Construction and Empirical Study of Digital Literacy Evaluation Model for Young Vocational Education Teachers Based on Multimodal Data

Shiying Yang[1], Jun Liu[1], Qi Dang[2]
[1]Office of Academic Affairs, Shaanxi Technical College of Finance and Economics, Xianyang 712000, Shaanxi, China
[2]School of Big Data and Artificial Intelligence, Shaanxi Technical College of Finance and Economics, Xianyang 712000, Shaanxi, China
Corresponding author: Jun Liu (e-mail: zhuanyun1007@163.com)

ABSTRACT With the increasing deployment of intelligent teaching environments and wireless digital infrastructures, efficient multimodal information acquisition and transmission have become essential for objective assessment in vocational education. To overcome the subjectivity and limited dimensionality of conventional evaluation methods, this study proposes a digital literacy evaluation model for young vocational education teachers based on a multimodal Transformer fusion network. Semantic representations of instructional texts are extracted using BERT, spatiotemporal behavioral features from classroom videos are learned through 3D Convolutional Neural Networks, and emotional acoustic characteristics are obtained from speech signals using OpenSMILE and temporal modeling. A cross-modal attention mechanism constrained by the Technological Pedagogical Content Knowledge (TPACK) framework is introduced to integrate heterogeneous information and dynamically optimize feature weighting across instructional scenarios. An end-to-end multi-task prediction architecture subsequently generates quantitative evaluations for six digital literacy dimensions with enhanced interpretability. Experimental results demonstrate approximately 92% classification accuracy on the test set, while quantitative scores for all six secondary dimensions exceed 85.5 in both theoretical instruction and practical training environments. The proposed framework establishes a high-precision and objective assessment methodology for teacher digital literacy and provides an effective engineering solution for multimodal data fusion, intelligent educational sensing, and distributed information processing. Furthermore, its multimodal perception and communication framework offers valuable insights for electromagnetic-enabled smart education systems, wireless sensing platforms, and next-generation intelligent information transmission technologies.

INDEX TERMS multimodal data fusion, digital literacy evaluation, Transformer network, intelligent sensing, electromagnetic-enabled education systems

I. INTRODUCTION

DIGITAL transformation is now a key driver for vocational education reform worldwide. As the primary young teachers in digital teaching play a decisive role in the process of talent cultivation and industry-fitting based on their digital literacy, establishing a reasonable evaluation system can not only play a significant guiding role in identifying teachers' deficiencies and offering theoretical support and implementation guidance for improving teacher training methods and enhancing vocational education effectiveness [1-2]. Under the impetus of policy and technology, exploring the composition of young teachers' digital literacy and evaluation mechanism in vocational education offers a critical methodological advancement in promoting the high-quality development of vocational education [3-4].

Most of the existing researches depend on self-report questionnaires or subjective scores from experts as its only data source. This kind of subjectivity and single-source data can't reflect the dynamic behavior feature of teachers in complex and real teaching scenes [5-6]. The technical challenges of Multimodal data fusion include significant feature distribution divergence across heterogeneous information streams and spatiotemporal alignment problem. Therefore, the mining of data value is insufficient [7-8]. There is no unified collection and processing specification for unstructured data such as practical training operation and teacher-student interaction in vocational education, which makes it difficult to quantify evaluation index [9-10]. In addition, traditional statistical method has less feature extraction ability in high-dimensional nonlinear and multimodal data, and cannot capture fine-grained changes in teaching behavior and emo-

tional interaction clues [11-12]. Traditional rule-based evaluation models depend on manually designed weights and have obvious subjectivity and lack of adaptability. Support vector machine and other shallow machine learning algorithms have poor generalization performance when dealing with high-dimensional sparse multimodal features, and it is difficult to model the complex nonlinear relationship between different modalities. Traditional early splicing or late voting strategies of multimodal fusion strategy ignored the information complementarity and contextual relationship between different modalities, causing key features to be buried in noise. Although existing deep learning models have achieved good performance in single-modal tasks, the cross-modal semantic alignment and attention allocation mechanism is still very coarse, and it is difficult to map the literacy dimension in specific vocational education scene. In contrast to traditional evaluation methods that rely on self-report questionnaires, the model proposed in this study is designed to completely replace these subjective instruments. By integrating objective multimodal data—comprising teaching texts, classroom videos, and speech signals—the framework eliminates personal bias and provides an evidence-based quantification of digital literacy [13-14].

In order to solve the problem of semantic alignment difficulty of multimodal data and evaluation subjectivity, this paper constructs a multi-head attention based multimodal Transformer fusion network. This paper uses BERT model to realize deep semantic encoding of lesson plan text, uses 3D convolutional neural network to extract spatiotemporal action features of classroom video, uses OpenSMILE toolkit to analyze affective acoustic parameters of speech signal. This paper introduces cross-modal attention module to calculate correlation matrix of text, video and audio feature, and dynamically change the weight of each modality to suppress noise interference. Finally input the fused high-order features into fully connected layer for nonlinear mapping, outputting the score of each dimension of digital literacy. This research solves the problem of evaluation based on single data source, establishes a data-driven objective evaluation framework, and provides new technical route and methodological basis for accurate assessment of vocational education teachers' digital literacy.

