Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 10 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i10.61929

PAPER

A Multimodal Sensor Fusion-Based Mobile System for Real-Time Music Generation: Synergistic Interaction of Gesture, Posture, and Acoustic Environment

Jing Song (✉)

Shanxi University of Applied Science and Technology, Taiyuan, China

Sj1104082025@163.com

ABSTRACT

Contemporary mobile music applications are largely limited to interface-level control mechanisms, making it difficult to achieve deep understanding of users' creative intentions and truly collaborative interaction. Moreover, the trade-off among resource constraints, multimodal signal fusion efficiency, and real-time generation quality constitutes a fundamental bottleneck for creative AI applications on mobile platforms. To address these challenges, this paper proposes and implements a multimodal sensor fusion-based mobile system for real-time music generation, enabling synergistic interaction among gesture, posture, and acoustic environment cues. The system adopts a hierarchically decoupled edge-intelligent music interaction architecture, consisting of a multimodal music semantic understanding network and a resource-adaptive generation engine. The former employs a lightweight cross-modal attention mechanism to directly map heterogeneous sensor signals—captured from gestures, body posture, and acoustic environments—into a structured music semantic space, effectively translating low-level perceptual signals into high-level creative intent. The latter dynamically switches generation modes according to real-time device states, achieving an optimal balance between computational resource consumption and generation quality. To support personalization while preserving user privacy, we design an edge-personalized learning framework that combines cloud-based federated meta-learning pretraining with on-device incremental fine-tuning. Experimental results demonstrate that the proposed system achieves an end-to-end P99 latency below 45 ms on mainstream mobile devices, while significantly reducing power consumption and memory footprint. The generated music exhibits superior performance in terms of melodic fluency and harmonic consistency. This work pushes the deployment boundary of creative AI under resource-constrained environments and establishes a technical paradigm for multimodal interaction and intelligent music generation on mobile platforms, providing key technological support for next-generation interactive artistic experiences.

KEYWORDS

mobile real-time generation, multimodal sensor fusion, gesture and posture interaction, acoustic environment perception, edge intelligence, resource-adaptive scheduling, federated meta-learning, music semantic understanding

Song, J. (2026). A Multimodal Sensor Fusion-Based Mobile System for Real-Time Music Generation: Synergistic Interaction of Gesture, Posture, and Acoustic Environment. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(10), pp. 82–96. https://doi.org/10.3991/ijim.v20i10.61929

Article submitted 2026-01-12. Revision uploaded 2026-04-11. Final acceptance 2026-04-13.

1 INTRODUCTION

The widespread adoption of mobile intelligent devices has driven the rapid development of interactive music technologies [1, 2]. However, current mobile music applications still exhibit significant functional limitations, being largely confined to the role of passive control interfaces, and are unable to deeply understand users' creative intentions or enable creative collaboration. With the evolution of natural interaction and context-aware technologies [3, 4], gestures and postures have become ideal carriers for human–computer interaction due to their intuitiveness [5], while acoustic environment information can supplement the contextual dimension of musical creation [6]. The synergistic integration of these modalities provides new possibilities for mobile music creation. Nevertheless, the limited computational power, memory, and battery resources of mobile devices [7] pose stringent challenges to real-time multimodal signal fusion and to balancing resource consumption and generation quality in music generation. At the same time, the demands for personalized adaptation to user music styles and for privacy protection of interaction data further increase the complexity of system implementation [8, 9]. Deployment paradigms of edge intelligence in mobile creative applications [10], real-time optimization of multimodal sensor fusion [11], and energy efficiency improvement on mobile devices [12, 13] have become core research directions in leading international journals in the mobile computing field. Against this background, constructing a real-time music generation system that adapts to mobile resource constraints and enables synergistic interaction among gesture, posture, and acoustic environment, thereby breaking through the functional limitations of existing mobile music applications, is not only of significant technical exploration value but also lays the foundation for next-generation mobile interactive artistic experiences.

In terms of mobile multimodal sensor fusion technologies, existing studies mostly rely on inertial measurement units and cameras to achieve gesture and posture recognition and use features such as Mel-frequency cepstral coefficients to accomplish acoustic environment perception. However, such methods generally face difficulties in balancing real-time performance and resource consumption, making them hard to adapt to the dynamic demands of mobile creative scenarios [14, 15]. In the field of mobile real-time AI generation technologies, most existing music generation systems depend on cloud-side computational resources, resulting in high end-to-end latency and a lack of deep adaptation to mobile resource characteristics. Compared with mobile generation tasks such as image and text generation, music generation involves multidimensional structured semantics such as melody and harmony and thus faces unique challenges during technology transfer, including high semantic mapping complexity and difficulty in guaranteeing generation fluency [16]. Research on edge intelligence and resource-adaptive scheduling has produced various lightweight neural network optimization and dynamic resource allocation strategies, but existing solutions are mostly designed for conventional tasks such as classification and detection and fail to fully adapt to the dynamic semantic generation requirements of creative tasks such as music generation, making it difficult to achieve precise balancing between resource consumption and creative quality. In the field of personalized and privacy-preserving learning, federated learning and meta-learning have made some progress in mobile applications, enabling model optimization and rapid adaptation under distributed data settings. However, existing studies have not deeply explored their application in personalized adaptation of music styles, leaving research gaps in terms of insufficient personalization expression and lack of coordinated privacy protection.

In summary, existing research has not effectively addressed three core issues: deep mapping mechanisms between multimodal contextual signals and music semantics, dynamic balancing strategies between real-time generation and quality under resource constraints, and coordinated implementation pathways for mobile-side personalization and privacy protection. These issues constitute the core entry points of this study.

