Using Self-Distillation Strategy to Enhance the Generalization Ability of Nursing Skill Scoring System for Low-Sample Tasks

Yongqi Zheng[1], Fangfang Zhang[1], Xiuxiu Wu[2], Jingya Zhang[3]
[1]Zhengzhou Health College Health Administration Department, Zhengzhou 450000, Henan, China
[2]Zhengzhou Health College Education Department, Zhengzhou 450000, Henan, China
[3]The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, Henan, China

Corresponding author: Yongqi Zheng (e-mail: 16627522596@163.com)

ABSTRACT Automated assessment of nursing skills is important for standardizing clinical training, but it often suffers from overfitting and poor generalization when labeled data are scarce. To address this limitation, this paper proposes a multimodal framework that integrates a cross-modal attention fusion module with a dynamic self-distillation strategy. Video, gesture, and text inputs are processed through dedicated encoders, including TimeSformer, BiLSTM, and RoBERTa, to capture spatiotemporal, kinematic, and semantic features. The cross-modal attention module enables fine-grained interaction among modalities. Self-distillation, with teacher parameters updated every five training cycles, progressively transfers knowledge to enhance robustness in low-sample conditions. The method is evaluated on a nursing-skill dataset aligned with the Nursing Operation Technical Specifications. Results show that the proposed method achieves RMSE of 2.98 using only 10 training samples, which decreases to 0.95 with 50 samples. The model also shows strong cross-validation stability, with RMSE fluctuation within 0.01–0.03, and real-time efficiency, with a response time of 17 ms. These results indicate that the framework provides a practical solution for data-scarce skill assessment in nursing education.

INDEX TERMS nursing skills, scoring system, low-sample learning, self-distillation, cross-modal attention

I. INTRODUCTION

NURSING skills assessment is a core component of medical education and clinical practice, and its accuracy directly impacts the nursing quality and patient safety [1,2]. Traditional assessment relies on manual scoring, which has drawbacks such as subjectivity, inefficiency, and high costs. In recent years, automated scoring systems based on multimodal data and deep learning [3,4] have emerged as a research hotspot. Still, existing methods are prone to overfitting in low-sample tasks due to data scarcity and lack of generalization ability. With the clear requirements of the Nursing Operation Technical Specifications for standardized nursing operations, the automated scoring system has become a key direction for improving evaluation efficiency and consistency. In particular, the model's performance has significantly decreased in fine-grained scoring dimensions, such as nursing operation standardization [5] and safety. Although multimodal fusion technology can improve the ability of feature interaction modeling, its reliance on large-scale labeled data still restricts its practical application. Therefore, it is urgent to explore a scoring modeling method that is both robust and interpretable in low-sample scenarios, thereby promoting the intelligence and standardization of nursing skill assessment.

This paper proposes a nursing skill scoring method based on self-distillation strategy and cross-modal attention fusion, aiming to solve the core challenge of insufficient model generalization ability in low-sample tasks. A dynamic self-distillation framework is designed to alleviate the overfitting problem under low-sample conditions by transferring knowledge from the teacher-student model. By constructing a CMA (Cross-Modal Attention) module, during the "venipuncture" stage, the nurse's hand movement trajectory is captured through video. Combined with the pressure feedback of the gesture sensor and the language prompts in the text record, the fine-grained interaction relationship between the three modalities of video, gesture, and text is dynamically captured, breaking through the traditional fusion method's reliance on modal alignment accuracy. A dynamically updated teacher-student self-distillation framework is constructed to update the teacher model parameters every 5 training cycles. The iterative knowledge transfer

theory is introduced to alleviate the problem of knowledge solidification that may be caused by the static teacher model and improve the robustness of the model under small samples. Experiments demonstrate that this method significantly outperforms advanced models, such as CrossViT (Cross-Attention Vision-Text Transformer) and UniViLM (Unified Video-Language Model), on a self-built nursing dataset, achieving an RMSE (Root Mean Square Error) of 2.98 in a low-sample scenario (10 samples in the training set). This study offers a generalizable solution for nursing skill assessment in resource-constrained settings and provides new insights for multimodal small-sample learning.

II. RELATED WORK

In recent years, multimodal data fusion technology has shown significant potential in medical skill assessment. Looking back on the research evolution path, it can be systematically sorted out into its technical evolution process from three dimensions: "modal expansion-fusion optimization-low sample adaptation." The first stage: modal expansion promotes the improvement of information integrity. Early research primarily relied on single-modality modeling [6,7], but the one-sidedness of the information limited it and could not fully capture the complexity of nursing operations. With the development of deep learning, cross-modal attention mechanisms have been introduced into the field of medical education [8,9]. By modeling the interactive relationship between visual, text, and sensor data, the fine-grained recognition ability in the medical field is improved. The second stage: Fusion optimization strengthens the cross-modal interaction modeling ability. VATT (Video-Audio-Text Transformer) processes multimodal inputs through a unified architecture [10] to achieve joint analysis of operation steps and communication content; MViT (Multiscale Vision Transformer) [11] uses hierarchical feature extraction to enhance the spatiotemporal modeling accuracy of video actions.

Phase 3: Low-sample adaptation to meet the challenge of data scarcity. Although multimodal fusion significantly improves model performance, existing methods generally have two limitations: first, they rely too heavily on large-scale labeled data and are difficult to adapt to scenarios with scarce samples in nursing teaching; second, alignment errors between modalities can easily lead to feature confusion and affect scoring stability. Therefore, it is urgent to design a lightweight and low-sample-friendly multimodal fusion framework. Low-sample learning is a key direction to alleviate the problem of scarce medical data. Traditional methods alleviate overfitting by pre-training models or expanding synthetic data, but are limited by domain adaptability and modality complexity in nursing scoring tasks. Self-distillation strategies are widely used in small sample scenarios because they do not require additional teacher models [12,13]. TimeSformer (Time-Space Transformer)-Distill significantly improves the generalization ability of video action recognition through iterative knowledge transfer and optimization of parameter updates using soft labels from earlier versions of the model. Combining mask reconstruction with distillation loss can reduce computational overhead while enhancing feature robustness [14]. However, existing studies mostly focus on a single modality and lack collaborative modeling and analysis of multi-source heterogeneous data [15,16]. Although the self-distillation strategy can alleviate overfitting in low-sample training, it relies on the output of the teacher model as a supervision signal, which may cause gradient conflicts between the student model and the true label. Especially when the teacher model has not fully converged, the noise in the soft label can mislead the parameter update direction, exacerbating model oscillation and training instability. Therefore, it is necessary to explore a teacher-student framework with a dynamic update mechanism, combined with a multimodal collaborative distillation strategy, to alleviate the long sequence modeling defects and gradient conflict risks of the attention mechanism, thereby improving the consistency and robustness of scoring under low samples.

III. METHOD

A. DATA COLLECTION AND PREPROCESSING

1) Equipment Deployment and Data Synchronization

This study uses a multimodal data acquisition system that integrates high-precision vision, inertial measurement, and audio equipment to ensure that multi-dimensional information of the entire nursing operation process is fully captured. Visual data acquisition is achieved through an industrial-grade camera (model: Basler ace acA2440-35uc) fixed above the ward, which supports 1080p@30fps resolution and frame synchronization triggering function, covering the full body movements of the nursing staff and the patient interaction area.

