Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 9 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i09.61739

PAPER

Interactive Mobile AI Video Creation: System Design Integrating Style Transfer and Narrative Control

Nan Wang, Ziwei Zhao (✉)

Hebei International Business Vocational College, Qinhuangdao, China

13933675555@139.com

ABSTRACT

Existing AI-based video stylization techniques face significant challenges when deployed on mobile devices, including high latency, insufficient temporal coherence, lack of narrative control, and poor resource adaptability. To address these limitations, this paper proposes a low-latency interactive AI video creation system specifically designed for mobile platforms. Without relying on newly developed foundational AI models, the proposed system leverages a hierarchical system architecture, lightweight model optimization, deep integration of narrative logic, and device-side collaborative scheduling to enable real-time video generation, temporal consistency, and controllable storytelling on resource-constrained mobile devices. This paper focuses on the technical implementation of the system's core innovative modules. Extensive multi-dimensional experiments demonstrate that the proposed system achieves significant advantages in performance, generation quality, and user interaction experience. Experimental results show that the per-frame processing latency on three representative mobile devices is reduced to as low as 28 ms, 32 ms, and 45 ms, respectively, with end-to-end latency consistently maintained below 200 ms. The system attains a temporal consistency score of 0.92 and a stylization quality rating of 8.6, significantly outperforming three baseline models. Moreover, the proposed system exhibits strong adaptability across heterogeneous computing capabilities, maintains low power consumption (≤200 mA/hr), and delivers high-quality creative outputs. These results indicate that the system provides an efficient and practical engineering solution for deploying AI video creation on mobile devices. This study contributes a feasible and effective pathway toward the large-scale application of intelligent video creation technologies on mobile platforms and holds substantial significance for advancing mobile AI-driven creative systems.

KEYWORDS

mobile AI, style transfer, real-time video generation, narrative logic, resource optimization

1 INTRODUCTION

The continuous upgrading of heterogeneous computing capabilities of mobile devices and the rapid iteration of AI generation technologies [1–4] are driving

Wang, N., Zhao, Z. (2026). Interactive Mobile AI Video Creation: System Design Integrating Style Transfer and Narrative Control. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(9), pp. 43–56. https://doi.org/10.3991/ijim.v20i09.61739

Article submitted 2026-01-28. Revision uploaded 2026-03-20. Final acceptance 2026-03-31.

intelligent video creation to migrate from the cloud to the device side. Mobile AI video creation has become a frontier research topic in the fields of interaction design and computer vision [5, 6]. However, existing technical solutions still face core bottlenecks in practical applications. Traditional style transfer models incur high computational overhead [7], making it difficult to adapt to the limited computing power, memory, and energy resources of mobile devices [8], resulting in real-time performance that fails to meet interactive requirements. Meanwhile, frame-by-frame stylization processing easily causes temporal flickering in videos [9, 10] and lacks effective support for users' narrative intentions, which constrains the improvement of interactive creative experiences on mobile devices. To address the above problems, this paper focuses on an engineering-oriented innovation pathway. Through system-level architecture design and multi-technology integrated optimization, the proposed approach overcomes resource constraints and experience bottlenecks and constructs a mobile AI video creation system that simultaneously achieves real-time performance, temporal coherence, and narrative capability. This work provides support for the practical deployment of device-side intelligent creation technologies and demonstrates both significant academic exploration value and industrial application potential.

Current research on mobile style transfer mainly focuses on lightweight optimization for single images or simple video style transformation, lacking deep integration of narrative logic and end-to-end collaborative design on the device side, which makes it difficult to adapt to complex interactive creative scenarios [11]. Although device–cloud collaborative solutions can alleviate the computational burden on mobile devices [12, 13], their availability is significantly reduced in weak-network or offline environments. Moreover, such solutions fail to fully exploit the computing potential of heterogeneous hardware on mobile devices, resulting in low hardware utilization efficiency [14, 15]. Based on the limitations of existing studies, this paper identifies four core innovation entry points. First, a hierarchical architecture with deep collaboration among interaction, computation, and devices is constructed to achieve precise adaptation between AI computation and mobile system characteristics. Second, an adaptive lightweight style transfer model is designed, relying on a dual-branch network structure, pre-computation caching strategies, and dynamic resolution adjustment to balance real-time performance and stylization quality. Third, a narrative-driven temporal consistency control mechanism is established to fundamentally address coherence issues in the video stylization process. Fourth, a device-side dedicated computation scheduling strategy is developed to maximize heterogeneous hardware utilization while achieving precise power consumption control.

The remainder of this paper is organized as follows. Section 2 describes the overall system architecture design and clarifies the functional positioning of each layer and the core workflow logic. Section 3 provides an in-depth analysis of the technical implementation details of the three core innovative modules and explains the key optimization strategies. Section 4 designs multi-dimensional experimental schemes and verifies system performance and effectiveness through comparative and ablation experiments. Section 5 summarizes the research work of this paper, analyzes existing limitations, and discusses future research directions. The main contributions of this paper can be summarized as three aspects. First, an interactive AI video creation system architecture for mobile devices is constructed, achieving deep coupling between narrative logic and style transfer. Second, a multi-dimensional device-side optimization technology framework is developed, effectively breaking through the bottleneck of real-time video generation under mobile device resource constraints. Third, a full-process system engineering implementation and rigorous experimental validation are completed, providing reliable technical support for the industrial deployment of mobile AI video creation technologies.

