Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 17 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i17.62957

PAPER

Optimization of Personalized Tourism Services via Multimodal Mobile Interaction

Duchuan Zhang(✉)

Zhengzhou Tourism College, Zhengzhou, China

zhangduchuan@zztrc.edu.cn

ABSTRACT

In mobile tourism scenarios, the inherent contradiction between the demand for real-time personalized services and the constraints of terminal computing power, battery life, and network fluctuations presents a significant challenge. Existing solutions commonly suffer from fragmented interaction entry points, delayed contextual responses, and excessive reliance on cloud computing. To address these issues, a three-tier collaborative optimization framework, encompassing on-device perception, edge coordination, and dynamic orchestration, was proposed. Within this framework, modality-level attention pooling was employed to achieve adaptively weighted fusion of visual, auditory, and sensory signals. A physical-digital fused mobile context tensor was constructed, and temporal encoding was utilized to capture the evolutionary patterns of user intent. Service scheduling was formulated as a constrained Markov decision process, and a resource-aware deep reinforcement learning algorithm was applied to derive an optimal edge-device collaborative service orchestration policy. To ensure deployability on mid-range mobile devices, complementary mechanisms of dynamic network pruning and pseudo-label distillation were incorporated. Experimental results on a real-world scenic area dataset demonstrated that the proposed method achieved a Top-1 intent recognition accuracy that was 9.7% higher than the next-best baseline. The end-to-end average response latency was reduced by 30.3% compared to reinforcement learning-based orchestration schemes of a similar type. Furthermore, the energy consumption per inference was decreased by 42% relative to the original model, with an accuracy degradation controlled within 3%. This work provides a viable technical paradigm for optimizing smart tourism services in mobile environments.

KEYWORDS

mobile multimodal interaction, context-aware computing, dynamic service orchestration, deep reinforcement learning, on-device lightweighting, smart tourism

Zhang, D. (2026). Optimization of Personalized Tourism Services via Multimodal Mobile Interaction. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(17), pp. 152–165. https://doi.org/10.3991/ijim.v20i17.62957

Article submitted 2026-04-17. Revision uploaded 2026-06-11. Final acceptance 2026-06-20.

1 INTRODUCTION

With the deep integration of mobile computing and smart tourism, intelligent terminals have become the primary vehicles for tourist interaction during travel, with multimodal interaction and personalized services increasingly recognized as crucial directions for enhancing the smart tourism experience [1, 2]. Tourism scenarios are characterized by significant spatiotemporal heterogeneity, wherein tourist behaviors switch frequently, and the availability of modalities such as voice, vision, touch, and location sensing undergoes dynamic changes. This imposes heightened demands on service real-time performance, continuity, and personalization. However, constrained by terminal computing power, battery life, and network fluctuations prevalent in scenic areas, traditional cloud-centric architectures tend to incur high communication latency and service interruptions, rendering them inadequate for stably supporting real-time interactions and computationally intensive applications [3, 4]. Therefore, the construction of a real-time multimodal personalized service framework that is aware of on-device resource constraints holds considerable value for advancing the deployment of edge intelligence in smart tourism.

Existing research still exhibits four principal shortcomings. At the multimodal interaction level, prevalent approaches predominantly rely on feature concatenation, decision-level fusion, or fixed weight assignment, which prove inadequate in addressing modality missing, noise interference, and channel occupancy conflicts. Furthermore, high-accuracy models incur substantial computational overhead, rendering them incapable of meeting the low-latency inference demands on the device side [5, 6]. At the context-aware intent recognition level, existing recommendation methods largely depend on historical check-ins, ratings, and visitation trajectories while insufficiently leveraging on-site contextual information such as weather, crowding levels, real-time location, device status, and immediate interactions. Effective modeling of the temporal evolution of tourist intent is also lacking [7, 8]. The long-tail distribution of tourism service categories further biases models toward high-frequency attractions and common demands, thereby diminishing the capacity to recognize low-frequency interests, rare requirements, and transient intents [9]. At the service orchestration level, existing approaches primarily target latency, throughput, and quality of service as optimization objectives, without incorporating terminal battery level, memory, computational load, and link status into a unified closed-loop decision framework. Although certain studies have attended to the synergistic optimization of edge service deployment, travel itinerary recommendation, and computational resource allocation, service migration, resource reconfiguration, and multi-node coordination in dynamic scenic area networks remain complex challenges [10, 11]. Moreover, cloud-edge-device task offloading and routing mechanisms exhibit strong dependence on network stability [12]. At the on-device deployment level, existing models lack systematic lightweighting solutions that comprehensively cover structural compression, parameter pruning, quantization inference, and hardware adaptation, making it difficult to simultaneously balance accuracy, latency, memory footprint, and energy consumption. Additionally, offline training and online incremental learning remain disconnected from each other, hindering models from continuously adapting to new tourist behaviors and scenarios while remaining susceptible to catastrophic forgetting [13, 14].

