Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 6 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i06.60865

PAPER

Learner Behavior Recognition and Dynamic Guidance Mechanism in Interactive Mobile Simulation Systems for Vocational Education

Yan Wang, Linli Wei (✉)

Hunan Vocational College Engineering, Hunan, China

13974838841@sohu.com

ABSTRACT

Mobile simulation technologies provide scenario-based solutions for practical instruction in vocational education. However, existing systems are commonly constrained by limited interaction perception modalities, delayed learner behavior recognition, and a lack of personalized guidance strategies, resulting in insufficient adaptability to the acquisition of complex vocational skills. To address these challenges, a learner behavior recognition and dynamic guidance mechanism integrating multimodal perception and imitation learning was proposed, and a three-layer edge-cloud collaborative architecture consisting of perception, understanding, and guidance was established. Multimodal interaction perception was achieved through the synchronous acquisition of user interface (UI) layout trees, device sensor data, and operation sequence data. A lightweight spatiotemporal attention network was designed to perform behavior encoding and recognition, while imitation learning was introduced to enhance recognition performance in long-tail operation scenarios. A context-aware guidance decision engine was constructed using reinforcement learning, enabling the dynamic generation of guidance strategies adapted to learners' skill levels and task contexts. Deployment efficiency on mobile devices was improved through dynamic computation offloading. Experimental results demonstrated that the proposed lightweight spatiotemporal attention model achieved a recognition accuracy of 94.2% for basic operations and a macro-F1 score of 0.876 for long-tail operations, with end-to-end latency consistently maintained at or below 325 ms. The dynamic guidance mechanism reduced average task completion time by 28.3% and decreased operation error rates by 41.2%. In addition, overall system deployment performance significantly outperformed that of existing mobile educational systems. These findings validate the effectiveness and superiority of the proposed approach and provide a practical and efficient technical framework for the deep integration of mobile computing technologies with vocational education, substantially enhancing vocational skill acquisition efficiency.

Wang, Y., Wei, L. (2026). Learner Behavior Recognition and Dynamic Guidance Mechanism in Interactive Mobile Simulation Systems for Vocational Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(6), pp. 25–38. https://doi.org/10.3991/ijim.v20i06.60865

Article submitted 2025-12-03. Revision uploaded 2026-01-27. Final acceptance 2026-02-02.

KEYWORDS
mobile simulation system, vocational education, behavior recognition, multimodal fusion, imitation learning, dynamic guidance, edge computing

1 INTRODUCTION

The core objective of vocational education lies in enabling scenario-based and personalized skill acquisition accompanied by timely feedback [1, 2]. However, traditional offline practical training models are constrained by limitations in equipment availability, physical space, and safety regulations, making it difficult to support large-scale, high-quality instructional delivery [3, 4]. Owing to advantages such as scenario reusability and controllable risk, virtual simulation technologies have been widely adopted as an essential supplement to practical teaching in vocational education [5–7]. With the widespread adoption of mobile intelligent terminals, mobile simulation systems, leveraging their inherent portability [8, 9], have become well suited to fragmented learning contexts [10]. As a result, the application boundaries of virtual simulation technologies in vocational education have been further extended, enabling more ubiquitous forms of vocational skill learning.

Despite the substantial practical value of mobile simulation systems, their deep integration into vocational education remains hindered by three critical technical bottlenecks. First, interaction perception remains limited in scope, as most existing approaches rely primarily on touch-coordinate data, while the semantic structure of UIs and device sensor information—both of which embed operational context—are largely neglected. This limitation results in insufficient robustness in learner behavior perception [11]. Second, significant latency persists in learner behavior recognition and guidance delivery, preventing alignment with the real-time requirements of vocational skill operations and limiting the timely correction of operational deviations [12, 13]. Third, scenario adaptability remains inadequate. The prevalence of long-tail operational scenarios and inter-learner variability in vocational skills exceeds the adaptation capabilities of existing systems, thereby impeding the realization of personalized guidance [13, 14]. Collectively, these challenges constrain the effectiveness of mobile simulation systems in supporting vocational skill acquisition and underscore the need for a novel technical framework. Accordingly, the central research problem is defined as the development of a lightweight, multimodal data fusion-based model capable of achieving real-time and accurate learner behavior recognition in mobile vocational simulation environments, alongside the design of a context-aware dynamic guidance mechanism that effectively balances recognition accuracy, response speed, and mobile resource constraints.

The primary objective of this study is to develop an integrated technical framework for multimodal perception, lightweight behavior recognition, and personalized guidance in mobile simulation environments for vocational education. Through low-latency and high-accuracy behavior recognition combined with dynamic guidance, improvements in learners' vocational skill acquisition efficiency are targeted. In pursuit of this objective, the major contributions of this study are summarized below. First, a multimodal mobile interaction perception paradigm is proposed that integrates user interface (UI) layout trees, device sensor data, and operation sequences, thereby overcoming the limitations of conventional approaches that rely

solely on touch-coordinate information and substantially enriching both the dimensionality and depth of behavior perception. Second, a lightweight behavior recognition model enhanced by imitation learning is designed for long-tail operational scenarios commonly observed in vocational skills. Real-time inference on mobile devices is achieved while maintaining high recognition accuracy, effectively addressing the recognition challenges posed by long-tail behaviors. Third, a context-aware dynamic guidance decision engine is constructed to dynamically generate personalized guidance strategies based on learners' states and task-specific contexts, thereby improving the adaptability and effectiveness of instructional guidance. Fourth, systematic empirical validation is conducted within representative vocational education simulation scenarios, providing reusable technical frameworks and empirical evidence to support the deep integration of mobile computing technologies with vocational education.