2 OVERALL SYSTEM ARCHITECTURE DESIGN

To address the core challenges faced by AI video creation on mobile devices, including limited computing resources, high requirements for real-time interaction, and insufficient adaptation of the creation process to hardware characteristics, this system is designed with lightweight, highly interactive, and strongly adaptive principles at its core. A three-level architecture with layered decoupling and collaborative linkage is constructed, and a closed-loop streaming pipeline workflow is designed to achieve integrated end-side operation throughout the entire creation process. This ensures both the quality and narrative coherence of AI video generation while fitting the operational characteristics and hardware constraints of mobile devices, solving issues such as network dependency and high latency in traditional end-cloud collaborative architectures, and adapting to interactive creation scenarios on mobile devices anytime and anywhere.

The system designed in this paper adopts a three-level architecture consisting of an interaction layer, a computation engine layer, and a device layer. Each layer is deeply aligned with the characteristics of mobile devices and achieves lightweight full-process operation through efficient collaboration. The interaction layer targets dual mobile platforms and follows a touch-first design principle. It supports three core input modes, including gestures, voice, and template-based inputs. A real-time preview module is integrated synchronously, which adaptively renders visual content based on device screen resolution to ensure consistency between preview effects and final generation results. The preview latency is strictly controlled within 50 ms to guarantee interaction fluency. The computation engine layer serves as the core carrier of innovations and adopts a modular design. It consists of three main modules: lightweight narrative planning, real-time style transfer rendering, and mobile device resource management and scheduling. Lightweight interfaces are used for inter-module linkage to reduce redundant data transmission and improve computational efficiency. The device layer focuses on mining mobile hardware characteristics and adapting to hardware constraints. For heterogeneous hardware across different platforms, a hardware capability perception model is constructed, while memory and energy consumption constraint modeling is conducted synchronously. The device layer monitors device operating states in real time and provides precise data support for dynamic adjustment of upper-layer modules.

The system adopts a closed-loop streaming pipeline workflow, enabling full-process automated operation with end-to-end total latency controlled within 200 ms, meeting real-time interaction requirements on mobile devices. After users input creation requirements through the interaction layer, the interaction layer completes preprocessing tasks such as material format standardization and instruction semantic parsing and transmits the data to the computation engine layer. Based on the preprocessed data, the lightweight narrative planning module constructs a narrative graph using a lightweight graph neural network to determine key frame positions, scene logical relationships, and segment arrangement order. Guided by the narrative graph, the real-time style transfer rendering module performs stylization processing while receiving dynamic configuration instructions from the resource management scheduler to adjust computational parameters. The stylized results are fed back to the interaction layer in real time for secondary user adjustments, forming a closed-loop interaction. This pipeline adopts a parallelized design, in which tasks such as video decoding, style computation, and encoding output are decomposed and executed in parallel to maximize end-to-end processing efficiency.

3 CORE MODULE TECHNICAL IMPLEMENTATION

The three core modules form a fully integrated technical system with layered support and collaborative linkage around the technical pain points of interactive AI video creation on mobile devices. The mobile-adaptive lightweight style transfer model addresses the balance between real-time performance and image quality on the device side, serving as the foundation for AI video visual generation. The visual coherence control fused with narrative logic achieves deep integration of style transfer and narrative expression, solving issues such as visual flickering and misalignment of narrative subjects caused by frame-by-frame stylization, thereby improving the content quality of video creation. The interactive-oriented mobile system's collaborative optimization ensures stable and efficient operation of the preceding modules from dimensions such as hardware adaptation, task scheduling, and power consumption control, serving as the underlying support for full-process interactive creation. All modules follow mobile, lightweight design requirements, and through dynamic data interaction and parameter linkage between modules, they achieve a full-chain technological breakthrough in end-side AI video creation from demand input to result generation.

3.1 Mobile adaptive lightweight style transfer model

The mobile adaptive lightweight style transfer model proposed in this paper takes lightweight structural design, hierarchical caching strategies, and dynamic adaptation mechanisms as its core, aiming to maximize the reduction of device-side computational overhead and latency while ensuring stylization quality. The framework structure is shown in Figure 1. The model adopts a dual-branch architecture consisting of a basic style backbone and a detail texture enhancement branch, achieving graded guarantees for real-time performance and visual quality. The basic style backbone is a mandatory execution branch. It replaces conventional convolution with depthwise separable convolution, decoupling spatial convolution and channel convolution, thereby reducing the number of parameters and computational cost to 1/8–1/5 of traditional convolution. Meanwhile, channel pruning is implemented based on L1 regularization to remove redundant feature channels while retaining core propagation paths. The total number of network layers is strictly controlled within 12 layers, and the single-frame forward propagation time does not exceed 30 ms. The detail texture enhancement branch is an on-demand activation module. It is only activated when users pause creation or specify high-precision regions or when device resources are sufficient. This branch adopts a lightweight attention mechanism to focus on image edges and texture features, supplementing detail information through small-range convolution operations. Moreover, it processes only key frames and adjacent frames to avoid high-frequency computation on full frames. The features of the two branches are integrated through a dynamic weight fusion module, with the fusion formula defined as:

$$F = \alpha F_b + (1-\alpha)F_d \tag{1}$$

where, F_b denotes the feature of the basic branch, F_d denotes the feature of the detail branch, and α is a dynamic weight coefficient that is adaptively adjusted according to the device resource state. When resources are constrained, α is

increased to above 0.8 to prioritize fluency, while under sufficient resources it is reduced to 0.5–0.6 to enhance detail performance.

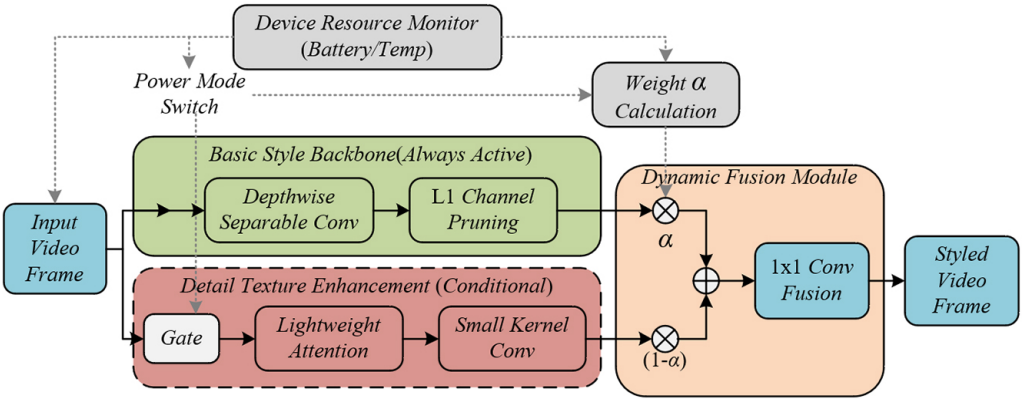

Fig. 1. Schematic diagram of the mobile adaptive lightweight style transfer model

To further reduce device-side computational pressure, the model designs a two-level strategy combining cloud-side precomputation and device-side cache reuse. When a style template is loaded on the device for the first time, deep style features are extracted through cloud-side or local preprocessing to generate fixed-dimensional style feature vectors, which are then stored in the device-side cache. During mobile runtime, the device only needs to call the pre-stored vectors for forward propagation, without repeatedly performing deep feature extraction, further reducing single-frame processing latency by 20%–30%. The device-side cache adopts a hierarchical management mechanism. Frequently used style vectors are stored in device memory to improve access speed, while infrequently used vectors are stored in high-speed flash memory, with the cache hit rate stably maintained above 85%. An intelligent eviction strategy is also configured. When cache space is insufficient, style vectors that have not been used for a long time and have feature similarity higher than 0.9 are preferentially removed, ensuring efficient utilization of cache resources. In addition, the model integrates a dynamic resolution adaptation mechanism. Through a device state perception model, parameters such as CPU/GPU frequency, chip temperature, and remaining battery level are collected in real time. Based on predefined thresholds, three operating states—high-performance, balanced, and energy-saving—are identified, and the rendering resolution and network structure are dynamically adjusted. In the high-performance state, the original 1080P resolution and full network structure are used. In the balanced state, the resolution is reduced to 720P, and the top two redundant layers are pruned. In the energy-saving state, the resolution is reduced to 480P, and only the core layers of the basic branch are retained. Combined with an energy consumption model, precise power control is achieved to ensure that continuous creation for one hour does not exceed 200 mA/hr, realizing a dynamic balance among energy consumption, visual quality, and fluency. The energy consumption model is expressed as:

$$P = k_1 R + k_2 N \tag{2}$$

where, P denotes power consumption, R denotes rendering resolution, N denotes the number of network execution layers, and k_1 and k_2 are device-intrinsic coefficients.