To address the aforementioned issues, a three-tier collaborative optimization framework encompassing on-device perception, edge coordination, and dynamic orchestration is constructed, with four core contributions. First, a modality-adaptive

weighted multimodal unified representation method is proposed, wherein modality-level attention pooling is employed to dynamically allocate computational weights, thereby accommodating the modality occupancy variations across different scenarios. Second, a lightweight spatiotemporal context-fused temporal intent decoding model is developed, in which a mobile context tensor integrating multi-dimensional physical information is defined, a cross-modal gated fusion unit is designed to achieve semantic alignment, and temporal encoding is incorporated to capture the evolutionary patterns of intent, thereby alleviating the long-tail distribution problem of services. Third, a resource-aware deep reinforcement learning-based service orchestration mechanism is introduced, wherein service optimization is formulated as a constrained Markov decision process, a multi-granularity action space and a multi-objective reward function are designed, and pre-filtering is incorporated to enhance localization robustness. Fourth, an on-device lightweight deployment scheme and an online self-evolution closed loop are devised, in which model compression is achieved through structured pruning and knowledge distillation, and incremental optimization is realized based on user feedback.

The remainder of this study is organized as follows: Section 2 elaborates on the overall framework architecture and core technical details; Section 3 validates the effectiveness and robustness of the proposed method through multiple sets of experiments; Section 4 concludes the study and discusses future research directions.

2 MULTIMODAL MOBILE SERVICE OPTIMIZATION FRAMEWORK FOR TOURISM SCENARIOS

2.1 Problem formalization and overall architecture

The core problem investigated in this work can be formalized as follows: subject to the hard constraints of mobile terminal computational load, remaining battery capacity, and network quality, an optimal sequence of personalized service actions for tourists is to be derived based on real-time multimodal interaction signals and spatiotemporal contextual information acquired by the device, thereby achieving joint optimization of service matching quality and system resource consumption. To address this problem, a three-tier collaborative optimization framework encompassing on-device perception, edge coordination, and dynamic orchestration is constructed. From bottom to top, the framework comprises four layers: the multimodal unified representation layer, the mobile context fusion layer, the lightweight intent decoding layer, and the resource-aware service orchestration layer. Guided by the logical thread of feature dimensionality reduction with fidelity preservation, context-enhanced intent modeling, and constraint-driven decision-making, the framework operates as follows: lightweight feature perception and rapid decision-making are performed on the device side; collaborative computation and model incremental updates are provided by edge nodes; and elastic scheduling of service resources is enabled through global dynamic orchestration. In this manner, personalized tourism service adaptation in complex scenic-area environments is supported while millisecond-level response performance is guaranteed. Figure 1 illustrates the proposed edge-device collaborative optimization framework for multimodal mobile services in tourism scenarios.

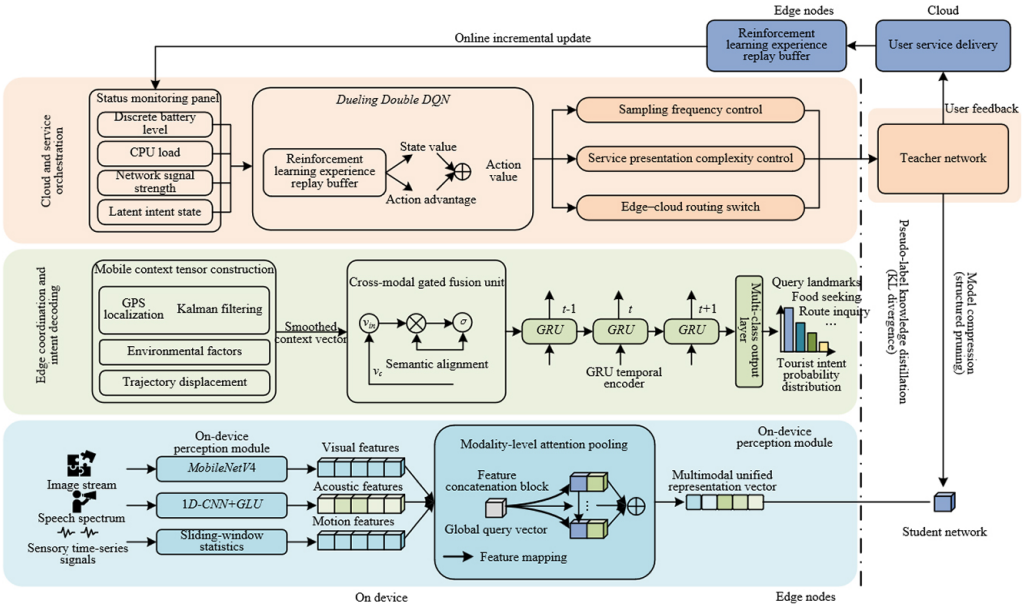

Fig. 1. Edge-device collaborative optimization framework for multimodal mobile services in tourism scenarios