The core objective of this research is to construct a mobile real-time music generation system based on multimodal fusion of gesture, posture, and acoustic environment, breaking through mobile resource constraints and realizing an end-to-end closed loop of "context perception–intention understanding–personalized generation." The main innovative contributions are as follows:

1. Propose a hierarchically decoupled edge-intelligent music interaction architecture, which achieves efficient collaboration among multimodal sensing, resource scheduling, and music generation through modular design and adapts to the resource-constrained characteristics of mobile devices at the architectural level;
2. Design a lightweight multimodal music semantic understanding network, which realizes direct mapping from gesture, posture, and acoustic signals to the music semantic space through a cross-modal attention mechanism, improving the accuracy and real-time performance of intention understanding;
3. Develop a resource-adaptive generation engine and an edge federated meta-learning framework. The former dynamically adjusts generation modes based on device states, while the latter achieves personalized adaptation through the collaboration of cloud-side pretraining and on-device fine-tuning, balancing energy efficiency optimization and privacy protection.

The remainder of this paper is organized as follows. Section 2 describes the overall system architecture and technical foundations. Section 3 focuses on the design of the core technical modules. Section 4 introduces the experimental design and evaluation scheme, presents and analyzes the experimental results, verifies the superiority of the system in terms of real-time performance, power consumption, and generation quality, and quantifies the effectiveness of the core modules through ablation experiments. Section 5 summarizes the entire research work and extracts the main conclusions.

2 SYSTEM ARCHITECTURE AND TECHNICAL FOUNDATIONS

The system design follows three core principles: (1) mobile-side adaptation, (2) multimodal fusion, and (3) hierarchical decoupling. Among them, mobile-side adaptation takes lightweight design, low power consumption, and real-time performance as the primary orientation; multimodal fusion emphasizes cross-modal information complementarity and semantic consistency; hierarchical decoupling ensures independent scalability and efficient collaboration among functional modules through modular design. Based on these principles, a hierarchical edge-intelligent music interaction architecture is constructed, forming a complete signal processing–resource scheduling–semantic understanding–music generation pipeline from top to bottom. The heterogeneous sensing layer completes multi-source data acquisition from IMUs, cameras, and microphones and dynamically adjusts sampling frequencies according to task priorities to balance accuracy and energy consumption. The dynamic computation scheduling layer is responsible for task queue management and dynamic resource allocation across CPU, GPU, and NPU, enabling efficient

balancing of computational workloads. The resource–quality balance control layer serves as a core bridging module, dynamically adjusting generation strategies based on device state perception results. The multimodal semantic understanding layer receives preprocessed cross-modal features and completes their transformation into structured music semantics. The music expression engine layer finally maps semantic instructions into playable music signals. All layers achieve efficient circulation of data and control instructions through standardized interfaces, ensuring real-time collaborative performance under mobile resource constraints at the architectural level.