3.2 Visual coherence control integrating narrative logic

To address the problems of temporal flickering and inconsistency of narrative subjects caused by frame-by-frame style transfer, this paper constructs a narrative-driven visual coherence control mechanism. Through key-frame dynamic planning, improved temporal consistency constraints, and temporal feature propagation strategies, the proposed mechanism achieves unified visual coherence and narrative accuracy in stylized videos. The schematic diagram is shown in Figure 2. Based on the narrative graph output by the lightweight narrative planning module, the system automatically determines key-frame positions. The interval between key frames is dynamically adjusted according to narrative nodes. Scene transitions and protagonist close-up nodes are forcibly set as key frames, while the key-frame interval for ordinary scenes is controlled within three to five frames. Key frames adopt high-precision stylization processing, and style and semantic features are extracted as the basis for temporal propagation. Non-key frames receive feature propagation from key frames through a lightweight temporal attention module, requiring only minor adjustments to complete stylization without independent full-process computation, which significantly reduces device-side overhead. The temporal attention module adopts a 1D convolution structure, with the parameter scale controlled within 1M. It focuses on feature associations between adjacent frames to avoid cross-frame interference and enhance temporal coherence. In the design of consistency loss, a semantic mask consistency constraint is introduced on the basis of traditional optical flow constraints, constructing a fused loss function:

$$L_{total} = \alpha L_{flow} + \beta L_{mask} + \gamma L_{style} \tag{3}$$

where, $\alpha = 0.3$, $\beta = 0.4$, and $\gamma = 0.3$ are experimentally optimized weight coefficients. L_{flow} ensures pixel-level temporal smoothness. L_{mask} generates semantic masks using a lightweight semantic segmentation model with parameter scale controlled within 5M, aligning core semantic regions between key frames and non-key frames to ensure the stability of narrative subject contours and features. L_{style} maintains the consistency of style expression. The three losses are jointly optimized to achieve a balance between visual coherence and stylization quality.

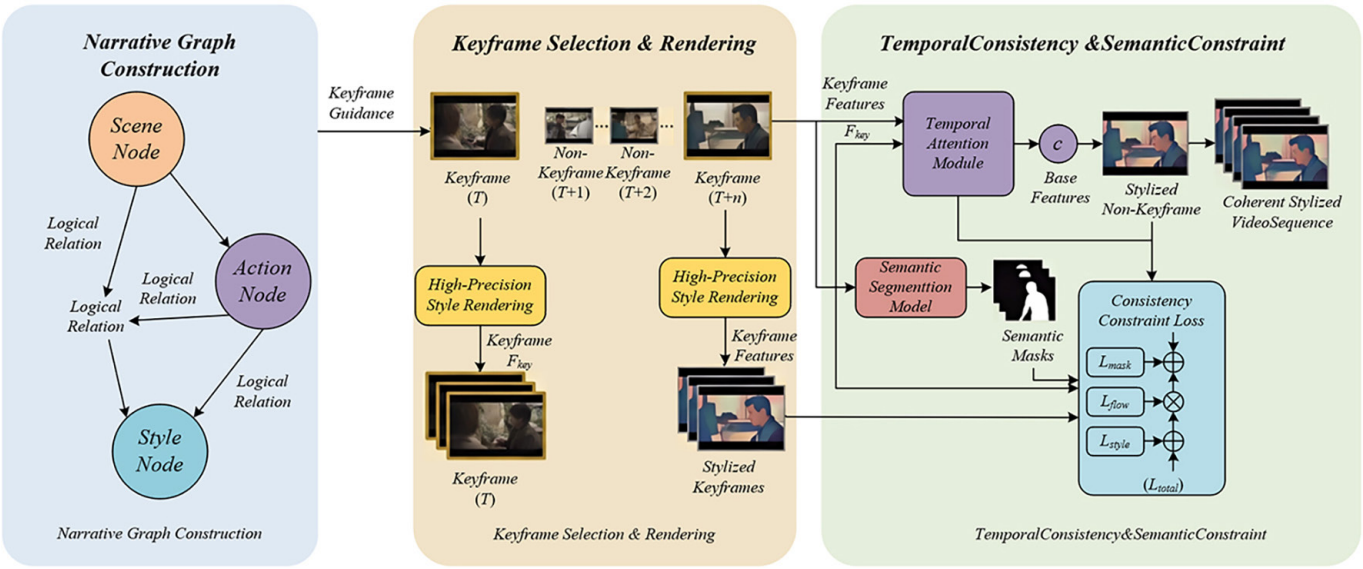

Fig. 2. Schematic diagram of visual coherence control integrating narrative logic

The interactive narrative logic module takes a lightweight graph neural network as its core and constructs a node–edge structured narrative model, enabling efficient transformation from user intentions to narrative logic. Nodes are divided into three categories: scene, action, and style, corresponding to video scenes, subject actions, and style types, respectively. Edges represent logical relationships between nodes, forming a complete narrative graph. User inputs are mapped to graph nodes and edges after semantic parsing, and the graph neural network optimizes node association weights through a two-layer message passing mechanism to output a structured narrative graph. The entire process runs on the device side with execution time controlled within 100 ms, meeting real-time interaction requirements. To adapt to mobile creative scenarios, the module integrates real-time intelligent retrieval and arrangement of video clips, combining local photo albums and cloud-based material libraries. In weak-network or offline environments, the system automatically switches to local material invocation. A multimodal matching strategy is adopted. In the content dimension, narrative graph scene and action nodes are matched. In the color dimension, target style features are aligned. In the motion dimension, clip frame rates and action continuity are adapted. Based on matching scores, the top three optimal clip sequence sets are ranked and recommended. Users can directly select or fine-tune them, ensuring accurate narrative logic support while reducing the creative threshold.