2.2 Modality-adaptive weighted multimodal unified representation layer

The multimodal inputs captured in real time by mobile devices comprise three categories of signals: image streams $I \in R^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels of the image, respectively; speech instructions transformed into two-dimensional mel-spectrogram features $T \in R^{F\times N}$ after mel-frequency conversion, where F denotes the number of mel-frequency bands and N denotes the number of time frames; and time-series signals from touch, gyroscope, and other sensors, denoted as $S \in R^{d\times t}$, where d is the sensor dimension and t is the temporal window length. For the visual branch, a MobileNetV4 backbone network optimized via neural architecture search is employed to extract compact visual features f_v, with small convolutional kernels and depthwise separable structures adopted to meet the low-latency requirements of the device side. For the acoustic branch, one-dimensional depthwise separable convolutions combined with gated linear units are utilized to extract semantic features f_a. Sensor signals are encoded as motion features f_s through sliding-window statistics. To accommodate the dynamically changing modality channel occupancy states characteristic of tourism scenarios, the three branch features are adaptively weighted and fused through a modality-level attention pooling mechanism, which replaces conventional static weighting schemes to adapt to the computational resource allocation demands of different interaction scenarios.

Modality-level attention pooling dynamically computes confidence weights for each modality through a learnable query vector. The weight computation formula is given by:

$$\alpha_k = \frac{exp\left(q^T \cdot \tanh(W_k f_k + b_k)\right)}{\sum_{j\in\{v,a,s\}} exp\left(q^T \cdot \tanh(W_j f_j + b_j)\right)} \tag{1}$$

where, $q \in R^{d_m}$ denotes the globally shared learnable query vector, and W_k and b_k denote the projection weight matrix and bias vector for the k-th modality, respectively,

with $k \in \{v, a, s\}$ corresponding to the visual, acoustic, and sensor modalities. Based on the modality weights, the branch features are mapped into a unified feature space via the projection matrix W_{proj} and then weighted and summed to obtain the final multimodal unified representation vector:

$$F_{multi} = \sum_k \alpha_k \cdot W_{proj} f_k \tag{2}$$

where, $F_{multi} \in R^{d_m}$, and d_m denotes the feature dimension of the unified representation. This mechanism enables automatic adjustment of the modality weight distribution according to the interaction scenario: when the user is in a crowded environment with both hands occupied, the weight of the acoustic modality is adaptively increased, while the image sampling frequency is simultaneously reduced to eliminate ineffective computational overhead; when the user raises the device to capture a landmark, the visual modality weight becomes dominant, ensuring the feature precision required for scene perception. In this manner, dynamic on-demand allocation of on-device computational resources is achieved, and an adaptive balance is maintained between feature representation fidelity and computational energy efficiency.

2.3 Lightweight temporal intent decoding layer with spatiotemporal context fusion

On the basis of multimodal interaction representation, on-site spatiotemporal constraints are further injected to enhance the scenario adaptability of intent recognition. To this end, a mobile context encoding that fuses physical states and behavioral characteristics is constructed. The tourist's current location, mapped to a discretized geographic unit, is denoted as Loc_t; environmental factors, calibrated by onboard sensors, are denoted as Env_t; the trajectory displacement vector within the past time window and the duration of stay are denoted as ΔP_t and $Stay_t$, respectively. The four categories of features are concatenated along the dimension and then mapped through a multi-layer perceptron to obtain a unified context vector $E_t = \text{MLP}([Loc_t; Env_t; \Delta P_t; Stay_t])$. To achieve deep semantic alignment between the interaction representation and the context encoding, a cross-modal gated fusion unit is designed, wherein the two types of features are projected into a joint latent space for adaptive fusion. The computation process is given by:

$$G_t = \sigma(W_g[F_{multi}; E_t] + b_g) \odot \tanh(W_u F_{multi} + W_v E_t) \tag{3}$$

$$H_t = LayerNorm(G_t + F_{multi}) \tag{4}$$

where, W_g and b_g denote the weight matrix and bias vector of the gating branch; W_u and W_v denote the projection matrices for the interaction features and context features, respectively; σ denotes the sigmoid activation function; and $\odot$ denotes the Hadamard product. This unit enables dynamic adjustment of the injection weight of contextual information according to the current scenario, thereby reinforcing the guiding role of spatiotemporal context in intent inference while preserving the core features of the interaction.

Tourist service intents exhibit dynamic evolutionary characteristics throughout the touring process, and the fused features at a single moment are insufficient for completely characterizing the temporal evolution patterns of intent. To address this, a gated recurrent unit is employed as the temporal encoder, performing temporal

modeling on the sequence of enhanced representations across consecutive time steps, thereby capturing intent transitions across different phases, from navigation and browsing to stationary experiential activities. The hidden state update is controlled by the update gate, which determines the fusion ratio between the historical state and the current candidate state, ultimately yielding the temporal hidden state h_t at the current moment. Based on this hidden state, the intent prediction module outputs the probability distribution over various service intents through a fully connected layer followed by a softmax activation, which is formalized as:

$$P(Intent_t = i \mid h_t) = Softmax(W_{out} h_t + b_{out}) \tag{5}$$

where, W_{out} and b_{out} denote the weight matrix and bias vector of the output layer. To address the issue of uneven service category distribution in tourism scenarios, wherein high-value rare intents are prone to being overwhelmed by common samples, focal loss is adopted as the loss function during training. By down-weighting the loss contribution of easily classified common services, the recognition accuracy for long-tail intents is improved, thereby ensuring the recommendation recall rate for scarce in-depth experiential services.

