Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 8 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i08.61247

PAPER

A Deep Learning–Driven Mobile Interaction System for Industrial Process Design: An Automotive Industry Case Study

Chutong Liu[1](✉), **Qingxin Guo**[2]

[1]Guangdong University of Technology, Guangzhou, China

[2]Hubei Polytechnic University, Huangshi, China

1112217002@mail2.gdut.edu.cn

ABSTRACT

To address the demands of intelligent transformation in the automotive industry under the paradigm of Industry 4.0 and to overcome limitations associated with experience-dependent process design, prolonged iteration cycles, and insufficient mobile interaction capabilities, a deep learning–driven mobile interaction system based on cloud–edge–end collaboration was proposed for automotive process design scenarios. A hierarchical collaborative architecture is established, in which the cloud layer provides knowledge support and computational assurance through a process knowledge graph and data augmentation based on artificial intelligence generated content (AIGC), the edge layer achieves a balance between responsiveness and efficiency via dynamic model scheduling and multi-terminal collaboration, and the mobile layer enhances interaction quality through adaptive rendering and multimodal fusion techniques. To address key technical bottlenecks, a two-stage lightweight model generation framework, an on-device multimodal scene understanding method, and a real-time simulation optimization scheme driven by Physics-informed neural networks (PINNs) were designed, forming a closed-loop mobile design interaction workflow. The effectiveness of the proposed system was validated using the joining process of automotive door outer panels as a representative case, supported by multi-dimensional comparative experiments and ablation studies. Through the deep integration of deep learning and industrial mobile interaction, the proposed approach enables a paradigm shift in automotive process design from tool-assisted operation toward intelligent co-creation, providing both technical support and practical guidance for mobile intelligent design in industrial applications.

KEYWORDS

automotive process design, deep learning, cloud–edge–end collaboration, mobile interaction, Industry 4.0, real-time simulation

Liu, C., Guo, Q. (2026). A Deep Learning–Driven Mobile Interaction System for Industrial Process Design: An Automotive Industry Case Study. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(8), pp. 172–187. https://doi.org/10.3991/ijim.v20i08.61247

Article submitted 2026-01-06. Revision uploaded 2026-02-18. Final acceptance 2026-02-20.

1 INTRODUCTION

The ongoing wave of Industry 4.0 has accelerated the transformation of the automotive industry toward intelligent and flexible manufacturing paradigms [1, 2]. As a core component of automotive production, process design directly determines both product quality and manufacturing efficiency [3–5]. Conventional automotive process design remains highly dependent on expert experience, while rigid and standardized design workflows result in prolonged iteration cycles [6]. In addition, inefficient data circulation across departments has been shown to constrain collaborative efficiency [7], making it increasingly difficult to accommodate the growing demand for product customization and rapid iteration in contemporary automotive development. In parallel, the widespread adoption of mobile computing technologies and their deepening integration into industrial scenarios [8–10] have introduced new opportunities for process design. Mobile terminals such as industrial tablets and augmented reality (AR) devices, enabled by their inherent portability [11], allow designers to operate directly within production workshops and prototype testing environments [12]. Such capabilities facilitate immediate linkage between design activities and manufacturing contexts, intensifying the demand for mobile, context-aware, and real-time interactive design tools. However, existing desktop-based process design tools are poorly suited to the portability constraints of mobile scenarios, while current mobile tools lack the intelligent capabilities required for core process design tasks. As a result, a widening gap has emerged between them.

Existing research related to the intelligence and mobility of automotive process design exhibits a pronounced dualistic trend, with effective integrative solutions remaining largely unexplored. On the one hand, desktop-based studies have primarily focused on the development of intelligent plug-ins for Computer-Aided X (CAX)/Computer-Aided Design (CAD) tools, where machine learning algorithms have been introduced to support process parameter recommendation and design scheme validation. Although such approaches can enhance desktop design efficiency, their extension to production sites remains constrained by hardware form factors, limiting real-time interaction between design activities and on-site operating conditions [13, 14]. On the other hand, mobile-oriented research has predominantly concentrated on user interface optimization and the implementation of basic interaction functionalities [15], with the primary objective of improving operational convenience [16, 17]. The generative and inferential capabilities afforded by deep learning are largely absent, rendering such systems incapable of supporting critical tasks such as intelligent process scheme generation and simulation-based verification. Overall, the powerful intelligence of deep learning has yet to be deeply integrated with the situational awareness and real-time interaction characteristics of mobile computing. Moreover, a systematic framework and corresponding technical solutions that simultaneously achieve intelligence and robust mobile adaptability—particularly for core automotive process design workflows—remain insufficiently addressed.

To address the aforementioned research gaps, a deep learning–driven mobile interaction system based on cloud–edge–end collaboration was proposed, specifically tailored for automotive process design scenarios. The core innovations are embodied across three dimensions. The first dimension concerns architectural innovation. The conventional cloud–end binary computing paradigm is transcended

through the introduction of edge computing nodes, forming a three-layer collaborative architecture. Through task scheduling and computational buffering at the edge layer, a dynamic balance is achieved between the limited computational resources of mobile devices and the real-time performance requirements of process design tasks. The second dimension focuses on interaction innovation. A multimodal fusion interaction mechanism is established, in which speech, touch, and gesture inputs are deeply integrated to accommodate the operational characteristics of mobile environments. Naturalistic and on-site interaction for process design is thereby enabled, significantly enhancing both the intuitiveness and efficiency of design operations. The third dimension addresses closed-loop innovation. A real-time closed-loop workflow encompassing on-device interaction, edge-side inference, and cloud-based simulation is constructed. Design requirements are initiated at the mobile end, lightweight model inference and data preprocessing are performed at the edge, and large-scale computation together with high-fidelity simulation support is provided by the cloud. Through this process, intelligent process scheme generation, on-site adjustment, simulation-based validation, and iterative optimization are achieved in an integrated manner. The proposed system is intended to bridge existing research gaps by facilitating a paradigm shift in automotive process design from traditional tool-assisted workflows toward intelligent co-creation while providing technical support for mobile intelligent design in industrial applications.