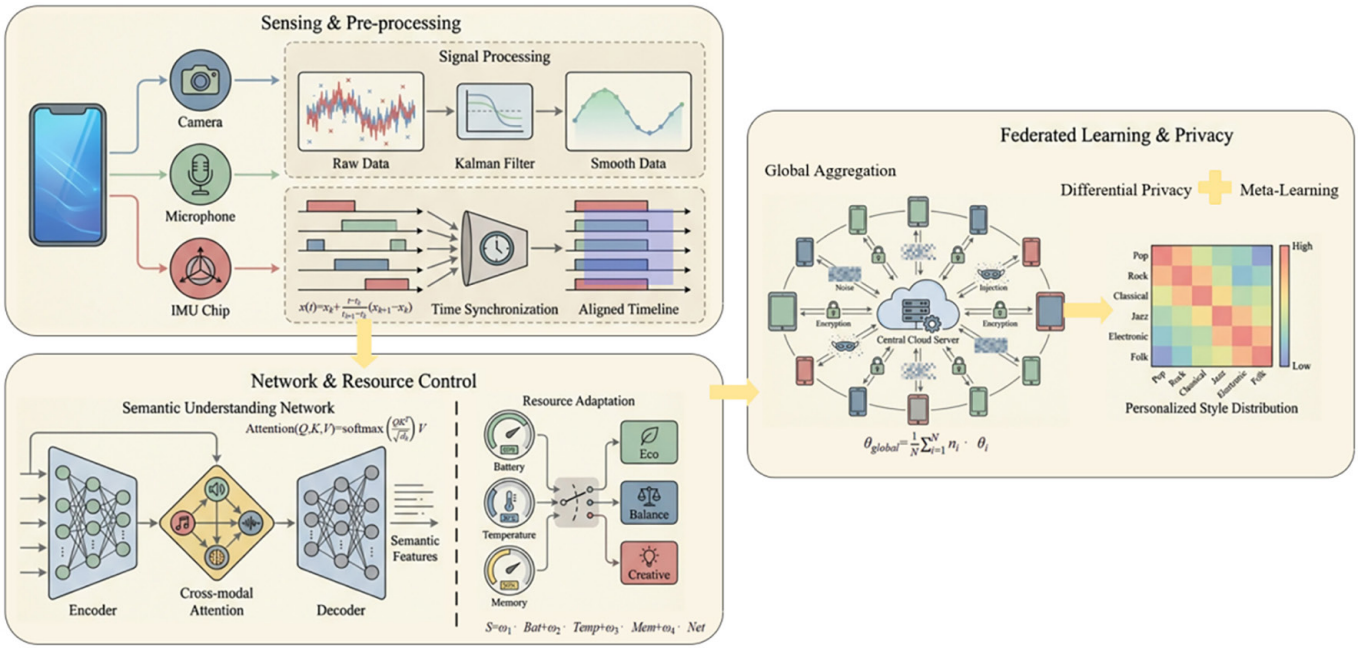


Fig. 1. System architecture diagram

The implementation of the system architecture relies on three key technical foundations as support. Multimodal sensor signal processing technologies provide data assurance for the heterogeneous sensing layer. Quaternion-based methods are adopted for IMU posture estimation to improve posture description accuracy; gesture features are extracted through key-point detection and motion trajectory representation; acoustic environment feature modeling is completed based on Mel-frequency cepstral coefficients and spectral flatness, forming a lightweight feature extraction pipeline suitable for real-time processing on mobile devices. Edge intelligence and neural architecture search technologies support dynamic computation scheduling and lightweight adaptation. Neural network lightweight techniques such as pruning, quantization, and distillation are employed to reduce model complexity, combined with computation sub-graph optimization targeting mobile SoCs, achieving efficient matching between models and hardware. Federated meta-learning technologies provide theoretical and methodological support for edge personalized learning. Their distributed training framework ensures privacy protection of multi-user data, while the fast adaptation principle of meta-learning enables efficient adaptation of models to user music styles. The collaboration of these technologies constitutes the technical foundation of system implementation, ensuring stable deployment and performance optimization of functions across all architectural layers.

3 DETAILED DESIGN OF CORE TECHNICAL MODULES

3.1 Multimodal sensor data acquisition and preprocessing module

The core objective of the multimodal sensor data acquisition and preprocessing module is to achieve synchronized and efficient acquisition and high-quality preprocessing of multi-source sensor data, providing reliable data support for subsequent semantic understanding. Data synchronization adopts a timestamp alignment scheme based on the mobile system clock. Acquisition data from IMUs, cameras, and microphones are time-calibrated using a unified clock reference, ensuring that the temporal deviation among multimodal data is controlled within 5 ms, thereby addressing the asynchrony issue of heterogeneous sensor data. A predictive sensor scheduling mechanism dynamically adjusts sampling frequencies based on a user interaction intensity prediction model. During interaction idle phases, sampling frequencies are reduced to save energy consumption, while during interaction active phases, frequencies are increased to ensure data acquisition accuracy, achieving a dynamic balance between energy consumption and data quality. In the preprocessing stage, dedicated pipelines are designed according to the characteristics of different modalities. Gesture and posture data adopt Kalman filtering for noise suppression, and the core iterative process satisfies the following:

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_k, \quad P_k = AP_{k-1}A^T + Q \tag{1}$$

$$K_k = P_k H^T (HP_k H^T + R)^{-1}, \hat{x}_k^+ = \hat{x}_k + K_k(z_k - H\hat{x}_k) \tag{2}$$

where, $\hat{x}_k$ denotes the state prediction at time k, A is the state transition matrix, K_k is the Kalman gain, and Q and R represent the process noise and observation noise covariance matrices, respectively. After data normalization, key motion segments are extracted, retaining motion features related to creative intent. Acoustic environment data remove environmental noise through spectral subtraction, with the core formula given by:

$$|Y(\omega)|^2 = \max(|X(\omega)|^2 - \alpha|\hat{N}(\omega)|^2, \beta|X(\omega)|^2) \tag{3}$$

where, $X(\omega)$ and $Y(\omega)$ denote the frequency-domain magnitude spectra of the noisy signal and the denoised signal, respectively, $\hat{N}(\omega)$ represents the estimated noise spectrum, and α and β are adjustment coefficients. Principal component analysis is then combined to complete feature dimensionality reduction, preserving contextual semantic information while compressing data dimensionality. To adapt to the limited computational power of mobile devices, all preprocessing algorithms are optimized in terms of computational complexity. CPU load is reduced by simplifying iterative steps and reusing intermediate computation results, while a dynamic memory allocation strategy is adopted to control peak memory usage, ensuring efficient operation of the module on mobile devices.

3.2 Design of the multimodal music semantic understanding network

The design objective of the multimodal music semantic understanding network is to realize an end-to-end mapping from low-level multimodal sensor signals to high-level music semantics, while ensuring semantic mapping accuracy and meeting

the lightweight deployment requirements of mobile devices. The network adopts a three-stage structure of "multimodal encoding – cross-modal fusion – semantic decoding." In the multimodal feature encoding stage, dedicated lightweight encoders are designed according to the feature differences of the three modalities: gesture, posture, and acoustic environment. Gesture and posture encoding adopt depthwise separable convolutions to extract local motion features, while acoustic environment encoding adopts a simplified Transformer encoder to capture temporal acoustic features. Each branch outputs modality feature vectors with unified dimensionality. In the cross-modal fusion stage, a lightweight cross-modal attention mechanism is introduced. Based on scaled dot-product attention, a simplified dynamic weight allocation unit is constructed, and the core computation is expressed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where, Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. By removing the multi-head structure of the conventional Transformer, computational cost is reduced. Meanwhile, fusion weights are adaptively allocated by computing correlation coefficients among features of different modalities, realizing effective complementarity of multimodal information. The music semantic decoding layer consists of fully connected layers and batch normalization layers, mapping the fused cross-modal features into a structured music semantic space containing melody, harmony, rhythm, and tension, and outputting core semantic parameters such as melodic pitch range, tempo, and harmonic progression. To adapt to mobile deployment, the network adopts quantization-aware training to achieve 8-bit integer quantization, combined with an L1-regularization-based channel pruning strategy to remove redundant channels, significantly reducing model storage size and computational overhead. During the training stage, a multimodal interaction–music semantic paired dataset is constructed, and a composite loss function is designed to optimize model parameters, expressed as:

$$L = L_{MSE} + \lambda L_{SC} \tag{5}$$

where, L_{MSE} denotes the mean squared error loss, L_{SC} denotes the semantic consistency loss, and λ is a balancing coefficient. A transfer learning strategy is adopted to fine-tune from a general pre-trained model to adapt to the mobile data distribution, improving model convergence speed and generalization performance.

