusively from the held-out teacher subset of each fold.

C. TRAINING PROCESS AND EVALUATION PROTOCOL

In the experiment, six-dimensional behavior labels serve as supervised targets, and the model is trained with a multi-label cross-entropy loss. Model training follows the data partitioning protocol defined in the experimental configuration section, and optimization focuses on improving convergence stability and generalization performance under fixed teacher-level splits. The Adaptive Moment Estimation optimizer is used with a learning rate of 0.001, dynamically adjusted based on validation performance. Model performance is evaluated using F1 score and AUC to ensure all label dimensions are effectively learned. An early stopping strategy halts training when validation performance ceases to improve, preventing overfitting.

V. EVALUATION RESULTS

A. MULTIMODAL FUSION AND DIMENSIONAL PERFORMANCE EVALUATION

This experiment examines university faculty digital leadership by comparing single- and multi-modal inputs in identifying six key dimensions: technology application (D1), value guidance (D2), organizational coordination (D3), feedback response (D4), resource integration (D5), and digital influence (D6), reflecting technical proficiency, value shaping, team and process management, responsiveness, resource allocation, and digital impact. Four single-modal data types are collected: classroom voice (M1), video (M2), courseware content (M3), and platform interaction logs (M4). Fifteen modal combinations are constructed: single-modal inputs capture individual features; bimodal inputs reveal cross-modal associations, e.g., voice-video consistency and voice-log coordination; trimodal inputs enhance feature complementarity across behavior, content, and interactions; full-modal inputs integrate all four types for comprehensive cross-modal fusion. Figure 3 shows predictive performance of these multimodal integrations, with metrics computed per teaching session.

In F1 results, the full-modal approach achieves the highest scores in technology application (0.88), value guidance (0.80), and digital influence (0.86), benefiting from complementary capture of language, visuals, content structure, and interactive behavior. Among trimodal approaches, video + PPT + log achieves 0.88 in organizational coordination, exceeding the full-modal's 0.85, highlighting the advantage of combining visual and structured content with behavioral records. For bimodal approaches, voice + video achieves 0.72 in feedback response, ranking first among bimodal methods, indicating a strong link between interaction frequency and language features.

In AUC results, full-modal scores highest in feedback response (0.87) and digital influence (0.90), showing that multi-source fusion forms a stable decision boundary. Video + PPT + log attains 0.91 in organizational coordination, while single-modal log reaches 0.88 in resource integration, demonstrating the discriminative power of behavioral data.

In accuracy, full-modal achieves 0.94 in resource integration, matched by PPT + log, indicating strong complementarity between structured content and behavioral features. Voice + log attains 0.71 in feedback response, highest among bimodal methods. Overall, full-modal performs best in most dimensions, but trimodal or bimodal combinations can compete in specific cases due to targeted information.

B. ROBUSTNESS OF MULTIMODAL MODELS

In multimodal digital leadership modeling, attention visualization reveals the model's focus on six-dimensional behavioral labels across different modalities. Classroom voice captures speech content and tone, reflecting language-driven features such as value orientation and technical explanation. Video captures facial expressions, posture, and movements, reflecting non-verbal cues like coordination and feedback. Courseware content represents teaching structure and information presentation, linked to value guidance and technology application. Platform interaction logs record operational behaviors and resource allocation, relating to feedback and resource integration. These four modalities map attention to the six labels across temporal or event sequences, forming multi-level alignments from language to vision and content to behavior. Figure 4 shows label attention distributions and temporal evolution, averaged over all teaching sessions.

In the classroom voice modality, technology application peaks at 25% mid-lesson during tool demonstrations, while value guidance accumulates over 70% in the first fifteen minutes, reflecting the opening value framework. In video, organizational coordination reaches 25% early, indicating nonverbal guidance in classroom order and task assignment, and feedback response starts at 30%, correlating with eye contact and gesture frequency. In courseware content, value guidance totals 53% in the first three pages, aligning with the visual presentation of core values, while attention to technology application drops to 12% mid-to-late lesson as content shifts to concepts. In platform interaction logs, digital influence reaches 31% during the course summary with online discussions and resource pushes, and resource integration accounts for 60% in the second half, reflecting frequent use of external materials. These patterns highlight how multimodal inputs capture digital leadership at different teaching stages and the complementary role of each modality in label recognition.