To adapt to complex lighting conditions, the camera is equipped with an adjustable LED (Light-Emitting Diode) fill light (color temperature 5000K) and uses H.265 encoding compression storage (bit rate 15Mbps). Hand motion capture uses the Dexmo F2 wearable data glove, which has a built-in six-axis inertial measurement unit (IMU (Inertial Measurement Unit), sampling rate 100Hz, $\pm$16g accelerometer, $\pm$2000°/s gyroscope) and force feedback sensor, which can record finger joint angles (accuracy $\pm$0.5°), grip strength (0–50N) and hand space trajectory in real time.

Voice collection utilizes a clip-on condenser microphone (model: Rode VideoMic NTG) with a sampling rate of 44.1 kHz and a signal-to-noise ratio of 72 dB. It is connected to the main control computer through an audio interface to synchronously record the conversation between the nurse and the patient.

The multimodal data synchronization mechanism is based on dual calibration of the hardware trigger and software timestamp. When the system starts, the central controller sends out sound and light synchronization signals simultaneously.

Optical trigger: The instantaneous light pulse (duration 50ms) of the LED array is captured by the camera as a keyframe marker.

Audio trigger: The 5kHz sine wave pulse (duration 50ms) is played through the speaker and recorded by the microphone as the audio starting point.

Inertial trigger: The IMU of the Dexmo glove detects the controller vibration signal (peak acceleration > 3g), marking the starting position of the sensor data.

The synchronization error is quantitatively evaluated by cross-correlation analysis. Defining the video frame timestamp sequence $T_v = \{t_{v,1}, t_{v,2}, \cdots, t_{v,N}\}$, audio sample timestamp $T_a = \{t_{a,1}, t_{a,2}, \cdots, t_{a,M}\}$, and sensor data timestamp $T_s = \{t_{s,1}, t_{s,2}, \cdots, t_{s,K}\}$, and their time domain alignment error Δt is determined by the offset corresponding to the peak cross-correlation coefficient:

$$\Delta t = \arg\max_{\tau} \frac{\sum_{i=1}^{L} (x(t_i) - \bar{x})\,(y(t_i + \tau) - \bar{y})}{\sqrt{\sum_{i=1}^{L} (x(t_i) - \bar{x})^2}\sqrt{\sum_{i=1}^{L} (y(t_i + \tau) - \bar{y})^2}} \tag{1}$$

In formula (1), x(·) and y(·) represent the normalized amplitudes of the two modal signals, and $\bar{x}$ and $\bar{y}$ represent their means.

2) Data Labeling and Preprocessing

This study employs a double-blind expert labeling strategy, with two senior nursing education experts (with over 10 years of clinical experience) independently completing the key action labeling. The labeling tool used is the open source platform Label Studio, which supports multimodal data synchronous labeling and version control.

The nursing scenario is shown in Figure 1.

Figure 1 is from the Flickr platform.

According to the Nursing Operation Technical Specifications [17,18], the criteria for key actions such as disinfection, puncture, and fixation are clearly defined, where the diameter of the disinfection area is ≥5cm and the puncture angle is 30°–45°. Experts mark the start and end time of each key action through an interactive timeline and assigned dimension labels.

Cohen's Kappa coefficient is used to quantify the consistency of annotations between experts:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

$$P_o = \frac{1}{T}\sum_{t=1}^{T} \delta\left(a_1^{(t)}, a_2^{(t)}\right) \tag{3}$$

$$P_e = \sum_{k=1}^{K} p_{1,k} p_{2,k} \tag{4}$$

In formula (2-4), Po is the observed consistency rate, Pe is the random consistency rate. p1,k and p2,k are the frequency distribution of expert 1/2 for label k. Based on the annotation results, a sliding window merging algorithm is used to remove redundant data. The minimum interval threshold is set to 2s. If the adjacent action time windows are less than 2s, they are merged into continuous event segments.

Based on the annotated timestamps, FFmpeg is used to extract key event video segments. The frame rate is reduced from 30fps to 10fps, and the resolution is adjusted to 480p. Adaptive histogram equalization is applied to each frame to suppress illumination differences:

$$I_{norm}(x,y) = clip\left(\frac{I(x,y) - \mu_{local}(x,y)}{\sigma_{local}(x,y) + \epsilon}, [0, 255]\right) \tag{5}$$

In formula (5), μ_{local} and σ_{local} are the mean and standard deviation of the local window (8×8 pixels). The original 100Hz gesture data is downconverted to 20Hz using bilinear interpolation:

$$G_{20}^{(k)} = \frac{1}{5}\sum_{i=5(k-1)+1}^{5k} G_{100}^{(i)} \tag{6}$$

Extracting gesture features, including acceleration mean and angle variance. The acceleration mean is:

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} \|a_i\|_2 \tag{7}$$

The angle is the variance:

$$\sigma_\theta^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(\theta_i - \bar{\theta}\right)^2 \tag{8}$$

In formula (8), θ_i is the angle of the finger joint.

Based on the stop word list in the nursing field (including 132 high-frequency filler words such as "en", "a", "ranhou", etc.), regular expressions are used to match and remove them. TF-IDF (Term Frequency-Inverse Document Frequency) weighted word frequency is used to filter key communication content, calculate word frequency and inverse document frequency, and calculate the score for each sentence:

$$Score(s_j) = \sum_{w \in s_j} f(w, s_j) \cdot IDF(w) \tag{9}$$

The top 20% of the sentences are retained as key communication segments.

Calculating the motion amplitude ratio of the video frames before and after cropping:

$$MAR = \frac{1}{T}\sum_{t=1}^{T} \frac{\|\nabla I_t\|_1}{\max(\|\nabla I_t\|_1)} \tag{10}$$

In formula (10), ∇I_t is the gradient amplitude of frame t.

The video redundancy is evaluated by calculating the motion amplitude ratio of the video frames before and

FIGURE 1. Nursing scenario

after cropping. The standard deviation change rate of the acceleration mean before and after downsampling is calculated to reflect the stability of the gesture features. The average TF-IDF score of key sentences and the Flesch-Kincaid readability index are calculated to reflect the text information density.

The preprocessing effect is shown in Table 1.

TABLE 1. Pretreatment effects

Modal	Index	Before pre-processing	After pre-processing
Video	Motion amplitude ratio	0.78 ± 0.12	0.45 ± 0.08
Gesture	Rate of change of mean standard deviation of acceleration	6%	2%
Text	TF-IDF key sentence score	0.32 ± 0.05	0.91 ± 0.07
Text	Flesch Kincaid readability index	4.2	4.8

The preprocessing process significantly optimizes the quality and efficiency of multimodal data. Video cropping reduces redundant actions through downsampling and illumination normalization while retaining key action details; after gesture data downsampling, feature stability (acceleration mean standard deviation = 2%) meets the scoring requirements; text cleaning improves the TF-IDF score of key sentences and the Flesch-Kincaid index, indicating that the semantic focus is more prominent.

B. RATING PREDICTION MODEL

1) Feature Extraction

Timesformer performs feature extraction as a video encoder, and its output sequence vector constitutes the feature representation of video [19,20]. The input video clip shape is V∈RT×H×W×C, T is 16, H is 480, W is 640, and C is 3. The video is segmented into spatiotemporal blocks using a 3D convolutional layer with a kernel size of (1,16,16) and a stride of (1,16,16). A learnable spatiotemporal position code Epos∈R(T×N+1)×D is added. The core innovation of TimeSformer is to split the video attention calculation into spatial attention and temporal attention to reduce the computational complexity.

The attention weights between spatial blocks are calculated within a single frame. The spatial attention formula is:

$$Attention_{spatial} = Softmax\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) V_s \qquad (11)$$

In formula (11), Q_s, K_s, and V_s are query, key, and value matrices, respectively.