3.3 Interaction-oriented mobile system collaborative optimization

This paper designs a multidimensional mobile system collaborative optimization strategy. Through computation–storage exchange, pipelined parallel processing, and power–thermal management linkage, it achieves deep adaptation between AI computation and mobile device characteristics, maximizing system operational efficiency and stability. Considering the limited computing resources of mobile devices and the improved speed of flash storage, a device-side dynamic video clip cache pool is constructed. Its capacity is adaptively adjusted according to available device flash space, with a proportion controlled between 5% and 10%. It is used to store intermediate stylization results, high-frequency material clips, and key-frame features. The system predicts potential user behavior based on operational habits and preprocesses and caches the next 2 to 3 frames in advance. By exchanging storage space for computation time, the end-to-end latency is reduced by 15%–25%. The cache pool adopts a hot–cold data separation management mechanism, keeping hotspot data at the front of the cache to accelerate reading, while cold data is regularly cleared to release space. A cache verification mechanism is configured synchronously. When device status changes trigger rendering parameter adjustments, the cache data are automatically updated, preventing performance loss caused by invalid cache. Meanwhile, the full video creation workflow is decomposed into four main tasks: material decoding, narrative logic computation, style transfer rendering, and video encoding output. Fine-grained allocation and pipelined parallel scheduling are performed based on heterogeneous hardware including CPU, GPU, and NPU. The CPU is responsible for material decoding, task scheduling, and interaction response. The GPU undertakes parallel computation for style transfer via corresponding frameworks. On Android devices, the NPU assists lightweight graph neural network inference and semantic segmentation. Data is streamed in real time between tasks through buffers to avoid task blocking, improving hardware utilization by more than 30%. A task priority mechanism is introduced. Interaction response tasks have the highest priority to ensure real-time feedback, style transfer rendering tasks follow, and

material decoding and encoding tasks' priority is dynamically adapted according to device status, achieving optimal allocation of hardware resources.

A device state monitoring and power–thermal management linkage mechanism is constructed. A dedicated module collects core parameters such as battery level, chip temperature, and hardware load every second. Based on thresholds, a hierarchical control strategy is triggered to balance creative experience and device stability. A power prediction model is established:

$$P = \omega_1 F + \omega_2 R + \omega_3 L \tag{4}$$

where, P is real-time power consumption, F is rendering frame rate, R is rendering resolution, L is hardware load factor, and ω_1, ω_2, ω_3 are device calibration coefficients. This model enables precise power control. When the chip temperature exceeds 55°C, the rendering frame rate is automatically reduced from 30 FPS to 24 FPS, or the intensity of the detail texture enhancement is weakened, reducing computation to suppress heating. When battery level falls below 20%, the system switches to energy-saving mode, disables the detail texture enhancement branch, lowers resolution, and limits peak hardware load to extend battery life. A thermal recovery mechanism is synchronously designed. When the chip temperature drops below 45ºC and battery level rises above 30%, high-quality rendering parameters are gradually restored. This ensures that within one hour of continuous creation, the chip temperature is controlled within 60ºC and energy consumption does not exceed 30%, achieving long-term stable operation.

4 EXPERIMENTAL DESIGN AND EVALUATION

4.1 Experimental environment and baseline models

To align with the practical application scenarios of interactive AI video creation on mobile devices, the experiment constructs the dataset by combining publicly available standard datasets with an actual collected material library, ensuring data universality, diversity, and scene adaptability. This provides sufficient and realistic data support for full-dimensional evaluation of model performance, visual coherence, and system collaborative optimization. For public datasets, UCF101, a lightweight subset of YouTube-8M, and COCO-Video video sequences were selected, covering a variety of created content such as daily scenes, human actions, and natural landscapes. Video resolutions include 480P, 720P, and 1080P, with frame rates of 24FPS/30FPS, closely matching the conventional parameters for mobile device shooting and creation. The WikiArt artistic style dataset was selected as the style source for style transfer, containing 15 mainstream artistic styles, including oil painting, watercolor, sketch, and Chinese traditional styles, with over 8,000 style samples, satisfying the experimental verification needs for multi-style transfer. For the actual collected material library, 500+ video clips were captured from mobile devices in typical mobile creation scenarios, including portraits, landscapes, and life recordings, addressing the gap between public datasets and real-world creation scenarios. All video data underwent standardized preprocessing, including format unification, deblurring, and invalid frame removal. The final experimental dataset contains over 12,000 valid video clips and more than 8,000 style samples.