The remainder of this study is organized below. Section 2 presents the overall system architecture and details the design of the core innovations, including multimodal perception, lightweight behavior recognition, and dynamic guidance. Section 3 describes the experimental design and reports the experimental results, followed by an in-depth analysis of the effectiveness of the proposed technical mechanism and a comparative discussion with state-of-the-art research. The final section summarizes the main conclusions, discusses the limitations of the study, and outlines directions for future research.

2 SYSTEM ARCHITECTURE AND CORE TECHNOLOGIES

2.1 Overall architecture design

A three-layer edge-cloud collaborative architecture organized around the "perception-interpretation-guidance" paradigm is adopted. The core design principle follows a strategy of lightweight sensing on mobile devices, precise computation at the edge and cloud, and multi-channel real-time feedback. Through hierarchical task allocation and collaborative scheduling, constraints imposed by limited computational and energy resources on mobile devices are balanced while ensuring the timeliness and accuracy of learner behavior recognition and guidance feedback. The mobile perception layer is dedicated to low-energy, low-latency data acquisition and preprocessing, providing high-quality inputs for subsequent analysis. The edge-cloud analysis layer is responsible for computation-intensive tasks, including behavior recognition and guidance decision-making. Precise analysis is achieved by leveraging the low transmission latency of edge nodes in conjunction with the superior computational capacity of cloud resources. The guidance feedback layer delivers decision outcomes to learners through the coordinated multi-channel output mechanism, thereby forming a closed-loop intervention process encompassing data acquisition, analysis, and feedback. The collaborative efficiency of the three-layer architecture is governed by dynamic computational task allocation. Computational workloads are adaptively distributed between local devices and edge/cloud resources based on task complexity and network conditions. For example, basic recognition tasks associated with simple operational scenarios are executed locally to minimize transmission latency, whereas imitation-based matching tasks for complex long-tail operations are offloaded to edge nodes, maintaining efficient system responsiveness across diverse scenarios.

2.2 Mobile perception layer: multimodal data fusion-based perception technologies

To overcome the limitations of conventional mobile simulation systems that rely solely on touch-coordinate data, a multimodal data fusion-based perception scheme is designed by integrating three core data modalities: UI layout trees, device sensor data, and operation sequences. Through this integration, a comprehensive operational context perception dimension is constructed. UI layout tree data are obtained in real time via mobile Application Programming Interfaces (APIs), enabling the extraction of interface control types, spatial positions, and state information, which are subsequently organized into a structured tree representation. This approach avoids the high energy consumption and latency associated with traditional screen-recording-based methods. Device sensor data are synchronously collected from gyroscopes, accelerometers, and touch sensors to capture the physical posture of the device and touch characteristics during learner interactions. Operation sequence data record the temporal order of control interactions and foreground-background application switching states, thereby forming continuous chains of operational behaviors. Multisource data synchronization is achieved using system timestamps with millisecond-level alignment. A timestamp deviation threshold of $\tau \leq 5$ ms is defined to ensure temporal consistency among UI layout changes, sensor signals, and operational behaviors. During preprocessing, targeted strategies are applied to enhance data quality. Sensor data are denoised using Kalman filtering, with the core state update equation defined as:

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_k + K_k(z_k - H\hat{x}_{k-1}) \tag{1}$$

where, $\hat{x}_k$ denotes the filtered state at time step k, A represents the state transition matrix, K_k is the Kalman gain, and z_k corresponds to the observed measurement. Through this formulation, sensor noise is effectively suppressed. UI layout tree features are encoded using one-hot representations of control types, followed by coordinate normalization and projection into a low-dimensional vector:

$$v = [t_{norm}, x_{norm}, y_{norm}, s_{norm}] \tag{2}$$

where, t_{norm} denotes the normalized type encoding, x_{norm} and y_{norm} represent normalized spatial coordinates, and s_{norm} corresponds to the normalized state. Operation sequence data are generated using a sliding window with a fixed length of $L = 20$, producing fixed-length input sequences to ensure dimensional consistency for subsequent model processing.

2.3 Core technologies of the edge-cloud analysis layer

To balance mobile resource constraints with the accuracy and real-time requirements of behavior recognition, a lightweight spatiotemporal attention network is designed to enable efficient fusion and precise extraction of multimodal features. Figure 1 illustrates the three-branch parallel encoding architecture, as well as the internal structures of the spatiotemporal attention fusion module and the intent filtering mechanism.