2 SYSTEM ARCHITECTURE DESIGN

A three-layer collaborative architecture based on cloud–edge–end collaboration is designed, in which functional boundaries and data flow mechanisms across layers are explicitly defined. The primary objective is to enable deep learning–driven, high–real-time process design interaction on resource-constrained mobile devices, thereby addressing the fundamental tension between insufficient computational supply and excessive interaction latency inherent in conventional architectures. Within this architecture, data circulation is treated as the core organizing principle. Design requirements are initiated at the mobile end, where on-site interaction is performed; task scheduling and lightweight model inference are undertaken by edge nodes; and large-scale computational resources together with knowledge support are provided by the cloud. Through this arrangement, a closed-loop system is formed, characterized by top-down computational resource allocation and bottom-up data feedback. The proposed three-layer collaborative paradigm overcomes the limitations of traditional cloud–end binary architectures. By introducing computational buffering and task offloading at the edge layer, latency caused by direct mobile-to-cloud communication is effectively mitigated while the computational burden associated with deploying full-scale models on mobile devices is alleviated. As a result, a dynamic balance between intelligent capability and real-time interaction performance is achieved, enabling effective adaptation to the multi-scenario process design requirements encountered in automotive manufacturing environments. Figure 1 illustrates the three-layer structure comprising the cloud, edge, and mobile layers, as well as the corresponding data flow directions, emphasizing both top-down computational resource distribution and bottom-up data feedback mechanisms.

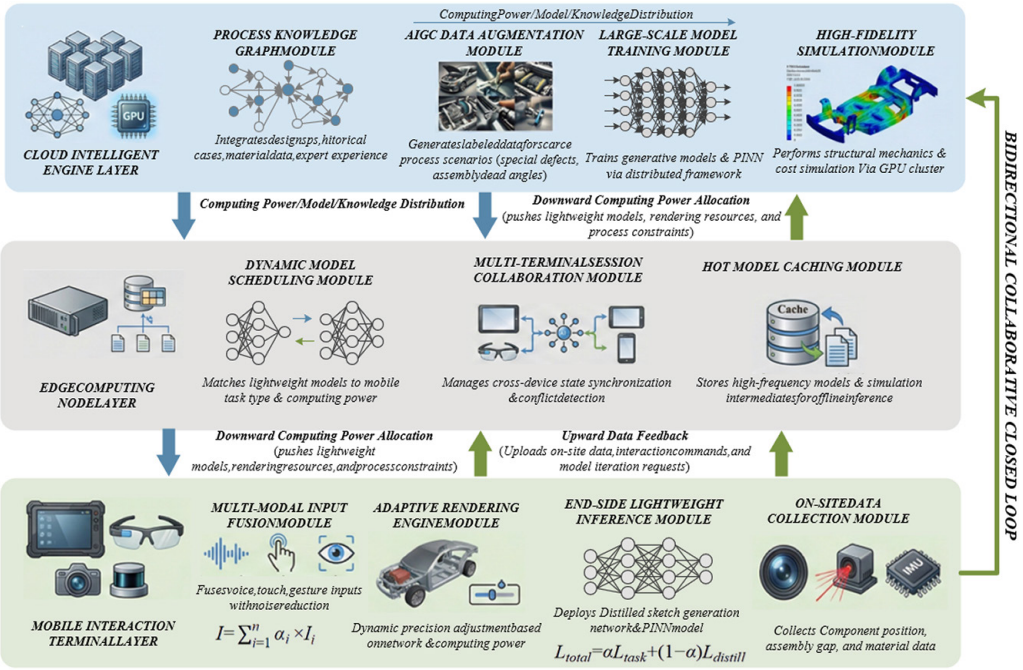

Fig. 1. Architecture of the cloud–edge–end collaborative mobile interaction system for automotive process design

The cloud-side intelligent engine is responsible for large-scale process data storage, extensive model training, and high-fidelity simulation computation, thereby providing the system with core computational resources and knowledge support. Innovation at the cloud layer is primarily concentrated in two aspects: the construction of a process knowledge graph and artificial intelligence generated content (AIGC)–based data augmentation. The process knowledge graph is constructed using a structured methodology that integrates automotive process design specifications, historical process cases, material parameters, and expert knowledge. Ontology modeling techniques are employed to define process entities, attributes, and relationships, resulting in a reusable and inference-capable knowledge base. This knowledge base supports real-time compliance verification and case matching for process schemes initiated at the mobile end. An incremental knowledge update mechanism is further designed, in which on-site design data fed back from edge nodes are utilized to dynamically enrich knowledge representations for underrepresented or sparse scenarios through a confidence-based evaluation model. The core formulation is expressed as:

$$C_k = \frac{N_{valid,k}}{N_{total,k}} \times \omega_k \tag{1}$$

where, C_k denotes the confidence level of the k-th category of scenario knowledge, $N_{valid,k}$ represents the number of valid feedback data instances, $N_{total,k}$ denotes the total number of feedback instances, and ω_k corresponds to the scenario weighting coefficient. When C_k exceeds a predefined threshold, the corresponding knowledge is incorporated into the process knowledge graph. AIGC-based data augmentation is implemented using a targeted generation strategy. For sparse process scenarios, such as special material defects and assembly blind zones, the generation process is controlled through constraint injection to ensure consistency between synthetic data

and real process conditions. The generated data are subsequently subjected to dual validation based on material properties and structural dimensions, after which they are used for training lightweight models deployed at the mobile end. This strategy effectively alleviates data scarcity in deep learning model training.

Edge computing nodes function as the collaborative hub between the cloud and the mobile end and are adaptable to deployment environments such as factory servers and in-vehicle domain controllers. Through task scheduling and data processing, a balance between real-time responsiveness and computational overhead is achieved. Core innovations at the edge layer are reflected in dynamic model loading and caching mechanisms, as well as multi-terminal session collaboration. Dynamic model loading and intelligent caching are realized through precise scheduling based on mobile-end task types and computational states. The category of the current design task is identified via a task feature extraction model, after which the corresponding lightweight model is matched and dynamically loaded. In parallel, a hotspot model identification criterion is established, whereby models with invocation frequencies exceeding a predefined threshold are cached locally. In the event of network disconnection, cached models support offline inference at the mobile end. Upon network recovery, design data and inference results are synchronized using an incremental synchronization algorithm, thereby ensuring continuity of the design workflow. Multi-terminal session collaboration is implemented through a distributed session management architecture, in which a unique session identifier is assigned to multiple mobile terminals associated with the same process design project. Interaction states and design data across terminals are synchronized in real time via a state synchronization protocol. A design conflict detection mechanism is further introduced: when concurrent modifications to the same process parameter are initiated by multiple terminals, priority is determined based on timestamps and permission levels, effectively preventing data synchronization conflicts. Through this mechanism, information silos commonly encountered in multi-role collaborative design under mobile scenarios are effectively mitigated.