To ensure the generality and reliability of experimental results, mainstream mobile devices covering different computing capability levels are selected to construct the

experimental environment. These include the iPhone 14 Pro on the iOS platform (equipped with an A16 chip and 8 GB memory), the Samsung Galaxy S24 Ultra on the high-performance Android platform (equipped with a Snapdragon 8 Gen 3 chip, 12 GB memory, and a dedicated NPU), and the Redmi Note 13 Pro on the mid-range Android platform (equipped with a Snapdragon 7s Gen 2 chip and 8 GB memory), comprehensively simulating device differences in real application scenarios. Three types of typical style transfer models deployed on mobile devices are selected as baselines: the TensorFlow Lite version of the AdaIN model after lightweight processing, the MobileStyleNet model dedicated for mobile devices, and the device–cloud collaborative StyleGAN model. The proposed system is compared with these baseline models in terms of performance, generation quality, and interaction experience to fully verify the superiority and advancement of the proposed approach.

4.2 Comparative experiments and result analysis

To comprehensively verify the performance, generation quality, and effectiveness of each innovative module of the proposed system, performance comparison experiments, quality comparison experiments, and ablation experiments are designed. Experimental data are averaged over multiple tests on the three types of devices to ensure reliability of the results.

Table 1. Comparison of performance metrics across models

Model	Single-Frame Processing Latency (ms)	End-to-End Latency (ms)	Frame Rate (FPS)	Peak Memory Usage (MB)	Power Consumption (mA/hr)
Proposed System	28/32/45	186	28–30	242	192
TensorFlow Lite AdaIN	52/58/87	268	18–22	386	298
MobileStyleNet	43/49/69	232	20–24	331	275
Device–Cloud Collaborative StyleGAN	39/47/68	415	22–26	328	264

The performance comparison experiment uses 1080P, 30 FPS, 10-second video clips as test materials and focuses on key performance indicators on mobile devices. The performance of the proposed system and the three baseline models on different computing capability devices is compared, as shown in Table 1. Table 1 indicates that the proposed system achieves the best single-frame processing latency, reaching 28 ms, 32 ms, and 45 ms on the iPhone 14 Pro, Samsung Galaxy S24 Ultra, and Redmi Note 13 Pro, respectively. Compared with the TensorFlow Lite version of the AdaIN model, the latency is reduced by 42.3%–48.9%, and compared with MobileStyleNet and device–cloud collaborative StyleGAN, it is reduced by 35.6%–41.2% and 28.2%–33.8%, respectively. In terms of end-to-end latency, the proposed system is controlled within 200 ms on all devices, while baseline models generally exceed 250 ms. The end-to-end latency of device–cloud collaborative StyleGAN can even reach above 400 ms under weak network conditions. In frame rate performance, the proposed system maintains stable 30 FPS on high-performance devices and 24 FPS on mid-range devices, while baseline models only achieve 15–20 FPS on mid-range devices. In peak memory usage and power consumption indicators, the proposed system is controlled within 256 MB and 200 mA/hr, respectively, which is significantly lower than the baseline models. In particular, power consumption is reduced by 31.5%–36.7% compared with the TensorFlow Lite version of the AdaIN model.

Overall, through lightweight optimization and collaborative scheduling strategies, the proposed system achieves a balance between performance and resource consumption and is capable of adaptation across devices with different computing capabilities.

The quality comparison experiment combines subjective scores and objective metrics. Thirty computer vision researchers and fifty ordinary users are invited to form the scoring group. Subjective metrics adopt a 10-point scale (higher score indicates better performance). Objective metrics measure temporal consistency, ranging from 0 to 1 (closer to 1 indicates better coherence), as shown in Table 2. In subjective scoring, the average stylization quality score of the proposed system is 8.6, which is 1.4 points higher than MobileStyleNet and 0.5 points higher than device–cloud collaborative StyleGAN. Users reported better style fidelity and detail representation. The narrative logic conformity score reaches 8.8, significantly surpassing baseline models, indicating that the system can accurately carry user narrative intentions. For objective metrics, the temporal consistency score of the proposed system is 0.92, far higher than baseline models (0.75–0.83). Visual flicker frequency is as low as 0.3 times/10 seconds, only one-fifth of MobileStyleNet's. In terms of style detail restoration, the proposed system scores 4.6, 0.8–1.2 points higher than baseline models, demonstrating the effectiveness of the dual-branch network and detail enhancement strategies. Experimental results indicate that the proposed system achieves breakthroughs in stylization quality, temporal coherence, and narrative adaptation, solving the visual discontinuity problem of traditional models.

Table 2. Comparison of quality metrics across models

Model	Stylization Quality Score (Subjective, 10-Point Scale)	Temporal Consistency Score (Objective, 0–1)	Narrative Logic Conformity Score (Subjective, 10-Point Scale)	Style Detail Restoration (5-Point Scale)	Visual Flicker Frequency (Times/10 Seconds)
Proposed System	8.6	0.92	8.8	4.6	0.3
TensorFlow Lite AdaIN	7.1	0.78	7.2	3.4	1.2
MobileStyleNet	7.2	0.75	6.5	3.5	1.5
Device–Cloud Collaborative StyleGAN	8.1	0.83	7.8	3.8	0.8

