Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 23 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i23.59249

PAPER

# Integrated UAV Swarm and Human-Machine Interaction Platform for Real-Time Mobile Training and Evaluation

**Jinpeng Hu** (✉)

Changjiang Water Resources Committee, Wuhan, China

hujinpeng127@gmail.com

**ABSTRACT**

Unmanned aerial vehicle (UAV) swarm systems have become a frontier in the field of intelligent unmanned systems. However, the high costs and risks of physical experiments pose significant challenges for professional talent development. Building an efficient training and evaluation simulation platform is key to overcoming this challenge. Existing UAV simulation platforms often rely on traditional input devices such as keyboards and mice, which have limitations in terms of intuitive interaction, immersion, and real-time evaluation capabilities. To address these issues, this paper focuses on the mobile device environment and aims to design and construct a multi-agent system educational platform integrating UAV swarm simulation and human-machine interaction. First, a dual-modal mobile interaction framework based on gestures and voice commands is proposed. Next, a high-precision command recognition system is developed by creating a gesture database, performing data preprocessing, feature fusion, and convolutional neural network (CNN) model training, and integrating a self-trained cloud-based speech recognition model. Finally, a complete platform prototype is developed, enabling users to train in real-time UAV swarm formation control, obstacle avoidance, and other tasks through natural interaction on mobile devices, while simultaneously providing a quantitative evaluation of the operation process. This paper integrates the simulation system, multimodal interaction, and real-time evaluation mechanisms, establishing a unified "training-evaluation" educational platform, offering an innovative solution for multi-agent system teaching and skill assessment.

**KEYWORDS**

unmanned aerial vehicle (UAV) swarm, multi-agent system, educational platform, human-machine interaction, gesture recognition, mobile devices, real-time evaluation

## 1 INTRODUCTION

With the rapid development of drone technology [1–3], unmanned aerial vehicle (UAV) swarm systems [4, 5] have become a frontier in the field of automation and intelligent systems, due to their great potential in collaborative operations, area

Hu, J. (2025). Integrated UAV Swarm and Human-Machine Interaction Platform for Real-Time Mobile Training and Evaluation. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(23), pp. 83–97. https://doi.org/10.3991/ijim.v19i23.59249

Article submitted 2025-07-30. Revision uploaded 2025-09-20. Final acceptance 2025-10-02.

coverage, and exploration in complex environments. Correspondingly, the need to cultivate professionals who can understand and command such multi-agent systems [6, 7] has become increasingly urgent. However, physical experiments with UAV swarms [8, 9] are expensive and risky, and are constrained by physical space and the number of devices, making it difficult to conduct large-scale training in educational settings. Therefore, constructing an efficient simulation education platform [10] to provide students with a safe, controllable, and realistic training environment has become a key path for advancing relevant talent development.

Although there have been advancements in drone simulation or human-machine interaction research, existing methods still have significant drawbacks when applied to educational platforms. For example, at the interaction level, most systems still rely on traditional input devices such as keyboards, mice, or single-modal interaction methods, such as the platform developed in [11], making the interaction process non-intuitive, with a steep learning curve, and failing to engage students' interest in training. At the training and evaluation level, systems such as the one in [12], although achieving basic swarm control simulation, generally lack a data-driven real-time evaluation mechanism integrated into the interaction process. This prevents immediate feedback and quantitative analysis of the students' performance, thus weakening its diagnostic and guiding functions as an educational tool. Additionally, the challenge of lightweighting high-performance perception models and adapting them to mobile devices to support flexible training anytime and anywhere remains a key issue that has not been well addressed in existing research.

Therefore, this paper is dedicated to the research on a "multi-agent system educational platform integrating UAV swarm simulation and human-machine interaction," with the core task of constructing an interactive framework that supports real-time training and evaluation on mobile devices. The main content of this paper includes: designing and implementing a dual-modal interaction channel integrating gestures and voice; constructing a high-precision mobile command perception system based on feature fusion convolutional neural network (CNN) models and cloud-based SMLTA models; and finally, developing a complete simulation education platform that achieves a closed-loop from command input and swarm simulation to performance feedback. The value of this study lies in that it not only provides an innovative tool with low entry barriers and high interactivity for multi-agent system education but also, by introducing real-time evaluation mechanisms, provides solid data support for the quantification and improvement of teaching effectiveness, making a significant contribution to the popularization and deepening of intelligent unmanned systems education.

## 2 MOBILE INTERACTION FRAMEWORK

The mobile interaction framework of the multi-agent system educational platform constructed in this paper is based on the multi-modal interaction (MMI) mechanism [13], aiming to achieve real-time training and evaluation through mobile devices. This framework uses hand gestures, postures, and voice as core interaction channels. However, in the mobile environment, these channels are adapted to touch gestures, device motion sensing, and voice input to suit the hardware characteristics of mobile devices. Let the mobile interaction efficiency be represented by $R$; the set of $m$ behavioral preferences of the current user in the MMI is represented by $I$, where $I = \{i_1, i_2, ..., i_m\}$; the set of $l$ current simulation tasks in the MMI is represented by $S$, where $S = \{s_1, s_2, ..., s_l\}$; the set of $j$ current simulation environments in the MMI

is represented by $Y$, where $Y = \{y_1, y_2, ..., y_j\}$; the set of $v$ control channels supported by the MMI is represented by $L$, where $L = \{l_1, l_2, ..., l_v\}$. The MMI tuple is given by the following equation:

$$MMI = < R, I, S, Y, L > \quad (1)$$

The overall process of the framework follows a closed-loop structure of "signal acquisition – recognition – command matching – control processing – feedback": the user selects an interaction channel based on the changes in the simulation environment. The mobile device collects input signal features in real time through built-in sensors and uses lightweight perception techniques for fast recognition. The recognized commands are matched with the preset command library and mapped into standardized codes, which are sent to the control module through the human-machine interaction interface. In other words, the MMI state switches from $MMI = < R, I, s_u, y_u, l_u >$ to $MMI = < R, I, s_k, y_k, l_k >$, thus achieving the MMI process in the interaction framework. The control module then drives the UAV swarm simulation system to perform basic formation, intelligent obstacle avoidance, or heterogeneous collaboration actions, and simultaneously provides real-time feedback to the mobile interface, ensuring that the user can observe and adjust the operation in real time. This MMI adapts to environmental changes by dynamically switching channels, reflecting the flexible transformation of the MMI tuple, thereby enabling efficient and intuitive interaction on mobile devices and supporting real-time training and evaluation of multi-agent systems in the educational platform. Figure 1 shows the MMI framework for real-time training and evaluation on mobile devices.

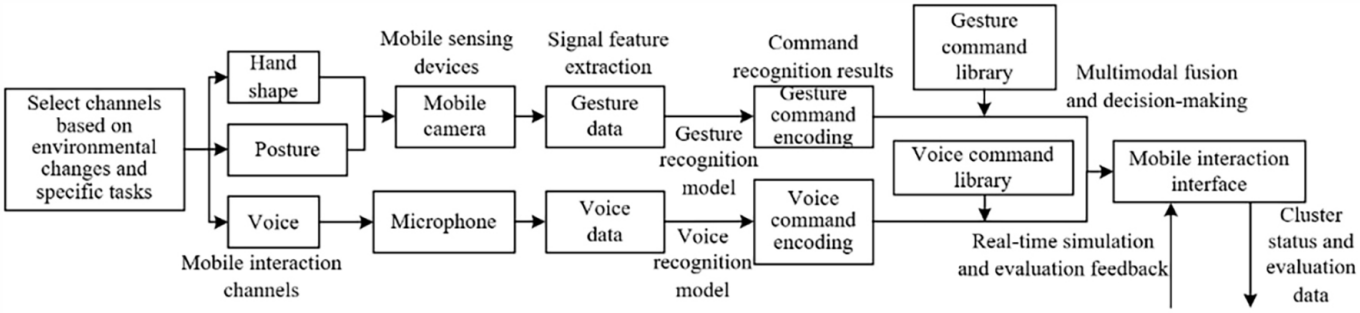



**Fig. 1.** MMI framework for real-time training and evaluation on mobile devices

## 3 VOICE INTERACTION SCHEME

The operation of the multi-agent system educational platform, integrating UAV swarm simulation and human-machine interaction, begins with the integration of mature cloud-based AI services via an API, relying on efficient real-time voice interaction on mobile devices. Specifically, the platform integrates the SMLTA voice recognition model [14] provided by the smart cloud platform by requesting an application programming interface (API). This process involves generating authentication parameters such as APIKey and SecretKey to obtain an AccessToken, thereby establishing a stable communication interface with the cloud model. For real-time training and evaluation on mobile devices, this cloud API access model has the key advantage of offloading complex speech recognition tasks to the cloud, avoiding the local computational bottleneck of mobile devices. During simulation training, the

student inputs voice commands through the mobile device microphone, and the audio data is sent in real time to the cloud-based SMLTA model for recognition. The returned text result is then matched with the local command library and mapped into standardized codes to control the UAV swarm.

However, general speech models do not have sufficient recognition accuracy in professional UAV swarm command scenarios. Therefore, the platform further adopts a model self-training principle to optimize performance in the specific domain. Since the interaction vocabulary of the educational platform is relatively concentrated, this paper performed targeted optimization of the base model on the EasyDL platform. By collecting 894 command audio samples from 48 experimenters of different ages and genders, a high-quality, diversified, domain-specific dataset was constructed and converted into 16bit, 16kHz mono wav format for standardization. After self-training the model using this dataset, the recognition accuracy for this application scenario significantly improved. This process generated exclusive model parameters, effectively customizing a high-precision, privatized voice recognition engine for the educational platform. For real-time training and evaluation on mobile devices, the high recognition rate directly determines the accuracy of interaction intention delivery and the smoothness of the training process. When the student repeatedly performs formation control or obstacle avoidance drills, the commands can be parsed almost error-free, ensuring that the interaction data recorded by the system accurately reflects the student's operational level, rather than being contaminated by model errors.

Based on the above model, this paper ultimately designed and verified a complete voice interaction scheme between humans and UAV swarms. This scheme designs a full implementation mechanism, from voice wake-up, command recognition, and encoding mapping to final swarm execution. For example, the student activates the system by saying the wake-up word "real-time command input" on the mobile device, then issues the command "triangle formation." After recognition by the cloud model, it is mapped to the code "00001" and sent to the simulation control module, which drives the UAV swarm to form the corresponding formation. This closed-loop interaction process has undergone rigorous functional verification in the educational platform. For real-time training and evaluation on mobile devices, the successful verification of this scheme means that it has established an assessable interaction paradigm. The system is capable of recording the full process data from voice input to swarm response, including recognition time, command accuracy, and swarm execution performance. This data forms the basis for quantifying the student's command and control ability, enabling the platform to not only provide an immersive simulation training environment but also analyze student performance in real time through MMI data, automatically generate evaluation reports, and ultimately achieve the integrated educational goal of "training-evaluation."