2.4 Resource-aware deep reinforcement learning service orchestration mechanism

After the temporal decoding of user intent is completed, the intent prediction results need to be further transformed into executable service delivery strategies, with terminal computing power, battery life, and network status incorporated into a unified decision-making framework. Mobile service orchestration in tourism scenarios is modeled as a constrained Markov decision process, through which joint optimization of service quality and resource consumption is achieved via sequential decision-making. The state space of this process is defined as $s_t = \{h_t, Bat_t, CPU_t, Net_t\}$, where h_t denotes the hidden state vector output by the temporal intent decoder, Bat_t denotes the discrete hierarchical representation of remaining terminal battery capacity, CPU_t denotes the current processor load rate, and Net_t denotes the real-time network signal strength. The state set comprehensively covers user intent and full-dimensional resource constraints on the device side. The corresponding action space encompasses three levels of decision granularity: sampling frequency adjustment for multimodal inputs, dynamic routing selection between the on-device lightweight model and the cloud-based high-precision model, and presentation complexity control of service delivery content, thereby enabling elastic scheduling across the entire pipeline from input and computation to output. To address the issue of location drift caused by complex terrain in scenic areas, which may interfere with the accuracy of contextual states, an adaptive Kalman filter is embedded at the state input stage to perform real-time smoothing correction on the position sequence. The filter gain and position update formulas are given by:

$$K_t = P_{t|t-1} C^T \left(C P_{t|t-1} C^T + R_t \right)^{-1}, \hat{L}_{t|t} = \hat{L}_{t|t-1} + K_t \left(Z_t - C \hat{L}_{t|t-1} \right) \tag{6}$$

where, K_t denotes the Kalman gain matrix, $P_{t|t-1}$ denotes the prior state covariance matrix, C denotes the observation matrix, R_t denotes the observation noise covariance, $\hat{L}_{t|t-1}$ denotes the prior position estimate, and Z_t denotes the position observation at the current time step. This filtering module effectively suppresses instantaneous localization noise, ensures the stability of contextual state inputs, and prevents erroneous service switching triggered by anomalous positioning data.

To guide the policy toward optimal trade-offs under multiple constraints, a multi-objective composite reward function is designed, simultaneously quantifying service benefits and resource costs. Its specific form is given by:

$$R(s_t, a_t) = \lambda_1 \cdot CTR(a_t) + \lambda_2 \cdot Dwell(a_t) - \lambda_3 \cdot I_{latency>\delta} \cdot Penalty - \lambda_4 \cdot \Delta Bat \quad (7)$$

In this formulation, $CTR(a_t)$ denotes the estimated click-through rate of the service content under the corresponding action, and $Dwell(a_t)$ denotes the estimated duration of user's stay on the service page; together, these two terms characterize the quality of service matching and the user experience benefit. $I_{latency>\delta}$ is an indicator function that takes the value of 1 when the end-to-end service response latency exceeds the threshold and 0 otherwise, with penalty denoting the timeout penalty factor. ΔBat denotes the estimated additional battery consumption incurred by executing the current action. The weight coefficients λ_1, λ_2, λ_3, and λ_4 correspond to the relative importance of each optimization term and can be dynamically adjusted according to the terminal operating status—for example, the weight of λ_4 is increased in low-battery scenarios to enforce energy constraints, while λ_3 is increased in high-latency-sensitive scenarios to ensure real-time response performance.

Given the characteristics of the state space, which comprises continuous resource variables, and the action space, which consists of discrete multi-granularity combinations, the dueling double deep Q-network algorithm is employed to solve for the optimal orchestration policy. By decoupling the state value function and the action advantage function, the stability and convergence efficiency of discrete action value estimation are improved. During training, a prioritized experience replay mechanism is introduced, wherein samples in the replay buffer are prioritized according to their temporal-difference errors, and samples with higher errors are assigned greater sampling probabilities, thereby accelerating policy iteration in rare critical scenarios. This orchestration mechanism enables adaptive adjustment of the service delivery mode according to the real-time terminal state: when network conditions are sufficient, routing to the cloud model is performed to deliver high-precision personalized services; in weak-network or low-battery environments, switching to on-device lightweight inference is executed with simplified service presentation forms. In this manner, dynamic optimal balance between service experience and resource consumption is achieved within millisecond-level response constraints.