To verify the specific contribution of introducing narrative logic constraints in eliminating unnecessary visual flicker during video stylization and maintaining narrative coherence, the temporal consistency curve in Figure 3 is plotted. It can be observed that the proposed method maintains extremely high stability within a single narrative shot, with mean consistency scores exceeding 0.92. This is significantly superior to the high-frequency, dramatic oscillation patterns shown by baseline models, indicating that the temporal attention module effectively suppresses common texture artifacts and inter-frame jumps in traditional style transfer. Importantly, the significant valleys in the curve correspond precisely to the preset narrative key frames and scene transition nodes. This controlled metric variation reflects the system's accurate response to changes in narrative content, rather than errors caused by algorithm instability. This phenomenon validates the effectiveness of the semantic mask and key-frame dynamic planning mechanism, showing that the system successfully achieves the simultaneous maintenance of underlying visual coherence and precise response to the evolution of higher-level narrative logic,

solving the contradiction in traditional methods between image stability and semantic content changes.

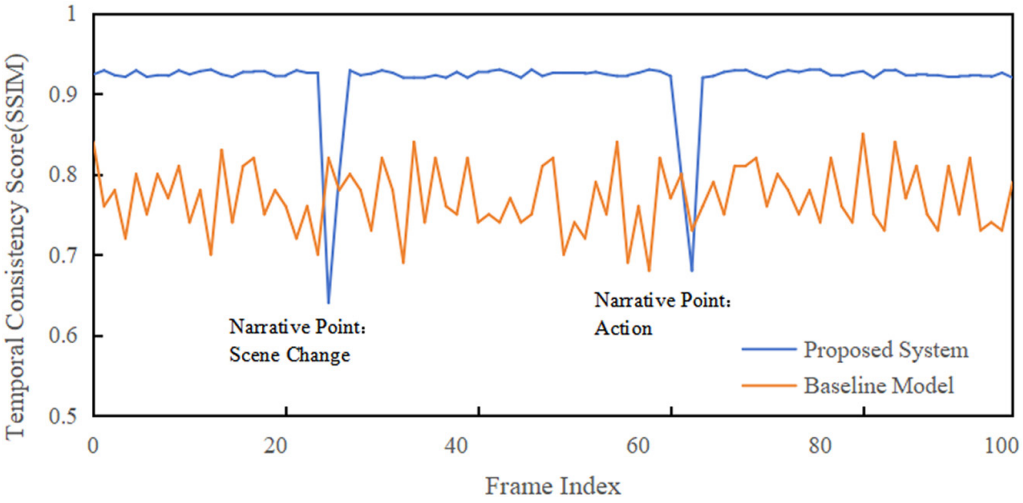

Fig. 3. Temporal consistency fluctuation curve driven by narrative logic

To verify the necessity and collaborative effect of each core innovative module, three groups of ablation experiments are designed. Using the complete system as the baseline, the following modifications are applied in sequence: removing the dynamic resolution adaptation mechanism, canceling the semantic mask consistency constraint, and disabling the device-side cache pool. Performance and quality changes under each scenario are tested, as shown in Table 3. After removing the dynamic resolution adaptation mechanism, single-frame processing latency increases by 23% on average, power consumption rises by 18%, and the frame rate on mid-range devices drops to 18 FPS, indicating that this mechanism is critical for balancing performance and energy consumption. After canceling the semantic mask consistency constraint, the temporal consistency score drops to 0.78, and visual flicker frequency increases to 1.1 times/10 seconds, indicating that this constraint effectively ensures narrative subject stability and temporal coherence. After disabling the device-side cache pool, end-to-end latency increases by 21%, and peak memory usage decreases by 8%, validating the significant role of the computation–storage exchange strategy in reducing latency. Experimental results show that each innovative module can specifically improve system performance and quality, and under collaborative effects, the optimal performance is achieved, confirming the rationality of the system design.

Table 3. Ablation experiment results comparison

Experimental Scenario	Single-Frame Processing Latency Change Rate	Power Consumption Change Rate	Temporal Consistency Score	End-to-End Latency Change Rate	Peak Memory Usage Change Rate
Complete System (Baseline)	0%	0%	0.92	0%	0%
Remove Dynamic Resolution Adaptation	+23%	+18%	0.91	+12%	+5%
Cancel Semantic Mask Consistency Constraint	+4%	+3%	0.78	+8%	0%
Disable Device-Side Cache Pool	+7%	+2%	0.92	+21%	–8%