The mobile interaction end encompasses form factors such as industrial ruggedized tablets and AR devices, providing naturalistic human–computer interaction interfaces while performing lightweight local inference and real-time rendering. Innovation at this layer is primarily focused on adaptive rendering engine optimization and multimodal input fusion. The adaptive rendering engine adopts a three-level precision dynamic adjustment strategy, in which high-precision, balanced, and lightweight modes are switched in real time based on available network bandwidth and terminal computational capacity. A rendering resource preloading mechanism is further introduced: based on a prediction model trained on historical design paths, core component rendering resources are proactively delivered by edge nodes, thereby reducing rendering latency. Multimodal input fusion is implemented through a weighted fusion model that enables deep integration of speech, touch, and gesture inputs. The core formulation is expressed as:

$$I = \sum_{i=1}^{n} \alpha_i \times I_i \tag{2}$$

where, I denotes the final intent result, α_i represents the weighting coefficient of the i-th modality, and I_i corresponds to the single-modality intent recognition result. The weighting coefficients α_i are dynamically adjusted according to the recognition accuracy of each modality under the current interaction context. To address noise interference inherent in mobile environments, lightweight on-device denoising

and intent correction modules are incorporated. Speech inputs are optimized using adaptive filtering algorithms, while gesture trajectory features are extracted to eliminate erroneous touch operations. Through these mechanisms, both the accuracy and real-time performance of interaction intent recognition are ensured under complex workshop conditions.

3 KEY TECHNICAL IMPLEMENTATION

3.1 Mobile-end process sketching and lightweight model generation framework

Mobile-end process sketching and model generation are constrained by a fundamental tension between the strong expressive capability of large-scale generative models and the limited computational resources and low-latency requirements of mobile devices. Conventional lightweight approaches often result in substantial degradation of generation accuracy, rendering them inadequate for the stringent standards of automotive process design. To address this challenge, a two-stage lightweight generation and distillation optimization framework was developed. Through collaborative task partitioning between the cloud and the mobile end, high-precision generation capability is preserved while mobile-side resource consumption is strictly controlled, enabling real-time interactive iteration in process design. During the cloud-side pre-generation stage, the process knowledge graph is employed as a constraint to regulate the output of the cloud-based large model through a knowledge-guided generation logic. Instead of producing complete three-dimensional models, only lightweight feature maps are generated, leading to a significant reduction in feature dimensionality relative to full models and substantially lowering cross-end transmission overhead. The generated feature maps encapsulate process constraint boundaries, key structural features, and material property information. In this manner, compliance with process design standards is ensured for subsequent mobile-end derivation while sufficient semantic support is provided for lightweight mobile-side networks.

The mobile-end derivation and optimization stage focuses on the design and refinement of an ultra-lightweight sketch derivation network, with the total parameter size strictly constrained to within 50 MB. Knowledge distillation techniques are employed to transfer generative capability and process cognition from the cloud-based large model. The distillation process is optimized using a dual-loss objective, formulated as:

$$L_{total} = \alpha L_{task} + (1-\alpha) L_{distill} \tag{3}$$

where, α is the weighting coefficient and is set to 0.3; L_{task} denotes the process design task loss, computed based on deviations between generated results and process specifications; and $L_{distill}$ represents the distillation loss, used to measure discrepancies between intermediate feature representations of the mobile-end network and the cloud-based large model. Through this formulation, accurate replication of cloud-side process design logic by the mobile-end network is achieved. To further optimize mobile-end performance, 8-bit Integer (Integer-8) quantization compression and structured layer pruning are applied. Pruning is conducted based on neuron contribution evaluation, whereby redundant neurons and connections with contribution values below a threshold of 0.01 are removed. During quantization, calibration datasets are employed to correct precision bias, resulting

in a significant reduction in inference latency. Through simple sketch-based input, designers are able to trigger real-time generation of detailed design schemes at the mobile end, supporting fine-grained operations such as wall thickness adjustment and weld point layout optimization. In this way, a real-time balance between creative expression and process constraints is achieved while both interaction efficiency and design accuracy are maintained.

3.2 Mobile-end process scene understanding based on multimodal perception

A central challenge in mobile process design lies in the disconnect between the design workflow and on-site operating conditions. Conventional scene understanding approaches rely on offline preprocessing at desktop platforms and are therefore unable to satisfy the requirements of real-time design validation in mobile environments. To address this limitation, a mobile-end multi-sensor fusion and lightweight scene inference approach was developed. Through real-time integration and rapid processing of multi-dimensional perceptual data, instantaneous on-device understanding of process scenes is achieved, providing direct support for on-site validation of design scheme adaptability. The system integrates mobile-end cameras, light detection and ranging (LiDAR), and inertial measurement unit (IMU) sensors to establish a multi-source synchronous data acquisition mechanism. Timestamp alignment algorithms are employed to eliminate latency discrepancies among heterogeneous sensors, enabling synchronized acquisition of key scene information, including component positioning, assembly clearances, and material characteristics. A feature-level fusion strategy is adopted to integrate multi-sensor data, with the core formulation expressed as:

$$F_{fusion} = \sum_{i=1}^{3} \sum \beta_i F_i \tag{4}$$

where, F_{fusion} denotes the fused feature representation; β_i represents the weighting coefficient associated with each sensor, with initial values set to 0.4 for the camera, 0.4 for LiDAR, and 0.2 for the IMU, and dynamically fine-tuned based on scene recognition accuracy; and F_i corresponds to features extracted from individual sensors. Through this strategy, perceptual limitations inherent to single-sensor configurations in complex workshop environments are effectively mitigated.