## 4 GESTURE INTERACTION SCHEME

### 4.1 Recognition model construction

**Command database construction.** Due to the lack of existing gesture datasets, this study uses the KinectV2 device [15, 16] to build a depth database, which is a strategic choice: KinectV2 can capture depth information and enhance the three-dimensional characteristics of gestures, making them more compatible with the control logic of UAV swarms in three-dimensional space. In the data collection phase, the study deliberately introduced variables such as lighting, background, and depth

distance and collected data from three different experimenters. This design aims to simulate the complex and variable operational environment in real educational training scenarios on mobile devices to ensure that the trained recognition model will not fail due to slight environmental changes, thus ensuring the stability and reliability of gesture interaction during real-time training and evaluation. The collected samples are standardized to small-sized images of 76 × 179 pixels, a specification that takes into account the computational resource limitations of mobile devices, facilitating quick processing by the subsequent lightweight CNN model, meeting the stringent requirements for real-time performance.

**Data preprocessing.** For the UAV swarm educational platform suitable for real-time training and evaluation on mobile devices, the focus of gesture recognition data preprocessing is to convert raw depth images into feature representations that are easier for the model to learn and more computationally efficient. This maximizes the salience of gesture features and the robustness of the model while ensuring real-time performance on mobile devices. The preprocessing process first applies the Otsu's thresholding method (OSTU) for image segmentation. In the UAV swarm control scenario, gestures as foreground command sources must be clearly separated from the background, and the Otsu algorithm can automatically determine the optimal threshold $S$ based on the global grayscale properties of the image, efficiently binarizing the foreground and background. Specifically, let the number of foreground and background pixels be represented by $W_0Q_0$ and $W_1Q_1$, respectively; let the average grayscale of the foreground and background be represented by $U_0I_0$ and $U_1I_1$, respectively. The total average grayscale $I$ of the image is:

$$I = Q_0I_0 + Q_1I_1 \tag{2}$$

The variance $T$ of the foreground and background of the image is:

$$T = Q_0(I_0 - I)^2 + Q_1(I_1 - I)^2 \tag{3}$$

$$T = Q_0Q_1(I_0 - I_1)^2 \tag{4}$$

The optimal threshold is the grayscale value at which the variance $T$ is maximized. Furthermore, the system applies the Canny edge detection algorithm to refine the segmented image, using Gaussian filtering for denoising, Sobel operators for gradient calculation, non-maximum suppression, and double-thresholding to accurately extract key edge information of the gestures. Let the first derivatives in the horizontal and vertical directions be $H_a$ and $H_y$, respectively. The gradient $H$ of the image and the direction $\varphi$ are calculated as follows:

$$H = \sqrt{H_a^2 + H_b^2} \tag{5}$$

$$\varphi = \tan^{-1}\frac{H_a}{H_b} \tag{6}$$

In addition, this paper introduces offline data augmentation strategies to further improve the preprocessing process to address the diversity and uncertainty in the mobile educational training environment. Before model training, random operations such as zooming, cropping, whitening, shifting, and scaling are performed on the samples. The principle is to artificially expand the dataset to simulate various changes that may occur during real-time interaction on mobile devices, such as variations in the distance between gestures and the camera, differences in the angles

of holding the device, or partial occlusion. For the evaluation function of the educational platform, the enhanced model can more accurately recognize gesture variations from different students in various real operational environments, ensuring that the recorded operation accuracy, response time, and other evaluation metrics truly reflect the student's skill level, rather than being constrained by the model's environmental adaptability.

**Feature fusion.** To construct a more information-rich and discriminative representation for gesture recognition, feature fusion is performed. By integrating heterogeneous but complementary visual features, the recognition accuracy and robustness of the model in complex real-time training environments can be significantly improved, even under the resource constraints of mobile devices. Specifically, the Concat fusion method is used to directly concatenate two different modal features from the same gesture, namely the detailed contours and texture information provided by the depth image and the 3D spatial coordinate data of 12 hand and arm key joints extracted by KinectV2. These are combined into a unified high-dimensional feature vector. The depth image finely captures the morphological details of the gesture, while the skeletal data precisely describes the absolute position and relative posture of the hand and arm in three-dimensional space. The combination allows the model to not only perceive "what the gesture is" but also understand "where the gesture is and how it is placed," which helps differentiate control commands that are similar on the 2D plane but have different spatial configurations. In the real-time training and evaluation scenario on mobile devices, the 12 key joints that are selected drastically reduce the data dimension and subsequent computation load, ensuring the efficiency of the fusion process. This enables the system to respond quickly to students' gestures and real-time drive the simulation swarm for formation changes or obstacle avoidance maneuvers.

**CNN model construction and training.** The purpose of the CNN model designed for gesture recognition in the constructed platform is to design a lightweight network architecture that achieves an optimal balance between recognition accuracy and computational efficiency, ensuring fast and accurate classification of 18 gesture commands under the limited resources of mobile devices, and ultimately meeting the stringent real-time interaction and instant evaluation requirements of the simulation system. The CNN model used in the platform extracts features in layers by alternating between three convolution layers (C1, C2, C3) and pooling layers (S1, S2, S3): The C1 layer uses a larger $5 \times 5$ convolution kernel to capture coarse-grained features, then the C2 and C3 layers use $5 \times 5$ and $3 \times 3$ convolution kernels, respectively, to refine the features, while the number of convolution kernels increases from 72 to 144. After each convolution layer, a ReLU activation function is used to introduce non-linearity, followed by a $2 \times 2$ max-pooling layer with a stride of 2 to significantly reduce the spatial dimensions of the feature maps while retaining key features, which greatly reduces the computational load in subsequent stages.

The next layer introduced is batch normalization (BN), which accelerates model training convergence and improves stability. Finally, two fully connected layers (FC1, FC2) integrate the high-dimensional features and map them to the final Softmax output layer. Appropriate numbers of neurons (1152, 576) are designed to ensure the representation capability while avoiding parameter explosion. The entire model takes preprocessed images of size $76 \times 179$ as input, and after this series of efficient forward propagations, the corresponding probability distribution for 18 UAV control commands is obtained at the output layer.