2.5 On-device lightweight deployment and online incremental evolution mechanism

To ensure efficient deployment of the entire framework on mainstream mid-range mobile devices, a systematic lightweighting scheme is constructed from two dimensions—structural compression and knowledge transfer—thereby minimizing on-device inference overhead within a controlled accuracy loss. In the dynamic network pruning strategy, the scaling factors of batch normalization layers are adopted as the quantitative basis for channel importance. During the training phase, an L1 sparse regularization constraint is imposed on all scaling factors across the entire network, with the sparse loss term defined as:

$$L_{sparse} = \sum_{\gamma \in \Gamma} |\gamma| \quad (8)$$

where, Γ denotes the set of scaling factors from all batch normalization layers in the network, and γ denotes the scaling coefficient corresponding to an individual

convolutional channel. After the sparse-regularized training is completed, channels with scaling factors below a pre-defined threshold are identified as redundant during the inference phase and are structurally pruned, thereby significantly compressing the model parameter count and floating-point operations while preserving the core feature representation capability. On this basis, an offline–online collaborative pseudo-label distillation mechanism is introduced, wherein unannotated interaction samples collected during the daytime are used by the cloud-based teacher network during nighttime Wi-Fi environments to generate soft labels. The on-device lightweight student network performs imitation learning by minimizing the Kullback–Leibler divergence between the outputs of the two networks, with the distillation loss formulated as:

$$L_{distill} = KL(p_{teacher} \| p_{student}) \tag{9}$$

where, $p_{teacher}$ denotes the class probability distribution output by the teacher network, and $p_{student}$ denotes the predicted distribution of the student network. This distillation mechanism does not require large-scale backpropagation training to be executed on the device side; rather, accuracy degradation caused by pruning is compensated through incremental knowledge transfer alone, thereby achieving synergistic optimization of model size and recognition accuracy.

To enable continuous iterative performance improvement of the system, a user feedback-driven online incremental self-evolution closed loop is constructed. Interactive behaviors of tourists, such as liking, bookmarking, or ignoring pushed services, are transmitted back to the reinforcement learning experience replay buffer as real reward signals. Based on newly collected samples, the Q-network parameters are fine-tuned through incremental updates, with the matching accuracy between intent prediction and service orchestration policies being continuously refined. This closed-loop mechanism does not require centralized batch retraining; instead, parameter iteration is accomplished solely through collaborative computation between the device side and edge nodes. As user usage duration increases, the personalization degree of service matching and resource utilization efficiency are continuously improved, forming a positive feedback cycle in which usage frequency and system performance mutually reinforce each other.

3 EXPERIMENTAL DESIGN AND RESULT ANALYSIS

3.1 Experimental setup

The experiments were conducted using a combination of publicly available datasets and field-collected datasets. For the field-collected dataset, three typical 5A-class scenic areas in China were selected as test sites, and 120 subjects of different age groups were recruited to participate in a seven-day in-travel interaction experiment. A total of 186,000 valid multimodal interaction samples were collected, encompassing four types of data: images, speech, inertial sensing, and location trajectories. These samples were annotated to form 12 categories of service intents, including guided tours, inquiries, check-ins, and food recommendations. As a supplement, the Travel-Multimodal standard tourism multimodal dataset was adopted as the public dataset, containing 23,000 annotated samples. All data were randomly divided into training, validation, and test sets at a ratio of 7:1:2.

For on-device testing, three tiers of mainstream mobile devices were selected, equipped with Snapdragon 8 Gen 3, Snapdragon 7 Gen 2, and Dimensity 700 processors, corresponding to high-end, mid-range, and low-end mobile terminals, respectively. Edge computing nodes were configured with ARM architecture-based edge

servers equipped with 16 GB of memory. Cloud computing nodes were configured with graphics processing unit servers equipped with NVIDIA A10 graphics cards and 32 GB of memory. All experiments were repeated 10 times under identical environmental conditions, and the average values were taken as the final results.

3.2 Intent prediction accuracy comparison experiment

This set of experiments was designed to validate the improvement effect of multimodal adaptive fusion and spatiotemporal context modeling on user intent recognition accuracy, with particular emphasis on the recognition performance for long-tail service categories. The intent recognition performance comparison of different methods is presented in Figure 2.