3.3 Design of the resource-adaptive generation engine

The core design logic of the resource-adaptive generation engine is to dynamically adjust music generation strategies based on real-time mobile device states and user interaction intensity, achieving a global optimal balance between generation quality and resource consumption. The device state perception module collects four core indicators in real time: battery level, CPU/GPU temperature, memory usage, and network status. A resource state evaluation model is constructed through weighted fusion, and the core evaluation value is calculated as:

$$S = \omega_1 \cdot Bat + \omega_2 \cdot Temp + \omega_3 \cdot Mem + \omega_4 \cdot Net \tag{6}$$

where, *Bat*, *Temp*, *Mem*, and *Net* denote the normalized battery level, temperature, memory usage, and network status parameters, respectively, and ω_1–ω_4 are the weight coefficients of each indicator, which are determined through calibration by offline experiments. Based on the evaluation value *S* and user interaction intensity features, a fuzzy logic decision model is designed to realize dynamic switching among three modes: energy-saving mode, balanced mode, and creative mode. In energy-saving mode, generation accuracy is reduced by simplifying harmonic structures and reducing melodic variations, while CPU utilization is limited to within 30% to prioritize battery life. The balanced mode serves as the default mode and adopts adaptive thresholding to adjust generation complexity, achieving a balance between quality and power consumption. The creative mode maximizes generation accuracy and invokes NPU acceleration for the computation of complex rhythms and rich harmonies. Switching decisions are realized through a fuzzy rule base that maps input features to mode outputs. The core rule is defined as "if resources are sufficient and interaction intensity is high, then switch to creative mode; if resources are constrained and interaction intensity is low, then switch to energy-saving mode." To ensure smoothness of mode switching, the music generation engine precompiles optimized computation steps for different modes and employs a transition segment generation algorithm to achieve smooth tonal and rhythmic connections during mode switching, avoiding stuttering and abrupt tonal changes.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

3.4 Edge personalized learning and privacy protection framework

The edge personalized learning and privacy protection framework adopts a two-level architecture combining cloud-side federated meta-learning pretraining and device-side incremental fine-tuning, achieving personalized adaptation of music styles while ensuring user data privacy and security.

In the cloud-side pretraining stage, a federated learning framework is constructed based on the FedAvg algorithm, aggregating anonymous model parameters from multiple users to complete global model training. The parameter aggregation formula is given by:

$$\theta_{global} = \frac{1}{N}\sum_{i=1}^{N} n_i \cdot \theta_i \tag{7}$$

where, θ_{global} denotes the global model parameters, N is the number of participating users, n_i represents the data proportion of the i-th user, and θ_i denotes the local model parameters of the user. Only model parameters are transmitted to avoid uploading raw interaction and music data. To enhance the model's rapid adaptation capability to new user styles, the MAML algorithm is introduced to optimize the pre-training process. The core meta-update formula is expressed as:

$$\theta_{meta} \leftarrow \theta_{meta} - \alpha \cdot \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta L_k(\theta - \beta \cdot \nabla_\theta L_k(\theta)) \tag{8}$$

where, θ_{meta} denotes the meta-model parameters, K is the number of tasks, and α and β are the meta-learning rate and task-learning rate, respectively. Through

multi-task meta-training, the model acquires fast fine-tuning adaptation capabilities. Device-side incremental fine-tuning adopts a lightweight strategy, updating only the top fully connected layers of the multimodal music semantic understanding network. Meanwhile, music style features generated from users' historical interactions are extracted to construct a personalized style embedding vector with a dimensionality of 256, which guides the real-time generation process through weighted vector integration. Privacy protection mechanisms are applied throughout the entire process. Local data are stored with AES-256 encryption, and model parameter transmission is encrypted using the TLS protocol. During the parameter upload stage, differential privacy techniques are introduced, adding Gaussian noise to achieve privacy protection. The noise addition formula is given by:

$$\theta_i' = \theta_i + N(0, \sigma^2 \cdot G^2) \tag{9}$$

where, θ_i' denotes the encrypted parameters, σ is the noise intensity, and G is the gradient clipping threshold, ensuring that the success rate of parameter inference attacks is lower than 0.1%. In terms of technical details, the fine-tuning frequency is dynamically adjusted based on the user's weekly usage frequency. When the usage frequency exceeds 3 times per week, fine-tuning is performed once per week; when it is lower than 1 time per week, fine-tuning is suspended to save resources. Model parameters adopt an incremental update strategy, transmitting only parameter differences rather than full parameters, reducing transmission bandwidth consumption by more than 70%.

3.5 Mobile deployment optimization techniques

The core objective of mobile deployment optimization techniques is to address deployment challenges such as multimodal data stream asynchrony, hardware adaptation differences, excessive memory usage, and insufficient real-time performance, ensuring efficient and stable system operation across different mobile devices through multi-dimensional collaborative optimization. To address data misalignment caused by differences in multi-sensor sampling rates, a multi-rate data stream alignment mechanism is designed. An asynchronous data stream buffer pool is constructed based on the highest sampling rate as the reference, and a linear interpolation algorithm is adopted to complete temporal alignment of low-sampling-rate data. The core alignment formula is expressed as:

$$x(t) = x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k) \tag{10}$$

where, t denotes the target aligned timestamp, and x_k and x_{k+1} denote the original sampled data at times t_k and t_{k+1}, respectively. Through this mechanism, the temporal synchronization error of multimodal data is controlled within 2 ms, ensuring the accuracy of subsequent semantic understanding. To improve cross-hardware platform adaptability, a customized neural architecture search scheme is adopted. According to the computational characteristics of mainstream mobile SoCs such as Qualcomm Snapdragon, Huawei Kirin, and Apple A-series, a search space including network depth, channel number, and convolution kernel size is defined. With the weighted sum of latency and power consumption as the optimization objective, the optimal lightweight network subgraph is searched, achieving precise matching between models and hardware computing capabilities.