To enable real-time inference at the mobile end, a lightweight convolutional neural network–long short-term memory (CNN–LSTM) fusion model was designed. Depth-wise separable convolutions are employed to replace conventional convolutional layers, reducing the parameter scale to within 8 MB. In parallel, a channel attention mechanism is incorporated to strengthen the extraction of salient features, achieving a balance between model light-weighting and recognition accuracy. Within the proposed architecture, spatial features extracted by the CNN and temporal features captured by the LSTM are deeply fused, enabling integrated execution of three core tasks: component recognition, assembly gap detection, and defect localization. By avoiding the cumulative inference latency associated with the serial coupling of multiple models, the on-device inference time is ultimately constrained to within 100 ms. To meet the requirements of AR–based enhancement and adaptation, a scene coordinate calibration mechanism was further designed. A three-dimensional coordinate system of the on-site environment is established

using spatial point cloud data acquired by LiDAR sensors. Model recognition results are then accurately mapped to virtual process annotations through precise coordinate alignment, enabling real-time overlay of virtual design schemes onto physical components in the production environment. Through AR devices, designers are able to intuitively assess the alignment between process design schemes and actual operating conditions, allowing immediate identification of assembly interference, excessive gaps, and related issues. In this manner, an integrated closed-loop workflow combining virtual design and on-site validation is established, effectively overcoming the long-standing limitation of mobile design systems being decoupled from real production contexts. Figure 2 illustrates the CNN–LSTM–based multi-sensor spatiotemporal feature fusion network architecture.

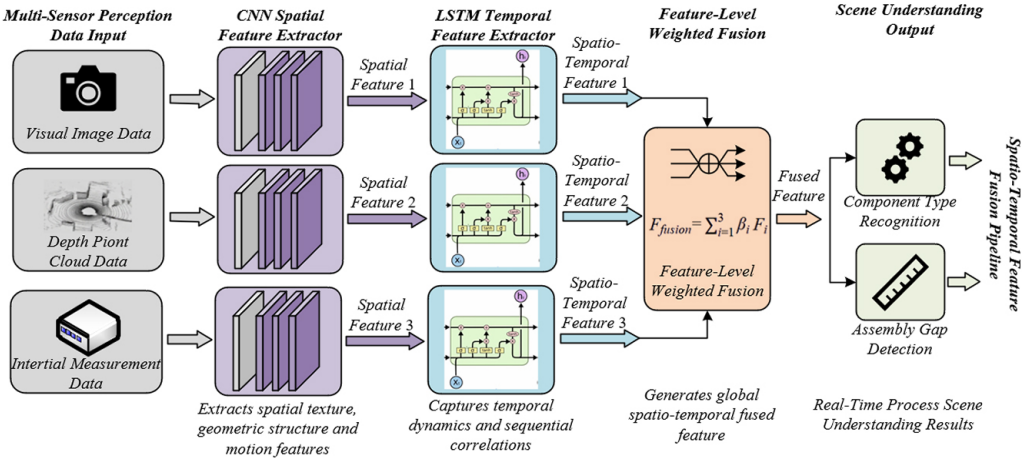

Fig. 2. Architecture of the multi-sensor spatiotemporal feature fusion network based on CNN–LSTM

3.3 Mobile-end real-time process feasibility simulation and interactive optimization

Conventional computer-aided engineering (CAE) simulations suffer from inherent computational inefficiency, typically requiring hour-level computation cycles. As a result, integration into a real-time mobile interaction loop is infeasible, and process optimization remains dependent on offline analysis, severely constraining the iterative efficiency of mobile design workflows. To address this limitation, a lightweight Physics-informed neural network (PINN) model combined with a cloud–end collaborative simulation strategy was developed. Through dynamic coupling between rapid on-device prediction and high-accuracy cloud-based computation, a balance between real-time responsiveness and simulation accuracy is achieved, enabling an immediate mobile-end modification–prediction–optimization closed loop. At the mobile end, a lightweight PINN model is deployed, with the network architecture simplified to six layers and the total parameter size constrained to within 12 MB. Predictive accuracy is enhanced by incorporating process-related physical constraints. The core loss function is defined as:

$$L_{PINN} = L_{data} + \lambda L_{physics} \tag{5}$$

where, L_{data} denotes the data-fitting loss, and $L_{physics}$ represents the physics-based constraint loss, which incorporates mechanical equilibrium equations and material constitutive relationships. The weighting coefficient λ is set to 0.6 to ensure that

model predictions remain consistent with fundamental physical laws. For modifications to key process parameters—such as weld point positioning and component wall thickness—the model is capable of outputting predicted trends in mechanical performance and manufacturing cost within 1 s, thereby satisfying the real-time interaction requirements of mobile process design.

A cloud–end collaborative simulation mechanism is implemented using a dynamic precision adaptation strategy. A precision decision threshold ε is set to 5%. When the relative error between the mobile-end PINN prediction and historical high-fidelity simulation data falls below ε, the predicted results are directly returned to the designer. When the error exceeds the threshold, or when a formal simulation report is required, high-fidelity cloud-based simulation is automatically triggered. At the edge layer, an intermediate result cache pool is constructed to store shared data and intermediate computational results generated during cloud-side simulations. For subsequent simulation tasks of similar types, cached results can be directly reused, thereby reducing redundant computation and significantly shortening cloud simulation response time. Mobile-end interactive optimization is realized using a multi-objective genetic algorithm. A slider-based visual interface is designed to support real-time adjustment and weight allocation across three core indicators: strength, weight, and cost. The population size of the algorithm is set to 50, and the number of iterations is constrained to within 20 generations to ensure rapid convergence of optimization results. The optimization objective function is defined as:

$$\min O = w_1 \frac{S}{S_0} + w_2 \frac{W}{W_0} + w_3 \frac{C}{C_0} \tag{6}$$

where, w_1, w_2, and w_3 denote the weighting coefficients of the respective indicators; S, W, and C represent the strength, weight, and cost of the current design scheme; and S_0, W_0, and C_0 correspond to the baseline reference values. Through dynamic updating of the optimal solution and real-time feedback, on-site interactive optimization of process parameters is achieved.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental environment configuration