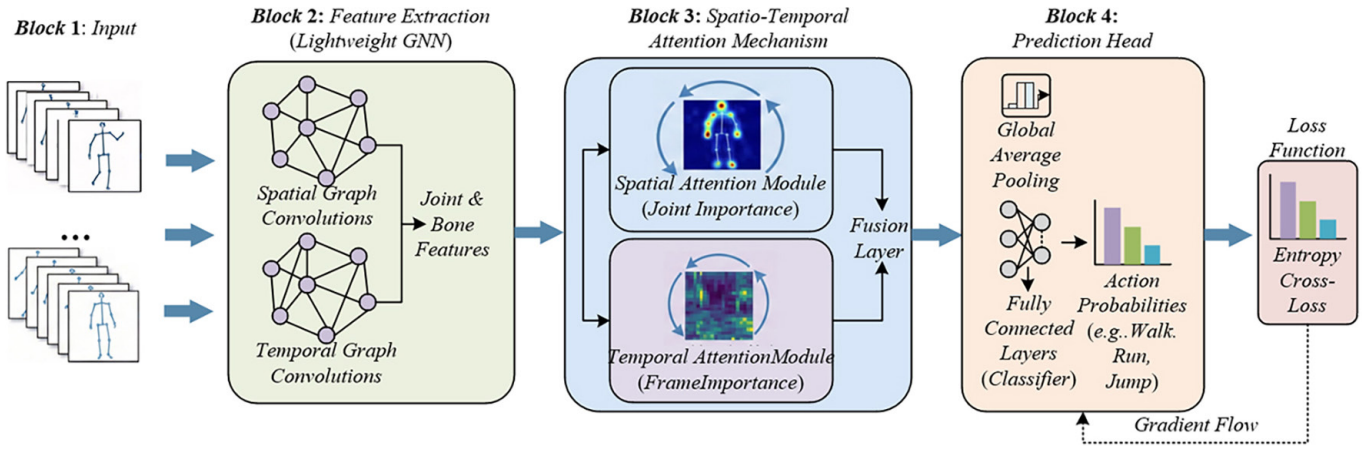

Fig. 1. Architecture of the behavior recognition model based on a lightweight GNN and a spatiotemporal attention mechanism

As shown in the figure, a three-branch parallel encoding architecture is adopted. The UI layout tree encoding branch is constructed using a lightweight graph neural network (GNN), in which controls are modeled as nodes and spatial relationships are represented as edges. A graph structure is thus formed, and spatial relational features among controls are captured through a single graph convolution layer:

$$H' = \sigma(\tilde{A}HW) \tag{3}$$

where, $\tilde{A}$ denotes the normalized adjacency matrix, H represents the node feature matrix, and W is the convolution weight. The sensor data encoding branch employs a one-dimensional convolutional neural network (1D-CNN). Temporal features are extracted using three convolutional layers with a kernel size of 3 and a stride of 1, producing an output feature dimension of 128. The operation sequence encoding branch is implemented using a bidirectional long short-term memory (BiLSTM) network with a hidden state dimension of 128, enabling the capture of temporal dependencies in operational behaviors. The core innovation lies in the spatiotemporal attention fusion module, in which adaptive attention weights are used to dynamically allocate the relative contributions of different modality features. The weights are computed as:

$$\alpha_i = \frac{\exp(W_i f/\sqrt{d})}{\sum_{j=1}^{3}\exp(W_j f/\sqrt{d})} \tag{4}$$

where, f_i denotes the feature representation of the i-th modality, W_i represents the weight matrix, and d denotes the feature dimensionality. In addition, an intent filtering mechanism is introduced to suppress noisy operational features with attention weights lower than a predefined threshold of $\theta = 0.3$. The behavior recognition head is implemented using two fully connected layers, which output the probability distribution over behavior categories. To further reduce the model size, knowledge distillation is incorporated. The distillation loss function is defined as:

$$L = \lambda L_{ce} + (1-\lambda)L_{kd} \tag{5}$$

where, L_{ce} denotes the cross-entropy loss, L_{kd} represents the Kullback-Leibler (KL) divergence-based loss, and $\lambda = 0.7$. The final model parameter count is constrained to 0.876 million, enabling real-time inference to be supported directly on mobile devices.

To address long-tail operational scenarios prevalent in vocational skills, a key-operation demonstration-enhanced framework is designed to improve recognition adaptability through the construction of a structured demonstration repository and a many-to-one matching model. Demonstration data are recorded by domain experts following critical skill execution procedures. Multimodal features and operational logic labels are extracted using a customized DemoParser tool, resulting in a demonstration repository comprising 40 distinct skill units. The core innovation lies in a many-to-one matching algorithm based on dynamic time warping (DTW) with cosine similarity fusion. First, DTW is employed to compute the temporal similarity between the real-time behavior sequence and each demonstration segment:

$$S_{dtw} = \min \sum_{i,j} |f_{r,i} - f_{d,j}| \tag{6}$$

where, $f_{r,i}$ denotes the real-time feature, and $f_{d,j}$ represents the demonstration feature. Subsequently, cosine similarity between feature vectors is computed as:

$$S_{cos} = \frac{f_r \cdot f_d}{\|f_r\| \cdot \|f_d\|} \tag{7}$$

The fused similarity score is then obtained as follows:

$$S = \alpha S_{cos} + (1-\alpha)(1 - \frac{S_{dtw}}{S_{dtw,max}}) \tag{8}$$

where, $\alpha = 0.6$ is the weighting coefficient. Recognition results are produced using a weighted fusion strategy. The fusion formula combining the base model output probability with the demonstration-matching similarity is given as:

$$P_{final} = \beta P_{base} + (1-\beta)S \tag{9}$$

The proposed framework effectively addresses the limitation whereby a single demonstration segment fails to cover diverse operational variants, thereby improving recognition accuracy in long-tail scenarios.