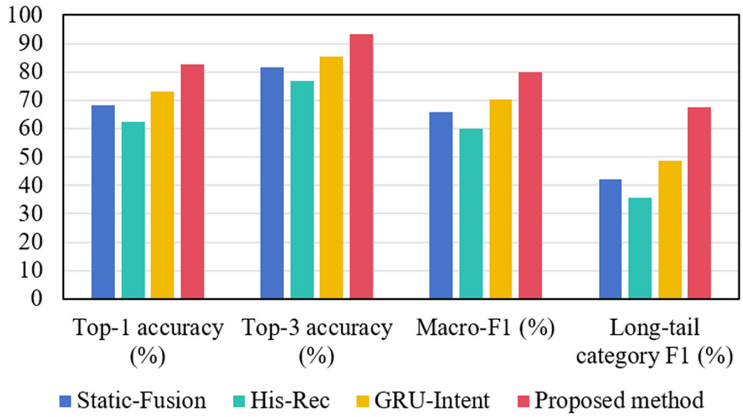

Fig. 2. Intent recognition performance comparison of different methods

As shown in Figure 2, the proposed method outperformed all comparison baselines across all intent recognition metrics. Specifically, a Top-1 accuracy of 82.6% was achieved, which is 9.7 percentage points higher than that of the next-best GRU-Intent method. The Macro-F1 score reached 79.8%, representing an improvement of 9.4 percentage points. Notably, the long-tail category F1 score was improved from 48.6% to 67.5%, corresponding to an increase of 18.9 percentage points, indicating that the fusion of contextual information and the targeted design of the loss function effectively alleviated the issue of uneven service category distribution and enhanced the recall capability for rare high-value intents. In comparison with the His-Rec method, which relies solely on historical behaviors, the proposed method, through the integration of instantaneous physical context and temporal behavioral features, enabled more precise capture of users' dynamic on-site demands, thereby avoiding the deviation between static preference profiles and real-time intents.

3.3 Service orchestration: a comprehensive performance and energy efficiency experiment

This set of experiments was designed to validate the comprehensive performance of the resource-aware reinforcement learning-based orchestration mechanism under multiple constraints, with particular emphasis on the balance among real-time performance, energy efficiency, and service quality under different network and battery scenarios. The experiments were conducted on mid-range mobile devices, and the performance was tested under normal network and weak-network conditions, as well as full-battery and low-battery scenarios. The results are presented in Table 1.

The experimental results demonstrate that, under normal network conditions, the average end-to-end latency of the proposed method was only 124 ms, which is 74.5% lower than that of pure cloud-based scheduling and 30.3% lower than that of the non-resource-aware DQN-Sched, fully satisfying the millisecond-level response requirements. The energy consumption per unit time was reduced by 21.8% compared to DQN-Sched, reflecting the energy efficiency advantage of resource-aware scheduling. Under weak-network conditions, the latency of the pure cloud-based method increased sharply to 1257 ms, whereas the proposed method, relying primarily on on-device local inference with cloud-based coordination as a supplement, exhibited only a modest increase in latency to 162 ms, demonstrating strong scenario robustness. Under low-battery conditions, through automatic frequency reduction and on-device routing switching, the energy consumption increase of the proposed method was constrained to within 8.2%, while the service quality composite score remained higher than all baseline methods, achieving an optimal trade-off between service experience and resource consumption.

Table 1. Comprehensive performance comparison of different service orchestration methods

Method	Average Latency (ms)	Energy Consumption Per Unit Time (mAh/min)	Service Quality Composite Score	Latency Under Weak Network (ms)	Energy Consumption Under Low Battery (mAh/min)
Cloud-Sched	486	1.82	82.3	1257	1.91
QL-Sched	215	1.36	71.5	231	1.42
DQN-Sched	178	1.24	76.8	195	1.30
Proposed method	124	0.97	83.7	162	1.05

3.4 Core module ablation experiment

This set of experiments was designed to quantify the contribution of each component to the overall performance by removing core modules one at a time, thereby validating the necessity of each module design. Based on the complete model, the modality-level attention, the cross-modal gated fusion unit, the preposed Kalman filter, the prioritized experience replay, and the long-tail loss optimization were sequentially removed, and the changes in accuracy, latency, and energy consumption were compared. The results are presented in Table 2.

Table 2. Ablation experiment results of the core modules

Model Variant	Top-1 Accuracy (%)	Macro-F1 (%)	Average Latency (ms)	Energy Per Inference (mJ)
Complete model	82.6	79.8	124	1.86
Removal of the modality-level attention	77.3	74.1	118	1.79
Removal of the cross-modal gated fusion unit	76.8	73.5	120	1.81
Removal of the preposed Kalman filter	79.2	76.4	122	1.84
Removal of the prioritized experience replay	79.7	77.1	131	1.92
Removal of the long-tail loss optimization	80.1	75.2	123	1.85

The ablation results demonstrate that all core modules made positive contributions to the final performance. Among these, the modality-level attention and the cross-modal gated fusion unit had the most significant impact on accuracy, with top-1 accuracy decreasing by 5.3 and 5.8 percentage points, respectively, after their removal. This validates that dynamic modality weighting and deep contextual fusion serve as the core pillars for improving intent recognition accuracy. After the removal of the preposed Kalman filter, accuracy decreased by 3.4 percentage points, indicating that location noise suppression effectively enhanced the stability of state inputs, which is crucial for the stable output of the reinforcement learning policy. After the removal of prioritized experience replay, latency and energy consumption increased slightly, while accuracy decreased by 2.9 percentage points, suggesting that this mechanism effectively improved the quality of policy convergence and the decision-making accuracy in rare scenarios. After the removal of long-tail loss optimization, Macro-F1 decreased by 4.6 percentage points, whereas top-1 accuracy decreased by only 2.5 percentage points, further confirming that its primary role lies in improving the overall recognition balance under class imbalance.