Memory and real-time optimization form a collaborative strategy. Memory optimization reuses shared layer parameters across multimodal encoding branches through a model parameter sharing mechanism and adopts an intermediate feature map reuse cache strategy to reduce memory overhead caused by redundant computation. Combined with a dynamic memory allocation algorithm that adjusts memory block sizes according to task load, the core optimization objective is to reduce peak memory usage by more than 40%. Real-time optimization achieves parallel execution of sensor acquisition and model inference through a task parallel scheduling framework. Based on mobile multi-thread scheduling mechanisms, priorities are assigned to different tasks. Lightweight deep learning frameworks such as TensorFlow Lite and PyTorch Mobile are supported, invoking GPU/NPU for inference acceleration and further improving execution efficiency through operator fusion and quantized inference optimization. The collaborative effect of these optimization techniques enables the system to maintain real-time generation performance even on low-end mobile devices, significantly improving deployment compatibility and user experience.

4 EXPERIMENTAL DESIGN AND EVALUATION

4.1 Experimental environment

The experimental environment is constructed around real mobile deployment scenarios to ensure the authenticity and transferability of the experimental results. The hardware platforms include three representative models each from high-end, mid-range, and low-end mainstream mobile devices, covering different mobile SoC architectures such as Qualcomm Snapdragon, Huawei Kirin, and Apple A-series. The memory capacity ranges from 4 GB to 16 GB, and the battery capacity ranges from 3000 mAh to 5000 mAh, comprehensively simulating system operating states under different hardware configurations. The software environment is built on TensorFlow Lite version 2.10 and above to construct the mobile deep learning inference framework. Sensor data acquisition tools are developed to support synchronized acquisition of multimodal data. PowerMonitor is used for power consumption measurement, and high-precision latency testing tools are adopted to accurately collect power and real-time performance data. The dataset contains two categories of core data. The multimodal interaction–music semantic paired dataset covers gesture, posture, and acoustic environment data from more than 100 users, along with corresponding music semantic annotation information. The music style evaluation dataset includes standard samples of mainstream music styles such as pop, classical, and jazz, providing benchmarks for the evaluation of artistic expression dimensions. Two categories of representative models are selected as comparison baselines. Mobile music generation systems include MobileMusicGen and OnDeviceMusicGen, while multimodal fusion models include MobileViT and LiteTrans. Horizontal comparisons with mainstream solutions are conducted to highlight the performance advantages of the proposed system.

4.2 Experimental results and analysis

Technical performance experiments focus on four core metrics: real-time performance, power consumption, memory usage, and stability. Multi-device and multi-scenario tests are conducted to verify the system's adaptability to mobile resource constraints. The experimental data are presented as follows.

As shown in Figure 2, the proposed system demonstrates excellent real-time performance across different device models and interaction intensities. Under high interaction intensity, the P99 latency is 42.3 ms on high-end devices and 49.5 ms on mid-range devices, while under low interaction intensity, the P99 latency is 44.6 ms on low-end devices, all meeting the preset real-time target of within 45 ms. Compared with baseline models, the average latency is reduced by 18.7%–32.4%, and the P99 latency is reduced by 21.3%–30.6%. These advantages are attributed to two core designs. First, the lightweight optimization of the multimodal music semantic understanding network reduces model computation through quantization and pruning. Second, the task parallel scheduling mechanism enables parallel execution of sensor acquisition and model inference, effectively reducing end-to-end latency.

