Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 12 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i12.62255

PAPER

Multimodal Perception-Driven Optimization of Mobile Interaction Interfaces and Supply Chain Efficiency

Jia Liu[1] (✉), Tian Gao[2]

[1]Zhejiang Technical Institute of Economics, Hangzhou, China

[2]Zhejiang Lijiu Environmental Technology Co., Ltd, Hangzhou, China

230069@zjtie.edu.cn

ABSTRACT

In mobile supply chain operations, constraints imposed by mobile terminal resources, fluctuations in operator cognitive load, and inadequate interface adaptability severely limit operational efficiency and increase the risk of operational errors. To address these challenges, a mobile interaction interface optimization and efficiency enhancement model was proposed, grounded in cloud-edge collaborative multimodal perception. A lightweight temporal fusion network, integrating time-division cross-modal attention and knowledge distillation, was developed to enable efficient processing of multimodal data at the edge and accurate prediction of operator intent, thereby accommodating the resource constraints of mobile devices. A mechanism for estimating instantaneous cognitive load, driven by multimodal proxy indicators, was constructed. Combined with meta-reinforcement learning, a load-aware dynamic interface reconfiguration strategy was formulated, enabling real-time adaptation between the interface and operator state. A multimodal intent-driven Markov decision process-based workflow preloading model was established, and an online Bayesian optimization framework was incorporated to form a closed-loop optimization system, effectively improving supply chain efficiency. This study provides an innovative solution for the application of mobile interaction technologies in supply chain scenarios and promotes the deep integration of mobile multimodal interaction with industrial environments.

KEYWORDS

multimodal perception, cloud-edge collaboration, mobile interaction interface, cognitive load, supply chain efficiency, closed-loop optimization

1 INTRODUCTION

The rapid advancement of mobile technologies has driven a transformation of supply chain operations toward intelligent and mobile paradigms. Mobile terminals such as industrial tablets and smart glasses have become core carriers for on-site operations, including warehouse picking, inventory counting, and goods receiving, significantly breaking the spatial limitations of traditional operations [1–4] and

Liu, J., Gao, T. (2026). Multimodal Perception-Driven Optimization of Mobile Interaction Interfaces and Supply Chain Efficiency. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(12), pp. 140–153. https://doi.org/10.3991/ijim.v20i12.62255

Article submitted 2026-02-07. Revision uploaded 2026-05-02. Final acceptance 2026-05-09.

enhancing operational flexibility. However, the complexity of mobile supply chain operation scenarios, combined with the inherent characteristics of mobile terminals [5–9], has introduced a series of core technical bottlenecks that urgently require resolution. Computational power, energy consumption, and storage capacity of mobile devices are limited, making it difficult to support complex multimodal fusion computations. Operator cognitive load fluctuates significantly during dynamic operational processes, whereas fixed interaction interfaces fail to adapt to real-time changes in operator states. Additionally, the pronounced asynchrony of multimodal data renders it challenging to simultaneously achieve high precision and real-time performance in operation intent prediction. Consequently, low operational efficiency and high error rates are observed, severely restricting the deep application of mobile technologies in supply chain scenarios. Therefore, the development of a multimodal interaction interface optimization solution that accommodates mobile device resource constraints, cognitive load perception, and operational efficiency enhancement is necessary. Such a solution can not only address practical pain points in mobile supply chain operations but also enrich research outcomes in mobile multimodal interaction and adaptive interface optimization, providing crucial technical support for efficient interaction in industrial mobile scenarios, in alignment with the core research orientation of mobile technology journals toward lightweight, real-time, and scenario-adaptive design.

Existing related research has been conducted around three core directions: mobile multimodal fusion [10, 11], mobile interface adaptation [12, 13], and supply chain mobile operation optimization [14, 15]. Nevertheless, notable limitations remain, rendering these approaches insufficient for practical application requirements. In the domain of mobile multimodal fusion, most existing studies adopt cloud-centric fusion architectures [16, 17], with insufficient consideration of mobile device resource constraints. Even among lightweight fusion models, effective temporal correlation and dynamic triggering mechanisms are lacking, resulting in inadequate real-time performance of operation intent prediction. In mobile interface adaptation, existing solutions are mostly designed based on fixed rules or single touch modalities, without incorporating operator cognitive load as a core basis for interface adjustment, thus struggling to balance interface adaptability with operational fluency. In supply chain mobile operation optimization, research efforts have predominantly focused on optimizing the operational process itself, without achieving deep coupling among multimodal interaction, interface optimization, and process preloading. An effective closed-loop optimization mechanism is absent, leading to inflexibility in adapting to differences in operator habits and dynamic changes in warehouse environments.

To address the aforementioned research gaps and practical challenges, the core contributions of this work are as follows. In terms of theory and methodology, a cloud-edge collaborative lightweight multimodal temporal fusion architecture is proposed. Efficient processing of multimodal data at the edge and accurate intent prediction are achieved through the synergy of time-division cross-modal attention and knowledge distillation. A cognitive load estimation method driven by multimodal proxy indicators is constructed, and combined with meta-reinforcement learning, a load-aware dynamic interface reconfiguration strategy is developed. Furthermore, a process preloading and closed-loop optimization model driven by multimodal intent is established. Closed-loop enhancement of interaction and efficiency is realized using a Markov decision process and online Bayesian optimization.