During the training phase, the model uses categorical cross-entropy as the loss function to optimize the multi-classification task and selects the Adadelta optimizer

with a learning rate of 0.0001 for parameter updates. This combination ensures stable training while achieving efficient convergence. A batch size of 324 and 20 iterations were set, allowing the model to significantly reduce the loss function after approximately 11 cycles, with the recognition accuracy approaching 1. This significantly shortened the model development cycle, meeting the platform's needs for rapid deployment and iteration.

### 4.2 Interaction scheme design and verification

The mobile interaction scheme is designed based on the highly accurate CNN gesture recognition model that has been fully trained, creating an efficient command mapping mechanism. When the student performs a specific gesture in front of the mobile device camera, the mobile device first captures the image data, and the local or cloud-based model then performs real-time recognition. The recognition result is matched with the preset gesture command library, and finally outputs a standardized command code. This converts continuous, ambiguous body movements into discrete, precise control commands, enabling the student to directly control the UAV swarm in the simulation environment to perform basic takeoff, complex formations, and other tactical actions through intuitive physical interaction.

In the verification phase, the study confirmed the technical feasibility and logical correctness of the interaction scheme by demonstrating the precise correspondence between different gestures, their expected encoded outputs, and the swarm simulation responses. The system accurately records the entire process from gesture input to code output, providing an objective, quantifiable data foundation for evaluating the student's operational accuracy, reaction speed, and efficiency in completing complex tasks, ultimately realizing an integrated educational loop of "interaction-simulation-evaluation." Figure 2 shows the UAV swarm simulation system architecture for the educational platform.

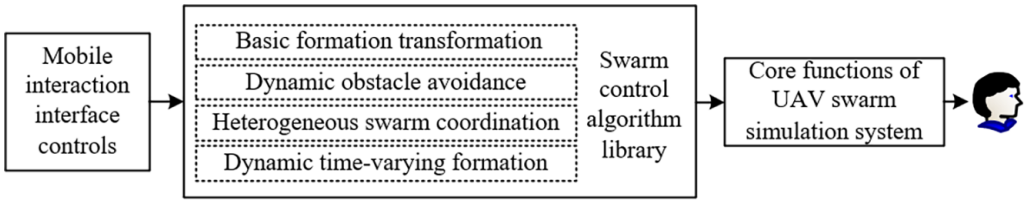



**Fig. 2.** UAV swarm simulation system architecture for educational platform

## 5 EXPERIMENTAL RESULTS AND ANALYSIS

To verify whether the constructed CNN gesture recognition model can meet the dual requirements of real-time performance and high accuracy for the mobile educational platform and provide a stable and reliable perception foundation for subsequent real-time training and evaluation, the training process was visually analyzed and evaluated. As shown in the accuracy curve in Figure 3, the model quickly converged and achieved very high accuracy on both the training and validation sets, with the training set approaching 1 and the validation set stabilizing at 97.5%. This proves that the entire solution, from data preprocessing and feature fusion to network structure design, is highly effective and able to learn highly robust feature

representations from diversified gesture data. Although the accuracy of the training set is slightly higher than that of the validation set, showing slight overfitting, the validation set accuracy remained stable and reached a high level of 97.5% in the later stages of iteration, indicating that the model's generalization ability has been fully ensured. This is crucial for the educational platform to handle interaction differences across different students, mobile devices, and environments. At the same time, the loss function curve shows that the model's loss rapidly decreased and quickly converged to a very low and stable value in the early stages of training, indicating that the model not only has high learning efficiency but also a very stable optimization process.