To achieve the objective of minimal intervention with maximal benefit, a dynamic guidance decision engine based on lightweight reinforcement learning is constructed, with a focus on multidimensional context modeling and adaptive policy generation. Contextual information modeling integrates real-time interaction states, historical skill data, task scenario information, and environmental context, forming a high-dimensional state representation defined as $\mathbf{s}_t = [\mathbf{s}_{int}, \mathbf{s}_{his}, \mathbf{s}_{task}, \mathbf{s}_{env}]$, with an overall dimensionality of 256. The action space comprises three categories of decision variables: guidance timing, guidance method, and guidance granularity. These variables collectively yield 16 composite actions. A multi-objective reward function is designed as:

$$R_t = w_1 R_{eff} + w_2 R_{err} + w_3 R_{acc} - w_4 R_{int} - w_5 R_{delay} \tag{10}$$

where, R_{eff} denotes the reward for task efficiency improvement, R_{err} represents the reward for error rate reduction, R_{acc} corresponds to the guidance acceptance reward, R_{int} is the penalty for excessive intervention, and R_{delay} denotes the delay penalty. The weighting coefficients w_1–w_5 are dynamically adjusted within the range

of 0.1 to 0.4. The decision model is trained using the Proximal Policy Optimization (PPO) algorithm. Model quantization and layer pruning are applied to achieve lightweight deployment, ensuring that inference latency at edge nodes is maintained at or below 100 ms. Through this design, optimal guidance strategies are dynamically generated in response to real-time learner states. Figure 2 illustrates the closed-loop process of the state space, action space, reward function design, and decision generation within the reinforcement learning environment.

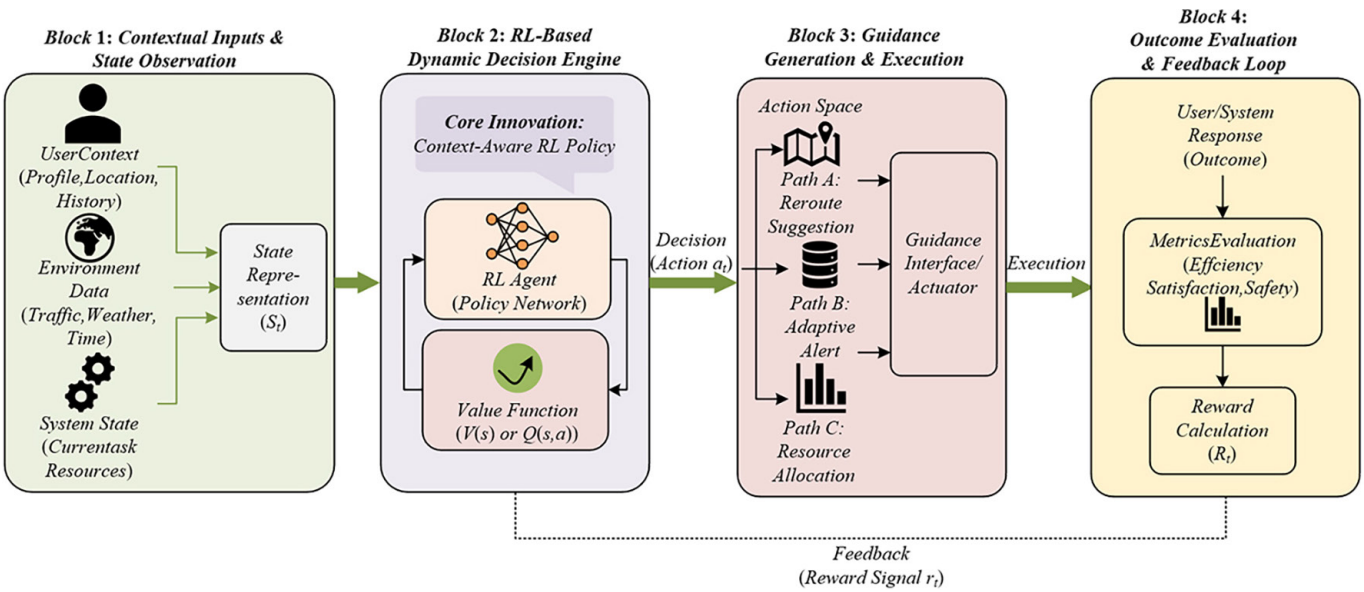

Fig. 2. Framework of the reinforcement learning-based context-aware dynamic guidance decision engine

2.4 Mobile deployment optimization strategies

To resolve the tension between mobile resource constraints and real-time system requirements, a deployment optimization strategy is proposed that combines model lightweighting, dynamic offloading, and resource management. The core innovation lies in adaptive task allocation and fine-grained, end-to-end resource regulation. At the model lightweighting level, joint optimization is applied to the perception-layer preprocessing models and the base behavior recognition model. INT8 quantization is employed to reduce parameter precision from 32-bit to 8-bit, resulting in a 75% reduction in model size. In addition, structured pruning is introduced with a pruning ratio of 30% applied to convolutional and fully connected layers. Key weights are selected using L1 regularization, defined as $R(W) = \lambda \sum_{i,j} |W_{ij}|$. Furthermore, computational logic is fused across adjacent convolutional and batch normalization layers to reduce memory access overhead during inference.