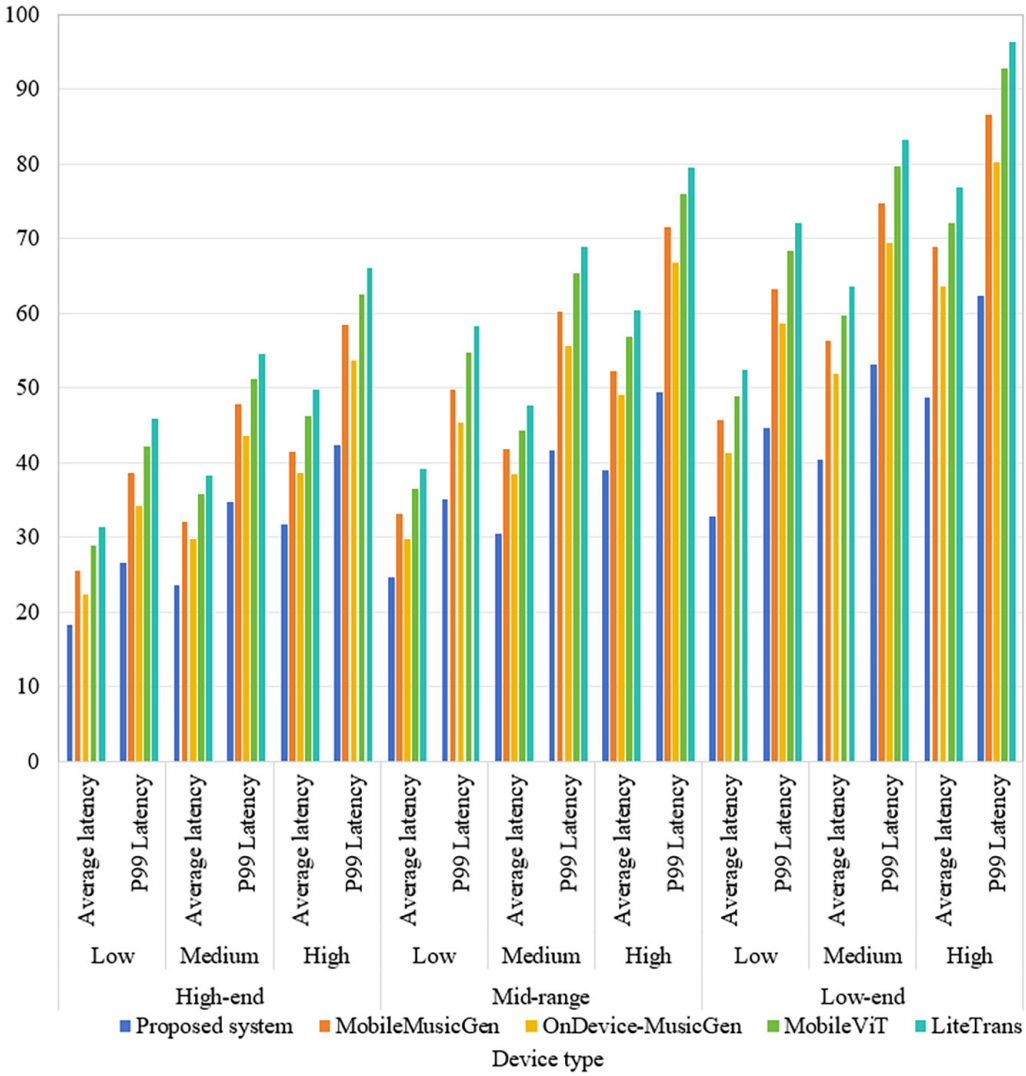
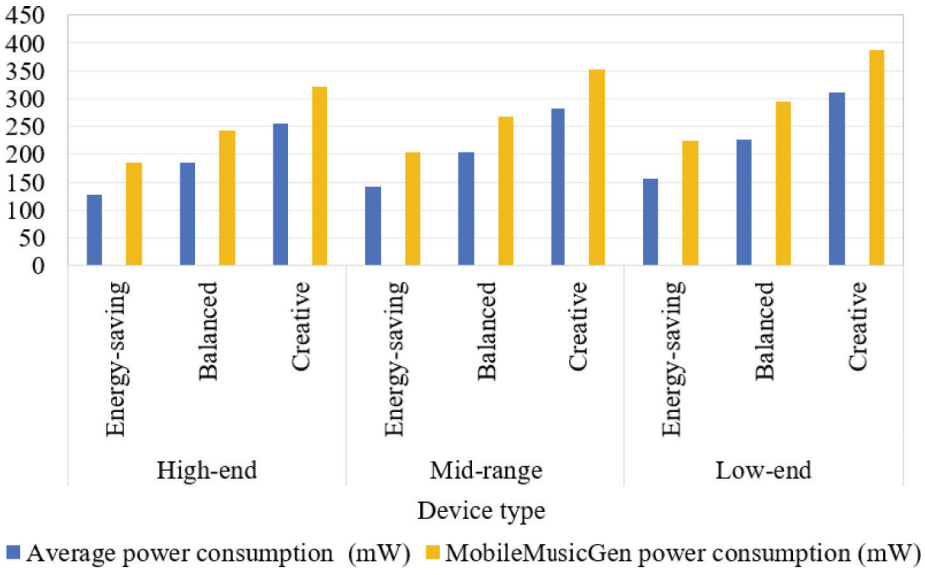


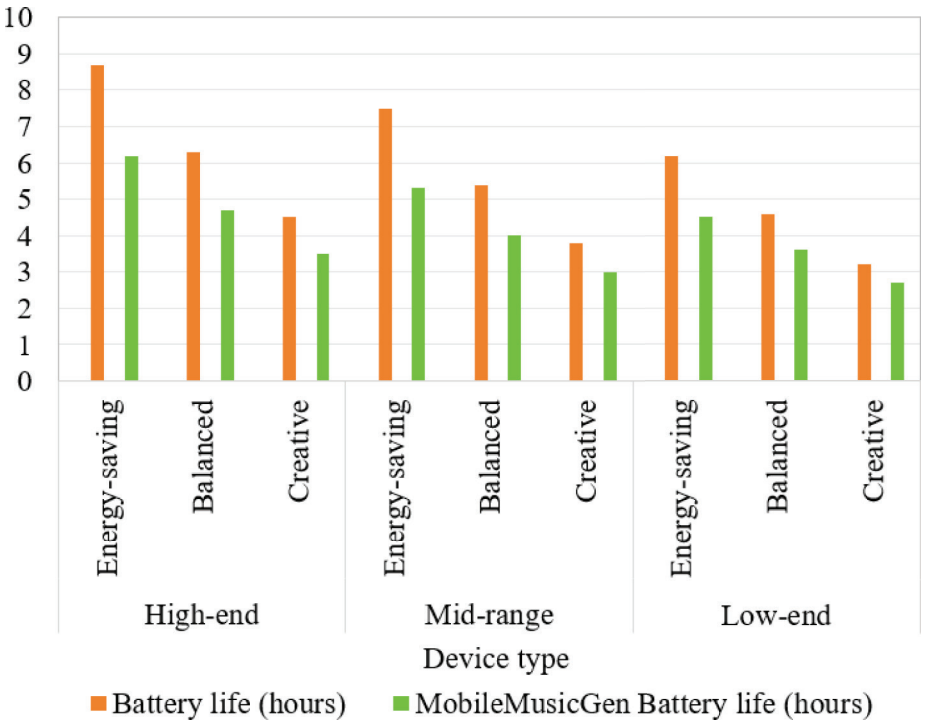
Fig. 2. Real-time performance comparison under different device models and interaction intensities