Fusion of spatial attention outputs across time steps, temporal attention:

$$Attention_{temporal} = Softmax\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t \qquad (12)$$

After 12 layers of spatiotemporal separation attention and multi-layer perceptron, the final embedding of [CLS] token is extracted as the global feature of the video. BiLSTM (Bidirectional Long Short-Term Memory) [21,22] models the time series to extract gesture features. The formula is:

$$\overrightarrow{h}_t = LSTM_{forward}(x_t, \overrightarrow{h}_{t-1}) \qquad (13)$$

$$\overleftarrow{h}_t = LSTM_{backward}(x_t, \overleftarrow{h}_{t+1}) \qquad (14)$$

In formula (13-14), $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the forward and backward hidden states, respectively.

Concatenating bidirectional hidden states generates context-aware features:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in R^{512} \qquad (15)$$

Mapped to the target dimension through the fully connected layer:

$$v_{gesture} = ReLU(W_g \cdot h_t + b_g) \qquad (16)$$

RoBERTa (Robustly Optimized BERT Approach) models semantic embedding to extract text features [23,24]. The input of RoBERTa is a triplet(E_{token}, $E_{segment}$, $E_{position}$), which is a Token embedding, a sentence embedding, and a position embedding. RoBERTa models text context dependencies based on a 12-layer Transformer Block, each layer of which contains multi-head self-attention and a feedforward network.

Multi-head self-attention is:

$$MultiHead(Q, K, V) = Concat(Attention(Q_i, K_i, V_i)_{i=1}^{h})W^O \qquad (17)$$

The feed-forward network is:

$$FFN(H) = ReLU(HW_1 + b_1)W_2 + b_2 \qquad (18)$$

Extracting the last layer output of [CLS] token as text global embedding:

$$v_{text} = H_{[CLS]} \in R^{768} \qquad (19)$$

To eliminate the scale differences between modes, L2 normalization is performed on the features of each mode:

$$v_{mod}^{norm} = \frac{v_{mod}}{\|v_{mod}\|_2} \tag{20}$$

2) Cross-Modal Attention Fusion

The CMA module uses a multi-head attention mechanism to focus on key information in one modality from the perspective of another modality, thereby establishing fine-grained interactions between modalities [25,26].

Video ↔ Gesture: Capturing the association between action and visual features. When performing the "disinfection" step, the video shows that the nurse is wiping the skin, and the gesture data provides information on whether the finger applies the correct pressure, constructing an attention mechanism:

$$\alpha_{vg}^{(t,\tau)} = Softmax\left(\frac{Q_{video}^{(t)}(K_{gesture}^{(\tau)})^T}{\sqrt{d_k}}\right) \tag{21}$$

Applying the attention weights to the value vector to obtain the weighted features:

$$f_{vg}^{(t)} = \sum_{\tau=1}^{100} \alpha_{vg}^{(t,\tau)} \cdot v_{gesture}^{(\tau)} \tag{22}$$

Video ↔ Text: Associating the operation steps with the communication content, it is used to analyze the consistency between visual actions and language descriptions. The formula is:

$$\alpha_{vt}^{(t,\tau)} = Softmax\left(\frac{Q_{video}^{(t)}(K_{text}^{(\tau)})^T}{\sqrt{d_k}}\right) \tag{23}$$

$$f_{vt}^{(t)} = \sum_{\tau=1}^{100} \alpha_{vt}^{(t,\tau)} \cdot v_{text}^{(\tau)} \tag{24}$$

Gesture ↔ Text: Analyzing the consistency between hand movements and verbal expressions, focusing on whether the language expression of the caregiver is appropriate when performing specific actions. It is calculated as:

$$\alpha_{gt}^{(t,\tau)} = Softmax\left(\frac{Q_{gesture}^{(t)}(K_{text}^{(\tau)})^T}{\sqrt{d_k}}\right) \tag{25}$$

The weighted gesture features are:

$$f_{gt}^{(t)} = \sum_{\tau=1}^{100} \alpha_{gt}^{(t,\tau)} \cdot v_{gesture}^{(\tau)} \tag{26}$$

To further integrate information, the original features of each modality are concatenated with their attention-weighted versions to form enhanced fusion features:

$$F(t) = [v_{video}(t); f_{vg}(t); v_{gesture}(t); f_{vt}(t); v_{text}(t); f_{gt}(t)] \tag{27}$$

Then, the feature dimension is compressed to 1792 dimensions through the fully connected layer, and the activation function is introduced:

$$F_{fusion}^{(t)} = ReLU(W_{fuse} \cdot F(t) + b_{fuse}) \tag{28}$$

Since the nursing skill score is a comprehensive evaluation of the entire operation process rather than a frame-by-frame score, global average pooling is used to aggregate the time dimension information:

$$f_{global} = \frac{1}{T}\sum_{t=1}^{T} F_{fusion}^{(t)} \tag{29}$$

Global average pooling not only retains the key information of the overall operation, but also reduces the complexity of subsequent rating prediction.

The global feature f_{global} is input into the multi-layer perceptron for rating prediction. The first layer is:

$$h_1 = ReLU(W_1 \cdot f_{global} + b_1) \tag{30}$$

The second layer is:

$$h_2 = ReLU(W_2 \cdot h_1 + b_2) \tag{31}$$

The output layer outputs the total score:

$$s_{total} = \sigma(W_o \cdot h_2 + b_o) \tag{32}$$

C. SELF-DISTILLATION STRATEGY

1) Teacher-Student Framework

The self-distillation strategy transfers knowledge from the teacher-student model and uses soft label supervision to improve the generalization ability of the nursing skill scoring system under low sample conditions and alleviate overfitting. The teacher-student framework [27,28] is shown in Figure 2.

This paper proposes a nursing skill scoring model optimization strategy based on self-distillation, which employs a teacher-student framework to enhance the model's generalization ability under low-sample training conditions. The teacher model is a snapshot of the model saved at the beginning of training, and the student model is the updated model from the current iteration. The two share the network structure but have independent parameters. During the training process, the student model simultaneously learns the true label (hard label) and the soft label output by the teacher model, and guides the parameter update through the mixed loss function. During the training process, the student model parameters are copied to the teacher model at fixed intervals (5 epochs) to achieve progressive knowledge transfer.

The core of CMA module is to calculate the attention weight across modes through the multi head attention mechanism, using the sequence output of one mode as query and the sequence output of the other mode as key and

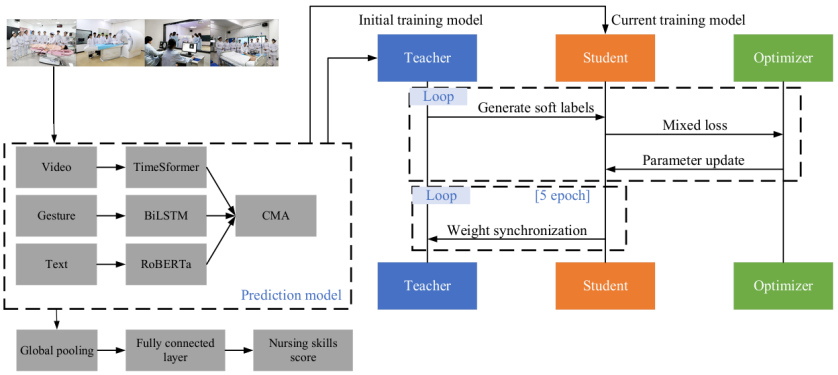

FIGURE 2. Teacher-student framework

value, so as to achieve fine-grained interaction among video, gesture and text modes. The enhanced fusion feature of each time step is composed of the original feature and all related cross modal weighted features, which integrates the complementary information of multimodal.