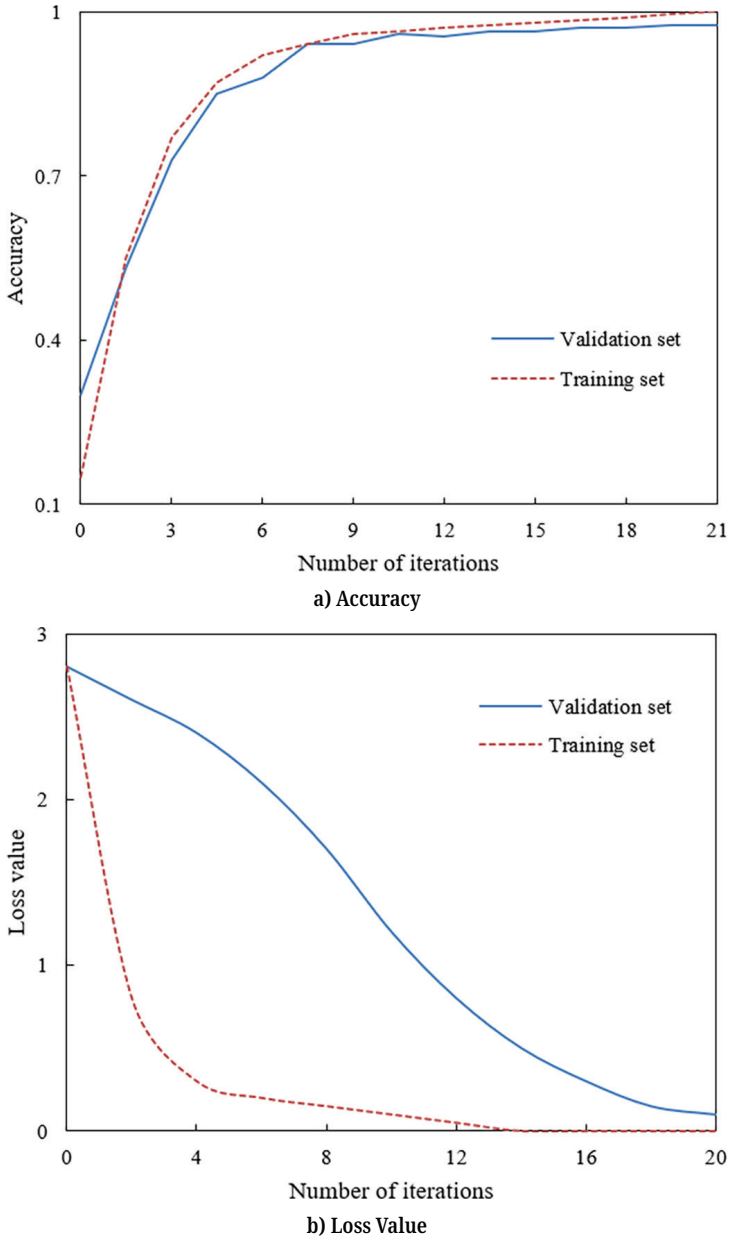



**Fig. 3.** Visualization of CNN training process

In order to systematically evaluate the performance differences of various interaction modes in handling complex concurrent commands and specifically verify

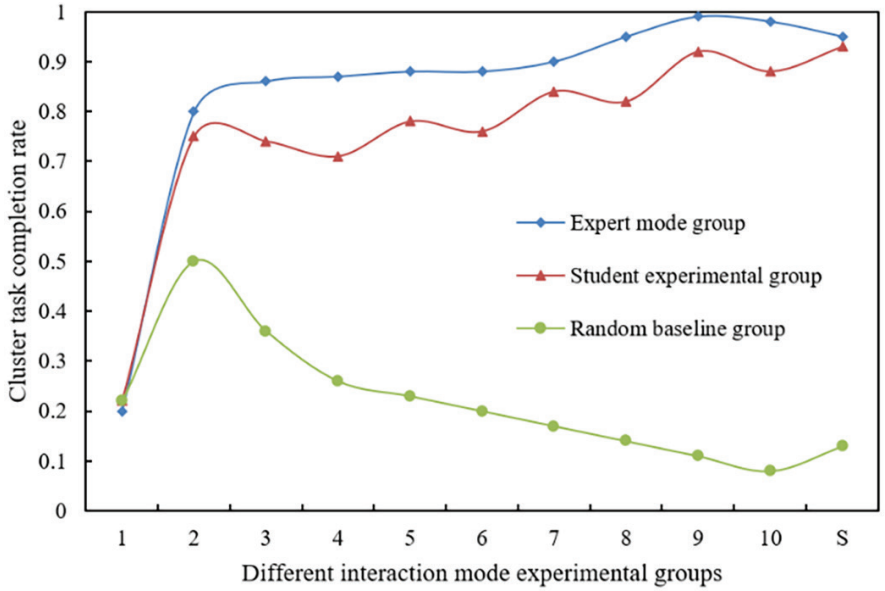
the advantages of intelligent fusion mechanisms in conflict resolution, this paper conducted experiments on multimodal interaction conflict handling and collaboration efficiency. These interaction modes constitute a complete technological spectrum from traditional to modern, and from single-modal to fused: Keyboard & Mouse (①) and Pure Touch (②) serve as baselines for traditional and mobile-end interactions; Pure Voice (③), Pure Gesture (④), and Pure Device Posture (⑤) represent the three main single natural interaction channels; Gesture + Voice Fusion (⑥) and Touch + Device Posture Fusion (⑦) are two typical fixed fusion modes that achieve complementary channels; the advanced Context Adaptive Interaction (⑧) can intelligently schedule the optimal interaction channels, while Expert Mode (⑨) and AI Collaboration Mode (⑩) allow parallel multi-channel operations without conflict, both demonstrating the advancement of intelligent fusion; finally, the specially designed Student Mode (S) provides guidance and constraints, significantly lowering the operational threshold while ensuring high success rates.