2 METHODOLOGY

2.1 Cloud-edge collaborative multimodal perception and lightweight temporal fusion model

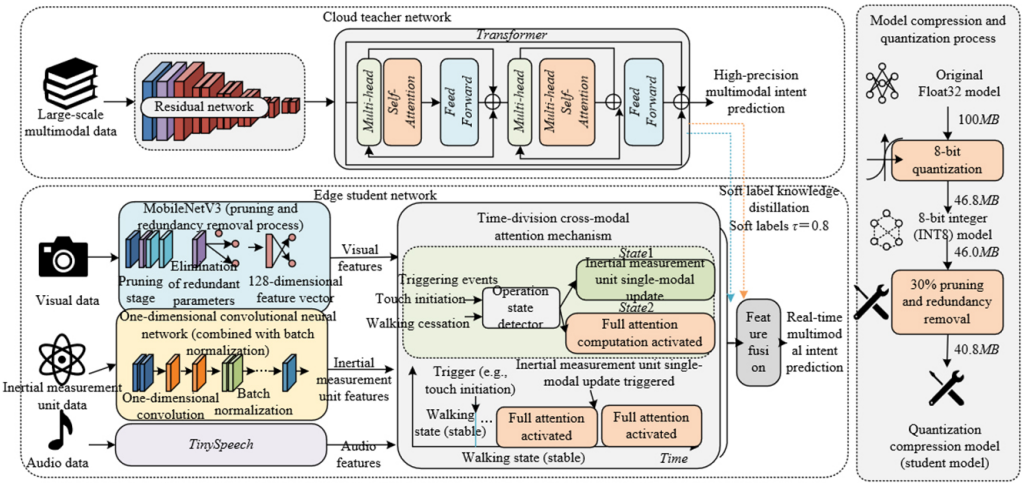

Fig. 1. Architecture of the cloud-edge collaborative multimodal perception and lightweight temporal fusion network

Multimodal data acquisition is conducted using an industrial-grade tablet, integrated with a camera, microphone, inertial measurement unit, and touchscreen sensor to capture visual, auditory, motion, and tactile data, respectively. Acquisition frequencies for each modality are adapted according to their inherent characteristics: visual data at 30 frames per second, inertial measurement unit data at 100 Hz, and tactile data at 50 Hz, thereby ensuring comprehensive capture of multi-dimensional operational information. The temporal misalignment problem caused by the asynchrony of heterogeneous data streams is addressed by combining hard time stamps with an adaptive sliding window. The size of the sliding window is dynamically adjusted based on operational states, ranging from 0.5 to 2 seconds. When a change in operational state is detected, the window is adaptively enlarged to preserve data integrity; when the state stabilizes, the window is reduced to lower computational overhead. All data alignment and lightweight preprocessing are performed on the edge side without reliance on high-speed network transmission, thereby effectively guaranteeing real-time responsiveness of the system. Figure 1 shows the architecture of the cloud-edge collaborative multimodal perception and lightweight temporal fusion network.

The lightweight temporal fusion network serves as the core component for efficient multimodal data processing and accurate intent prediction. An optimized student network is deployed on the edge side, comprising three branches: visual, motion, and auditory. In the visual branch, a lightweight MobileNetV3 network is adopted, from which redundant parameters of the fully connected layers are removed, focusing on the extraction of key spatial features such as cargo contours and hand gestures, ultimately producing a 128-dimensional feature vector. For the motion branch, a one-dimensional temporal convolutional network is employed, with a kernel size of 3×1, combined with batch normalization and dropout operations to suppress overfitting. Dynamic encoding of terminal posture and arm movement trajectories captured by the inertial measurement unit is performed, outputting 64-dimensional temporal features. The auditory branch is based on the TinySpeech lightweight

model, from which features of voice commands and environmental sound events are extracted, yielding a 32-dimensional feature vector. Multimodal feature fusion is achieved through a time-division triggering mechanism. The moment of operational state transition is determined by detecting abrupt changes in features from operational events such as touch initiation and scan completion. Complete cross-modal attention computation is triggered only at such moments; at all other times, only the inertial measurement unit motion features are updated to reduce computational load. Modality confidence is calculated from the variance of each modality's features, as expressed in the following equation:

$$c_i = \frac{\sigma_i}{\sum_{j=1}^{N} \sigma_j} \tag{1}$$

where, c_i denotes the confidence of the i-th modality, σ_i represents the variance of the features of the i-th modality, and N is the total number of modalities. Attention weights are obtained by normalizing the modality confidences, as given by:

$$\alpha_i = \frac{c_i}{\sum_{j=1}^{N} c_j} \tag{2}$$

Through this weight allocation, adaptive fusion of multimodal features is achieved. A teacher network with a residual network-Transformer architecture is deployed in the cloud, where the residual network extracts deep visual features and Transformer captures multimodal temporal dependencies to enable high-precision intent prediction. Daily edge operation logs are uploaded to the cloud, and the teacher network performs incremental training based on these logs. Knowledge is then transferred to the edge student network via soft label distillation with a temperature coefficient $\tau = 0.8$, and student network parameters are fine-tuned weekly to adapt to operator habits and environmental changes. The student network is optimized using 8-bit quantization compression and a pruning strategy with a 30% pruning rate, ensuring that the model size does not exceed 50 MB and that edge inference latency is maintained within 20 ms, thereby achieving a balance between accuracy and lightweight deployment.