The choice of five cycles as the update frequency is based on the balance between the stability of training and the timeliness of knowledge transfer: too low update frequency will lead to the solidification of teachers' model knowledge and can not provide a continuous evolution monitoring signal for students' model; However, too high update frequency may introduce noise due to excessive fluctuation of teacher model, which will interfere with the convergence of student model. The teacher model and the student model adopt the same network structure, but the parameters are independent. The dynamic updating process achieves the gradual and smooth evolution of knowledge by periodically copying the parameters of the student model to the teacher model, rather than a one-time hard replacement. This mechanism is particularly important under the condition of low samples: it not only avoids the noise caused by too fast update of the teacher model, but also prevents the knowledge solidification caused by too slow update. By synchronizing the parameters every five cycles, the teacher model can continuously provide relatively stable soft labels optimized by multiple rounds of student model as a monitoring signal, so as to promote the convergence of student model, effectively suppress the training oscillation or catastrophic forgetting that may be caused by the limited distribution of small sample data, and ensure the robustness of knowledge transfer.

2) Loss Function and Training Process

This paper adopts a self-distillation-based training strategy [29,30] to improve the generalization ability of the student model in low-sample tasks by jointly optimizing the hard label loss L_{hard} and the soft label loss L_{soft} generated by the teacher model. L_{hard} represents the mean square error based on real expert ratings, which is used to measure the difference between the model output and the manual annotation, and is defined as:

$$L_{hard} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \tag{33}$$

In formula (33), y_i is the true score; $\widehat{y}_i$ is the model prediction value; and N is the number of samples. L_{soft} is the KL (Kullback-Leibler) divergence loss between the output distribution of the student model and the teacher model, which is used to guide the student model to learn the knowledge contained in the teacher model. Its form is:

$$L_{soft} = \sum_{i=1}^{N} p_{teacher}(x_i)\log\frac{p_{teacher}(x_i)}{p_{student}(x_i)} \tag{34}$$

In formula (34), $p_{teacher}$ and $p_{student}$ are the output probability distributions of the teacher model and the student model for sample x_i, respectively.

The final loss function is a weighted combination:

$$L = \alpha \cdot L_{hard} + \beta \cdot L_{soft} \tag{35}$$

The training process adopts an iterative update strategy: first, the teacher model is fixed and the student model is trained with multimodal inputs; after every 5 training epochs, the student model parameters are copied to the teacher model to achieve dynamic evolution of knowledge; the whole process continues until the model performance converges on the test set, ensuring that the scoring system still has good generalization ability and scoring stability under limited samples.

IV. EXPERIMENT

A. DATASETS AND EVALUATION METRICS

To systematically evaluate the generalization ability of the nursing skill scoring model proposed in this paper under low sample conditions, this study designs a set of progressive training set expansion experiments. The experiment is based on a self-built nursing skill scoring dataset, which contains N=90 fully annotated multimodal samples. Each sample corresponds to a standard nursing operation process and covers video, gesture sensor data, and speech-to-text multimodal input. All samples are independently scored by two senior nursing experts, and the average value is

taken as the final score label. This study was approved by the Ethics Committee of The first affiliaten hospital of Zhengzhou university, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

The experimental setup is as follows: In the initial stage, the training set is set to only 10 samples, and the test set is the remaining 80 samples; then the number of training samples is gradually increased, and each time 1 sample is added, 1 sample in the test set is correspondingly reduced, forming a series of experimental configurations with progressive sample numbers. Finally, the number of training set samples changes from 10 to 89, and the test set decreases from 80 to 1, covering the complete evaluation path from extremely low sample to nearly full sample learning.

By varying the number of samples, it examines the performance evolution trend of the model under different sample sizes. It assesses whether the performance in the small sample area is stable and whether the improvement curve is steep, thereby verifying the method's practicality and robustness in resource-constrained scenarios.

All samples were scored independently by two senior nursing experts according to the technical specification for nursing operation, with a score range of 0-100 points. Finally, the average value was taken as the real score label of the sample. The score range reflects the comprehensive performance of nursing skills, of which more than 90 points are excellent, 80-89 points are good, 70-79 points are qualified, and less than 70 points are unqualified.

Multiple models are compared, and the results are presented in Table 2.

To effectively and accurately measure the prediction performance of nursing skill scores, three indicators are set, including RMSE, MAE (Mean Absolute Error), and PCC (Pearson Correlation Coefficient). RMSE measures the square root of the mean square error between the predicted score and the true score, and the formula is:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{36}$$

In formula (36), y_i is the expert score; $\widehat{y}_i$ is the model prediction score; and N is the number of test samples. MAE measures the average absolute error between the predicted score and the true score, and the formula is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{37}$$

PCC measures the linear correlation between the model prediction score and the expert score, with a value range of [-1, 1]. The closer to 1, the higher the consistency.

$$PCC = \frac{\sum(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})}{\sqrt{\sum(y_i - \bar{y})^2} \cdot \sqrt{\sum(\widehat{y}_i - \bar{\widehat{y}})^2}} \tag{38}$$

B. EXPERIMENTAL IMPLEMENTATION

This study trains and evaluates the proposed nursing skill scoring model and its comparative model in a unified experimental environment to ensure the comparability and reproducibility of the results. All experiments are run on a server equipped with an NVIDIA A100 $\times$ 4 GPU (Graphics Processing Unit), using the PyTorch 1.13 deep learning framework and the HuggingFace Transformers library to achieve efficient modeling of text modality and video-related modules.

The model parameter settings are shown in Table 3.

To adapt to low-sample training scenarios, the default batch size is set to 16 and automatically adjusted to 8 when the number of training samples is less than 30, thereby alleviating the issue of unstable gradient estimation with small samples. The optimizer uses AdamW with weight decay control. The initial learning rate is set to 1×10^{-4}, and the weight decay coefficient is set to 1×10^{-5} to balance the parameter update speed and overfitting risk.

V. RESULTS

A. SELF-DISTILLATION RESULTS

The model snapshot (epoch=10) saved at the beginning of teacher model training, when the training set is 10 samples, calculates the PCC of the output results of the teacher model and the student model to reflect the consistency of teacher-student predictions. The feature distribution of different score intervals (<60, 60- 70, 70-80, 80-90, $\geq$90) is visualized by t-SNE (t-Distributed Stochastic Neighbor Embedding) to analyze the difference in score expression between the teacher model and the student model. The results are shown in Figure 3.

The prediction consistency analysis between the teacher model and the student model under low-sample training conditions shows that the self-distillation strategy significantly promotes the effectiveness of knowledge transfer in low-sample scenarios (PCC increases from 0.60 at Epoch=10 to 0.96 at Epoch=100).