The experimental platform was constructed on a real-world industrial-grade software and hardware stack to accurately emulate cloud–edge–end collaborative deployment scenarios in automotive manufacturing environments, thereby ensuring the industrial applicability and reliability of the experimental results. At the mobile end, two types of devices were employed: industrial ruggedized tablets and AR smart glasses. The tablets were equipped with an Intel Core i7-12700H processor, 16 GB LPDDR5 memory, and an Adreno 730 Graphic Processing Unit (GPU), running the Android 13 operating system. The AR devices consisted of Microsoft HoloLens 2, configured with the Windows Holographic operating system. All mobile-end model inference was executed using a quantized TensorFlow Lite framework. Edge nodes were deployed on dedicated factory servers, configured with Intel Xeon Gold 6448Y processors, 64 GB DDR5 memory, and NVIDIA A10 GPUs. Task scheduling was implemented using Kubernetes Edge, while model caching strategies were constructed based on Redis. At the network layer, adaptive switching between 5G and gigabit Wi-Fi was supported. The cloud layer was implemented using a computing cluster comprising four NVIDIA A100 GPUs (80 GB memory each). Large-scale

model training was conducted via the PyTorch Distributed framework. The process knowledge graph was constructed using Neo4j, and high-fidelity structural mechanics simulations were performed using ANSYS 2024 R1.

4.2 Comparative experimental results and analysis

To comprehensively evaluate the overall performance of the proposed system, three representative mainstream solutions were selected as baselines. A multi-dimensional quantitative evaluation framework was constructed across performance efficiency, design quality, and interaction experience. Statistical results are reported as mean ± standard deviation, obtained from 50 repeated experimental trials. Differences among methods were assessed using independent-samples t-tests to determine statistical significance. All experimental data were collected under an industrial-grade software and hardware environment, ensuring both the reliability of the results and their applicability to real-world automotive manufacturing scenarios.

Table 1 presents a comparative analysis of performance efficiency metrics across different solutions. The results indicate that scheme generation latency is reduced by 47.8% relative to Baseline 3, while interaction response latency is reduced by 46.4%. These improvements are primarily attributed to dynamic model loading and intelligent caching mechanisms implemented at the edge layer. By predicting mobile-end task types and access frequencies, high-frequency lightweight models are proactively cached at edge nodes, thereby avoiding cross-end transmission of full-scale models and redundant mobile-end inference. In addition, the adaptive rendering engine dynamically adjusts rendering precision according to available network bandwidth, further reducing latency overhead.

Table 1. Comparison of performance efficiency metrics across different solutions

Solution	Scheme Generation Latency (ms)	Interaction Response Latency (ms, 95th Percentile)	Design Iteration Cycle (min)	Mobile-End Model Inference Energy Consumption (mWh)	Cloud Data Transmission Volume (MB)	Edge Cache Hit Rate (%)
Baseline 1	1850 ± 120	420 ± 35	45.2 ± 3.8	—	890 ± 65	—
Baseline 2	1280 ± 95	380 ± 40	38.5 ± 4.2	850 ± 60	520 ± 48	—
Baseline 3	920 ± 80	280 ± 30	32.1 ± 3.5	620 ± 55	310 ± 32	62.3 ± 4.1
Proposed system	480 ± 50*	150 ± 20*	18.3 ± 2.2*	380 ± 45*	120 ± 21*	89.7 ± 3.2*

Note: "—" indicates that the corresponding functionality is not supported. Values marked with "*" indicate statistically significant differences compared with all baseline solutions ($p < 0.05$).

The design iteration cycle is shortened by 59.5% compared with Baseline 1, which is mainly ascribed to the real-time closed-loop workflow integrating mobile-end interaction, edge-side inference, and cloud-based simulation. Through this mechanism, scheme generation, adjustment, and validation are completed directly on-site, eliminating frequent switching between desktop and mobile platforms and enabling immediate linkage between design activities and production scenarios. Mobile-end inference energy consumption is reduced by 38.7% relative to Baseline 3. This reduction is attributable to model quantization and optimized task scheduling, through which redundant neurons are removed and computational resources are dynamically allocated, effectively minimizing unnecessary energy consumption and accommodating the endurance constraints of mobile devices. Furthermore, cloud data transmission volume is reduced by 61.3% compared with Baseline 3, while

the edge cache hit rate is increased by 44.0%. These results further substantiate the effectiveness of the edge-collaborative architecture in optimizing data transmission efficiency and computational resource reuse.

Table 2. Comparison of design quality metrics across different solutions

Solution	Process Scheme Compliance Rate (%)	Simulation Prediction Accuracy (%)	Solution Diversity Entropy	Defect Recognition Accuracy (%)	Process Parameter Optimization Degree (%)
Baseline 1	82.3 ± 2.5	92.1 ± 1.8	1.2 ± 0.3	78.5 ± 3.6	15.2 ± 2.3
Baseline 2	78.5 ± 3.2	—	0.8 ± 0.2	72.3 ± 4.1	8.7 ± 1.9
Baseline 3	85.6 ± 2.8	88.5 ± 2.3	1.5 ± 0.4	84.2 ± 3.3	22.5 ± 2.7
Proposed system	94.2 ± 1.6*	95.8 ± 1.2*	2.3 ± 0.5*	92.6 ± 2.4*	35.8 ± 3.1*

Note: "—" indicates that the corresponding functionality is not supported. Values marked with "*" indicate statistically significant differences compared with all baseline solutions ($p < 0.05$).