Operational context and intent prediction are generated based on the output of the aforementioned lightweight temporal fusion network, enabling real-time generation of the current operational context category and a probability vector of the next operation intent. Six typical operational context categories are predefined in the system, covering major supply chain mobile operation scenarios such as one-handed scanning while walking and two-handed quantity entry while stationary. The intent probability vector outputs the top-three operation intents, including high-frequency actions such as clicking a confirmation button and raising the terminal to capture shelf images, with a refresh rate of 5 to 10 times per second, providing stable driving input for subsequent interface optimization and process preloading. To reduce the misclassification rate of similar intents, contextual prior constraints are introduced into the intent prediction process.

2.2 Instantaneous cognitive load estimation driven by multimodal proxy indicators

Instantaneous estimation of cognitive load is achieved using non-invasive multimodal proxy indicators. Based on cognitive load theory, three types of indicators that

are strongly relevant to mobile supply chain operations and capable of real-time acquisition are selected, thereby avoiding the interference of traditional invasive physiological indicators with the operational process. Input effort indicators are captured via the touchscreen sensor, including pressure variation and inter-click interval. Pressure variation is measured by its standard deviation; a standard deviation not less than 0.1 N indicates high input effort, while an inter-click interval not less than 1.5 seconds is determined as a state of high cognitive load. Physiological state indicators are captured via the near-infrared channel of the camera, measuring the rate of change of pupil diameter. A rate of change not less than 10% per second is considered high cognitive load. Body stability indicators are derived from the micro-jitter amplitude of the terminal detected by the inertial measurement unit, evaluated by the standard deviation of jitter. A standard deviation not less than 0.5° reflects a decline in operator body stability, corresponding to a high cognitive load. The dynamic changes of cognitive load are synergistically captured by these three types of indicators, constructing a comprehensive cognitive load characterization system from the three dimensions of input effort, physiological state, and body stability, thereby providing reliable inputs for subsequent real-time estimation.

Cognitive load estimation is performed using a lightweight random forest regressor. The model consists of 30 decision trees, each with a depth controlled within 8, balancing estimation accuracy with edge deployment efficiency. Input features are the standardized results of the three types of proxy indicators, with each indicator mapped to the interval [0, 1] using min-max normalization to eliminate dimensional effects. The normalization formula is given by:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

where, x is the original indicator value; x_{min} and x_{max} are the minimum and maximum values of that indicator, respectively; and x_{norm} is the normalized feature value. Model training is performed using 5-fold cross-validation to optimize decision tree split thresholds and node feature selection strategies. L1 regularization is introduced to suppress overfitting, with the regularized loss function expressed as:

$$L = L_0 + \lambda \sum_{i=1}^{M} |w_i| \tag{4}$$

where, L_0 is the base regression loss, λ is the regularization coefficient, w_i is the model parameter, and M is the total number of parameters. After optimization, model training time is controlled within 10 minutes, edge inference latency does not exceed 30 ms, and cognitive load estimation accuracy exceeds 88%, thereby satisfying the real-time and accuracy requirements for instantaneous cognitive load estimation in mobile supply chain operations.

2.3 Load-aware dynamic interface reconfiguration driven by meta-reinforcement learning

Load-aware dynamic interface reconfiguration is achieved within a meta-reinforcement learning framework, with its core objective being the construction of individualized interface adjustment strategies that adapt to operator cognitive load and operation intent. The state space is defined as a 64-dimensional three-component state vector, integrating current interface element layout features, the operation intent

prediction probability vector, and the cognitive load level, thereby comprehensively capturing the dynamic relationships among interface state, operational demand, and operator psychophysiological state. The action space is designed to include eight types of interface adjustment primitives tailored to mobile supply chain operation scenarios, covering button size scaling, voice confirmation pop-up triggering, menu collapse and expansion, visual guidance line overlay, input mode switching, pop-up delay adjustment, adaptive screen brightness, and cache preloading. Under high-load conditions, core buttons are enlarged by 20% to 50%, voice confirmation pop-ups are automatically triggered, non-core menus are collapsed, and color-coded visual guidance lines are added. Under low-load conditions, voice pop-ups are hidden, touch input is retained, and full menus are expanded, enabling precise alignment between interface adjustments and load states. The reward function, used to guide policy optimization, balances operational efficiency and accuracy and is calculated as:

$$R = 0.6 \times \frac{T_0 - T}{T_0} + 0.4 \times \frac{E_0 - E}{E_0} \tag{5}$$

where, T_0 and E_0 denote the baseline task completion time and operation error rate, respectively; and T and E represent the task completion time and error rate under the current policy. The weight allocation reflects an optimization objective that prioritizes efficiency while accounting for accuracy. Training is performed using a cloud-edge collaborative paradigm. In the cloud, offline pre-training is conducted using the proximal policy optimization algorithm with 10^5 historical interaction trajectories from 100 operators, generating a base interface adjustment policy. On the edge side, rapid fine-tuning is performed based on the first 10 operation data instances from the current operator, with fine-tuning time controlled within five minutes, thereby achieving individualized policy adaptation and ensuring that interface adjustments conform to the operational habits of different operators. Figure 2 shows the technical roadmap of cognitive load-driven meta-reinforcement learning interface reconfiguration and Bayesian closed-loop optimization.