Meanwhile, the OpenCL acceleration interface of the mobile GPU is invoked to offload convolution and matrix multiplication operations to GPU execution. Through these measures, the local inference latency of the base recognition model is constrained to within 183 ms, thereby satisfying real-time performance requirements. The dynamic computation offloading mechanism is governed by a joint decision strategy that considers both network quality and task complexity. A decision function is defined as $D = \gamma B + (1 - \gamma) C$, where B denotes network bandwidth, C represents task computational load, and $\gamma = 0.4$. When $D < D_{th}$, computation-intensive

tasks—such as demonstration matching and reinforcement learning-based decision making—are offloaded to edge servers. Leveraging the low-latency transmission characteristics of 5G and Wi-Fi 6 networks, end-to-edge data interaction latency is maintained at or below 30 ms. Resource utilization is controlled through the Linux kernel Control Groups (CGroup) mechanism, by which isolated resource domains are defined. The CPU peak utilization of model inference processes is limited to 28%, GPU utilization is constrained to 40%, and memory consumption is capped at 185 MB. In addition, dynamic voltage and frequency scaling is applied to reduce battery power consumption during continuous operation by 25%, thereby preventing interference with other device functionalities.

2.5 Guidance feedback layer: multichannel real-time feedback implementation

To accommodate the diverse guidance requirements of vocational education scenarios, a multichannel collaborative feedback mechanism is designed. The core innovation lies in noise-adaptive regulation and synchronized multichannel triggering, ensuring efficient delivery of guidance information. Visual guidance is implemented through the mobile UI rendering engine using multi-level feedback strategies. UI control highlighting is achieved by overlaying semi-transparent masks, with Red-Green-Blue (RGB) color values employed to enhance visual salience. Step-by-step prompt cards are presented as floating windows that support gesture-based navigation. Virtual agent operation demonstrations are delivered via pre-rendered animation sequences, enabling rapid playback while reducing real-time rendering overhead. Auditory guidance integrates an adaptive volume regulation module. Environmental noise intensity N is continuously captured through the device microphone. When $N > 60$ dB, automatic volume adjustment is triggered according to:

$$V_{adj} = V_{base} \times (1 + kN) \tag{11}$$

where, V_{base} denotes the baseline volume and $k = 0.005$. In parallel, speech enhancement algorithms are applied to suppress background noise. Haptic guidance is implemented via the device vibration motor API. Vibration signals with different frequencies are generated according to the severity of operational deviation: minor deviations trigger short vibrations at 100 Hz, whereas severe deviations trigger longer vibrations at 200 Hz. Synchronization of multichannel feedback is achieved through system timestamp alignment. A triggering deviation threshold of no more than 5 ms is enforced to ensure temporal consistency among visual, auditory, and haptic feedback, enabling learners to rapidly perceive guidance information and adjust their operational behavior accordingly.

3 EXPERIMENTS AND EVALUATION

3.1 Experimental design and dataset construction

The experiments were designed to systematically validate the core performance of the proposed technical framework. The primary objectives include evaluating the accuracy and real-time performance of the multimodal fusion recognition model, assessing the effectiveness of imitation learning in improving recognition of long-tail

operational behaviors, examining the personalization capability and intervention effectiveness of the dynamic guidance mechanism, and verifying system-level deployment performance on mobile devices. Two representative vocational education scenarios were selected: an electrical circuit maintenance simulation and a mechanical equipment disassembly and assembly simulation. The former comprises 12 basic operations and 8 long-tail fault-handling operations, whereas the latter includes 15 basic operations and 6 long-tail personalized operations. Both scenarios are closely aligned with practical vocational training requirements and collectively cover both common and individualized characteristics of skill execution. To ensure coverage of mainstream mobile hardware environments, three mid-range Android smartphones with different configurations were selected as experimental devices. The edge server was configured with an Intel Core i7-12700H processor and 32 GB of memory. Network conditions were established using 5G and Wi-Fi 6 to guarantee low-latency computation offloading and data transmission, thereby ensuring the generalizability and reliability of the experimental results.

Dataset construction emphasized multimodality and scenario adaptability, with particular attention given to data innovation and annotation rigor. During data collection, 60 participants were recruited and evenly divided into three groups according to skill proficiency: novices, intermediate learners, and domain experts. While completing simulation tasks in both experimental scenarios, multimodal interaction data—including UI layout trees, sensor data, and operation sequences—were synchronously collected. The resulting dataset consisted of 1,200 complete task sequences and 5,000 individual operation segments. Demonstration data were recorded by five domain experts following standardized procedures for critical skill execution, resulting in a structured demonstration repository containing 40 key skill units. Data annotation was conducted through a collaborative process involving domain experts and researchers. Annotation categories included 15 classes of basic operations, 14 classes of long-tail operations, and 3 classes of erroneous operations, as well as labels for operation deviation types and guidance acceptance levels. To ensure annotation quality, inter-annotator consistency was evaluated using Cohen's Kappa coefficient, with results indicating Kappa values of at least 0.85. These findings confirm a high level of annotation reliability and provide a robust data foundation for subsequent experimental evaluation.

3.2 Core experimental results and quantitative analysis

In this subsection, experimental results are presented according to the major technical innovations, with quantitative comparisons combined with statistical significance testing to systematically demonstrate the performance advantages of the proposed approach. All experiments were conducted using five-fold cross-validation to reduce random variance, and statistical significance was verified through two-way analysis of variance (ANOVA), ensuring the reliability of the conclusions.