The following key mechanisms are revealed: The slow growth of PCC (0.60$\rightarrow$0.75) in the early stage of training (Epoch=10~50) reflects the initial absorption process of the student model to the teacher model's knowledge. At this time, the student model has not yet fully captured the high-order semantic associations in the teacher model output distribution, and is limited by the instability of parameter updates under small samples; after Epoch=50, PCC accelerates (0.75$\rightarrow$0.96), indicating that the student model gradually strengthens the collaborative modeling ability of multi-source heterogeneous data through CMA. The soft labels generated by the teacher model provide implicit supervision signals beyond hard labels, guiding the student model to focus on the key features of scoring; The dynamic update mechanism of the iterative self-distillation framework effectively alleviates the problem of knowledge solidification that may be caused by the static teacher model, allowing the student model to still achieve optimal recon-

TABLE 2. Model comparison information

Model	Modal input	Integration method	Is self-distillation used?
This paper model	Video+gesture+text	CMA (Cross Modal Attention) Fusion	Yes
CrossViT	Video+Text	Dual stream Transformer interaction	No
UniViLM	Video+Text	Unified Transformer Architecture	No
TimeSformer-Distill	Video	–	Yes
VATT	Video+Audio+Text	Multimodal Transformer	No
MViT	Video	Multi scale feature extraction	No
VideoMAE (Video Masked Autoencoder)	Video	Self-supervised pre-training+fine-tuning	No

TABLE 3. Model parameter settings

Parameters	Setting value / range	Description
Hardware platform	NVIDIA A100 $\times$ 4	Supports multi-card parallel training
Framework	PyTorch 1.13	Deep learning main framework
Batch size	Default 16, adjusted to 8 in low sample stage	Reduce batch size in small sample stage to stabilize training
Optimizer	AdamW	Adaptive optimizer with weight decay
Initial learning rate	1×10^{-4}	Applicable to most Transformer models
Weight decay	1×10^{-5}	Control model complexity to prevent overfitting
Maximum training rounds	100	Prevent overtraining
Temperature parameter	2	Used in self-distillation to smooth the output distribution of teacher model
Loss function weight α	0.7	Balance hard label and soft label loss
Loss function weight β	0.3	Control the strength of knowledge transfer

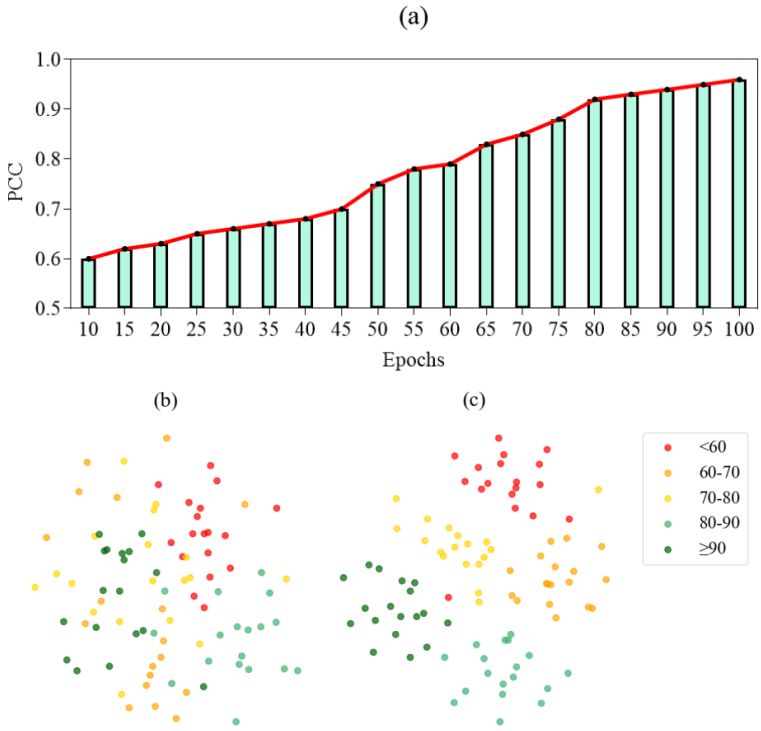

FIGURE 3. Prediction consistency and t-SNE visualization. Figure 3(a) Teacher-student prediction consistency; Figure 3(b) t-SNE at Epoch=10; Figure 3(c) t-SNE at Epoch=100

struction of the feature space through progressive knowledge transfer under low sample conditions. This paper verifies the unique advantages of self-distillation in low-sample tasks. Even if there are only 10 training samples, the soft labels of the teacher model can still convey global distribution information through KL divergence loss, reduce the risk of overfitting of local features by the student model, and finally form a highly consistent prediction distribution in the later stage of training.

Figure 3(b) and Figure 3(c) show the difference in multi-

modal feature distribution between the early stage of training (Epoch=10) and the late stage of training (Epoch=100), reflecting the evolution of feature representation during model learning. In the early stage of training, feature points in different scoring intervals overlap, indicating that the model has not yet effectively learned the discriminative features related to the score. At this point, the student model is in the parameter initialization stage, and its feature extraction capability is limited by random weight initialization and a limited number of gradient updates. Although the self-distillation strategy has introduced knowledge transfer from the teacher model, the early student model still lacks a deep understanding of multimodal interactions, resulting in blurred boundaries between rating intervals in the feature space. In addition, the risk of overfitting under low sample conditions further exacerbates the chaos of feature distribution, making it difficult for samples in the same rating interval to form stable clusters.

In the later stages of training, the feature points in different scoring intervals show a clear separation, indicating that the model has gradually optimized the feature representation through continuous training and knowledge transfer. CMA dynamically models the fine-grained interaction between the three modalities of video, gesture, and text, which strengthens the extraction of scoring-related features. Through self-distillation, knowledge transfer enables the soft label supervision of the teacher model to guide the student model in focusing on key scoring dimensions and suppressing noise interference. This evolution process verifies the effectiveness of this method in improving the generalization ability of the model through self-distillation and CMA in low-sample scenarios, and provides a theoretical basis for the actual deployment of the nursing skill scoring system.

The loss changes during the training process are shown in Figure 4.

The loss change data of the self-distillation training process clearly shows the optimization dynamics of the model under low-sample conditions: in the early stage (Epoch 1-10), the hard label loss drops from 1.45 to 1.20, the soft label loss drops from 2.30 to 2.00, and the total loss drops from 1.71 to 1.44, reflecting the underfitting of the model to the true label in the initial stage and the significant difference between the predicted distributions of the teacher and student models; the high loss at this stage is mainly due to the fact that the model parameters have not fully learned the data characteristics and the knowledge transfer effect has not yet appeared. As the training progresses (Epoch 20-50), the hard label loss drops rapidly from 0.95 to 0.42, the soft label loss drops synchronously from 1.70 to 0.70, and the total loss drops from 1.18 to 0.50, indicating that the model gradually captures the discriminative features related to the score through CMA, and the soft label supervision in the self-distillation strategy begins to effectively guide the student model to focus on key features and suppress noise interference. In the late stage of training (Epoch 60-100), all loss indicators enter a stable convergence stage, and the hard label loss eventually drops to 0.18, the soft label loss drops to 0.22, and the total loss converges to 0.19. The change in loss verifies the effectiveness of the self-distillation framework in low-sample scenarios - the teacher model provides implicit supervision signals beyond hard labels through dynamically updated soft labels, enabling the student model to still build robust feature representations with limited data. It is worth noting that the decrease in L_soft (90.4%) is significantly greater than that in L_hard (87.6%), indicating that knowledge transfer contributes more significantly to improving the model's generalization ability. The smooth downward trajectory of the total loss further confirms the rationality of the loss weight (α = 0.7, β = 0.3), which not only retains the dominant role of the true label but also fully utilizes the knowledge distillation advantages of the teacher model.

B. LOW SAMPLE PERFORMANCE COMPARISON

When the training set consists of 10 samples, the score prediction performance of various models is shown in Figure 5.