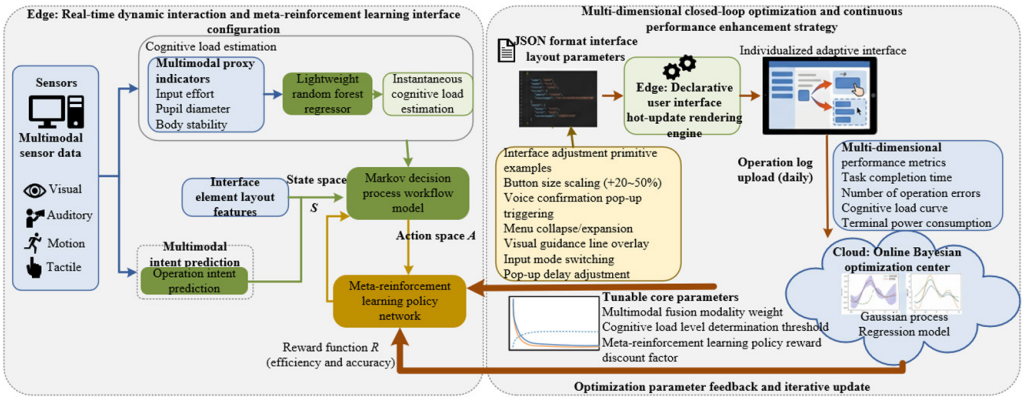

Fig. 2. Technical roadmap of cognitive load-driven meta-reinforcement learning interface reconfiguration and Bayesian closed-loop optimization

Low latency in interface updates is critical to ensuring operational fluency. Real-time interface reconfiguration is achieved using declarative user interface hot-update technology. Interface layout parameters are defined in JSON format, specifying core attributes such as position, size, and display state of interface elements. After receiving interface adjustment instructions on the edge side, the interface is

redrawn in real time by a lightweight rendering engine, allowing updates to be performed without application restart. To further reduce update latency, an incremental update strategy is adopted, in which only the changed interface elements are updated, avoiding the resource consumption and increased latency caused by full interface redrawing. After optimization, the end-to-end update latency is controlled within 50 ms, effectively preventing interface adjustments from interfering with operator workflows. Consequently, continuity and fluency in mobile supply chain operations are ensured, while the resource constraints of mobile terminals are accommodated.

2.4 Process preloading and closed-loop optimization driven by multimodal intent

Process preloading driven by multimodal intent is realized through modeling based on a Markov decision process. For three types of typical supply chain tasks—single-item picking, pallet receiving, and shelf inventory counting—the operational workflow is decomposed into a sequence of consecutive operation steps as states, and specific operator actions are defined as actions, thereby constructing a complete Markov decision process model. State transition probabilities are corrected using the prior probabilities output by the multimodal intent prediction module, with the correction formula given by:

$$P'_{ij} = P_{ij} \times (1 + \alpha \cdot c) \tag{6}$$

where, P_{ij} denotes the original transition probability from state i to state j; α is the transition probability enhancement coefficient set to 0.3; and c is an indicator variable of intent confidence. When the intent prediction confidence is not less than 0.8, $c = 1$; otherwise, $c = 0$. The preloading strategy is triggered based on the intent prediction confidence. When the intent prediction confidence reaches 0.7 or above, resources required for the corresponding operation, including the barcode scanning recognition model, numeric keypad layout, camera resources, and memory buffer, are automatically preloaded on the edge side. Preloading time is controlled within 100 ms, effectively eliminating delays caused by page switching and resource acquisition, significantly reducing operator waiting time, and enhancing operational continuity.

Closed-loop optimization is achieved through online Bayesian optimization, with the core objective being the minimization of task completion time and operation error rate, while keeping terminal power consumption within a reasonable range. The optimization process focuses on three core tunable parameters: the modality weight coefficient for multimodal fusion, the cognitive load level determination threshold, and the reward discount factor γ in the meta-reinforcement learning policy. Upon completion of each full operation cycle on the edge side, the operation log of that cycle is uploaded to the cloud. Each log contains task completion time, number of operation errors, cognitive load variation curve, and terminal power consumption data. An online Bayesian optimization model is constructed in the cloud based on a Gaussian process. The mean function and covariance function of the Gaussian process are given by:

$$\mu(x) = k(x, X)K(X, X)^{-1}y \tag{7}$$

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right) \tag{8}$$

where, x is the vector of optimization parameters, X is the set of historical parameters, y represents the performance metric corresponding to those parameters, σ_f^2 is the signal variance, and l is the length scale. During the iterative process, the iteration step size is dynamically adjusted according to parameter sensitivity, ensuring the rationality and efficiency of parameter updates. After 30 to 50 operation cycles, the optimization parameters converge to the optimal configuration, enabling precise adaptation to the specific warehouse environment and operator habits, thereby achieving continuous improvement in interaction performance and operational efficiency.

3 EXPERIMENTAL VALIDATION

3.1 Experimental environment and dataset

Experiments were conducted in a simulated warehouse environment, where a cloud-edge collaborative experimental platform was established to validate the performance of the proposed model. An Android 13 industrial tablet was used as the mobile terminal, equipped with a Snapdragon 870 processor, 8 GB random access memory, and 128 GB storage, and integrated with a 13-megapixel rear camera, an inertial measurement unit, a touchscreen pressure sensor, and a near-infrared camera, thereby meeting the real-time acquisition requirements for multimodal data. The cloud server was configured with an Intel Xeon E5-2690 central processing unit, an NVIDIA RTX 3090 graphics processing unit, and 64 GB random access memory, supporting model training, online distillation, and Bayesian optimization. The simulated warehouse environment occupied an area of 100 m^2, containing 50 shelves and 200 types of goods, and was divided into a picking zone, an inventory counting aisle, and a receiving area, replicating real-world mobile supply chain operation scenarios. The experimental dataset was derived from multimodal operation data collected from 100 operators aged 20–45, including both novice and experienced operators. Each operator performed 100 typical tasks: 40 picking, 30 inventory counting, and 30 receiving tasks. A total of 10^4 operation logs was collected, containing raw multimodal data, operation labels, and subjective ratings from the National Aeronautics and Space Administration Task Load Index (NASA-TLX) scale.