Table 2 presents a comparative analysis of design quality metrics across different solutions. The process scheme compliance rate is increased to 94.2%, primarily driven by the support of the cloud-based process knowledge graph. By integrating design specifications, historical cases, material parameters, and expert knowledge into a structured knowledge base, real-time compliance verification of generated schemes is enabled, allowing non-compliant parameter combinations to be accurately avoided. In addition, AIGC-based targeted generation of rare scenario data enhances model adaptability to special material defects and extreme operating conditions, thereby reducing compliance deviations caused by data scarcity. Simulation prediction accuracy is improved by 8.2% relative to Baseline 3. This improvement is attributed to the optimized design of the mobile-end PINN, in which physical constraints—such as mechanical equilibrium equations and material constitutive relationships—are explicitly incorporated. As a result, consistency with cloud-based high-fidelity simulations is maintained even under second-level inference latency. The solution diversity entropy is increased by 53.3% compared with Baseline 3, which can be ascribed to the combined effect of cloud-based AIGC data augmentation and multimodal input fusion. Through this synergy, generated solutions span a broader range of process combinations and parameter configurations, providing designers with richer creative alternatives and mitigating the limitations of homogeneous solution generation. Furthermore, defect recognition accuracy and process parameter optimization degree are increased by 9.9% and 59.1%, respectively, relative to Baseline 3. These results further substantiate the system's core capabilities in process quality control and parameter optimization.

To evaluate operational stability and adaptability under typical disturbance conditions in automotive manufacturing workshops, robustness tests were conducted with respect to two key industrial constraints: network bandwidth fluctuation and terminal computational capability variation. Experimental results are presented in Figures 3 and 4. From the network bandwidth robustness results, it is observed that, even under offline conditions, the proposed system is able to rely on edge-cached models to achieve a scheme generation latency of 620 ms and an interaction response latency of 210 ms. In contrast, Baseline 3 is unable to operate when bandwidth drops to zero due to the absence of offline inference capability. Under low-bandwidth conditions of 10 Mbps, scheme generation latency is reduced by 53.6% and interaction response latency by 54.8% relative to Baseline 3. Even at a bandwidth of 100 Mbps, latency reductions of 47.8%–46.4% are consistently maintained. These results substantiate the critical role of edge-side dynamic caching and adaptive

rendering mechanisms in mitigating network dependency. Terminal computational capability robustness tests further demonstrate that, on low-capability devices, mobile-end inference time is reduced by 51.1% compared with Baseline 3, while AR overlay accuracy is improved by 65.4%. With increasing computational capability, inference latency is further reduced to 110 ms, and AR overlay accuracy remains stable within 0.9 mm. By contrast, Baseline 3 continues to exhibit an inference latency of 320 ms on high-capability devices, with overlay accuracy limited to 3.5 mm. These findings indicate that lightweight model design combined with dynamic computational resource scheduling effectively accommodates heterogeneous mobile terminals with varying performance profiles.