II. EVALUATION MODEL CONSTRUCTION BASED ON MULTIMODAL TRANSFORMER

A. MULTIMODAL DATA PREPROCESSING AND FEATURE EXTRACTION MECHANISM

The differences in original signals from multimodal data in terms of dimension, scale and semantic space cause great differences in the original signal. If these signals are directly input into the model, it will lead to the problem of gradient explosion or feature overload [15-16]. Establishing a unified feature representation space is an indispensable process before evaluating the effective operation of the model. This paper uses mature deep neural network to encode three types of core data text, video and audio, so that the feature vector extracted by the above method has high discriminative ability and semantic information [17-18]. As shown in Figure 1, the overall structure divides the independent extraction path and final fusion interface of each modality very clear.

Figure 1 shows that the architecture is divided into three parts: input layer, feature extraction module and fusion interface. After word segmentation and padding, the text branch inputs the BERT encoder to generate semantic vectors. The spatiotemporal action feature of video branch is extracted by frame sampling and C3D (Convolutional 3D Neural Network). The audio branch extracts acoustic parameters by OpenSMILE and models temporal dynamics by LSTM (Long Short-Term Memory). Finally, the output of each branch is mapped to the unified space of 512 dimensions by linear projection layer, and then enters the cross-modal attention mechanism to calculate the mapping relationship of query key-value. The information complementarity is obtained by weighted summation and residual connection. Finally, the feature is output after layer normalization. This paper ensures the accurate end-to-end quantitative evaluation of heterogeneous data based on semantic alignment.

For teaching text data, which mainly includes lesson plans, teaching reflection logs and online discussion logs, due to the long sequence and high context dependence of this type of data, this paper chooses a pretrained language model as the text encoder. After word segmentation, the input text will be fed into the multi-layer Transformer structure to get bidirectional context encoding, and then output vector corresponding to the sequence start marker is chosen as the global semantic representation of the whole text, which integrates semantic information of lexicon and sentence level. To fit the subsequent fusion network, the high dimensional output vector is rescaled to a same dimension by linear projection layer. This process can also capture the logical structure and terminology used by teachers in digital instructional design.

For classroom video data, the research focus is on the logical information of teachers' body language, use of teaching aids and teacher-student interactions. Traditional three-dimensional convolutions can only capture spatial information, but can't capture temporal information. Therefore, this paper chooses three-dimensional convolutional neural network to extract the spatiotemporal feature of classroom video. The video is cut into fixed length segments, each of which contains consecutive frames. The three-dimensional convolutional neural network outputs the spatiotemporal feature maps by performing convolution in the height and width of each video simultaneously in three dimensions. Connected to the backend of network, global average pooling layer compresses the multidimensional feature maps into one-dimensional feature vector. To translate these physical actions into literacy scores, the model utilizes a predefined mapping library where specific spatiotemporal features—such as the precision of tool manipulation and the frequency

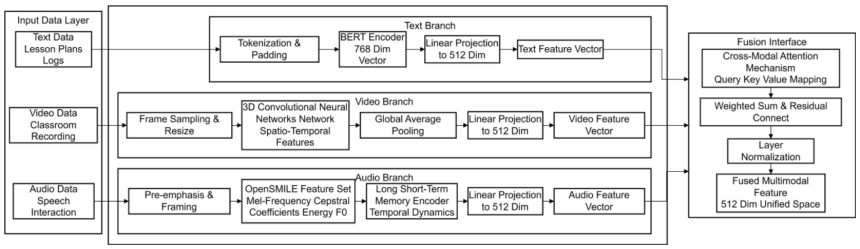

FIGURE 1. Overall architecture of the multimodal Transformer evaluation model

of digital interface interactions—are weighted against specific literacy indicators like "Operational Standardization" and "Digital Tool Proficiency." This vector quantifies the operational standardization and dynamic interaction times of teachers in the training process, reflecting the teacher's ability to apply digital technology in "learning by doing" scenario.

For audio data, the research focus is on the teacher's intonation, emotion and clarity. Due to the nonstationary characteristics of speech signal, it is necessary to extract the time-frequency domain feature. This paper chooses an open-source toolset to extract acoustic feature, including low level descriptor such as fundamental frequency, energy and Mel-frequency cepstral coefficients [19-20]. Then choose long short term memory network to model the sequential acoustic feature and capture the dynamic change of speech emotion. The final hidden state of the Long Short-Term Memory (LSTM) network is used as the final sentiment representation vector of the audio segment. This feature reflects the teacher's emotional engagement and communicative impact during the digital teaching process.

B. DESIGN OF CROSS-MODAL ATTENTION FUSION MECHANISM

Single-modal features can only reflect a local aspect of teachers' digital literacy and cannot capture the deep semantic relationships between multi-source information [21-22]. Cross-modal attention mechanisms achieve information complementarity and enhancement by dynamically calculating the correlation between different modal features. The core of this mechanism lies in constructing query, key, and value vectors, and generating weight matrices through a scaled dot product attention algorithm, thereby guiding the model to focus on the modal combination that contributes the most to the evaluation results [23-24].