Twenty operators were selected for the experiment, including 10 novices and 10 experienced operators, and were randomly divided into an experimental group and a control group. The proposed model was applied to the experimental group, whereas the control group was evaluated using three types of baseline methods in comparative experiments. The three baseline methods consisted of a pure touchscreen fixed interface, a fixed multimodal interface, and a reinforcement learning-based interface without load awareness, thereby fully demonstrating the performance advantages of the proposed model.

3.2 Experimental results and analysis

Validation of lightweight multimodal fusion network performance. This experiment validated the effectiveness of the time-division cross-modal attention mechanism and knowledge distillation through four sets of comparisons. The compared methods included the proposed fusion network, full-modal attention fusion, single-modal fusion, and a fusion network without knowledge distillation. The experimental results are shown in Table 1.

Table 1. Performance comparison of the lightweight multimodal fusion network and other methods

Method	Intent Prediction Accuracy (%)	Inference Latency (MS)	Model Size (MB)	Average Terminal Power Consumption (W)
Proposed fusion network	93.2 ± 1.1	18.3 ± 2.5	46.8 ± 3.2	2.3 ± 0.3
Full-modal attention fusion	91.5 ± 1.3	45.7 ± 4.2	89.6 ± 5.8	3.8 ± 0.5
Single-modal fusion	82.5 ± 1.8	16.9 ± 2.1	38.7 ± 2.9	2.1 ± 0.2
Fusion network without knowledge distillation	83.5 ± 1.6	19.1 ± 2.7	48.2 ± 3.5	2.4 ± 0.3

As shown in Table 1, the proposed fusion network achieved the best overall performance, with an intent prediction accuracy of 93.2%, which was significantly higher than that of single-modal fusion and the fusion network without knowledge distillation, representing improvements of 10.7% and 9.7%, respectively. Its accuracy was also slightly higher than that of full-modal attention fusion, an advantage primarily attributable to the application of knowledge distillation, which allows the edge-side student network to inherit the high-precision characteristics of the cloud-based teacher network. In terms of inference latency and model size, the proposed fusion network achieved an inference latency of only 18.3 ms and a model size of 46.8 MB, substantially outperforming full-modal attention fusion, which exhibited an inference latency of 45.7 ms and a model size of 89.6 MB, making it difficult to accommodate the resource constraints of mobile terminals. Compared with single-modal fusion, the proposed fusion network achieved a significant improvement in accuracy without a notable increase in inference latency or power consumption, demonstrating the advantage of the time-division cross-modal attention mechanism, in which complete fusion computation is triggered only at moments of operational state transition, thereby effectively reducing computational load and power consumption. Consequently, average terminal power consumption was controlled at 2.3 W, satisfying the battery life requirements of mobile devices. These experimental results indicate that the synergistic design of time-division cross-modal attention and knowledge distillation achieves a balance between multimodal fusion accuracy and mobile device adaptability, thereby validating the effectiveness of this innovative module.

Accuracy validation of the cognitive load estimation mechanism. This experiment compared the performance of the proposed cognitive load estimation method, single-indicator estimation, and invasive electroencephalography-based estimation, while also analyzing the correlation with the NASA-TLX subjective scale. The experimental results are shown in Table 2.

Table 2. Performance comparison between the cognitive load estimation method and other methods

Method	Estimation Accuracy (%)	Estimation Latency (ms)	Correlation with the National Aeronautics and Space Administration Task Load Index (NASA-TLX)	Operational Interference
Proposed method	89.3 ± 1.5	27.6 ± 3.1	0.87 ± 0.04	None
Single-indicator estimation	73.6 ± 2.2	25.8 ± 2.8	0.62 ± 0.06	None
Invasive electroencephalography-based estimation	90.1 ± 1.2	42.5 ± 4.3	0.89 ± 0.03	Severe

As shown in Table 2, the proposed cognitive load estimation method achieved an accuracy of 89.3%, which is close to the 90.1% achieved by the invasive estimation method. Its correlation with the NASA-TLX subjective scale reached 0.87, indicating

that the proposed method accurately captures dynamic changes in cognitive load and aligns well with subjective cognitive load perception. In terms of estimation latency, the proposed method achieved only 27.6 ms, significantly lower than the 42.5 ms of the invasive method, thereby satisfying the real-time requirements of instantaneous cognitive load estimation. Compared with single-indicator estimation, the proposed method improved accuracy by 15.7% and increased the correlation by 0.25. This improvement is primarily attributable to the synergistic characterization of cognitive load from three dimensions—input effort, physiological state, and body stability—using three types of multimodal proxy indicators, thereby avoiding the limitations of single-indicator approaches. Furthermore, the proposed method adopts a non-invasive design that does not interfere with on-site operator workflows, whereas invasive estimation requires the wearing of electroencephalography equipment, which severely impairs operational fluency and is difficult to adapt to mobile supply chain scenarios. These experimental results validate the rationality of the selected multimodal proxy indicators and the design of the lightweight random forest regressor, realizing accurate, real-time, and non-invasive cognitive load estimation.