Under the extremely low sample condition of only 10 samples in the training set, the nursing skill scoring model based on CMA fusion and self-distillation strategy proposed in this paper shows a significant performance advantage over the comparison model. The experimental results show that the model in this paper achieves the best performance, with RMSE=2.98 and PCC=0.89, which is significantly lower than that of advanced methods such as CrossViT (RMSE=4.15, PCC=0.72) and MViT (RMSE=4.30, PCC=0.70), indicating its strong generalization ability in low-sample scenarios. This advantage mainly stems from two points: First, the CMA module dynamically models the fine-grained interaction between video, gesture and text through the cross-modal attention mechanism, breaking through the traditional early fusion method's reliance on modal alignment accuracy and avoiding the problem of feature confusion under low samples; second, the self-distillation strategy effectively alleviates the risk of overfitting of the student model to hard labels through the soft label supervision of the teacher model, and gradually optimizes the parameter space through iterative knowledge transfer (updating the teacher model every 5 Epochs), so that the score distribution converges to the expert score. Comparing TimeSformer-Distill (RMSE=3.42, PCC=0.83) and VideoMAE (RMSE=3.61, PCC=0.81), it can be seen that although single-modality self-distillation or self-supervised pre-training can partially improve low-sample performance, the lack of multimodal collaborative modeling and dynamic knowledge transfer mechanism makes its PCC and RMSE still inferior to the model in this paper. In addition, VATT and MViT perform poorly under low samples (PCC<0.75), further verifying the dependence of traditional multimodal architectures on large-scale data. By applying self-distillation, this method provides a feasible solution for automated evaluation in scenarios with scarce

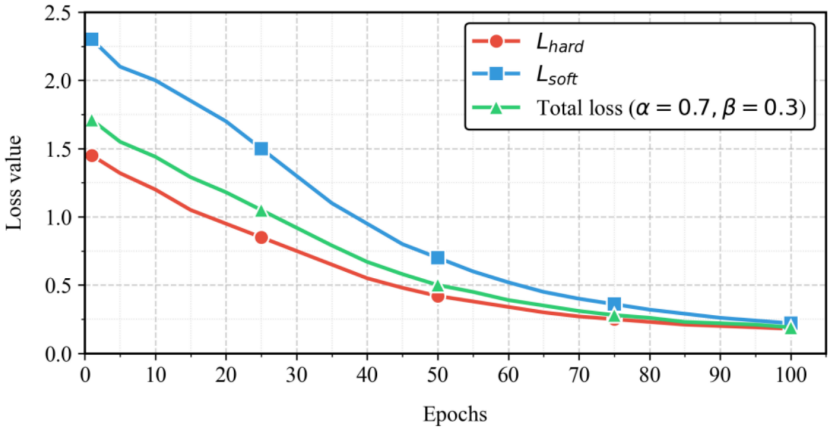

FIGURE 4. Changes in loss

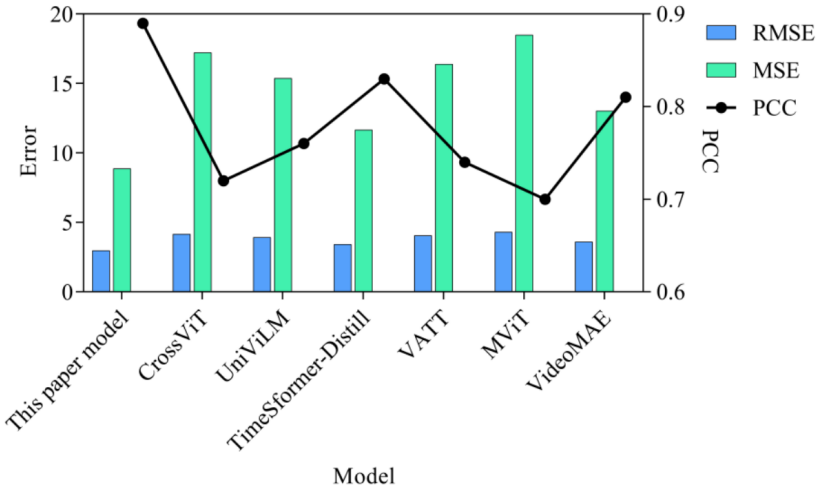

FIGURE 5. Low sample performance comparison

samples in nursing teaching.

C. SAMPLE DYNAMIC CHANGE SCENARIOS

The models with better performance in Figure 5 are selected for analysis, including UniViLM, TimeSformer-Distill, VideoMAE and the model in this paper. First, the training set is 10 samples; then, the training set is increased by 1 and the test set is reduced by 1, and the number of training set samples varies from 10 to 89. The results are shown in Figure 6.

Experimental results show that the proposed model shows significant advantages in the process of gradual growth of training samples from 10 to 89. Under extremely low sample conditions (N=10), the RMSE (2.98) of the proposed model has obvious advantages over UniViLM (3.92), TimeSformer-Distill (3.42), and VideoMAE (3.61), which verifies the effectiveness of the self-distillation strategy and cross-modal attention fusion. As the sample size increases, the performance of all models improves, but the proposed model always maintains the lowest RMSE: when N=50, the RMSE difference is still obvious (0.95 vs TimeSformer-Distill: 1.05). It is worth noting that the RMSE of this model first exceeded 1.0 (0.99) when N=44, which is earlier than other models, showing its fast convergence characteristics. When the sample size is $\geq$77, the RMSE of this model stably converges to 0, proving that this model has stronger stability under both extremely low sample and low sample conditions. The performance advantage mainly comes from: 1) The dynamic self-distillation framework effectively suppresses overfitting through the knowledge transfer of the teacher model; 2) The CMA module realizes the fine-grained alignment of video-gesture-text features and improves feature utilization. Experiments have confirmed that the proposed method can maintain a low score prediction error in the range of 10-89 samples.

D. ABLATION EXPERIMENT RESULTS

The ablation experiment is used to analyze the performance impact of each component on the low-sample nursing skill score prediction. The ablation experiment results are shown

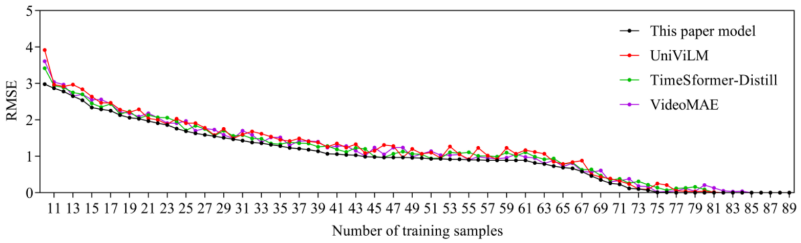

FIGURE 6. Sample dynamic change results

in Table 4.

The ablation experiment results show that multimodal fusion and CMA modules play a key role in the low-sample nursing skill scoring task. The RMSE of model 1 (video modality only) is 4.12, and the RMSE drops to 3.78 after adding gesture modality (model 2), indicating that BiLSTM effectively extracts the discriminative features of hand movements. Further introduction of text modality (model 3) reduces the RMSE to 3.23, verifying the supplementary value of communication content to scoring. The addition of the CMA module (model 4) brings significant improvement (RMSE=3.05). The core of this is to dynamically model the implicit association between video, gesture, and text through the cross-modal attention mechanism, breaking through the dependence of the early fusion method on the accuracy of modal alignment. Experimental data show that the CMA module is irreplaceable for multimodal collaborative modeling in low-sample scenarios and is significantly better than single-modal supplementation. This result verifies the effectiveness of the CMA module proposed in this paper under low-sample conditions.

The self-distillation strategy (model 5) further reduces RMSE to 2.98 and improves PCC to 0.89 based on model 4, revealing its key optimization capabilities in low-sample scenarios. Comparing model 4 with model 5, it can be seen that self-distillation effectively alleviates the problem of local overfitting of the student model to hard labels through the soft label supervision of the teacher model. The self-distillation and CMA modules have significant synergistic effects. CMA is responsible for multimodal interaction modeling, while self-distillation optimizes the feature space through global distribution consistency constraints. The combined effect of the two increases the PCC from 0.84 in model 3 to 0.89 in model 5.