Let the feature vectors of text, video, and audio be H_t, H_v, and H_a, respectively, all with dimension d. To eliminate the difference in dimensions and introduce nonlinear transformation, each modal feature is first mapped to the same latent space through a linear projection layer, generating query vector Q, key vector K, and value vector V. As shown in equation (1), the projection process is achieved through learnable weight matrices W_Q, W_K, and W_V:

$$Q = H_t W_Q, \quad K = H_v W_K, \quad V = H_a W_V \tag{1}$$

$W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are the projection matrices. Here, text is used as the query subject, and video and audio are used as context references, aiming to explore the specific manifestation of text semantics in audiovisual behavior [25-26]. This asymmetric attention design conforms to the logical law of "instructional design guiding teaching behavior" in vocational education. Based on the generated Q, K, and V, cross-modal attention weights are calculated [27-28]. As shown in equation (2), the scaling dot product attention mechanism is adopted, the transpose of Q and K is multiplied by the dot product operation, and divided by the scaling factor$\sqrt{d}$ to prevent gradient vanishing. Then, the attention distribution matrix A is obtained by normalization through the Softmax function:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{2}$$

Each element a_{ij} in matrix A represents the association strength between the i-th text semantic unit and the j-th audiovisual behavior unit [29-30]. A high weight means that the audiovisual behavior closely supports the corresponding instructional design concept and reflects the consistency of the teacher's digital literacy; a low weight may indicate that the instructional behavior is out of sync with the design or that there is noise interference [31-32]. The cross-modal fusion feature H_{fusion} is obtained by weighting and summing the value vector V using the attention weight matrix A. As shown in equation (3), the fusion feature not only retains the information of the original modality, but also incorporates the cross-modal interaction context:

$$H_{fusion} = AV \tag{3}$$

To further enhance the model's expressive power, residual connections and layer normalization are introduced. As shown in equation (4), the fused features are added to the original query features and then normalized to obtain the final output representation H_{out}:

$$H_{out} = \text{LayerNorm}(H_{fusion} + H_t) \tag{4}$$

Residual connections ensure the effective propagation of gradients in deep networks, avoiding information loss; layer normalization stabilizes feature distribution and accelerates model convergence [33-34].

C. DIGITAL LITERACY QUANTITATIVE SCORING MAPPING STRATEGY

High-order feature vector H_{out} is rich in semantic and behavioral interaction information after cross-modal attention fusion. However, its dimension is high and abstract, and cannot be used as evaluation results directly. Therefore, this study designs MLP as scoring mapping module to nonlinearly map fused features into specific digital literacy dimension space, and then completing the latent feature space to interpretable quantitative indicator space mapping, making evaluation results conform to vocational education and based on data. The mapping network includes two fully connected layers, and ReLU (Rectified Linear Unit) activation function is added in the middle of two layers. The first fully connected layer compresses the input feature dimension from d to the hidden layer dimension h, aiming to extract the core elements that are highly related to literacy evaluation and suppress redundant noise [35-36]. The second fully connected layer maps the hidden layer features to the output dimension C, where C represents the number of secondary indicators in the digital literacy evaluation index system. As shown in equation (5), the final output scoring vector Y is normalized by the Softmax function to ensure that the sum of the scores of each dimension is 1, which is convenient for subsequent conversion into a percentage system or grade system:

$$Y = \text{Softmax}\left(W_2 \cdot \text{ReLU}(W_1 \cdot H_{out} + b_1) + b_2\right) \tag{5}$$

W_1 and W_2 are learnable weight matrices, and b_1 and b_2 are bias terms. The application of the Softmax function gives the output values probabilistic meaning, intuitively reflecting the teacher's relative performance level in each competency dimension. To meet practical evaluation needs, the normalized probability values are multiplied by 100 to obtain specific scores in the 0-100 range.

This study extracts the response intensity values of text, video unimodal features, and multimodal fusion features on six core competency dimensions. By comparing and analyzing the differences in capturing complex teaching behaviors from different data sources, it evaluates the ability of cross-modal attention mechanisms to mine the complementarity of heterogeneous information, resulting in the findings shown in Figure 2.

Figure 2 shows that, in the teaching implementation dimension, the response strength of the multimodal fusion feature reached 0.92, which is higher than that of the text monomodal feature (0.60) and the video monomodal feature (0.80). In the social responsibility dimension, the fusion feature score improved to 0.75, which far exceeded that of the text monomodal feature (0.50) and the video monomodal feature (0.60). The large jump in this performance is due to the cross-modal attention mechanism successfully calibrating the operational norms in video stream with the instructing semantic information in audio stream. That is, avoiding the single context understanding or action recognition based on one data source context, and then allow model to evaluate the teacher in a comprehensive way in complicated training process. This demonstrates the technical superiority of multimodal fusion method in extracting high-dimensional nonlinear literacy feature.

D. MODEL OPTIMIZATION AND LOSS FUNCTION DESIGN

To ensure the convergence speed and generalization ability of the multimodal Transformer fusion network in the digital literacy assessment task, this study designed a composite loss function and adopted an adaptive optimization strategy [37-38]. Since digital literacy assessment has both classification (level determination) and regression (specific score) characteristics, a single loss function is difficult to simultaneously take into account the clarity of the class boundary and the accuracy of numerical prediction [39-40]. Therefore, a weighted composite objective function consisting of cross-entropy loss and mean squared error loss is constructed. Cross-entropy loss is used to constrain the model's discrimination boundary for literacy level division and enhance the inter-class separation degree in the feature space; mean squared error loss is used to minimize the Euclidean distance between the predicted score and the expert-annotated true value to ensure the granularity of the quantification results [41-42]. As shown in formula (6), the total loss L_{total} is defined as the linear sum of the two losses:

$$L_{total} = \lambda_1 L_{CE} + \lambda_2 L_{MSE} \tag{6}$$

Wherein, L_{CE} represents cross-entropy loss, L_{MSE} represents mean squared error loss, and λ_1 and λ_2 are balance coefficients, which were determined to be 0.6 and 0.4 respectively through grid search experiments, to emphasize the accuracy of grade determination while retaining the regression accuracy of the score. This composite loss mechanism effectively alleviates the gradient vanishing problem, enabling the model to quickly locate the optimal solution region in the early stage of training, and finely adjust the weights in the later stage to fit complex nonlinear relationships [43-44].