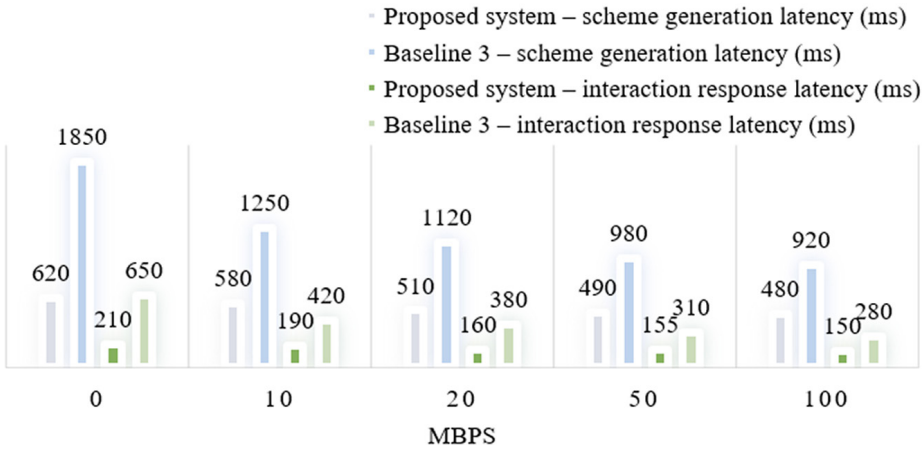

Fig. 3. Network bandwidth robustness test results

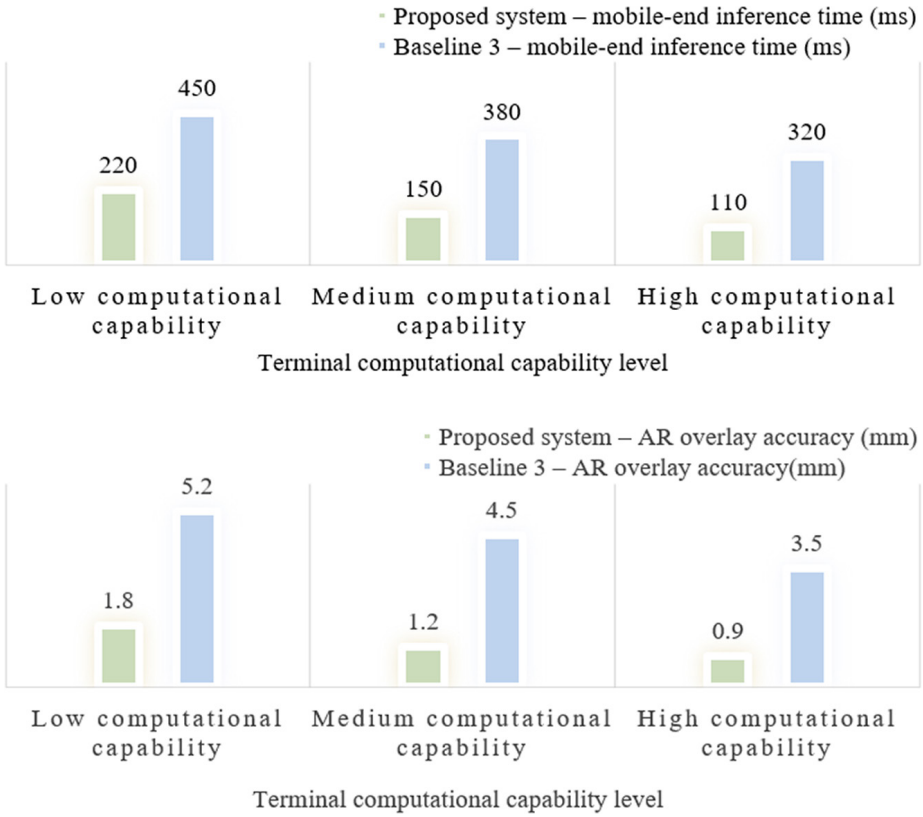

Fig. 4. Robustness test results of the terminal computational capability

Table 3. Comparison of interaction experience metrics across different solutions

Solution	Multimodal Interaction Accuracy (%)	User Satisfaction (System Usability Scale (SUS) Score)	Misoperation Rate (%)	AR Overlay Accuracy (mm)	Interaction Learning Cost (min)
Baseline 1	—	72.3 ± 4.5	8.5 ± 1.6	—	35.2 ± 4.3
Baseline 2	81.2 ± 3.8	68.5 ± 5.2	12.3 ± 2.1	—	22.6 ± 3.8
Baseline 3	87.6 ± 3.2	76.8 ± 4.8	7.2 ± 1.4	3.5 ± 0.6	18.4 ± 3.2
Proposed system	95.3 ± 2.1*	88.6 ± 3.5*	2.8 ± 0.9*	1.2 ± 0.4*	10.5 ± 2.5*

Note: "—" indicates that the corresponding functionality is not supported. Values marked with "*" indicate statistically significant differences compared with all baseline solutions ($p < 0.05$).

Table 3 presents a comparative analysis of interaction experience metrics across different solutions. Multimodal interaction accuracy is increased to 95.3%, primarily attributable to the lightweight on-device denoising and intent correction modules. To address workshop noise and touch interference, adaptive filtering algorithms are employed to optimize speech input, while gesture trajectory features are extracted to eliminate erroneous operations. Combined with a dynamic weighting strategy for multimodal fusion, reliable interaction performance is achieved under complex industrial conditions. User satisfaction reaches a score of 88.6 on the System Usability Scale (SUS). Feedback indicates that the principal advantages are concentrated in naturalistic interaction and strong on-site adaptability. Voice and gesture commands are executed without cumbersome operational steps, while AR technology enables precise overlay of virtual process schemes onto physical components. Overlay accuracy is maintained within 1.2 mm, allowing intuitive assessment of scheme suitability and effectively addressing the long-standing limitations of rigid interaction and contextual disconnection in traditional mobile tools.

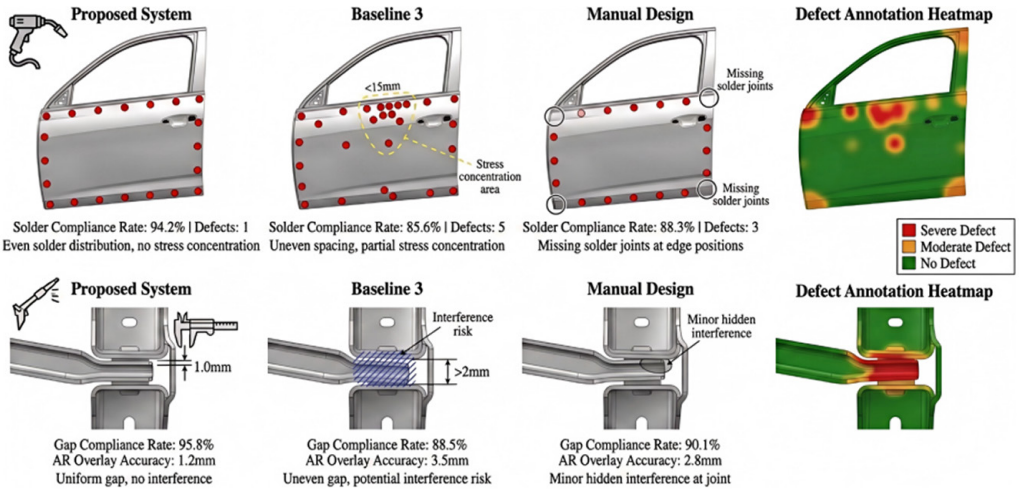

Fig. 5. Comparative visualization of representative process design cases

To systematically evaluate the topology optimization capability and defect avoidance mechanisms of the proposed framework under varying geometric constraints, comparative experiments were conducted using representative automotive body welding and chassis assembly scenarios. The visualized results presented in Figure 5 demonstrate that the proposed system achieves significantly superior performance

in both spatial distribution uniformity and dimensional control accuracy when compared with random baseline algorithms and conventional manual design approaches. In contrast to baseline algorithms, where unstable variance leads to pronounced stress concentration, and manual methods, where cognitive oversights frequently occur in boundary regions, the proposed model consistently maintains a process compliance rate exceeding 94%, with no evident geometric interference observed. The integrated defect heatmaps further confirm that potential engineering conflicts can be proactively identified and mitigated through deep learning–based representations. These results indicate that the proposed mobile interaction paradigm effectively bridges the cognitive gap between digital planning and physical execution, thereby ensuring high-fidelity manufacturing quality.

5 DISCUSSION AND CONCLUSION

Through the integration of a cloud–edge–end collaborative architecture and a set of core technical innovations, a substantial advancement has been achieved in the domain of industrial process design, where mobility and intelligence are deeply integrated. The advantages and innovative contributions of the proposed system are clearly distinguished from existing studies. Prior research in this area has largely exhibited a binary separation between desktop-based and mobile-based solutions: intelligent desktop plug-ins remain constrained by hardware form factors and cannot be extended to production sites, while mobile tools generally lack the deep learning capabilities required to support complex process design tasks. By contrast, the three-layer collaborative framework introduced in this study effectively bridges this technological divide, enabling deep integration between the generative and inferential capabilities of deep learning and the real-time interaction and contextual awareness characteristics of mobile computing. Key challenges associated with limited mobile computational resources, network instability, and environmental noise are systematically addressed through edge-side dynamic caching, lightweight model optimization, and multimodal denoising and correction mechanisms. As a result, a reusable technical paradigm for industrial mobile design is established. Beyond substantial improvements in the real-time performance and efficiency of automotive process design, a paradigm shift is facilitated from traditional desktop-based, tool-assisted workflows toward on-site intelligent co-creation, providing a novel solution for the intelligent upgrading of process design under the Industry 4.0 framework.