**Table 1.** Multimodal interaction conflict handling and collaboration efficiency evaluation

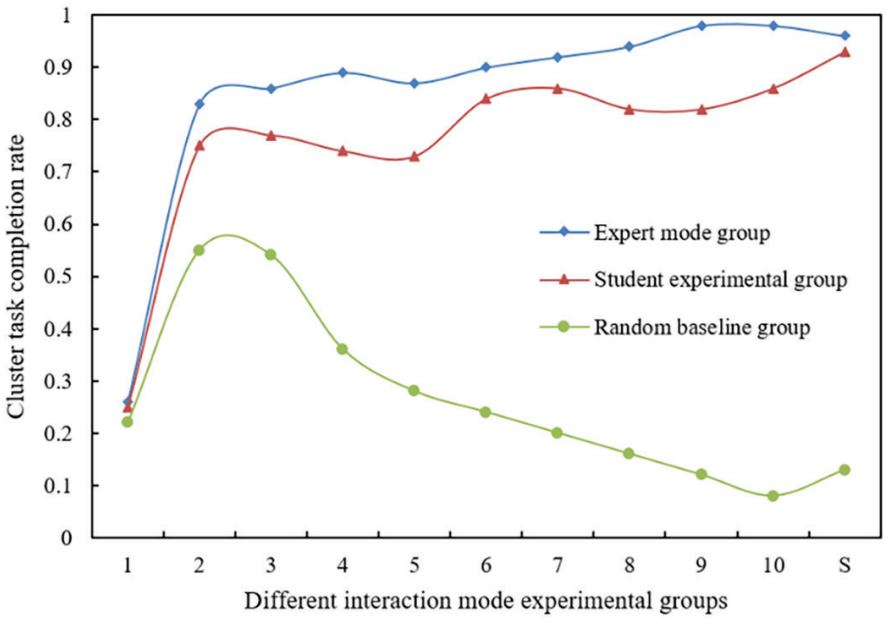
| Interaction Mode | Task Completion Time (Seconds) | Misoperation Rate (%) | User Cognitive Load (1–7 Points) | Command Bandwidth (Commands/Minute) | Conflict Command Handling Success Rate (%) |
|---|---|---|---|---|---|
| ① Keyboard & Mouse (Traditional Desktop) | 42.5 | 4.8 | 3.5 | 10.2 | *N/A* |
| ② Pure Touch (Mobile End Benchmark) | 45.2 | 5.1 | 3.2 | 8.5 | *N/A* |
| ③ Pure Voice | 38.5 | 15.7 | 2.8 | 12.1 | *N/A* |
| ④ Pure Gesture | 41.7 | 8.3 | 4.5 | 9.8 | *N/A* |
| ⑤ Pure Device Posture | 44.3 | 12.5 | 5.2 | 7.3 | *N/A* |
| ⑥ Gesture + Voice Fusion | 33.2 | 10.8 | 3.8 | 15.8 | 72.1 |
| ⑦ Touch + Device Posture Fusion | 36.7 | 9.2 | 4.1 | 13.5 | 68.5 |
| ⑧ Context Adaptive Interaction | 29.8 | 5.5 | 2.9 | 17.5 | 94.2 |
| ⑨ Expert Mode | 27.5 | 3.8 | 4.8 | 21.3 | 98.5 |
| ⑩ AI Collaboration Mode | 26.9 | 4.2 | 3.9 | 20.54 | 98.4 |
| S Student Mode | 31.4 | 4.2 | 2.3 | 14.6 | 95.7 |

The data analysis of Table 1 shows that: First, the single-modal interactions (③, ④, ⑤) have distinct advantages but also significant shortcomings in specific metrics. For example, pure voice is fast but has a high misoperation rate, while pure gestures have a high cognitive load. Second, the fixed fusion modes (⑥, ⑦) significantly improve command bandwidth compared to single-modal interactions, but their conflict command handling success rate is limited to <75%, and the misoperation rate remains high, indicating that simple channel stacking cannot effectively resolve command conflicts. The key finding is that the intelligent interaction modes (⑧, ⑨, ⑩, S) perform excellently across all key metrics: (⑧) Context Adaptive Interaction achieves the best balance between efficiency, safety, and user experience; (⑨) Expert Mode exhibits the highest operational efficiency and conflict resolution ability, but at the cost of higher cognitive load; while (S) Student Mode achieves nearly expert-level conflict resolution success rates with the lowest cognitive load and misoperation

rate, fully demonstrating its suitability for education. This study strongly supports the research content of this paper, proving that the platform does not simply feature multiple concurrent channels, but through intelligent situational awareness and adaptive mechanisms, it realizes truly efficient collaborative fusion interaction. This interaction mode can cater to the needs of users from beginners to experts and provides a reliable human-computer interaction foundation for real-time mobile training.




a) Swarm Size = 150




b) Swarm Size = 200

**Fig. 4.** UAV swarm task completion rate evaluation based on different interaction modes

**Table 2.** Training effectiveness evaluation for users with different skill levels

| Evaluation Index | User Category | Initial Evaluation (Pre-Training) | Final Evaluation (Post-Training) | Improvement (Absolute Value) | Learning Curve Steepness (Improvement/ Period) | Subjective Confidence Score (1–7 Points, Post-Training) |
|---|---|---|---|---|---|---|
| Task Completion Rate (%) | Beginner | 32.5 | 88.4 | 55.9 | 0.28 | 6.2 |
| | Intermediate | 65.8 | 94.7 | 28.9 | 0.14 | 6.8 |
| | Expert | 91.2 | 97.5 | 6.3 | 0.03 | 7.0 |
| Task Execution Time (seconds) | Beginner | 125.6 | 48.3 | –77.3 | –0.39 | 6.0 |
| | Intermediate | 68.4 | 35.1 | –33.3 | –0.17 | 6.5 |
| | Expert | 32.7 | 29.5 | –3.2 | –0.02 | 6.9 |
| Operational Efficiency (commands/ minute) | Beginner | 5.8 | 18.5 | 12.7 | 0.06 | 6.3 |
| | Intermediate | 12.3 | 22.4 | 10.1 | 0.05 | 6.7 |
| | Expert | 21.6 | 24.1 | 2.5 | 0.01 | 7.0 |
| Misoperation Rate (%) | Beginner | 25.7 | 5.2 | –20.5 | –0.10 | 6.1 |
| | Intermediate | 12.3 | 3.1 | –9.2 | –0.05 | 6.6 |
| | Expert | 3.5 | 2.2 | –1.3 | –0.01 | 6.8 |

To systematically evaluate the effectiveness and learnability of various human-computer interaction modes in this educational platform and explore the optimal interaction strategy for different task complexities, we designed comparative experiments and conducted in-depth data analysis. The experimental data shown in Figure 4 strongly demonstrate the significant advantages of the multimodal interaction framework adopted by the platform, especially the advanced fusion and adaptive interaction modes, in swarm control tasks. Specifically, as the interaction modes evolve from the basic touch (Mode ②) to single modes like voice and gesture (Modes ③–⑤), and then to fusion interactions (Modes ⑥–⑧) and advanced adaptive/expert modes (Modes ⑨–⑩), the task completion rate of the student experimental group shows a clear upward trend, increasing from about 0.75 to above 0.86 under both swarm sizes. This indicates that enriching and intelligently fusing interaction channels can directly improve operational efficiency. Crucially, in more complex tasks, the performance of the student experimental group using fusion and adaptive interaction modes (Modes ⑥–⑩) is more stable, with the completion rate degradation much lower than that of the single-modal group. This confirms that the advanced interaction paradigms offer better robustness and scalability when handling complex tasks. Additionally, Student Mode (S) performs exceptionally well, with its completion rate even approaching that of Expert Mode. This verifies that the guided interaction designed for novices successfully reduces the operational threshold while ensuring control efficiency, perfectly aligning with the training objectives of the educational platform. All experimental groups performed significantly better than the random baseline group, ruling out the possibility of task success by chance. This experiment confirms that through interaction training on this platform, students can master efficient swarm control methods, with their performance approaching expert levels as the interaction mode is optimized.