The comparison results of baseline models are shown in Table 5.

The experimental results show that the complete model proposed in this paper (RMSE=2.98, PCC=0.89) is superior to all baseline models under very low sample size (10 training samples). Traditional methods such as logistic regression and SVM have poor performance (rmse>4.8) because they can not effectively model multimodal time series. Although the single-mode model is slightly improved, it is still far less than the multimode fusion model, indicating that multisource information complementarity is very important for the scoring task.

E. GENERALIZATION ABILITY AND SYSTEM PERFORMANCE

A 10-fold cross-validation is performed on 90 samples, and the results of each fold are analyzed to measure the generalization ability of the nursing skill scoring system. The feature extraction, cross-modal fusion, and self-distillation prediction modules are integrated into a unified framework. A system service layer has been developed, comprising a scoring engine, result interpretation module, and alarm module. Through model quantization and TensorRT acceleration, a microservice architecture is deployed to support high concurrent access, and the response time and resource consumption performance of different systems are compared. The results are shown in Figure 7.

The generalization ability of 90 samples is strictly evaluated through the ten fold cross validation. The results show that the proposed system shows excellent stability and robustness in different partitions within the data set. In all 10-fold tests, the RMSE of the proposed system is consistently maintained in the range of 0.01- 0.03, with an average of 0.016 ± 0.007, which is lower than that of the comparison models (CrossViT:

0.058 ± 0.009; UniViLM: 0.067 ± 0.008). The proposed system achieves the lowest error in all folds, with a small fluctuation range, which verifies the key role of the self-distillation strategy and cross-modal attention fusion in improving model stability. Further analysis shows that in the most challenging 1st fold test, the RMSE of this system (0.03) is still 50% better than the best performing comparison model TimeSformer-Distill (0.06), proving its strong adaptability to unconventional operations. The range between folds is 0.02. This stability is particularly important in clinical nursing scenarios. In the system performance test, this system showed excellent operating efficiency and resource utilization. The response time reaches 17ms, which fully meets the real-time scoring requirements (clinical requirements $\leq$ 200ms). The CPU (Central Processing Unit) utilization is only 22.5%, significantly lower than the comparison system (35.0%-47.3%), which means that more concurrent evaluation tasks can be supported under the same hardware conditions. This stable performance, combined with the high accuracy in the 10-fold verification, proves the applicability of this system in real clinical environments.

TABLE 4. Ablation experiment results

Modules	Model 1	Model 2	Model 3	Model 4	Model 5
TimeSformer	✓	✓	✓	✓	✓
BiLSTM		✓	✓	✓	✓
RoBERTa			✓	✓	✓
CMA				✓	✓
Self-Distillation					✓
RMSE	4.12	3.78	3.23	3.05	2.98
MSE	18.98	15.51	11.33	9.12	8.88
PCC	0.78	0.82	0.84	0.86	0.89

TABLE 5. baseline model comparison results

Model type	Modal combination	RMSE	MAE	PCC
Logistic regression (baseline)	Video+gesture+text	5.12	4.45	0.62
SVM (baseline)	Video+gesture+text	4.87	4.21	0.65
Single mode CNN (video)	Video	4.35	3.78	0.72
Single mode bilstm (gesture)	Gesture	4.56	3.95	0.68
Monomode Roberta (text)	Text	4.89	4.32	0.64

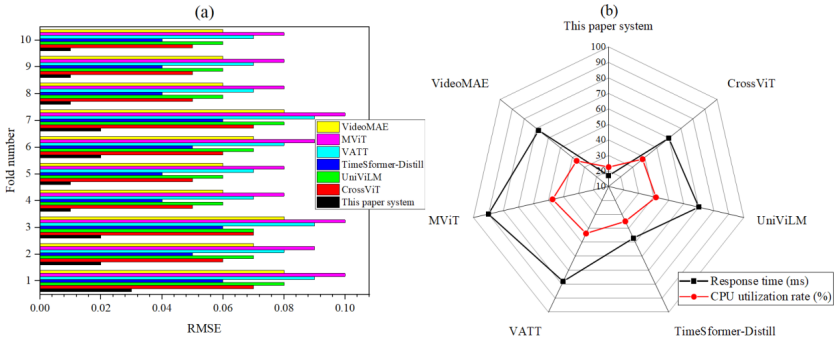

FIGURE 7. Generalization ability and system performance. Figure 7(a) 10-fold cross validation results; Figure 7(b) System performance

F. SUPERPARAMETRIC SENSITIVITY ANALYSIS

To verify the rationality of the key superparameter selection of teacher model update frequency, we compared the effects of different update frequencies (1, 3, 5, 10, 20 cycles) on the performance of the model under the setting of 10 training samples. The results are shown in Table 6.

TABLE 6. Superparametric Sensitivity Analysis

Update frequency (cycle)	RMSE	PCC
1	3.15	0.86
3	3.02	0.88
5	2.98	0.89
10	3.11	0.87
20	3.25	0.85

The analysis results show that the model achieves the best RMSE and PCC performance when the update frequency is 5 cycles. Too low frequency (20 cycles) leads to knowledge transfer lag and performance degradation; If the frequency is too high (1 cycle), the training stability will be affected by the increase of teacher model noise. Therefore, it is reasonable and effective to choose 5 cycles as the update frequency.

VI. CONCLUSION

This study innovatively integrates the self-distillation strategy with CMA to successfully address the challenge of the nursing skill scoring system's insufficient generalization ability under low sample conditions, achieving excellent performance with an RMSE of 2.98 and a PCC of 0.89 using only 10 training samples. This paper introduces the dynamically updated teacher-student framework into the field of nursing scoring, and avoids knowledge solidification through periodic parameter refresh. The CMA module designed in this paper realizes the collaborative optimization of video, gesture, and text modalities in an interpretable way, and reduces RMSE through feature fusion. In practical applications, the system can maintain high generalization performance in cross-validation results while maintaining 17ms real-time response. In summary, the nursing skill scoring method based on self-distillation and CMA proposed in this paper shows excellent generalization performance and stability in resource-constrained scenarios, provides

a generalizable technical path for low-sample multimodal learning, and has important clinical application value.

AUTHOR CONTRIBUTIONS

Conceptualization –Yongqi Zheng, Fangfang Zhang, Xiuxiu Wu and Jingya Zhang; methodology – Yongqi Zheng, Fangfang Zhang, Xiuxiu Wu and Jingya Zhang; investigation – Yongqi Zheng, Fangfang Zhang and Jingya Zhang; writing-original draft preparation – Yongqi Zheng, Fangfang Zhang, Xiuxiu Wu and Jingya Zhang. 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 received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of the first affiliaten hospital of Zhengzhou university, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] L. Cui, Y. Dong, S. Zhang, and X. Shi, "The value of the '7E'instructional model in the teaching of nursing students in nursing clinical probation," BMC Medical Education, vol. 24, no. 1, pp. 1222, 2024, doi: 10.1186/s12909-024-06157-9.

[2] K. Leighton, S. Kardong-Edgren, A. M. McNelis, C. Foisy-Doll, and E. Sullo, "Traditional clinical outcomes in prelicensure nursing education: An empty systematic review," Journal of Nursing Education, vol. 60, no. 3, pp. 136-142, 2021, doi: 10.3928/01484834-20210222-03.