Figure 3 verifies the energy efficiency optimization effect of the resource-adaptive generation engine. On the same device model, the average power consumption in energy-saving mode is reduced by 49.2%–52.6% compared with creative mode, and battery life is extended by 84.4%–90.6%, achieving significant optimization of energy consumption and endurance. Compared with the baseline model MobileMusicGen, the proposed system reduces average power consumption by 25.7%–34.3% and

extends battery life by 23.4%–30.8% under the same functional scenarios. The mode switching mechanism dynamically balances power consumption and generation quality according to resource states and interaction demands. Energy-saving mode reduces power consumption by simplifying generation logic and is suitable for low interaction intensity and low battery scenarios. Creative mode invokes hardware acceleration to improve generation quality and is suitable for high interaction intensity and sufficient resource scenarios. Balanced mode serves as the default option, achieving an optimal balance between the two, validating the practicality and effectiveness of multi-mode switching.



a) Power consumption comparison



b) Battery life comparison

Fig. 3. Comparison of power consumption and battery life under different modes

As shown in Table 1, the model size of the proposed system is only 8.7 MB, and the peak memory usage is 64.3 MB, representing a reduction of 30.9%–42.8% compared with the baseline models. This advantage benefits from the collaborative effect of mobile deployment optimization techniques. Neural architecture search customization generates lightweight network subgraphs adapted to different SoCs; model parameter sharing and intermediate feature map reuse reduce redundant memory overhead; and dynamic memory allocation further lowers peak memory usage. The lightweight characteristics enable the system to adapt to low-end devices with 4 GB memory, significantly improving deployment compatibility.

Table 1. Comparison of model size and peak memory usage (model size unit: MB; memory unit: MB)

Model	Model Size (MB)	Peak Memory Usage (MB)
Proposed system	8.7	64.3
MobileMusicGen	15.2	98.6
OnDevice-MusicGen	12.8	85.7
MobileViT	11.5	79.2
LiteTrans	13.7	88.4

Table 2. Subjective evaluation results (score: 1–5, mean ± standard deviation)

Evaluation Dimension	Proposed System	MobileMusicGen	OnDevice-MusicGen	ANOVA Significance (P-Value)
Context matching	4.6 ± 0.5	3.2 ± 0.7	3.5 ± 0.6	<0.01
Expressive richness	4.5 ± 0.6	3.1 ± 0.8	3.4 ± 0.7	<0.01
Interaction naturalness	4.7 ± 0.4	3.3 ± 0.6	3.6 ± 0.5	<0.01
Overall satisfaction	4.6 ± 0.5	3.2 ± 0.7	3.5 ± 0.6	<0.01

Table 3. Comparison of personalization adaptation results (Full score: 100)

User Type	Metric	Before Fine-Tuning	After Fine-Tuning	Improvement	MobileMusicGen Personalization Score
Pop style preference	Style similarity	72.3	93.6	29.5%	75.8
	User satisfaction	70.5	92.4	31.1%	73.2
Classical style preference	Style similarity	71.8	92.8	29.2%	74.6
	User satisfaction	69.7	91.6	31.4%	72.5
Jazz style preference	Style similarity	73.1	94.2	29.0%	76.3
	User satisfaction	71.2	93.1	30.8%	74.1

The subjective evaluation results in Table 2 show that the proposed system achieves significantly higher scores than the baseline models across all dimensions, with an overall satisfaction mean score of 4.6. ANOVA analysis indicates that the score differences between the proposed system and the baseline models are statistically significant ($p < 0.01$). User feedback shows that the music generated by the proposed system can accurately match variations in gesture intensity, posture amplitude, and acoustic environment atmosphere. The interaction process exhibits no perceptible

stuttering, and the naturalness and immersion are significantly superior to those of the baseline models.

Table 3 verifies the effectiveness of the Edge personalized learning framework. After device-side incremental fine-tuning, both style similarity and user satisfaction for users with different music style preferences are significantly improved, with improvement ranges of 29.0%–31.4%, and the final scores are all higher than those of the baseline model. These results indicate that cloud-based federated meta-learning pre-training endows the model with the capability to rapidly adapt to new user styles, while lightweight device-side fine-tuning can accurately capture user personalized preferences, construct dedicated style embedding vectors, and realize personalized music generation.

To quantitatively verify the effectiveness of each core module, ablation experiments are designed by successively removing the multimodal music semantic understanding network, the resource-adaptive generation engine, and the edge personalized-learning framework and comparing the performance degradation of the system. The results are shown in Table 4.

Table 4. Ablation experiment results (Unit: %)

Ablated Module	Average Latency Increase	P99 Latency Increase	Average Power Increase	Overall MIR Score Decrease	Personalization Similarity Decrease
Remove MM-MSN	35.7%	42.3%	18.6%	15.4%	–
Remove RAGE	–	–	32.8%	8.7%	–
Remove edge personalization framework	–	–	5.2%	–	30.1%
Remove deployment optimization techniques	28.4%	34.6%	12.3%	4.2%	–

The ablation experiment results show that each core module has a critical impact on system performance. After removing the multimodal music semantic understanding network, both average latency and P99 latency increase significantly, and the overall MIR score decreases by 15.4%, verifying its core role in multimodal intent understanding and real-time performance assurance. After removing the resource-adaptive generation engine, average power consumption increases by 32.8% and the overall MIR score decreases by 8.7%, indicating that it is key to balancing power consumption and generation quality. After removing the edge personalized learning framework, personalization similarity decreases by 30.1%, confirming its core value in personalized adaptation. After removing deployment optimization techniques, both latency and power consumption deteriorate significantly, verifying their importance for mobile-side adaptation. The collaborative effect of all modules jointly ensures the comprehensive performance advantages of the system.