This section evaluates model robustness by measuring output stability under masking perturbations applied to all modal inputs. These perturbations simulate common input degradation in real-world scenarios, and resulting score fluctuations across six digital leadership dimensions reflect the model's adaptability and stability. Figure 5 shows the standard deviation trends of each dimension under varying masking ratios.

As the masking ratio increases from 0% to 50%, score stability across the six dimensions generally decreases, showing a positive correlation between perturbation intensity and predictive uncertainty. Feedback response exhibits the largest increase in standard deviation, from 0.0495 to 0.1051, indicating high dependence on modal completeness. Digital influence and technology application also show strong volatility, rising to 0.0952 and 0.0911, reflecting reliance on fine-grained behavioral and semantic cues. In contrast, value guidance and resource integration remain relatively stable, with maximum deviations of 0.0707 and 0.0669, suggesting tolerance to input perturbations. Organizational coordination shows moderate sensitivity. Overall, increasing input perturbations mainly affect dimensions requiring detailed analysis or instantaneous behavioral responses, highlighting where stability protection should be prioritized.

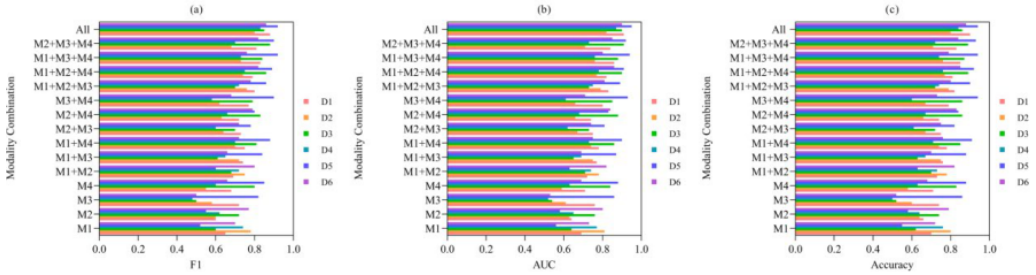

FIGURE 3. Sample-level comparison of F1, AUC, and accuracy of multimodal combinations in predicting six digital leadership dimensions. (a) F1 comparison; (b) AUC comparison; (c) Accuracy comparison

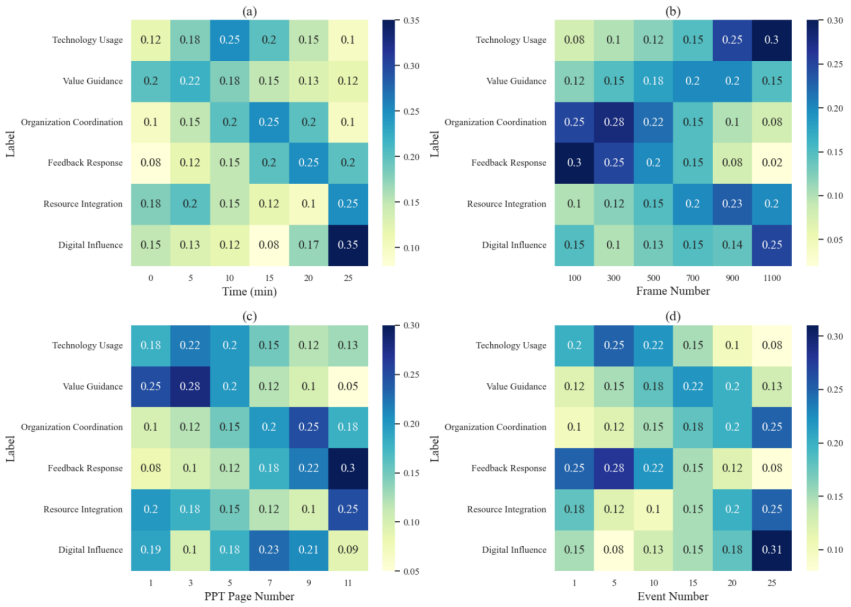

FIGURE 4. Sample-level averaged multimodal attention distribution heatmap across all teaching sessions. (a) Classroom voice attention distribution; (b) Video image attention distribution; (c) Courseware content attention distribution; (d) Platform interaction log attention distribution