In terms of optimization algorithm selection, the AdamW optimizer is used instead of the traditional Adam algorithm. AdamW avoids the failure of L2 regularization under adaptive learning rate by decoupling the weight decay and gradient update process, thereby improving the generalization performance of the model [45]. With the cosine annealing learning rate scheduling strategy, the learning rate decays periodically with the number of iterations in a cosine curve. As shown in formula (7), the learning rate η_t of the current Epoch is calculated as follows:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{cur}}{T_{\max}}\pi\right)\right) \tag{7}$$

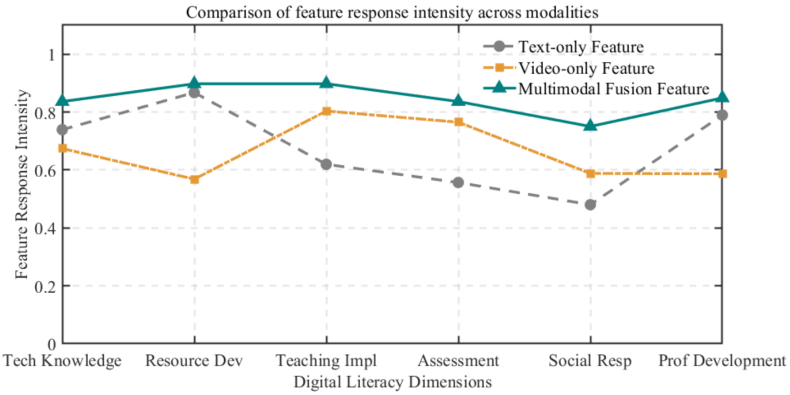

FIGURE 2. Comparison of literacy dimension distribution before and after multimodal feature fusion

Where $\eta_{\max}$ and $\eta_{\min}$ are the upper and lower bounds of the learning rate, respectively, T_{cur} is the current iteration step, and $T_{\max}$ is the maximum iteration period. This strategy allows the model to explore the global optimum with a larger step size in the early stages of training, and then fine-tune local minima with a smaller step size in the later stages, effectively avoiding getting trapped in local optima.

III. EXPERIMENTAL DESIGN AND DATA SOURCES

A. DATASET CONSTRUCTION AND SAMPLE SELECTION

This study randomly selected three high-level vocational colleges in western China as data collection bases. Considering the diversity and representativeness of the sample, three professional clusters of equipment manufacturing, electronic information and finance and commerce are selected respectively. Research subjects are defined as full-time young teachers who meet the dual criteria of having less than five years of professional service and being 35 years of age or younger. This standardized boundary ensures consistency with common academic and policy definitions of "young" practitioners in vocational education. This group of teachers is facing the stage of career adaptation and development, and the research value of their digital literacy development process is typical. Time period is one semester (16 weeks). Various teaching situations are considered, including theoretical courses, practical training guidance and online tutoring, etc. Multimodal data collection system deployed in intelligent classroom is used to collect classroom teaching video, audio and teacher operation log simultaneously. At the same time, the corresponding lesson plan text, teaching reflection log and student evaluation data are mined from school teaching management platform. All data are strictly anonymized. Delete sensitive personal information such as name and employee number, retain only behavioral characteristic tags related to research, and meet the requirements of ethical review.

The paper deletes the invalid segments, including unclear video, missing audio and segments less than 20 minutes; the data interruption caused by equipment failure is deleted; the teacher sample who is absent for more than two weeks during the experiment is deleted. Total 1248 valid multimodal teaching samples are obtained, involving 156 young teachers meeting these refined criteria, 78 males and 78 females, with a balanced number of males and females. In terms of age distribution, 26.9% are 22-25 years old, 57.1% are 26-30 years old, and 16.0% are 31-35 years old. In terms of teaching experience, 19.9% are less than 1 year, 43.6% are 1-3 years, and 36.5% are 3-5 years. The covered professional categories are engineering (39.7%), humanities (30.1%) and arts (30.2%). The structure of professional titles is mainly teaching assistant (60.3%) and lecturer (39.7%).

B. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

The hardware platform is configured as follows: two NVIDIA A100 80GB GPU accelerators are used to parallel processing high-dimensional multimodal data; the central processor uses Intel Xeon Gold 6248R@3.0GHz; 512GB DDR4 memory is used to guarantee the batch processing of large-scale data. The software environment is based on the Ubuntu 20.04 operating system. PyTorch 1.12.0 deep learning framework is used to build the model. CUDA 11.3 acceleration library is used to accelerate the low-level computation process. This paper load the pre-trained model BERT-Base and weight of C3D network from the official repository to ensure the fairness of the feature extraction baseline.

In the training process, this paper uses end-to-end supervised learning strategy. This paper chooses AdamW algorithm as the optimizer, and initializes the learning rate as 1×10^{-4}. It introduces cosine annealing scheduling strategy to adjust the learning rate of the model. This strategy can make the model reach a relatively fast convergence speed and a better search ability of global optimal solution. The batch size is set to 32 to balance memory usage and gradient estimation accuracy. The number of training iterations is set to 50, and the Dropout ratio is uniformly set to 0.3 to enhance the model's generalization ability. The weight decay coefficient is set to 1×10^{-2}, and the gradient clipping threshold is set to 1.0 to prevent gradient explosion.