3.5 On-device deployment performance measurement experiment

This set of experiments was designed to validate the deployment feasibility of the light-weighting scheme on mobile devices, with the compression effectiveness, runtime efficiency, and accuracy loss of different light-weighting strategies being compared. The original model, pruning-only, distillation-only, and the proposed joint scheme were selected for comparison, and the tests were conducted on mid-range mobile devices. The results are presented in Table 3.

Table 3. On-device deployment performance comparison of different light-weighting schemes

Scheme	Parameters (M)	Operations (G)	Inference Latency (ms)	Energy per Inference (mJ)	Accuracy Degradation Rate (%)
Original model	28.7	4.62	214	3.21	–
Pruning only	15.3	2.35	147	2.24	4.7
Distillation only	26.1	4.18	196	2.93	1.2
Proposed joint scheme	14.8	2.27	124	1.86	2.8

The experimental results demonstrate that the proposed joint light-weighting scheme compressed the model parameter count to 51.6% of the original model, reduced floating-point operations to 49.1%, and decreased the energy consumption per inference by 42.1% compared to the original model, while the accuracy degradation rate was controlled at 2.8%, meeting the design target of within 3%. In comparison with the individual light-weighting methods, pruning only achieved significant compression but incurred substantial accuracy loss, whereas distillation only incurred minimal accuracy loss but offered limited compression efficiency. Through the synergistic mechanism of volume compression via pruning and accuracy compensation via distillation, the proposed joint scheme achieved a superior trade-off between energy efficiency and accuracy. Further tests on three tiers of devices revealed that the per-frame inference latencies on high-end, mid-range, and

low-end devices were 86 ms, 124 ms, and 217 ms, respectively, all of which satisfied the basic requirements for real-time interaction, demonstrating broad device compatibility.

3.6 Real-scenario user validation experiment

This set of experiments was conducted in real scenic area environments to validate the practical application effectiveness and subjective user experience of the proposed method. A total of 60 subjects were recruited and randomly divided into three groups, with each group using the corresponding mobile application to complete a two-hour scenic tour. Objective interaction data were collected, and a 5-point subjective satisfaction questionnaire was administered. The results are presented in Figure 3.

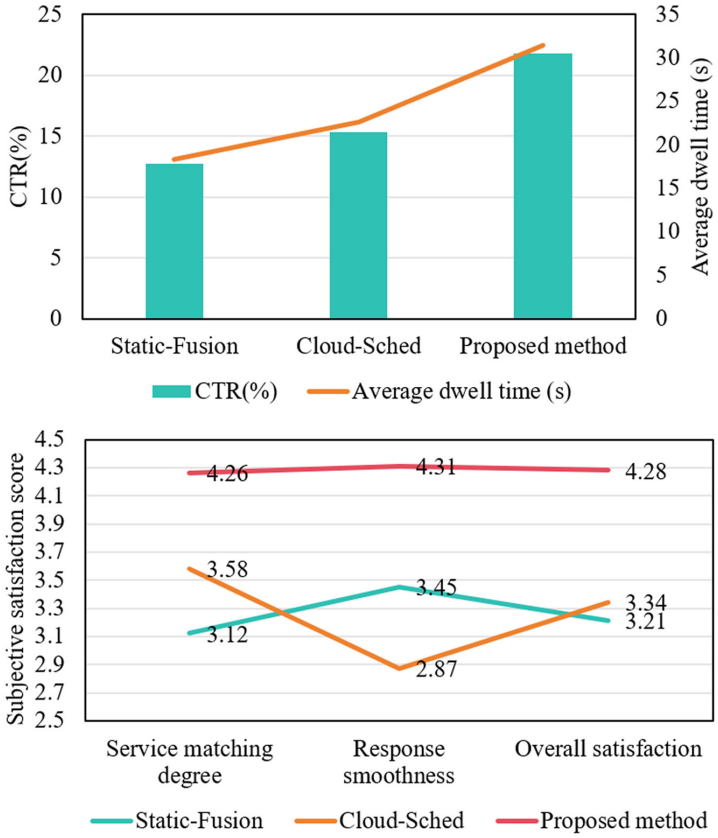

Fig. 3. Comparison of user experiment results in real scenarios

The real-scenario test results demonstrate that the service click-through rate of the proposed method was improved by more than 42.5% compared to the baseline methods, and the average user dwell time was increased by 39.4%. These objective data confirm that the accuracy of service matching and the appeal of the content were significantly superior to those of the comparison schemes. In terms of subjective ratings, the proposed method exhibited its most pronounced advantage in response smoothness, with a score of 4.31, primarily attributed to the low-latency characteristics of on-device inference, which avoided the loading lag issues associated with cloud-based scheduling. Both service-matching degree and overall

satisfaction scores exceeded 4.2, indicating that the personalized recommendations based on multimodal context-aware perception were effective in aligning with tourists' on-site needs. Based on the combined objective metrics and subjective feedback, the proposed method demonstrated strong deployment effectiveness and user acceptance in real complex scenic-area environments.