6 REFERENCES

[1] S. M. Hussain, K. Mohamad Yusof, S. Ashfaq Hussain, R. Asuncion, and S. Ghouse, "Integration of 4G LTE and DSRC (IEEE 802.11p) for enhancing vehicular network performance in IoV using optimal cluster-based data forwarding (OCDF) protocol," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 14, pp. 111–124, 2021. https://doi.org/10.3991/ijim.v15i14.19201

[2] E. D. Spyrou, V. Kappatos, and A. Anagnostopoulou, "Human—Artificial Intelligence teaming for automotive applications: A review," *International Journal of Transport Development and Integration*, vol. 8, no. 2, pp. 215–224, 2024. https://doi.org/10.18280/ijtdi.080201

[3] M. Müller, M. Lehmann, and H. Kuhn, "Measuring sequence stability in automotive production lines," *International Journal of Production Research*, vol. 59, no. 24, pp. 7336–7356, 2021. https://doi.org/10.1080/00207543.2020.1790685

[4] Y. Chen, Y. Ding, J. Jin, and D. Ceglarek, "Integration of process-oriented tolerancing and maintenance planning in design of multistation manufacturing processes," *IEEE Transactions on Automation Science and Engineering*, vol. 3, no. 4, pp. 440–453, 2006. https://doi.org/10.1109/TASE.2006.872105

[5] X. W. Wang and T. C. Wang, "Enhanced rule generation in product design through rough set theory and ant colony optimization," *Precision Mechanics & Digital Fabrication*, vol. 1, no. 1, pp. 11–23, 2024. https://doi.org/10.56578/pmdf010102

[6] C. Kong, J. Lei, and X. Zhou, "Research on the intelligent design system for automotive panel die based on geometry and knowledge driven," *The International Journal of Advanced Manufacturing Technology*, vol. 133, no. 1, pp. 765–790, 2024. https://doi.org/10.1007/s00170-024-13728-z

[7] R. Lipman, M. Palmer, and S. Palacios, "Assessment of conformance and interoperability testing methods used for construction industry product models," *Automation in Construction*, vol. 20, no. 4, pp. 418–428, 2011. https://doi.org/10.1016/j.autcon.2010.11.011

[8] X. Li and J. Wan, "Proactive caching for edge computing-enabled industrial mobile wireless networks," *Future Generation Computer Systems*, vol. 89, pp. 89–97, 2018. https://doi.org/10.1016/j.future.2018.06.017

[9] H. Hassan, J. M. Martínez, and C. Domínguez, "m-IC: A mobile device based multimedia learning methodology for industrial computing," *International Journal of Engineering Education*, vol. 31, no. 6, pp. 1678–1687, 2015.

[10] T. Kolossváry and P. Németh, "Systems engineering based sustainability improvement in automotive product development," *Journal of Engineering Management and Systems Engineering*, vol. 2, no. 1, pp. 53–60, 2023. https://doi.org/10.56578/jemse020103

[11] G. Reinhart and C. Patron, "Integrating augmented reality in the assembly domain-fundamentals, benefits and applications," *CIRP Annals*, vol. 52, no. 1, pp. 5–8, 2003. https://doi.org/10.1016/S0007-8506(07)60517-4

[12] Y. Wang, L. Zheng, and Y. Wang, "Event-driven tool condition monitoring methodology considering tool life prediction based on industrial internet," *Journal of Manufacturing Systems*, vol. 58, pp. 205–222, 2021. https://doi.org/10.1016/j.jmsy.2020.11.019

[13] C. Kong, J. Lei, J. Chao, and X. H. Zhou, "Research on automatic analysis and layout system of the stamping process for automotive panels and its key technologies," *The International Journal of Advanced Manufacturing Technology*, vol. 129, no. 9, pp. 4101–4120, 2023. https://doi.org/10.1007/s00170-023-12540-5

[14] R. A. Waraich, M. D. Galus, C. Dobler, M. Balmer, G. Andersson, and K. W. Axhausen, "Plug-in hybrid electric vehicles and smart grids: Investigations based on a microsimulation," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 74–86, 2013. https://doi.org/10.1016/j.trc.2012.10.011

[15] I. Qasim, M. W. Anwar, F. Azam, H. Tufail, W. H. Butt, and M. N. Zafar, "A model-driven mobile HMI framework (MMHF) for industrial control systems," *IEEE Access*, vol. 8, pp. 10827–10846, 2020. https://doi.org/10.1109/ACCESS.2020.2965259

[16] M. Aljamal, A. Alsarhan, M. Aljaidi, A. Mughaid, W. Q. Al Jamal, and A. E. Twairesh, "Securing the mobile future: An extensive analysis of the threat landscape from mobile devices user perspectives," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 8, pp. 140–158, 2025. https://doi.org/10.3991/ijim.v19i08.51959

[17] S. L. Jeng, W. H. Chieng, and Y. Chen, "Web-based human-machine interfaces of industrial controllers in single-page applications," *Mobile Information Systems*, vol. 2021, no. 1, p. 6668843, 2021. https://doi.org/10.1155/2021/6668843

7 AUTHORS

Chutong Liu is a doctoral candidate at the School of Art and Design, Guangdong University of Technology. His main research interests focus on art and design (E-mail: 1112217002@mail2.gdut.edu.cn).

Qingxin Guo obtained a Master's degree from East China Normal University and currently works at the School of Computer Science, Hubei Polytechnic University. His main research interests include Java programming language and database (E-mail: 207101@hbpu.edu.cn).

Imprint

iJIM – International Journal of Interactive Mobile Technologies
http://www.i-jim.org

Indexing
International Journal of Interactive Mobile Technologies is indexed in **Elsevier Scopus, Ulrich, EBSCO, Google Scholar, and DBLP**.

Publication Frequency
Bimonthly

Publisher
International Federation of Engineering Education Societies (IFEES)
IFEES-GEDC Secretariat
c/o George Mason University
Volgenau School of Engineering
4400 University Drive, MS 4A3
Fairfax, VA 22030
USA

Editorial Office
CTI Global
Frankfurt, Vienna, New York, Bengaluru, Hong Kong
Madrider Strasse 4
60327 Frankfurt am Main
Germany