Validation of interactive performance of meta-reinforcement learning-driven dynamic interface reconfiguration. This experiment validated the effectiveness of load awareness and the meta-reinforcement learning strategy through four sets of comparisons. The compared methods included the proposed meta-reinforcement learning-based interface, a fixed interface, a reinforcement learning-based interface without load awareness, and a manually optimized interface. The experimental results are shown in Table 3.

Table 3. Comparison of interactive performance of the meta-reinforcement learning-driven dynamic interface reconfiguration and other methods

Method	Task Completion Time (s)	Operation Error Rate (%)	National Aeronautics and Space Administration Task Load Index (NASA-TLX) Score	Operation Satisfaction Score	Interface Switching Latency (ms)
Proposed meta-reinforcement learning-based interface	48.2 ± 3.5	2.3 ± 0.8	3.2 ± 0.7	4.6 ± 0.4	42.7 ± 5.3
Fixed interface	72.5 ± 4.8	8.9 ± 1.5	7.6 ± 0.9	2.8 ± 0.5	–
Reinforcement learning-based interface without load awareness	59.6 ± 4.2	4.7 ± 1.1	5.1 ± 0.8	3.7 ± 0.6	45.3 ± 5.8
Manually optimized interface	52.8 ± 3.9	3.1 ± 0.9	4.3 ± 0.8	4.2 ± 0.5	68.5 ± 6.2

As shown in Table 3, the interactive performance of the proposed meta-reinforcement learning-based interface was significantly superior to that of the other compared methods. In terms of task completion time, the proposed meta-reinforcement learning-based interface achieved only 48.2 s, which was 33.5% shorter than that of the fixed interface and 19.1% shorter than that of the reinforcement learning-based interface without load awareness. Its performance was also slightly better than that of the manually optimized interface. This improvement is primarily attributable to the ability of the meta-reinforcement learning strategy to achieve dynamic interface adaptation based on cognitive load and operation intent, with action primitives such as button scaling and voice pop-up triggering under high-load conditions effectively reducing operational difficulty and input cost. In terms

of operation error rate, the proposed meta-reinforcement learning-based interface achieved only 2.3%, representing a 74.2% reduction compared with the fixed interface and a 51.1% reduction compared with the reinforcement learning-based interface without load awareness, indicating that load awareness and dynamic interface adjustment effectively reduce operational errors.

Regarding subjective experience, the proposed meta-reinforcement learning-based interface achieved a NASA-TLX score of 3.2, which was 57.9% lower than that of the fixed interface, and an operation satisfaction score of 4.6, reflecting the improvement in user experience enabled by interface adaptability. For interface switching latency, the proposed meta-reinforcement learning-based interface achieved 42.7 ms, which was lower than the 68.5 ms of the manually optimized interface and close to that of the reinforcement learning-based interface without load awareness, benefiting from the application of declarative user interface hot-update and incremental update strategies, thereby ensuring the smoothness of interface adjustments. These experimental results demonstrate that the meta-reinforcement learning-driven load-aware dynamic interface reconfiguration strategy achieves precise matching between the interface and operator state, significantly improving interactive performance and user experience.

Validation of efficiency improvement through process preloading and closed-loop optimization. This experiment compared the proposed complete model, a model without preloading, a model without closed-loop optimization, and Baseline 3, with a focus on validating the effectiveness of Markov decision process-driven preloading and online Bayesian optimization. The experimental results are shown in Table 4.

Table 4. Comparison of efficiency improvement through process preloading and closed-loop optimization and other methods

Method	Task Completion Time (s)	Efficiency Improvement (%)	Number of Cycles to Parameter Convergence	Efficiency Fluctuation Range (%)
Proposed complete model	47.8 ± 3.4	27.5 ± 2.1	42 ± 5	±1.2
Model without preloading	62.3 ± 4.1	8.9 ± 1.7	–	±1.3
Model without closed-loop optimization	53.6 ± 3.8	18.7 ± 1.9	–	±5.4
Baseline 3	59.7 ± 4.3	12.3 ± 1.8	–	±4.8

As shown in Table 4, the proposed complete model achieved an efficiency improvement of 27.5%, which was substantially higher than that of the model without preloading (8.9%), the model without closed-loop optimization (18.7%), and Baseline 3 (12.3%). The task completion time of the proposed complete model was the shortest, at only 47.8 s. The relatively low efficiency improvement of the model without preloading indicates that the Markov decision process-driven process preloading strategy effectively eliminates delays caused by page switching and resource acquisition by preloading resources such as barcode scanning models and numeric keypads, thereby reducing operator waiting time and improving operational efficiency. The model without closed-loop optimization achieved a lower efficiency improvement than the proposed complete model and exhibited an efficiency fluctuation range of ±5.4%, which was significantly higher than the ±1.2% observed for the proposed complete model. This finding indicates that online Bayesian optimization, through iterative updates of core parameters such as modality weights and cognitive load thresholds, enables rapid adaptation of the model to operator habits and warehouse environments, thereby achieving stable performance improvement. The proposed complete model required 42 cycles for parameter convergence, which falls within the expected range

of 30 to 50 cycles, demonstrating that closed-loop optimization converges quickly to the optimal configuration without requiring prolonged iteration to achieve efficiency gains. These experimental results validate the synergistic effect of process preloading and online Bayesian closed-loop optimization, demonstrating that they effectively improve supply chain efficiency while ensuring performance stability.