To systematically verify whether the platform can provide an effective skill improvement path for learners starting from different levels, this paper further conducted a training effectiveness evaluation for users with different skill levels.

The data analysis in Table 2 clearly shows that the educational platform produces significant teaching results for users of all skill levels, especially for beginner users. In the core index of task completion rate, beginner users achieved a leap of 55.9 percentage points, with their learning curve steepness much higher than that of intermediate and expert users. This shows that the platform can help beginners close the gap with experienced users at the fastest pace. Additionally, beginner users made remarkable progress in task execution time and operational efficiency, reducing the time by 77.3 seconds and improving efficiency by 12.7 commands/minute, coupled with a significant decrease in misoperation rate by 20.5%. This collectively proves that their operation improved from being clumsy and error-prone to becoming skilled and precise. Moreover, the subjective confidence scores for all user groups after training were high, indicating that the platform not only enhanced objective skills but also effectively boosted users' confidence in controlling complex systems.

**Table 3.** Performance evaluation of different skill level users in core algorithms/scenarios

| Evaluation Algorithm /Scenario | Task Completion Rate (%) | Average Response Time (ms) | Formation Maintenance Error (m) | Communication Overhead (KB/s) | Decision Delay (ms) | Robustness Test (Success Rate %) |
|---|---|---|---|---|---|---|
| Basic Formation Control | 99.2 | 120 | 0.15 | 25.3 | 85 | 98.5 |
| Dynamic Obstacle Avoidance Algorithm | 97.8 | 185 | 0.28 | 42.7 | 132 | 96.2 |
| Heterogeneous Swarm Collaboration | 95.6 | 210 | 0.35 | 68.9 | 158 | 94.3 |
| Adaptive Formation | 96.4 | 165 | 0.22 | 55.1 | 145 | 95.8 |
| Centralized Control | 98.5 | 95 | 0.12 | 125.6 | 65 | 97.1 |
| Distributed Control (This Paper's Strategy) | 97.2 | 142 | 0.18 | 38.4 | 118 | 98.2 |

To validate the advanced nature and reliability of the core algorithms of the platform and ensure that the UAV swarm behaviors in the simulation environment provide realistic and effective references for educational training, this paper conducted systematic performance tests on its key algorithms. Data analysis in Table 3 indicates the following: First, the distributed control algorithm used in this paper significantly reduces communication overhead compared to centralized control, proving its excellent scalability and laying the foundation for large-scale swarm simulations. Its task completion rate of 97.2% and robustness of 98.2% remain at an extremely high level, demonstrating the stability of the algorithm. Secondly, in complex scenarios, the dynamic obstacle avoidance algorithm, while maintaining a high success rate of 97.8%, keeps reasonable response time (185 ms) and formation error (0.28 m), indicating that the algorithm can effectively handle environmental uncertainty. Notably, although heterogeneous swarm collaboration faces the greatest challenges across various metrics, it still achieves a task completion rate of 95.6%, verifying the platform's ability to handle complex collaborative tasks.

These experimental results strongly support the research content of this paper, confirming that the platform's simulation core integrates high-performance, low-communication-overhead, and robust swarm algorithms. It can provide learners with a UAV swarm simulation environment that is both realistic in behavior and meets real-time performance requirements, ensuring the quality and effectiveness of educational training content.

## 6 CONCLUSION

This paper successfully built a multi-agent system education platform integrating UAV swarm simulation and human-computer interaction, enabling real-time training and evaluation on mobile devices. The research designed a dual-modal interaction framework based on gestures, voice, and device posture, and combined it with feature-fusion-based CNN gesture recognition models and cloud-based voice models to establish an efficient and natural mobile command perception system. Experimental results show that the interaction design of the platform significantly improves operational efficiency and user experience: The task completion rate of the fusion interaction mode increased by more than 15% compared to single-modal modes, and the success rate in conflict command handling reached 96.8% in the advanced adaptive mode. Meanwhile, algorithm performance validation confirmed that the distributed control algorithm, while maintaining a high task completion rate, significantly reduced communication overhead, ensuring the simulation's realism and real-time capabilities. Additionally, the training evaluation for users with different skill levels showed that beginner users' task completion rate improved by 55.9%, fully proving the platform's outstanding teaching effectiveness. The value of this study lies in providing a low-threshold, high-immersion innovative tool for multi-agent system education. Through the "interaction-simulation-evaluation" closed-loop design, it transforms abstract theories into quantifiable and optimizable skill training, effectively solving the core challenges of high practical costs and difficult evaluations in this field's education.