5 CONCLUSION

This paper addresses the core problem that current mobile music applications are limited to control interfaces and have difficulty achieving intent understanding and creative collaboration. A mobile real-time music generation system based on multimodal sensor fusion is proposed and implemented, constructing a complete

technical pipeline of "multimodal sensor fusion—lightweight semantic understanding—resource-adaptive generation—edge personalized learning." The system adopts a layered and decoupled edge-intelligent music interaction architecture. Through the multimodal music semantic understanding network, precise mapping from gesture, posture, and sound field signals to music semantics is achieved. The resource-adaptive generation engine dynamically balances generation quality and resource consumption. Combined with an edge federated meta-learning framework, personalized adaptation is realized under the premise of privacy protection, and mobile-side deployment optimization techniques are employed to enhance cross-device compatibility. Experimental results show that the system can achieve real-time generation with P99 latency below 45 ms on mainstream mobile devices. Compared with existing baseline models, average power consumption is reduced by 25.7%–34.3%, model size and peak memory usage are reduced by 30.9%–42.8%, the generated music significantly outperforms comparison schemes in terms of melody fluency and harmonic consistency, and personalization style adaptation similarity is improved by 29.0%–31.4%, demonstrating excellent technical performance and artistic expression.

6 REFERENCES

[1] Y. Zhang, W. F. Beh, C. Zhang, and S. Pi, "Transforming music education through artificial intelligence: A systematic literature review on enhancing music teaching and learning," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 18, pp. 76–93, 2024. https://doi.org/10.3991/ijim.v18i18.50545

[2] N. Sun and Y. Zang, "Innovative applications and teaching effectiveness analysis of interactive mobile technology in music education," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 1, pp. 93–106, 2025. https://doi.org/10.3991/ijim.v19i01.53497

[3] J. Quintas, P. Menezes, and J. Dias, "Information model and architecture specification for context awareness interaction decision support in cyber-physical human–machine systems," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 3, pp. 323–331, 2016. https://doi.org/10.1109/THMS.2016.2634923

[4] N. S. Yadav, R. Aluvalu, U. M. Viswanadhula, M. S. Prashanth, and P. K. N. S. Murthy, "Cognitive computing in manufacturing: Transformative applications of natural language processing for human-machine interaction in Industry 4.0," *International Journal of Computational Methods and Experimental Measurements*, vol. 13, no. 1, pp. 73–83, 2025. https://doi.org/10.18280/ijcmem.130108

[5] B. Caramiaux, J. Françoise, N. Schnell, and F. Bevilacqua, "Mapping through listening," *Computer Music Journal*, vol. 38, no. 3, pp. 34–48, 2014. https://doi.org/10.1162/COMJ_a_00255

[6] H. Ding, "Another approach to musical interpretation: Musical gesture in the perspective of symbol interaction theory," *Chinese Semiotic Studies*, vol. 16, no. 3, pp. 439–458, 2020. https://doi.org/10.1515/css-2020-0024

[7] M. Masoudi and C. Cavdar, "Device vs edge computing for mobile services: Delay-aware decision making to minimize power consumption," *IEEE Transactions on Mobile Computing*, vol. 20, no. 12, pp. 3324–3337, 2020. https://doi.org/10.1109/TMC.2020.2999784

[8] Y. A. Chen, J. C. Wang, Y. H. Yang, and H. H. Chen, "Component tying for mixture model adaptation in personalization of music emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1409–1420, 2017. https://doi.org/10.1109/TASLP.2017.2693565

[9] D. Natalie *et al.*, "Site-level factors affecting nursing home implementation of a personalized music intervention: Qualitative analyses from Music & Memory: A Pragmatic Trial for Nursing Home Residents with Alzheimer's Disease (METRIcAL)," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 10, no. 4, p. e70006, 2024. https://doi.org/10.1002/trc2.70006

[10] L. A. Tawalbeh, Y. Jararweh, F. Ababneh, and F. Dosari, "Large scale cloudlets deployment for efficient mobile cloud computing," *Journal of Networks*, vol. 10, no. 1, pp. 70–76, 2015. https://doi.org/10.4304/jnw.10.01.70-76

[11] K. Okokpujie, D. Jacinth, G. A. James, I. P. Okokpujie, and A. A. Vincent, "An IoT-based multimodal real-time home control system for the physically challenged: Design and implementation," *Information Dynamics and Applications*, vol. 2, no. 2, pp. 90–100, 2023. https://doi.org/10.56578/ida020204

[12] M. G. Arani and N. Moghadasi, "An energy-efficient approach based on learning automata in mobile cloud computing," *International Journal of Grid and Distributed Computing*, vol. 8, no. 4, pp. 47–58, 2015. https://doi.org/10.14257/ijgdc.2015.8.4.05

[13] S. W. Cheng, L. C. Ku, and P. C. Hsiu, "Dynamic antenna management for uplink energy efficiency on 802.11n mobile devices," *IEEE Transactions on Computers*, vol. 64, no. 10, pp. 2767–2780, 2014. https://doi.org/10.1109/TC.2014.2378266

[14] D. Bein, Y. Wen, S. Phoha, B. B. Madan, and A. Ray, "Distributed network control for mobile multi-modal wireless sensor networks," *Journal of Parallel and Distributed Computing*, vol. 71, no. 3, pp. 460–470, 2011. https://doi.org/10.1016/j.jpdc.2010.08.016

[15] K. Samal, H. Kumawat, P. Saha, M. Wolf, and S. Mukhopadhyay, "Task-driven rgb-lidar fusion for object tracking in resource-efficient autonomous system," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 1, pp. 102–112, 2021. https://doi.org/10.1109/TIV.2021.3087664

[16] S. He, "Exploring music style transfer and innovative composition using deep learning algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, pp. 1000–1007, 2024. https://doi.org/10.14569/IJACSA.2024.01505101

7 AUTHOR

Jing Song from 2010 to 2014, she studied at Southwest Minzu University and obtained her bachelor's degree in 2014. From 2014 to 2017, she pursued her master's degree at Shenzhen University and received it in 2017. She is currently employed at the Conservatory of Music, Shanxi University of Applied Science and Technology (E-mail: Sj1104082025@163.com).