To validate the micro-scale computational scheduling mechanism and resource consumption characteristics of the proposed lightweight temporal fusion architecture under the constrained hardware environment of mobile terminals, multimodal feature responses and dynamic load synergy data were extracted and analyzed from continuous typical operation segments. As shown in the experimental observations in Figure 3, when the operator was in a steady state such as walking, high-dimensional visual and auditory feature extraction remained in a dormant state, with only the basic sequence updates of the low-power inertial sensor being retained. Consequently, the real-time power consumption of the terminal was stably constrained to an extremely low baseline level. Upon the detection of abrupt changes in operational state, such as raising the terminal for scanning or screen confirmation, complete cross-modal attention computation was immediately triggered on the edge side. The activation levels of all modalities exhibited a pulsed joint surge, accompanied by the generation of a very narrow spike in efficient computational power consumption. This alternating operation pattern of on-demand activation and sparse computational distribution directly confirms at the physical level that the time-division triggering mechanism completely eliminates redundant computational load during non-critical periods. While ensuring accurate capture of dynamic operation intent in supply chain scenarios, this underlying mechanism substantially resolves the conflict between the computational bottleneck of industrial mobile devices and the requirement for long battery life, thereby establishing a critical hardware adaptation foundation for efficiency improvement in high-frequency mobile interaction scenarios.

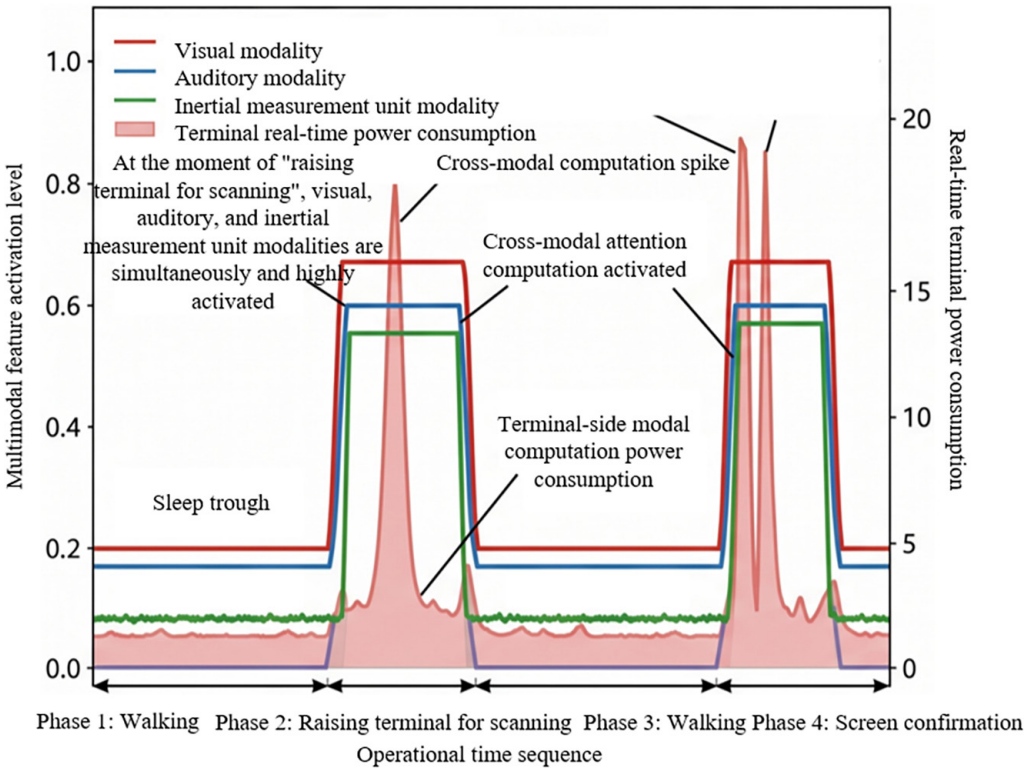

Fig. 3. Synergistic analysis of multimodal response and dynamic computational load

4 CONCLUSION

To address the efficiency bottlenecks and operational errors caused by mobile terminal resource constraints, fluctuations in operator cognitive load, and inadequate interface adaptability in mobile supply chain operations, a mobile interaction interface optimization and supply chain efficiency enhancement model based on multimodal perception was proposed. A cloud-edge collaborative, lightweight multimodal interaction architecture was constructed, in which efficient processing of multimodal data at the edge and accurate prediction of operation intent were achieved through the synergy of time-division cross-modal attention and knowledge distillation. Instantaneous cognitive load estimation driven by multimodal proxy indicators was combined with meta-reinforcement learning to formulate a load-aware dynamic interface reconfiguration strategy, enabling real-time adaptation between the interface and operator states. Process preloading driven by a Markov decision process and online Bayesian closed-loop optimization were employed to achieve significant improvement in operational efficiency and stable performance. This study enriches the research outcomes in mobile multimodal interaction and adaptive interface optimization, providing a lightweight and adaptive innovative solution for the deep application of mobile technologies in industrial supply chain scenarios and offering important technical support for the digital transformation of supply chains. Regarding the limitations of the model in complex task adaptation, multi-terminal compatibility, and robustness in extreme environments, future work will incorporate federated learning, transfer learning, and affective computing to expand application scenarios, optimize multimodal data acquisition accuracy and model adaptability, promote the deep integration of mobile interaction technologies with industrial supply chains, and better meet the demands of practical industrial applications.