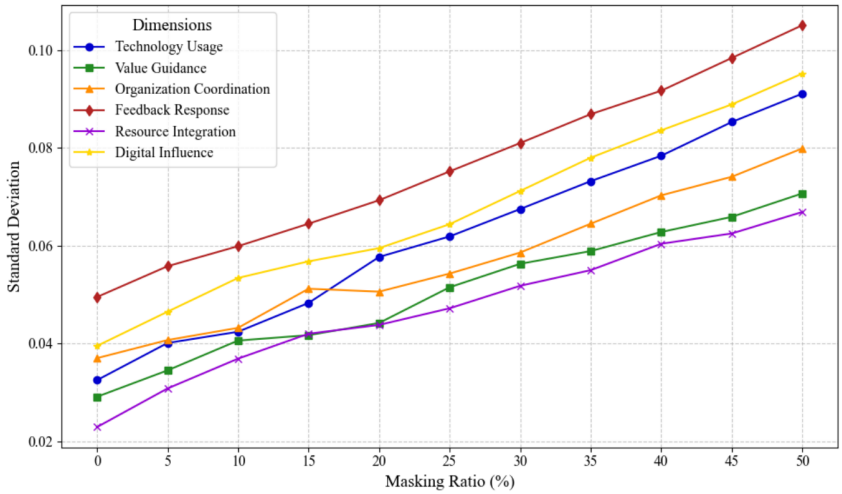

FIGURE 5. Stability of six-dimensional digital leadership scores under uniform masking perturbation

C. DYNAMIC EVOLUTION OF BEHAVIOR DURING THE TEACHING PROCESS

This section presents a case analysis of stage-wise digital leadership within individual teaching sessions. Teacher performance varies by teaching style and technology mastery. Teachers A, B, and C are randomly selected, with scores averaged over all behavioral segments of a ses-

sion. Teacher A excels in technology use and classroom interaction but is weaker in value guidance and feedback. Teacher B emphasizes interactive guidance and idea exchange but is less proficient in technology application and resource integration. Teacher C demonstrates balanced digital leadership across technology, coordination, and feedback. Analysis across teaching stages—introduction, explanation, discussion, and summary—reveals variations in technology use, value guidance, coordination, and feedback. Figure 6 shows normalized scores (0–1) of the three teachers across six digital leadership dimensions to illustrate intra-teacher temporal variation. In the introduction phase, Teacher A scores 0.9 in technology application, emphasizing the use of tools to engage students; Teacher B scores 0.85 in value guidance, focusing on guiding student thinking; Teacher C shows balanced performance across all dimensions. In the explanation phase, Teacher A maintains high technology application, while Teacher B leads in organizational coordination and feedback response, highlighting attention to interaction and student feedback; Teacher C remains balanced. During discussion, Teacher B scores 0.9 in coordination and feedback, controlling student discussions, while Teacher A's technology application drops to 0.8; Teacher C stays balanced. In the summary phase, Teacher A excels in technology application and digital influence, Teacher C in resource integration, and Teacher B in value guidance and feedback response. Overall, Teacher A prioritizes technology application, Teacher B emphasizes interaction and feedback, and Teacher C maintains balanced development across dimensions.