Table 1 reports the performance comparison between the proposed lightweight spatiotemporal attention-based model and three categories of baseline models. It is observed that the proposed model achieves an accuracy of 94.2% and an F1-score of 0.938 on basic operation recognition tasks. For long-tail operation recognition, a macro F1-score of 0.876 and a micro F1-score of 0.891 are obtained. Compared with the unimodal long short-term memory (LSTM), the generic multimodal model, and the MobileNet+LSTM baseline, improvements in macro F1-score for long-tail operations of 21.3%, 14.7%, and 8.9%, respectively, are achieved. The total number of

model parameters is limited to 0.876 million, and the average local inference latency is maintained at 183 ms. These results significantly outperform the baseline models in terms of parameter scale and latency, thereby fully satisfying the real-time inference requirements of mobile deployment scenarios.

Table 1. Performance comparison of multimodal fusion-based behavior recognition models

Model	Basic Operation Accuracy (%)	Basic Operation F1-Score	Long-Tail Operation Macro F1	Long-Tail Operation Micro F1	Parameters ($\times 10^4$)	Local Inference Latency (ms)
Unimodal LSTM	85.3	0.846	0.663	0.689	62.8	165
Generic multimodal model	89.7	0.892	0.729	0.754	156.3	287
MobileNet + LSTM	92.1	0.915	0.787	0.803	112.5	224
Proposed model	94.2	0.938	0.876	0.891	87.6	183

To further validate the necessity of multimodal fusion and the spatiotemporal attention module, ablation experiments were conducted, with the results summarized in Table 2. When the UI layout tree encoding, sensor data encoding, or spatiotemporal attention module is individually removed, the macro F1-score for long-tail operation recognition is observed to decrease by more than 10% in all cases. The most pronounced degradation occurs when the UI layout tree encoding is excluded, indicating that the structured semantic information embedded in UI layout trees provides critical support for recognizing complex interface operations. When the spatiotemporal attention module is removed, the interference rate of accidental touch operation recognition increases by 32.6%. This result confirms the effectiveness of the intent filtering mechanism in suppressing noise and further demonstrates that adaptive fusion of multimodal features constitutes a core factor in enhancing recognition robustness.

Table 2. Ablation study results of the multimodal fusion model

Model Variant	Long-Tail Operation Macro F1	Accidental Touch Recognition Interference Rate (%)
Full model	0.876	18.3
Without the UI layout tree encoding	0.744	21.5
Without the sensor data encoding	0.761	25.8
Without the spatiotemporal attention module	0.772	50.9

Table 3. Performance comparison of imitation learning-based enhancement strategies

Method	Long-Tail Operation Accuracy (%)	Matching Accuracy in Operational Variant Scenarios (%)	Average Matching Latency (ms)
Without imitation learning	78.5	–	–
Imitation learning + one-to-one matching	84.7	75.3	68.2
Imitation learning + many-to-one matching	89.3	92.1	44.8

The enhancement effect of imitation learning on long-tail operation recognition is summarized in Table 3. In the absence of imitation learning, recognition accuracy for long-tail operations is limited to 78.5%. After imitation learning is introduced, recognition accuracy increases to 89.3%, representing an absolute improvement of 10.8%. The most pronounced gains are observed in rare fault-handling operations. A comparison between the proposed many-to-one matching algorithm and a conventional one-to-one matching strategy shows that matching accuracy achieved by the many-to-one algorithm increases by 16.8% under operational variant scenarios, while average matching latency is reduced by 23.4 ms. These results demonstrate that the proposed matching algorithm can effectively accommodate operational variability while maintaining high matching efficiency. Boundary condition analysis indicates that when the number of demonstration segments reaches or exceeds 30, fluctuations in the macro F1-score for long-tail operations remain within 1.5%, suggesting that performance converges to a stable level. This threshold provides quantitative guidance for demonstration repository construction in subsequent system deployment, enabling recognition performance to be maintained while reducing the cost of demonstration data collection.

Figure 3 compares the performance differences between the proposed dynamic guidance strategy and two baseline guidance approaches. An overall guidance acceptance rate of 86.7% is achieved by the proposed approach. Compared with static rule-based guidance and single-channel guidance, average task completion time is reduced by 28.3% and 19.5%, respectively, while operational error rates are reduced by 41.2% and 27.8%. From the perspective of personalized adaptation, the dynamic guidance strategy demonstrates significant improvements across learners with different skill levels. For novice learners, operational error rates are reduced by 48.5%, and task completion time is shortened by 32.1%. For intermediate learners, error rates are reduced by 35.7%, and task completion time is shortened by 24.6%. These results indicate that dynamic guidance strategies grounded in learner state profiling are capable of accurately aligning with the needs of learners at different skill levels and delivering personalized interventions. The enhancement effect is particularly pronounced for novice learners with limited skill foundations.