4 CONCLUSION AND FUTURE DIRECTIONS

To address the core contradiction between the demand for real-time personalized services and terminal resource constraints in mobile tourism scenarios, a three-tier collaborative optimization framework encompassing on-device perception, edge coordination, and dynamic orchestration was constructed, forming a complete technical pipeline that integrates multimodal interactive perception, spatiotemporal context-aware intent decoding, and resource-constrained service orchestration. Within this framework, modality-level attention pooling was employed to achieve adaptive fusion of multi-channel signals, a cross-modal gated fusion unit was leveraged to accomplish semantic alignment between interaction features and physical context, a temporal encoding network was incorporated to capture the dynamic evolutionary patterns of user intent, and service scheduling was formulated as a constrained Markov decision process, through which multi-objective optimal trade-offs between service quality and resource consumption were realized via a deep reinforcement learning algorithm. Multiple comparative experiments and real scenic-area test results demonstrated that the proposed method significantly outperformed mainstream baseline methods in terms of intent recognition accuracy, system response speed, and operational energy efficiency and could be stably and efficiently deployed on mid-range mobile devices, thereby providing an engineering-feasible technical paradigm for real-time personalized optimization of smart tourism services in mobile scenarios.

Future research can be further deepened and extended from multiple dimensions. First, federated learning can be introduced to construct a distributed incremental training mechanism, enabling collaborative model updates across terminals while preserving user privacy and security, thereby enhancing the personalized adaptation capability across different scenarios and user groups. Second, a multi-user collaborative global context-aware mode can be explored, in which aggregated group movement trajectories and interaction data are leveraged to optimize the global scheduling of service resources in scenic areas, improving service provision efficiency under large-scale visitor flow conditions. Third, the multimodal interaction framework can be extended to incorporate emerging interaction modalities such as augmented reality, and spatial perception technologies can be integrated to build a more immersive mobile tourism service experience, driving the evolution of smart tourism toward higher-dimensional natural interaction.

5 REFERENCES

[1] S. Luo, "Rural tourism management cloud service platform based on interactive mobile embedded systems," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 13, pp. 130–147, 2024. https://doi.org/10.3991/ijim.v18i13.49065

[2] H. Jin, "Integration of mobile interaction technology in the tourism industry and its impact on tourism consumption patterns," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 1, pp. 140–154, 2025. https://doi.org/10.3991/ijim.v19i01.53495

[3] Q. V. Pham *et al.*, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020. https://doi.org/10.1109/ACCESS.2020.3001277

[4] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2131–2165, 2021. https://doi.org/10.1109/COMST.2021.3106401

[5] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020. https://doi.org/10.1162/neco_a_01273

[6] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, no. 8, pp. 2939–2970, 2022. https://doi.org/10.1007/s00371-021-02166-7

[7] Z. Wang, W. Höpken, and D. Jannach, "A survey on point-of-interest recommendations leveraging heterogeneous data," *Information Technology & Tourism*, vol. 27, no. 1, pp. 29–73, 2025. https://doi.org/10.1109/TKDE.2025.3551292

[8] J. Chen, W. Jiang, J. Wu, K. Li, and K. Li, "Dynamic personalized POI sequence recommendation with fine-grained contexts," *ACM Transactions on Internet Technology*, vol. 23, no. 2, p. 32, 2023. https://doi.org/10.1145/3583687

[9] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023. https://doi.org/10.1109/TPAMI.2023.3268118

[10] H. T. Malazi *et al.,* "Dynamic service placement in multi-access edge computing: A systematic literature review," *IEEE Access*, vol. 10, pp. 32639–32688, 2022. https://doi.org/10.1109/ACCESS.2022.3160738

[11] J. P. Esper, L. de S. Fraga, A. C. Viana, K. V. Cardoso, and S. L. Correa, "+Tour: Recommending personalized itineraries for smart tourism," *Computer Networks*, vol. 260, p. 111118, 2025. https://doi.org/10.1016/j.comnet.2025.111118

[12] Y. Sahni, J. Cao, L. Yang, and S. Wang, "Distributed resource scheduling in edge computing: Problems, solutions, and opportunities," *Computer Networks*, vol. 219, p. 109430, 2022. https://doi.org/10.1016/j.comnet.2022.109430

[13] Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki, and A. S. Hafid, "A comprehensive survey on TinyML," *IEEE Access*, vol. 11, pp. 96892–96922, 2023. https://doi.org/10.1109/ACCESS.2023.3294111

[14] M. De Lange *et al.*, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022. https://doi.org/10.1109/TPAMI.2021.3057446

6 AUTHOR

Duchuan Zhang holds an MBA from Henan University of Economics and Law and currently serves at Zhengzhou Tourism College. His research focuses on tourism development planning and leisure tourism (E-mail: zhangduchuan@zztrc.edu.cn).