[3] F. Imrie, S. Denner, L. S. Brunschwig, K. Maier-Hein, and M. Van Der Schaar, "Automated ensemble multimodal machine learning for healthcare," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 6, pp. 4213-4226, 2025, doi: 10.1109/JBHI.2025.3530156.

[4] S. M. Yu, M. J. Lan, J. Feng, T. T. Wang, and X. M. Hu, "Construction and implementation of the automated management system for outpatient nursing," Chinese Journal of Nursing Management, vol. 24, no. 2, pp. 261-265, 2024, doi: 10.3969/j.issn.1672-1756.2024.02.019.

[5] J. Vierula, E. Haavisto, M. Hupli, and K. Talman, "The assessment of learning skills in nursing student selection: A scoping review," Assessment & Evaluation in Higher Education, vol. 45, no. 4, pp. 496-512, 2020, doi: 10.1080/02602938.2019.1666970.

[6] A. Nakazawa, Y. Mitsuzumi, Y. Watanabe, R. Kurazume, S. Yoshikawa, and M. Honda, "First-person video analysis for evaluating skill level in the humanitude tender-care technique," Journal of Intelligent & Robotic Systems, vol. 98, pp. 103-118, 2020, doi: 10.1007/s10846-019-01052-8.

[7] M. Alastalo, L. Salminen, T. Vahlberg, and H. Leino-Kilpi, "Subjective and objective assessment in skills evaluation: A cross-sectional study among critical care nurses," Nordic Journal of Nursing Research, vol. 43, no. 1, Art. no. 20571585221089145, 2023, doi: 10.1177/20571585221089145.

[8] G. Zhang, X. Shen, Y. D. Zhang, Y. Luo, J. Luo, D. Zhu, et al., "Cross-modal prostate cancer segmentation via selfattention distillation," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5298-5309, 2021, doi: 10.1109/JBHI.2021.3127688.

[9] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, et al., "CKD-TransBTS: Clinical Knowledge-Driven Hybrid Transformer With Modality-Correlated Cross-Attention for Brain Tumor Segmentation," IEEE Transactions on Medical Imaging, vol. 42, no. 8, pp. 2451-2461, 2023, doi: 10.1109/tmi.2023.3250474.

[10] M. He, W. Du, Z. Wen, Q. Du, Y. Xie, and Q. Wu, "Multi-granularity aggregation transformer for joint video-audiotext representation learning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 6, pp. 2990-3002, 2022, doi: 10.1109/TCSVT.2022.3225549.

[11] B. S. Zhang and H. Fan, "Improved MobileViT algorithm for small samples," Journal of Computer Engineering & Applications, vol. 60, no. 22, pp. 251-251, 2024, doi: 10.3778/j.issn.1002-8331.2307-0250.

[12] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. -4403, 2021, doi: 10.1109/TPAMI.2021.3067100.

[13] S. An, Q. Liao, Z. Lu, and J. H. Xue, "Efficient semantic segmentation via self-attention and self-distillation," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 15256-15266, 2022, doi: 10.1109/TITS.2021.3139001.

[14] P. Qi, L. Chai, and X. Ye, "Unsupervised industrial anomaly detection based on feature mask generation and reverse distillation," Chinese Journal of Information Fusion, vol. 1, no. 2, pp. 160-174, 2024, doi: 10.62762/cjif.2024.734267.

[15] X. Ge, Y. Yang, L. Peng, et al., "Spatio-temporal knowledge graph based forest fire prediction with multi source heterogeneous data," Remote Sensing, vol. 14, no. 14, pp. 3496, 2022, doi: 10.3390/rs14143496.

[16] C. Buchanan, M. L. Howitt, R. Wilson, R. Booth, T. Risling, and M. Bamford, "Predicted influences of artificial intelligence on nursing education: Scoping review," JMIR nursing, vol. 4, no. 1, pp. 23933-23933, 2021, doi: 10.2196/preprints.23939.

[17] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 4, pp. 447-453, 2020, doi: 10.1016/j.jksuci.2019.09.004.

[18] F. Gu, M. H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," ACM Computing Surveys (CSUR), vol. 54, no. 8, pp. 1-34, 2021, doi: 10.1145/3472290.

[19] N. Pang, S. Guo, M. Yan, and C. Chan, "A Short Video Classification Framework Based on Cross-Modal Fusion," Sensors, vol. 23, no. 20, pp. 8425-8425, 2023, doi: 10.3390/s23208425.

[20] L. Wang, M. Cao, Y. Zhong, and X. Yuan, "Spatial-temporal transformer for video snapshot compressive imaging," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 7, pp. 9072-9089, 2022, doi: 10.1109/TPAMI.2022.3225382.

[21] R. P. Singh and L. D. Singh, "Dyhand: Dynamic hand gesture recognition using BiLSTM and soft attention methods," The Visual Computer, vol. 41, no. 1, pp. 41-51, 2025, doi: 10.1007/s00371-024-03307-4.

[22] Y. Song, M. Liu, F. Wang, J. Zhu, A. Hu, and N. Sun, "Gesture recognition based on a convolutional neural network– bidirectional long short-term memory network for a wearable wrist sensor with multi-walled carbon nanotube/cotton fabric material," Micromachines, vol. 15, no. 2, pp. 185, 2024, doi: 10.3969/j.issn.1671-1815.2022.04.028.

[23] Y. Wu, J. Huang, C. Xu, H. Zheng, L. Zhang, and J. Wan, "Research on named entity recognition of electronic medical records based on RoBERTa and radical-level feature," Wireless Communications and Mobile Computing, vol. 2021, no. 1, pp. 2489754-2489754, 2021, doi: 10.1155/2021/2489754.

[24] M. E. Chatzimina, H. A. Papadaki, C. Pontikoglou, and M. Tsiknakis, "A comparative sentiment analysis of Greek clinical conversations using BERT, RoBERTa, GPT-2, and XLNet," Bioengineering, vol. 11, no. 6, pp. 521, 2024, doi: 10.3390/bioengineering11060521.

[25] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," IEEE transactions on knowledge and data engineering, vol. 35, no. 4, pp. 3279-3298, 2021, doi: 10.1109/TKDE.2021.3126456.

[26] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," Neural Computing and Applications, vol. 34, no. 16, pp. 13371-13385, 2022, doi: 10.1007/s00521-022-07366-3.

[27] M. Tzelepi, N. Passalis, and A. Tefas, "Probabilistic online self-distillation," Neurocomputing, vol. 493, pp. 592-604, 2022, doi: 10.1016/j.neucom.2021.12.101.

[28] N. Paluru, H. Ravishankar, S. Hegde, and P. K. Yalavarthy, "Self distillation for improving the generalizability of retinal disease diagnosis using optical coherence tomography images," IEEE Journal of Selected Topics

in Quantum Electronics, vol. 29, no. 4: Biophotonics, pp. 1-12, 2023, doi: 10.1109/JSTQE.2023.3240729.

[29] S. Lee, W. Im, and S. E. Yoon, "Multi-resolution distillation for self-supervised monocular depth estimation," Pattern Recognition Letters, vol. 176, pp. 215-222, 2023, doi: 10.1016/j.patrec.2023.11.001.

[30] C. A. Ong, F. S. Tay, Y. L. Then, and C. McCarthy, "An evaluation of a pre-trained transformer-based self-distillation model (DINOv2) for cross-domain plant species identification," Neural Computing and Applications, vol. 37, no. 26, pp. 21969-21995, 2025, doi: 10.1007/s00521-025-11499-6.