IV. EMPIRICAL RESEARCH

A. MODEL CONVERGENCE AND STABILITY ANALYSIS

To verify the convergence characteristics and robustness of the multimodal Transformer fusion network during training, this study iteratively trained the model for 50 epochs and independently repeated the experiment five times to record the accuracy change trajectory on the test set for each run. By monitoring the fluctuations of the loss function values on the training and validation sets in real time and comparing the overlap of the accuracy curves under multiple independent runs, the search efficiency and stability of the optimization algorithm in the heterogeneous feature space are evaluated, resulting in the model training convergence and stability analysis results shown in Figure 3.

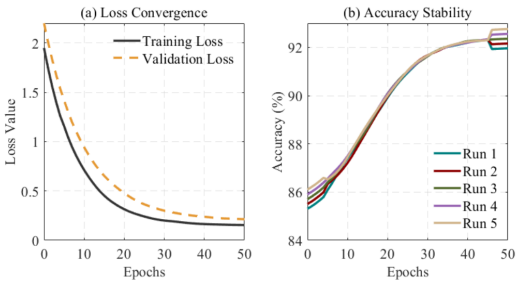

FIGURE 3. Model training convergence and stability analysis: (a) Loss convergence curve; (b) Accuracy stability comparison

As shown in Figure 3(a), the training loss value rapidly decreases to approximately 0.7 at the 10th epoch, then smoothly converges to approximately 0.15 at the 50th epoch, verifying that the loss value decreases monotonically. This smooth and oscillating decline curve is attributed to the AdamW optimizer combined with the cosine annealing learning rate scheduling strategy effectively avoiding local optima traps. Simultaneously, the cross-modal attention mechanism suppresses interference from early noise features, ensuring the accuracy of gradient update directions. As shown in Figure 3(b), the five independent experiments accuracy curve presents an S-shaped growth with the increase of iteration times, and it reaches about 92% at the 50th epoch finally. The extremely small dispersion also indicates that the complementarity of multimodal features enhances the robustness of model to small perturbation.

B. CASE STUDY ON ADAPTABILITY OF THE EVALUATION MODEL TO SPECIFIC VOCATIONAL EDUCATION SCENARIOS

To further verify the adaptability and differentiation recognition ability of the evaluation model in typical vocational education teaching scenario, this paper selects two typical kinds of core courses, namely theoretical teaching instruction and practical training guidance. For these two scenarios, the multimodal data is collected and input into trained Transformer fusion network for independent evaluation separately. Based on the distribution of quantitative score of teachers on six secondary dimensions of digital literacy, it is explored that whether the model can distinguish the typical "work-study integration" behavioral characteristics of vocational education, such as operational standardization and immediate feedback mechanism of practical training, and get Figure 4.

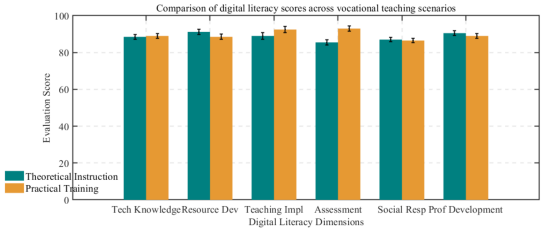

FIGURE 4. Digital literacy scores in different vocational education teaching scenarios

Figure 4 shows that in the practical training scenario, the teaching implementation dimension scored 92.5 points, and the assessment and diagnosis dimension scored 93.0 points, higher than the 89.0 and 85.5 points of the theoretical teaching scenario, respectively. This difference in scores is attributed to the cross-modal attention mechanism's weighted reinforcement of frequently occurring tool usage actions, safety standard checks, and real-time teacher-student interactions in the video stream, enabling the model to accurately identify high-weight skill operation features in the practical training process. In the resource development dimension, the theoretical teaching scenario scored 91.2 points, slightly higher than the practical training scenario's 88.5 points. This is due to the increased contribution of the lesson plan structure and multimedia courseware logic density in the text modality, proving that the model dynamically adjusts the modal weights according to the scenario context, achieving refined adaptation and objective evaluation of diversified teaching forms in vocational education.

C. KEY FEATURE CONTRIBUTION ANALYSIS

To reveal the internal decision-making logic of the multimodal Transformer fusion network and enhance the interpretability of the evaluation results, this study extracts the weight parameters in the cross-modal attention mechanism and quantitatively analyzes the contribution of each underlying behavioral feature to the final digital literacy score. By statistically analyzing the mean and standard deviation of each feature weight in five independent experiments, the key indicators that have the most significant impact on model prediction are selected, as shown in Table 1.

TABLE 1. Statistical analysis of the contribution weights of key multimodal behavioral features

Key Behavior Feature	Modality	Mean Attention Weight	Std. Dev.
Tool Operation Standardization	Video	0.852	0.031
Instructional Clarity	Audio	0.798	0.042
Safety Protocol Adherence	Video	0.765	0.038
Interaction Frequency	Video/Audio	0.712	0.055
Resource Structure Logic	Text	0.689	0.047
Feedback Timeliness	Audio	0.643	0.051
Digital Tool Proficiency	Video	0.598	0.062
Reflection Depth	Text	0.521	0.058

Table 1 shows that the standardization of tool operation ranks first with an average attention weight of 0.852, followed by the clarity of instructions and compliance with safety procedures with weights of 0.798 and 0.765, respectively. The weight distribution of above pattern is due to the inherent attribute of "work-study integration" of vocational education, that is, the model can complete the automatic allocation of higher priority for the modality of video and audio reflecting practical skills and occupational safety consciousness in feature fusion, and the resource structure logic and reflection depth of text modality are 0.689 and 0.521, which means that although a certain basis of instructional design is indispensible, in the evaluation of dynamic teaching scenario, the real-time interactive behavior and operation standardization play more decisive role in differentiating the level of digital literacy, and it also indicates that the model focuses on the accurate extraction of vocational education-specific behavioral characteristic, and the evaluation result has high robustness.