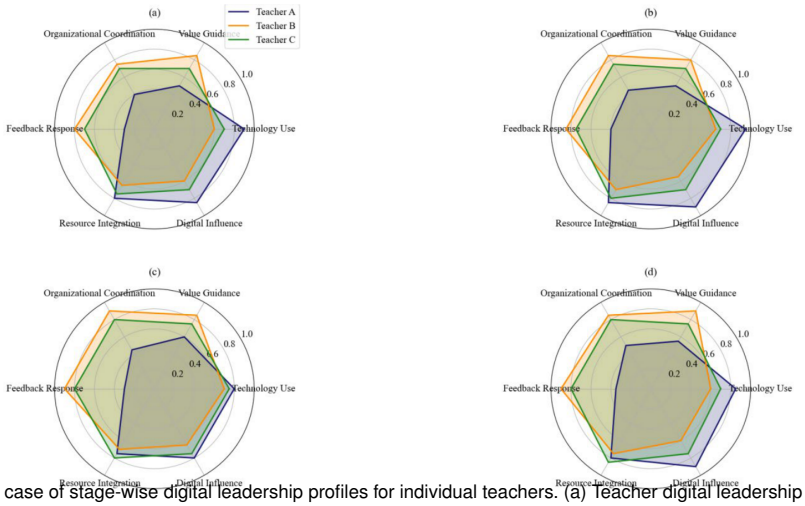

FIGURE 6. Illustrative case of stage-wise digital leadership profiles for individual teachers. (a) Teacher digital leadership performance in the introduction phase; (b) Teacher digital leadership performance in the explanation phase; (c) Teacher digital leadership performance in the discussion phase; (d) Teacher digital leadership performance in the summary phase

D. DIFFERENCES IN DIGITAL LEADERSHIP ACROSS TEACHER GROUPS

This section reports group-level differences in digital leadership across teacher experience and gender. New teachers (<5 years) are exploratory and experiment with diverse teaching tools but lack technology proficiency and coordination skills. Mid-career teachers (6–15 years) have strong teaching experience, organizational coordination, and emotional regulation, but are less innovative in digital technology use. Senior teachers with sixteen to twenty years of teaching experience focus on knowledge transfer and classroom quality but are conservative in their use of digital tools. Experienced teachers with more than twenty years of teaching experience combine deep instructional experience with advanced technical capabilities, demonstrating strong digital influence and adaptability in complex classroom environments. Figure 7 illustrates differences among teacher groups across digital leadership dimensions. Gender-based analysis is conducted within experience-aligned groups with consistent male-to-female sample proportions. All scores are rescaled to 0–10 for cross-group comparison.

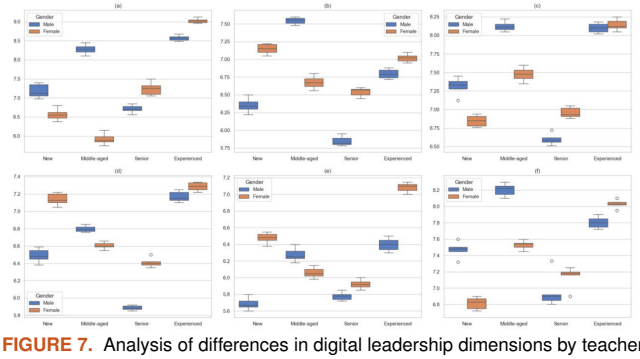

FIGURE 7. Analysis of differences in digital leadership dimensions by teacher category and gender. (a) Technology application; (b) Value guidance; (c) Organizational coordination; (d) Feedback response; (e) Resource integration; (f) Digital influence

Teachers show significant differences across digital leadership dimensions. Experienced teachers score higher in technology application, with females at 8.96–9.13 and males at 8.49–8.68, while new teachers score lower (males 6.98–7.40, females 6.38–6.80), indicating room for improvement. In organizational coordination, experienced males score 8.02–8.18 versus 7.45 for new teachers, reflecting

stronger coordination skills. Female teachers perform well in feedback response across both experience levels. In digital influence, experienced females reach 8.1, slightly above 7.9 for males, indicating greater teaching impact. Overall, seniority and gender significantly affect digital leadership performance, with experienced teachers performing more effectively.

VI. CONCLUSION

This paper introduces a multimodal behavioral analysis system based on the Perceiver IO architecture, designed to assess university teachers' digital leadership. By integrating multimodal data—including classroom interactions, digital tool usage, voice, video, courseware content, and platform logs—the system provides a comprehensive and contextually relevant evaluation. Experimental results show that this multimodal fusion approach improves assessment accuracy and behavioral pattern interpretation, achieving F1 scores of 0.88, 0.80, and 0.86 in technology application, value guidance, and digital influence, respectively. In practice, the system offers precise feedback for enhancing teachers' digital leadership, supporting personalized teaching improvements, and helping educators identify strengths and weaknesses in areas such as CAD/CAM integration, and leadership of digital production workflows. Ultimately, it contributes to developing digitally fluent, industry-ready graduates.