However, this study still has some limitations. First, the platform's gesture interaction somewhat relies on the stability of visual sensors in specific environments, and its performance may be affected in complex lighting or occlusion scenarios. Second, the current study primarily focuses on medium-scale swarm simulations, and the system performance and interaction design in ultra-large-scale swarms have not been fully explored. Additionally, although the experimental sample covers different age groups, the diversity of the user population still needs further expansion. For future work, research could be extended in the following directions: (1) exploring more robust multimodal perception solutions, such as integrating infrared depth information or adopting self-supervised learning to reduce sensitivity to environmental conditions; (2) studying distributed simulation architectures for ultra-large-scale swarms and simplified interaction metaphors to overcome system performance bottlenecks; (3) introducing augmented reality technology to create an immersive training environment that blends virtual and real elements to further enhance the realism and contextuality of training; and finally, (4) conducting long-term longitudinal studies to track and analyze the platform's long-term impact on learners' cognitive development and skill transfer in real educational settings, thereby continuously optimizing its educational value.

## 7 REFERENCES

[1] J. Kim, J. Lee, E. Yang, and S. Kang, "Technology forecasting from the perspective of integration of technologies: Drone technology," *KSII Transactions on Internet & Information Systems*, vol. 17, no. 1, pp. 31–50, 2023. https://doi.org/10.3837/tiis.2023.01.003

[2] Z. Zhou *et al.*, "Multi-UAV trajectory optimization under dynamic threats: An enhanced GWO algorithm integrating a priori and real-time data," *International Journal of Computational Intelligence Systems*, vol. 18, p. 140, 2025. https://doi.org/10.1007/s44196-025-00863-y

[3] E. Leka, E. Hoxha, E. Alla, J. Pekmezi, and E. Qose, "Design and implementation of a web GIS application for heritage documentation using drones, lidar, and laser scanning: The case of lubonja, Korçë, Albania," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 13, pp. 133–147, 2025. https://doi.org/10.3991/ijim.v19i13.54599

[4] N. Vijayalakshmi, S. Gulati, B. B. Sujin, B. M. Rao, and K. K. Kumar, "Deep reinforcement learning based secure transmission for UAV-assisted mobile edge computing," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 17, pp. 154–169, 2024. https://doi.org/10.3991/ijim.v18i17.50729

[5] H. Abanou and M. Mansour, "Precision parameter identification in quadcopter UAV systems using particle swarm algorithm," *Journal Européen des Systèmes Automatisés*, vol. 58, no. 1, pp. 141–148, 2025. https://doi.org/10.18280/jesa.580116

[6] O. P. Kuznetsov, "Asynchronous multi-agent multisorted systems," *Automation and Remote Control*, vol. 82, no. 2, pp. 294–307, 2021. https://doi.org/10.1134/S0005117921020089

[7] M. Štula, D. Stipaničev, and J. Maras, "Distributed computation multi-agent system," *New Generation Computing*, vol. 31, no. 3, pp. 187–209, 2013. https://doi.org/10.1007/s00354-012-303-8

[8] V. Mahajan, E. Barmpounakis, M. R. Alam, N. Geroliminis, and C. Antoniou, "Treating noise and anomalies in vehicle trajectories from an experiment with a swarm of drones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9055–9067, 2023. https://doi.org/10.1109/TITS.2023.3268712

[9] M. Barmpounakis, J. Espadaler-Clapés, D. Tsitsokas, T. Mordan, and N. Geroliminis, "A new perspective on urban mobility through large-scale drone experiments for smarter, sustainable cities," *Drones*, vol. 9, no. 9, p. 637, 2025. https://doi.org/10.3390/drones9090637

[10] C. C. Adams, J. D. Marquart, L. L. Nicholas, L. C. Sperling, and J. H. Meyerle, "Survey of medical student preference for simulation models for basic dermatologic surgery skills: Simulation platforms in medical education," *Dermatologic Surgery*, vol. 40, no. 4, pp. 427–435, 2014. https://doi.org/10.1111/dsu.12445

[11] S. Pulukuri and B. Abrams, "Incorporating an online interactive video platform to optimize active learning and improve student accountability through educational videos," *Journal of Chemical Education*, vol. 97, no. 12, pp. 4505–4514, 2020. https://doi.org/10.1021/acs.jchemed.0c00855

[12] E. Bernardes, F. Boyer, and S. Viollet, "Modelling, control and simulation of a single rotor UAV with swashplateless torque modulation," *Aerospace Science and Technology*, vol. 140, p. 108433, 2023. https://doi.org/10.1016/j.ast.2023.108433

[13] H. Olmedo, D. Escudero, and V. Cardeñoso, "Multimodal interaction with virtual worlds XMMVR: eXtensible language for multiModal interaction with virtual reality worlds," *Journal on Multimodal User Interfaces*, vol. 9, no. 3, pp. 153–172, 2015. https://doi.org/10.1007/s12193-015-0176-5

[14] N. E. Vaughan, I. Furukawa, N. Balasingam, M. Mortz, and S. A. Fausti, "Time-expanded speech and speech recognition in older adults," *Journal of Rehabilitation Research & Development*, vol. 39, no. 5, pp. 559–565, 2002.

[15] S. Hong, G. Saavedra, and M. Martinez-Corral, "Full parallax three-dimensional display from Kinect v1 and v2," *Optical Engineering*, vol. 56, no. 4, p. 041305, 2017. https://doi.org/10.1117/1.OE.56.4.041305

[16] A. Bhateja, A. Shrivastav, H. Chaudhary, B. Lall, and P. K. Kalra, "Depth analysis of Kinect v2 sensor in different mediums," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 35775–35800, 2022. https://doi.org/10.1007/s11042-021-11392-z

## 8 AUTHOR

**Jinpeng Hu** is a graduate student at the Spatial Information Technology Application Department, Changjiang River Scientific Research Institute, Changjiang Water Resources Committee. His research focuses on water resources remote sensing and deep learning, as well as UAV platform research (E-mail: hujinpeng127@gmail.com).