D. CONSISTENCY ANALYSIS OF EVALUATION RESULTS AND EXPERT SCORES

To verify the validity and objectivity of the output results of the multimodal Transformer evaluation model, this study conducted a paired correlation analysis between the quantitative scores generated by the model and the average scores of independent blind reviews by three domain experts. By calculating the Pearson correlation coefficient and plotting the scatter plot, the consistency between the automatic machine evaluation and human subjective judgment in terms of numerical trends was examined, resulting in the results shown in Figure 5.

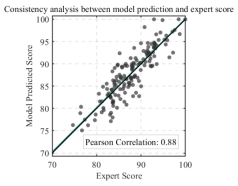

FIGURE 5. Consistency between model prediction scores and expert human scores

Figure 5 shows that the scatter points are closely clustered around the regression fitted line, which highly coincides with the $y = x$ reference line representing ideal consistency, with a coefficient of determination of 0.88. This high linear correlation is attributed to the cross-modal attention mechanism successfully aligning multi-source heterogeneous features such as operational norms in videos, instructional logic in audio, and instructional design in text. This enables the model to achieve a semantic understanding of complex teaching behaviors that approaches the cognitive level of human experts, effectively eliminating evaluation biases caused by single-modal data lacking context or limited perspective. This establishes the reliability and scientific basis for this automated evaluation tool to replace subjective human judgment in vocational education scenarios.

V. CONCLUSION

This study overcomes the problems of single data source and subjective bias in the evaluation of digital literacy of young vocational education teachers. Constructs multimodal Transformer fusion network based on multi-head attention mechanism, in which BERT is used to extract text semantics, 3D-CNN is used to extract spatiotemporal behavior in video, and OpenSMILE combined with LSTM is used to extract emotional feature of audio. The whole process of cross-modal attention mechanism realizes the alignment of heterogeneous data and eliminates semantic gap. The study has verified that the model can accurately capture vocational education-specific behavioral characteristic and has high robustness.

AUTHOR CONTRIBUTIONS

Conceptualization –Shiying Yang, Jun Liu and Qi Dang; methodology – Shiying Yang, Jun Liu and Qi Dang; investigation – Shiying Yang, Jun Liu and Qi Dang; writing-original draft preparation – Shiying Yang, Jun Liu and Qi Dang. 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 work was supported by Shaanxi Provincial Department of Education 2023 Education and Teaching Reform Project "Research on Teacher Development and Evaluation in Higher Vocational Colleges under the Context of Digital Transformation" (No. 23GZ011);

Shaanxi Provincial "14th Five-Year Plan" Education Science Planning Project 2025 "Research and Practice on Enhancing Digital Literacy of Young Teachers in Vocational Education" (No. SGH25Y3790);

Shaanxi Provincial Department of Education 2021 Education and Teaching Reform Project "Research on the Development Mechanism of Scale, Structure, Quality, and Efficiency of Higher Vocational Education in Shaanxi" (No. 21GG003)

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Wahjusaputri and I. Nastiti T, "Digital literacy competency indicator for Indonesian high vocational education needs," Journal of Education and Learning (EduLearn), vol. 16, no. 1, pp. 85-91, 2022, doi: 10.11591/edulearn.v16i1.20390.

[2] W. Jia and X. Huang, "Digital literacy and vocational education: Essential skills for the modern workforce," International Journal of Academic Research in Business and Social Sciences, vol. 13, no. 5, pp. 2195-2202, 2023, doi: 10.6007/IJARBSS/v13-i5/17080.

[3] D. Jatmoko, S. Suyitno, S. Rasul M, M. Nurtanto, N. Kholifah, A. Masek, and R. Nur H, "The factors influencing digital literacy practice in vocational education: A structural equation modeling approach," European Journal of Educational Research, vol. 12, no. 2, pp. 1109-1121, 2023, doi: 10.12973/eu-jer.12.2.1109.

[4] T. Nguyen L A and A. Habók, "Tools for assessing teacher digital literacy: a review," Journal of Computers in Education, vol. 11, no. 1, pp. 305-346, 2024, doi: 10.1007/s40692-022-00257-5.

[5] A. Azman N, I. Hamzah M, and H. Baharudin, "Digital teaching strategies of Islamic Education teachers: A case study in primary schools," International Journal of Learning, Teaching and Educational Research, vol. 24, no. 3, pp. 562-585, 2025, doi: 10.26803/ijlter.24.3.27.

[6] M. Korkmaz and O. Akçay A, "Determining digital literacy levels of primary school teachers," Journal of Learning and Teaching in Digital Age, vol. 9, no. 1, pp. 1-16, 2024, doi: 10.53850/joltida.1175453.

[7] C. Buchan M, J. Bhawra, and R. Katapally T, "Navigating the digital world: development of an evidence-based digital literacy program and assessment tool for youth," Smart Learning Environments, vol. 11, no. 1, pp. 8-32, 2024, doi: 10.1186/ s40561-024-00293-x.

[8] E. Yeşilyurt and R. Vezne, "Digital literacy, technological literacy, and internet literacy as predictors of attitude toward applying computer-supported education," Education and information technologies, vol. 28, no. 8, pp. 9885-9911, 2023, doi: 10.1007/s10639-022-11311-1.