5 REFERENCES

[1] J. Liu, "Strategy and practice for improving supply chain management through mobile interaction technology," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 4, pp. 178–192, 2025. https://doi.org/10.3991/ijim.v19i04.54219

[2] C. Li and Y. Gong, "Integration of mobile interaction technologies in supply chain management for S2B2C E-Commerce platforms," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 3, pp. 227–241, 2025. https://doi.org/10.3991/ijim.v19i03.53953

[3] J. Danker, I. Strnadová, M. Tso, J. Loblinzk, T. M. Cumming, and A. J. Martin, "'It will open your world up': The role of mobile technology in promoting social inclusion among adults with intellectual disabilities," *British Journal of Learning Disabilities*, vol. 51, no. 2, pp. 135–147, 2023. https://doi.org/10.1111/bld.12500

[4] H. Y. Liu, Z. Wu, and L. N. Yu, "Optimization of emergency stockpile siting: A review of models, influencing factors, and future research directions," *Journal of Engineering Management and Systems Engineering*, vol. 3, no. 4, pp. 226–235, 2024. https://doi.org/10.56578/jemse030404

[5] M. Sharifzadeh, M. C. Garcia, and N. Shah, "Supply chain network design and operation: Systematic decision-making for centralized, distributed, and mobile biofuel production using mixed integer linear programming (MILP) under uncertainty," *Biomass and Bioenergy*, vol. 81, pp. 401–414, 2015. https://doi.org/10.1016/j.biombioe.2015.07.026

[6] S. Dabic-Miletic, "The challenges of integrating AI and robotics in sustainable WMS to improve supply chain economic resilience," *Journal of Industrial Intelligence*, vol. 2, no. 2, pp. 119–131, 2024. https://doi.org/10.56578/jii020205

[7] P. Xing, M. Wang, and J. Yao, "Optimal service quality and pricing for App service supply chain with network externality based on four different scenarios," *Kybernetes*, vol. 52, no. 9, pp. 3425–3450, 2023. https://doi.org/10.1108/K-06-2021-0470

[8] J. C. Nan *et al.*, "Miniaturized inverted ultra-wideband multiple-input multiple-output antenna with high isolation," *Electronics Letters*, vol. 57, no. 3, pp. 100–102, 2021. https://doi.org/10.1049/ell2.12072

[9] M. A. S. Mustafa, "Predictive reliability-driven optimization of spare parts management in aircraft fleets using AI, IoT, and digital twin technologies," *Journal of Engineering Management and Systems Engineering*, vol. 4, no. 3, pp. 218–236, 2025. https://doi.org/10.56578/jemse040305

[10] X. Nie, L. Yi, L. T. Yang, X. Deng, F. Fan, and H. Zhang, "Tensor dynamic fusion-based modality-imbalanced multimodal federated learning in mobile edge computing for consumer applications," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 6570–6582, 2024. https://doi.org/10.1109/TCE.2024.3470243

[11] J. Wang, Y. Yan, and G. Zhao, "Task recommendation method combining multimodal cognition and collaboration in mobile crowdsensing systems," *Computer Networks*, vol. 229, p. 109796, 2023. https://doi.org/10.1016/j.comnet.2023.109796

[12] M. Miraz, M. Ali, and P. S. Excell, "Cross-cultural usability evaluation of AI-based adaptive user interface for mobile applications," *Acta Scientiarum Technology*, vol. 44, p. e61112, 2022. https://doi.org/10.4025/actascitechnol.v44i1.61112

[13] J. Nasreddine, L. Nuaymi, and X. Lagrange, "Adaptive power control algorithm with stabilization zone for third generation mobile networks," *Annals of Telecommunications*, vol. 61, nos. 9–10, pp. 1193–1211, 2006. https://doi.org/10.1007/BF03219888

[14] J. Yang *et al.*, "Cost performance optimization of waste heat recovery supply chain by mobile heat storage vehicles," *Energy Reports*, vol. 6, pp. 137–146, 2020. https://doi.org/10.1016/j.egyr.2020.05.009

[15] P. Zhang and Y. Dong, "RETRACTED ARTICLE: Strategy transformation of big data green supply chain by using improved genetic optimization algorithm," *Soft Computing*, 2023. https://doi.org/10.1007/s00500-023-08911-5

[16] M. Martínez-Romero *et al.*, "A cloud-based platform for harmonized COVID-19 data: Design and implementation of the rapid acceleration of diagnostics (RADx) data hub," *JMIR Public Health and Surveillance*, vol. 11, no. 1, p. e72677, 2025. https://doi.org/10.2196/72677

[17] R. Janssen, R. van de Molengraft, H. Bruyninckx, and M. Steinbuch, "Cloud-based centralized task control for human domain multi-robot operations," *Intelligent Service Robotics*, vol. 9, no. 1, pp. 63–77, 2016. https://doi.org/10.1007/s11370-015-0185-y

6 AUTHORS

Jia Liu graduated from Zhejiang University of Technology in 2015 with a Master's degree in Applied Economics, and now she is working at the Zhejiang Technical Institute of Economics. She has published four papers. Her research interests include international logistics, supply chain management, and international trade (E-mail: 230069@zjtie.edu.cn).

Tian Gao graduated from Zhejiang College of Zhejiang University of Technology, majoring in Environmental Design. He is a Senior Engineer at Zhejiang Lijiu Environmental Technology Co., Ltd. His research focuses on leveraging mobile interaction technologies to enhance operational efficiency and sustainability in industrial settings (E-mail: andygao413@outlook.com).