AUTHOR CONTRIBUTIONS

Conceptualization – Xin Wang and Biyi Jiang; methodology – Xin Wang and Biyi Jiang; investigation – Xin Wang; writing-original draft preparation – Xin Wang and Biyi Jiang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by the research results of the Heilongjiang Higher Education Teaching Reform

Project on "Digital Ideology and Politics" in 2022: "Research on the Effectiveness of Deep Learning in Ideological and Political Theory Courses of College Students". (Project No. SJGSZD2022001)

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. A. Kelder, J. Crawford, I. Al Naabi, and L. To, "Enhancing digital productivity and capability in higher education through authentic leader behaviors: A cross-cultural structural equation model," Education and Information Technologies, vol. 1, no. 1, pp. 1-17, 2025, doi: 10.1007/s10639-025-13422-x.

[2] H. Antonopoulou, P. Matzavinou, I. Giannoukou, and C. Halkiopoulos, "Teachers' Digital Leadership and Competencies in Primary Education: A Cross-Sectional Behavioral Study," Education Sciences, vol. 15, no. 2, pp. 1-16, 2025, doi: 10.3390/educsci15020215.

[3] T. C. Yang, "Assessment of the digital competencies of university instructors through use of the machine learning method," SN Social Sciences, vol. 3, no. 2, pp. 25-41, 2023, doi: 10.1007/s43545-023-00617-7.

[4] F. Tatari, M. Ebrahimi, and H. Tadayon, "Design, implementation, and evaluation of faculty member evaluation process using electronic 360-degree method," BMC Medical Education, vol. 25, no. 1, pp. 941-953, 2025, doi: 10.1186/s12909-025-07384-4.

[5] N. H. Hamzah, N. M. Radzi, and I. M. Omar, "Digital leadership competencies through a systematic literature review: bridging educational development and organizational success," Education, vol. 7, no. 24, pp. 37-64, 2025, doi: 10.35631/ijmoe.724003.

[6] E. Karimi, "Examining The Influence of Transformational Leadership on Digital Integration in Schools: The Role of Digital Competencies," Journal of Educational Leadership and Management, vol. 3, no. 1, pp. 90-105, 2024.

[7] M. Kamran, "Educational Leadership in a Digital Era: Navigating Challenges and Opportunities," AL-ĪMĀN Research Journal, vol. 3, no. 01, pp. 188-199, 2025, doi: 10.63283/54wxa275.

[8] M. Jing, Z. Guo, X. Wu, Z. Yang, and X. Wang, "Higher education digital academic leadership: Perceptions and practices from Chinese University Leaders," Education Sciences, vol. 15, no. 5, pp. 606-633, 2025, doi: 10.3390/educsci15050606.

[9] R. Pu, R. K. Dong, and S. Jiang, "Toward the Education for Sustainable Development (ESD): Digital leadership and knowledge-sharing behavior on the higher education institutional change," Education and Information Technologies, vol. 3, no. 1, pp. 1-23, 2024.

[10] Z. Cheng, A. Caliskan, N. B. K. Dinh, and C. Zhu, "A systematic review of digital academic leadership in higher education," International Journal of Higher Education, vol. 13, no. 4, pp. 38-52, 2024, doi: 10.5430/ijhe.v13n4P38.

[11] J. Jameson, N. Rumyantseva, M. Cai, M. Markowski, R. Essex, and I. McNay, "A systematic review and framework for digital leadership research maturity in higher education," Computers and Education Open, vol. 3, no. 1, pp. 100115-100142, 2022, doi: 10.1016/j.caeo.2022.100115.

[12] T. Karakose, H. Polat, T. Tülübaş, and M. Demirkol, "A review of the conceptual structure and evolution of digital leadership research in education," Education Sciences, vol. 14, no. 11, pp. 1166-1185, 2024, doi: 10.3390/educsci14111166.