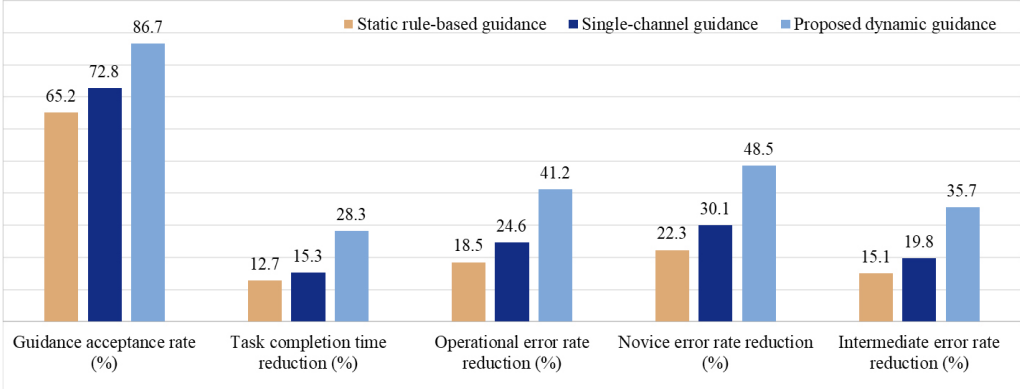

Fig. 3. Performance comparison of guidance mechanisms

Table 4 presents the deployment performance of the system on three mid-range Android smartphones with different hardware configurations, together with a comparative analysis against existing mainstream mobile education systems. The average end-to-end latency observed on the three devices is 298 ms, 312 ms, and 325 ms, respectively, all of which remain below the real-time threshold of 350 ms. Peak CPU utilization is maintained at or below 28%, memory consumption does not

exceed 185 MB, and battery power consumption during one hour of continuous operation remains within 15%. In comparison with the AMI system and the mobile-adapted version of LearnAct, the proposed system achieves an average reduction of 34.6% in end-to-end latency and a 27.3% decrease in memory usage, demonstrating a clear advantage in deployment efficiency. Validation of the dynamic offloading strategy under complex long-tail operational scenarios indicates that, after offloading is applied, average end-to-end latency is further reduced by 42 ms, while CPU utilization decreases by 12.5%. These results confirm that the edge-cloud collaborative dynamic offloading mechanism effectively alleviates mobile resource constraints and further optimizes overall system efficiency. User experience evaluation results indicate a mean System Usability Scale (SUS) score of 82.3, corresponding to an excellent usability level. In addition, 83.3% of learners report satisfaction with the multichannel guidance mechanism. Among them, 65.0% express a preference for the combined guidance mode integrating UI highlighting and voice prompts. In noisy workshop environments, this mode achieves an information transmission accuracy of 92.1%, demonstrating strong adaptability to the complex conditions characteristic of vocational education settings.

Table 4. Comparison of mobile deployment performance

Device/System	End-to-End Latency (ms)	Peak CPU Utilization (%)	Memory Usage (MB)	Battery Consumption Over 1 h (%)
Smartphone A	298	25	168	12
Smartphone B	312	28	176	13
Smartphone C	325	26	185	15
AMI system (Smartphone A)	456	38	231	19
LearnAct (Smartphone A)	472	42	245	21

4 CONCLUSIONS AND FUTURE WORK

In response to key challenges in mobile vocational education simulation environments—including delayed learner behavior recognition, insufficient personalization of guidance, and limited resource adaptability—a lightweight behavior recognition model integrating multimodal perception and imitation learning was developed, alongside a context-aware dynamic guidance mechanism. Together, these components formed an end-to-end technical framework encompassing perception, recognition, and guidance. Experimental validation demonstrated that the multimodal fusion recognition model achieved strong performance across both basic and long-tail operation recognition tasks, with the macro F1-score for long-tail operations improved by up to 21.3% relative to baseline models. At the same time, the model parameter count was constrained to 0.876 million and local inference latency was limited to 183 ms, achieving an effective balance between recognition accuracy and real-time performance. With the incorporation of imitation learning, recognition accuracy for long-tail operations was further increased to 89.3%, and the many-to-one matching algorithm was shown to effectively accommodate operational variant scenarios. The dynamic guidance strategy attained an acceptance

rate of 86.7%, with operational error rates reduced by 48.5% for novice learners and 35.7% for intermediate learners, highlighting strong personalization and adaptation capabilities. System deployment on mainstream mid-range mobile devices maintained end-to-end latency at or below 325 ms and peak CPU utilization at or below 28%, with dynamic offloading further optimizing resource utilization.

These results collectively demonstrate that the proposed technical framework outperforms existing approaches in terms of recognition accuracy, real-time performance, resource adaptability, and guidance effectiveness, thereby enabling more efficient vocational skill acquisition. The primary technical contribution lies in the establishment of a multimodal mobile interaction perception paradigm that overcomes the limitations of traditional single-modality sensing. Through lightweight model design and imitation learning-based enhancement strategies, a balanced trade-off among accuracy, real-time performance, and resource consumption is achieved, offering a novel technical pathway for the deep integration of mobile computing technologies with vocational education. From a practical perspective, the proposed framework provides a reusable technical solution for the development of mobile virtual simulation-based educational systems and effectively enhances the efficiency of vocational skill learning.