[9] D. Villar-Onrubia, L. Morini, I. Marín V, and F. Nascimbeni, "Critical digital literacy as a key for (post) digital citizenship: an international review of teacher competence frameworks," Journal of E-Learning and Knowledge Society, vol. 18, no. 3, pp. 128-139, 2022.

[10] E. Avinç and F. Doğan, "Digital literacy scale: Validity and reliability study with the rasch model," Education and information Technologies, vol. 29, no. 17, pp. 22895-22941, 2024, doi: 10.1007/s10639-024-12662-7.

[11] M. Gümüş M and V. Kukul, "Developing a digital competence scale for teachers: validity and reliability study," Education and information technologies, vol. 28, no. 3, pp. 2747-2765, 2023, doi: 10.1007/s10639-022-11213-2.

[12] N. Khan, A. Sarwar, B. Chen T, and S. Khan, "Connecting digital literacy in higher education to the 21st century workforce," Knowledge Management & E-Learning, vol. 14, no. 1, pp. 46-61, 2022, doi: 10.34105/j.kmel.2022.14.004.

[13] G. Barboutidis and E. Stiakakis, "Identifying the factors to enhance digital competence of students at vocational training institutes," Technology, knowledge and learning, vol. 28, no. 2, pp. 613-650, 2023, doi: 10.1007/s10758-023-09641-1.

[14] M. Dong L, C. Jiao B, and Zhang Q L.Research on Evaluation Index of Digital Literacy of Mathematics Teachers in Primary and Secondary Schools, "Modern Educational Technology," 2025; 35(4):35-43, doi: 10.3969/j.issn.1009-8097.2025.04.004.

[15] H. Choudhary and N. Bansal, "Barriers affecting the effectiveness of digital literacy training programs (DLTPs) for marginalised populations: a systematic literature review," Journal of Technical Education and Training, vol. 14, no. 1, pp. 110-127, 2022, doi: 10.30880/jtet.2022.14.01.010.

[16] N. Tytova and K. Mereniuk, "Digital literacy of future teachers in the realities of large-scale military aggression (Ukrainian experience)," Futurity Education, vol. 2, no. 3, pp. 50-61, 2022.

[17] A. Orakova, F. Nametkulova, G. Issayeva, S. Mukhambetzhanova, M. Galimzhanova, and G. Rezuanova, "The relationships between pedagogical and technological competence and digital literacy level of teachers," Journal of Curriculum Studies Research, vol. 6, no. 1, pp. 1-21, 2024, doi: 10.46303/jcsr.2024.2.

[18] L. Naz F, A. Raheem, U. Khan F, and W. Muhammad, "An effect of digital literacy on the academic performance of universitylevel students," Journal of Positive School Psychology, vol. 6, no. 8, pp. 10720-10732, 2022.

[19] N. Kholifah, I. Kusumawaty, M. Nurtanto, F. Mutohhari, D. Isnantyo F, and H. Subakti, "Designing the structural model of students' entrepreneurial personality in vocational education: An empirical study in Indonesia," Journal of Technical Education and Training, vol. 14, no. 3, pp. 1-17, 2022, doi: 10.30880/jtet.2022.14.03.001.

[20] S. Getenet, C. Haeusler, P. Redmond, R. Cantle, and V. Crouch, "First-year preservice teachers' understanding of digital technologies and their digital literacy, efficacy, attitude, and online learning engagement: Implication for course design," Technology, Knowledge and Learning, vol. 29, no. 3, pp. 1359-1383, 2024, doi: 10.1007/s10758-023-09724-z.

[21] B. Espejo Villar L, L. Lázaro Herrero, and G. Álvarez-López, "UNESCO strategy and digital policies for teacher training: The deconstruction of innovation in Spain," Journal of New Approaches in Educational Research, vol. 11, no. 1, pp. 15-30, 2022, doi: 10.7821/naer.2022.1.812.

[22] C. Weninger, "Digital literacy as ideological practice," Elt Journal, vol. 77, no. 2, pp. 197-206, 2023, doi: 10.1093/elt/ccad001.

[23] A. Morgan, R. Sibson, and D. Jackson, "Digital demand and digital deficit: conceptualising digital literacy and gauging proficiency among higher education students," Journal of Higher Education Policy and Management, vol. 44, no. 3, pp. 258-275, 2022, doi: 10.1080/1360080X.2022.2030275.

[24] P. Reddy, K. Chaudhary, B. Sharma, and S. Hussein, "Essaying the design, development and validation processes of a new digital literacy scale," Online Information Review, vol. 47, no. 2, pp. 371-397, 2023, doi: 10.1108/OIR-10-2021-0532.

[25] Y. Yim I H and J. Su, "Artificial intelligence literacy education in primary schools: a review," International Journal of Technology and Design Education, vol. 35, no. 5, pp. 2175-2204, 2025, doi: 10.1007/s10798-025-09979-w.

[26] O. Derevyanchuk, N. Ridei, N. Tytova, V. Ravlyuk, and O. Yatsko, "Implementation of the STEM project "modeling of spatial images of stellated Poyhedra" in the professional training of future vocational education teachers," Edelweiss Applied Science and Technology, vol. 9, no. 1, pp. 89-105, 2025, doi: 10.55214/25768484.v9i1.3623.

[27] Z. Abiddin N, I. Ibrahim, and A. Aziz S A, "Advocating digital literacy: Community-based strategies and approaches," Academic Journal of Interdisciplinary Studies, vol. 11, no. 1, pp. 198-211, 2022, doi: 10.36941/ajis-2022-0018.