[13] T. Karakose and T. Tülübaş, "Digital leadership and sustainable school improvement-a conceptual analysis and implications for future research," Educational Process: International Journal, vol. 12, no. 1, pp. 7-18, 2023, doi: 10.22521/edupij.2023.121.1.

[14] N. M. Ghamrawi and R. Tamim, "A typology for digital leadership in higher education: The case of a large-scale mobile technology initiative (using tablets)," Education and Information Technologies, vol. 28, no. 6, pp. 7089-7110, 2023, doi: 10.1007/s10639-022-11483-w.

[15] T. Xu, W. Deng, S. Zhang, Y. Wei, and Q. Liu, "Research on recognition and analysis of teacher-student behavior based on a blended synchronous classroom," Applied Sciences, vol. 13, no. 6, pp. 3432-3454, 2023, doi: 10.3390/app13063432.

[16] Z. Xuan, "DRN-LSTM: a deep residual network based on long short-term memory network for students behaviour recognition in education," Journal of Applied Science and Engineering, vol. 26, no. 2, pp. 245-252, 2022.

[17] J. Mo, R. Zhu, H. Yuan, Z. Shou, and L. Chen, "Student behavior recognition based on multitask learning," Multimedia tools and applications, vol. 82, no. 12, pp. 19091-19108, 2023, doi: 10.1007/s11042-022-14100-7.

[18] G. Tuerhong, F. Fu, and M. Wushouer, "Adaptive multimodal transformer based on exchanging for multimodal sentiment analysis," Scientific Reports, vol. 15, no. 1, pp. 27265-27278, 2025, doi: 10.1038/s41598-025-11848-4.

[19] Y. Shi, J. Cai, and L. Liao, "Multi-task learning and mutual information maximization with crossmodal transformer for multimodal sentiment analysis," Journal of Intelligent Information Systems, vol. 63, no. 1, pp. 1-19, 2025, doi: 10.1007/s10844-024-00858-9.

[20] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," Advances in neural information processing systems, vol. 34, no. 1, pp. 14200-14213, 2021.

[21] S. Nindam, S. H. Na, and H. J. Lee, "MultiFusedNet: A Multi-Feature Fused Network of Pretrained Vision Models via Keyframes for Student Behavior Classification," Applied Sciences, vol. 14, no. 1, pp. 230-250, 2023, doi: 10.3390/app14010230.

[22] G. Li, F. Liu, Y. Wang, Y. Guo, L. Xiao, and L. Zhu, "A convolutional neural network (CNN) based approach for the recognition and evaluation

of classroom teaching behavior," Scientific Programming, vol. 2021, no. 1, pp. 6336773-6336781, 2021, doi: 10.1155/2021/6336773.

[23] J. Hewstone and R. Araya, "Neural network-based approach to detect and filter misleading audio segments in classroom automatic transcription," Applied Sciences, vol. 13, no. 24, pp. 13243-13265, 2023, doi: 10.3390/app132413243.

[24] L. Zhou, X. Liu, X. Guan, and Y. Cheng, "CSSA-YOLO: Cross-Scale Spatiotemporal Attention Network for Fine-Grained Behavior Recognition in Classroom Environments," Sensors, vol. 25, no. 10, pp. 3132-3163, 2025, doi: 10.3390/s25103132.

[25] K. Fatehi, M. T. Torres, and A. Kucukyilmaz, "An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments," Speech Communication, vol. 167, no. 1, pp. 103151-103181, 2025, doi: 10.1016/j.specom.2024.103151.

[26] X. Qi, Y. Wen, P. Zhang, and H. Huang, "MFGCN: Multimodal fusion graph convolutional network for speech emotion recognition," Neurocomputing, vol. 611, no. 1, pp. 128646-128662, 2025, doi: 10.1016/j.neucom.2024.128646.

[27] J. Lei, J. Wang, and Y. Wang, "Multi-level attention fusion network assisted by relative entropy alignment for multimodal speech emotion recognition," Applied Intelligence, vol. 54, no. 17, pp. 8478-8490, 2024, doi: 10.1007/s10489-024-05630-8.