Future research will be advanced along three dimensions: technological extension, application expansion, and theoretical deepening. At the technological level, the integration of augmented reality and virtual reality with two-dimensional graphical user interface (GUI)-based guidance will be explored to enhance situational immersion. Federated learning and differential privacy techniques will be incorporated to address privacy protection challenges associated with cross-institutional data sharing. In parallel, further optimization of model architectures will be pursued to improve adaptability to low-specification mobile devices. At the application level, the proposed technical framework will be extended to a broader range of vocational education scenarios, including medical operations and intelligent manufacturing. Through deep integration with vocational education curricula, standardized mobile simulation-based educational products will be developed. At the theoretical level, educational psychology theories will be incorporated to conduct in-depth analyses of the alignment between guidance mechanisms and learners' cognitive load. By optimizing guidance granularity and timing, the effectiveness of educational interventions will be further enhanced, thereby contributing to the refinement of theoretical foundations and the practical implementation of mobile intelligent education systems.

5 REFERENCES

[1] B. Muchlas, P. Budiastuti, M. Khairudin, B. Santosa, and B. Rahmatullah, "The use of personal learning environment to support an online collaborative strategy in vocational education pedagogy course," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 2, pp. 24–41, 2023. https://doi.org/10.3991/ijim.v17i02.34565

[2] J. Kim, J. H. Park, and S. Shin, "Effectiveness of simulation-based nursing education depending on fidelity: A meta-analysis," *BMC Medical Education*, vol. 16, no. 1, p. 152, 2016. https://doi.org/10.1186/s12909-016-0672-7

[3] Y. Wang and L. Feng, "Vocational education in the era of big data: Course design and optimization strategy based on educational technology," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 22, pp. 143–158, 2024. https://doi.org/10.3991/ijim.v18i22.52447

[4] M. G. Landers, "The theory-practice gap in nursing: The role of the nurse teacher," *Journal of Advanced Nursing*, vol. 32, no. 6, pp. 1550–1556, 2000. https://doi.org/10.1046/j.1365-2648.2000.01605.x

[5] Y. Sun *et al.*, "Towards strong continuous consistency in edge-assisted VR-SGs: Delay-differences sensitive online task redistribution," *Computer Networks*, vol. 258, p. 111003, 2025. https://doi.org/10.1016/j.comnet.2024.111003

[6] A. Z. Sampaio, M. M. Ferreira, D. P. Rosário, and O. P. Martins, "3D and VR models in Civil Engineering education: Construction, rehabilitation and maintenance," *Automation in Construction*, vol. 19, no. 7, pp. 819–828, 2010. https://doi.org/10.1016/j.autcon.2010.05.006

[7] N. Szántó, G. D. Monek, and S. Fischer, "Development of an immersive, digital twin-supported smart reconfigurable educational platform for manufacturing training: A proof of concept," *Journal of Engineering Management and Systems Engineering*, vol. 3, no. 4, pp. 199–209, 2024. https://doi.org/10.56578/jemse030402

[8] S. Radosavljevic, V. Radosavljevic, and B. Grgurovic, "The potential of implementing augmented reality into vocational higher education through mobile learning," *Interactive Learning Environments*, vol. 28, no. 4, pp. 404–418, 2020. https://doi.org/10.1080/10494820.2018.1528286

[9] E. Lkhagvasuren, K. Matsuura, K. Mouri, and H. Ogata, "Dashboard for analyzing ubiquitous learning log," *International Journal of Distance Education Technologies*, vol. 14, no. 3, pp. 1–20, 2016. https://doi.org/10.4018/IJDET.2016070101

[10] J. Khlaisang and T. Teo, "An innovation-based virtual flipped learning system in a ubiquitous learning environment the 21st century skills of higher education learners," *Educational Technology & Society*, vol. 27, no. 1, pp. 100–116, 2024.

[11] C. Cockrell, D. Larie, and G. An, "Preparing for the next Pandemic: Simulation-based deep reinforcement Learning to discover and test multimodal control of systemic inflammation using repurposed immunomodulatory agents," *Frontiers in Immunology*, vol. 13, p. 995395, 2022. https://doi.org/10.3389/fimmu.2022.995395

[12] D. Triboan, L. Chen, F. Chen, and Z. Wang, "Semantic segmentation of real-time sensor data stream for complex activity recognition," *Personal and Ubiquitous Computing*, vol. 21, no. 3, pp. 411–425, 2017. https://doi.org/10.1007/s00779-017-1005-5

[13] M. A. Fadhel *et al.*, "Navigating the metaverse: Unraveling the impact of artificial intelligence—a comprehensive review and gap analysis," *Artificial Intelligence Review*, vol. 57, no. 10, p. 264, 2024. https://doi.org/10.1007/s10462-024-10881-5

[14] B. Sepehri, A. I. Almulhim, M. A. Adibhesami, S. Makaremi, and F. Ejazi, "Artificial intelligence role in promoting Saudi Arabia's smart cities: Addressing SDGs for socio-cultural challenges," *Социологическое Обозрение*, vol. 23, no. 4, pp. 20–47, 2024. https://doi.org/10.17323/1728-192x-2024-4-20-47

6 AUTHORS

Yan Wang holds a Bachelor's degree and is an Associate Professor. Her research focuses on vocational education. She works at the Department of Modern Economy and Trade, Hunan Vocational College of Engineering (E-mail: 13787411035@sohu.com).

Linli Wei holds a Bachelor's degree and is a Lecturer. Her research focuses on vocational education and student management. She works at the Department of Modern Economy and Trade, Hunan Vocational College of Engineering (E-mail: 13974838841@sohu.com).