[28] K. Budiarto M, R. Karsidi, and A. Rahman, "E-Learning platform for enhancing 21st century skills for vocational school students: A systematic literature review," Electronic Journal of E-Learning, vol. 22, no. 5, pp. 76-90, 2024, doi: 10.34190/ ejel.22.5.3417.

[29] T. Junaedi A, P. Panjaitan H, I. Yovita, K. Veronica, N. Renaldo, and J. Jahrizal, "Advancing digital and technology literacy through qualitative studies to bridging the skills gap in the digital age," Journal of Applied Business and Technology, vol. 5, no. 2, pp. 123-133, 2024, doi: 10.35145/jabt.v5i2.170.

[30] Demir, Kürat, and M. Yavuz, "Evaluation of Teachers' Motivation in Terms of Research Literacy," Research on Education & Psychology (REP), vol. 9, no. 2, 2025, doi: 10.54535/rep.1757746.

[31] S. Nurhayati, A. Fitri, R. Amir, and Z. Zalisman, "Analysis of the implementation of training on digital-based learning media to enhance teachers' digital literacy," Al-Ishlah: Jurnal Pendidikan, vol. 16, no. 1, pp. 545-557, 2024, doi: 10.35445/ alishlah.v16i1.4029.

[32] I. Iskandar, S. Sumarni, R. Dewanti, and A. Asnur M N, "Infusing Digital Literacy in Authentic Academic Digital Practices of English Language Teaching at Universities," International Journal of Language Education, vol. 6, no. 1, pp. 75-90, 2022, doi: 10.26858/ijole.v6i1.31574.

[33] I. Kazanidis and N. Pellas, "Harnessing generative artificial intelligence for digital literacy innovation: A comparative study between early childhood education and computer science undergraduates," Ai, vol. 5, no. 3, pp. 1427-1445, 2024, doi: 10.3390/ai5030068.

[34] H. Tinmaz, M. Fanea-Ivanovici, and H. Baber, "A snapshot of digital literacy," Library Hi Tech News, vol. 40, no. 1, pp. 20-23, 2023, doi: 10.1108/LHTN-12-2021-0095.

[35] M. Shvardak, M. Ostrovska, N. Bryzhak, A. Predyk, and L. Moskovchuk, "The Use of Digital Technologies in Professional Training of Primary School Teachers," International Electronic Journal of Elementary Education, vol. 16, no. 3, pp. 363-376, 2024, doi: 10.26822/iejee.2024.337.

[36] S. Korir and P. Ngetich, "The Role of Digital Literacy Integration in TVET Curricula for Future Workforce Preparedness in Kenya," Africa Journal of Technical and Vocational Education and Training, vol. 11, no. 1, pp. 2-10, 2026, doi: 10.69641/ afritvet.2026.111198.

[37] S. Andriyani, S. Haq M, M. Sholeh, et al., "Teacher Digital Literacy and Merdeka Curriculum Readiness: The Role of Learning Communities in

Preparing 21st-Century Competencies in Secondary Schools," Journal of Innovation and Research in Primary Education, vol. 5, no. 1, pp. 147-157, 2026, doi: 10.56916/jirpe.v5i1.2692.

[38] H. Nakamura, "Digital Literacy and Critical Thinking Skills among Secondary Students," Global Dialogues in Humanities and Pedagogy, vol. 4, no. 10, pp. 24-35, 2026.

[39] Q. Sultonova D, "Youth Of Uzbekistan And Digital Literacy: Current Status, Development Factors, And Policy-Ecosystem Pathways," TLEP-International Journal of Multidiscipline, vol. 3, no. 2, pp. 33-40, 2026.

[40] I. Rikmasari, L. Yuliandini, S. Sugiarti, et al., "Reconstructing Teacher Professionalism: Integrating Digital Literacy and Innovative Pedagogy in the Era of Technological Disruption," Journal of Innovation and Research in Primary Education, vol. 5, no. 1, pp. 1232-1243, 2026, doi: 10.56916/jirpe.v5i1.3011.

[41] M. YU and S. FAN, "Exploration of the Talent Training Mode of Vocational Undergraduate Universities in the Era of Digital Intelligence," US-China Education Review, vol. 16, no. 1, pp. 5-10, 2026, doi: 10.17265/2161-623X/2026.01.002.

[42] F. Damayanti, A. Musliman, A. Stiawan R, et al., "Integration of Artificial Intelligence Based Digital Media in Science Learning Activities to Improve Critical Thinking Skills of Vocational High School Students," Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education), vol. 14, no. 1, pp. 12-20, 2026, doi: 10.24815/jpsi.v14i1.93.

[43] A. Zarqan I, Y. Nugraha D, G. Sitompul, et al., "Technology-Based Startup Ideas and MVP Development: A Digital Literacy Perspective among High School Students," Journal of Management and Business Analytics, vol. 2, no. 1, pp. 93-107, 2026.

[44] A. Andika, "Digital Literacy and Self-Efficacy Development Strategies for Technostress Coping Among Older Lecturers in Higher Education Institutions," Jurnal Rumpun Manajemen dan Ekonomi, vol. 3, no. 1, pp. 191-199, 2026.

[45] N. Delanoy, "Leveraging Strategic Partnerships to Enhance Digital and AI Literacy Among Preservice and In-service Teachers," Journal of Digital Life and Learning, vol. 6, no. 1, pp. 204-225, 2026, doi: 10.51357/jdll.v6i1.362.