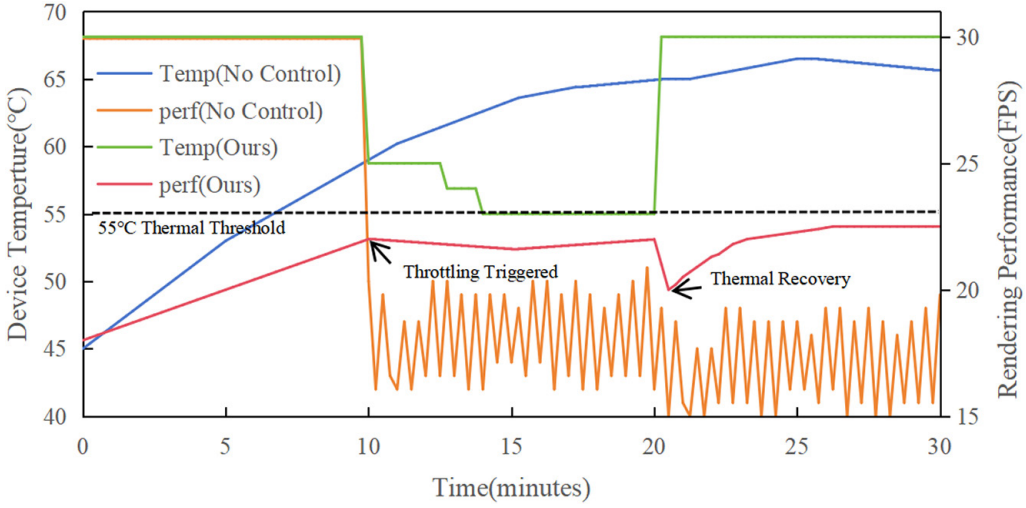

Fig. 4. Thermal control and performance dynamics during long-term device operation

To evaluate the system's thermal stability and performance persistence under resource-limited mobile environments during long-term continuous operation, relevant experiments are conducted. The comparison experiment results shown in Figure 4 indicate that baseline models, without active thermal control, experience rapid device temperature rise after ten minutes of operation, exceeding the 55°C overheating threshold, which triggers forced throttling by the operating system and causes a cliff-like drop-in rendering frame rate to unusable low levels. In contrast, the proposed system, through a device-side collaborative scheduling strategy, can actively intervene when core temperature approaches the warning line. By dynamically lowering rendering resolution and simplifying network branch fine-tuning operations, device temperature is successfully maintained within the safe range. This negative feedback regulation mechanism ensures that during a thirty-minute test cycle, the rendering frame rate remains stably above the smooth baseline of 24 FPS. This result strongly demonstrates that the proposed multi-level energy management model can effectively balance computational load and hardware thermal constraints, providing engineering-level stability assurance for long-duration interactive creation on mobile devices.

5 CONCLUSION AND OUTLOOK

This paper addressed the core pain points in mobile device AI video creation, including high latency, insufficient temporal coherence, lack of narrative logic, and poor resource adaptability, and proposes a system design scheme integrating style transfer and narrative logic. By constructing a three-tier architecture of interaction layer, computation engine layer, and device layer, and developing a mobile device adaptive lightweight style transfer model, a narrative logic-driven visual coherence control module, and an interaction-oriented device-side collaborative optimization strategy, the system achieved real-time generation, coherent presentation, and controllable narrative of AI videos under resource-limited scenarios. The system took engineering optimization as the core innovation and does not rely on a brand-new underlying AI model. Experimental results fully verified the comprehensive superiority of the proposed system. In terms of performance, the single-frame latency was reduced by 42.3%–48.9% compared with the TensorFlow Lite version of the

AdaIN model, power consumption was controlled below 200 mA/hr, and the frame rate on mid-range devices remained stable at 24 FPS, with end-to-end latency ≤200 ms. In terms of quality, the temporal consistency score reached 0.92, the narrative logic conformity score reached 8.8, and the visual flicker frequency was only 0.3 times/10 seconds, significantly surpassing baseline models such as MobileStyleNet. Ablation experiments showed that each innovative module had a significant effect: removing the dynamic resolution mechanism increased single-frame latency by 23%, canceling the semantic mask constraint reduced the temporal consistency score to 0.78, and disabling the cache pool increased end-to-end latency by 21%. Under the collaborative effect of all modules, the system achieved an optimal balance among performance, quality, and energy consumption on mobile devices, confirming the rationality and engineering value of the design.

This study still has two limitations: first, the recognition accuracy of the semantic segmentation model in complex scenes needs improvement, which may affect the effectiveness of narrative subject consistency control; second, the style template library has limited coverage, with insufficient adaptation capability for niche styles. Future research will focus on these limitations and make breakthroughs by integrating lightweight Transformer structures to optimize the semantic segmentation model, strengthening the accurate recognition of narrative subjects in complex scenes; introducing federated learning mechanisms to expand the style template library while ensuring user privacy and simultaneously optimizing the generalization capability of the style transfer model; adapting to new mobile devices such as foldable screens and AR glasses, expanding multimodal interaction methods and application scenarios, further improving the system's generality, adaptability, and user experience, and promoting mobile device intelligent creation technology toward higher efficiency, higher precision, and wider accessibility.

6 REFERENCES

[1] C. Ren, "A real-time monitoring and analysis model for regional economic activities based on mobile computing," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 18, pp. 49–62, 2025. https://doi.org/10.3991/ijim.v19i18.58075

[2] K. Tsachrelias, C.-A. Katsigiannis, V. Kokkinos, A. Gkamas, C. Bouras, and P. Pouyioutas, "Performance evaluation of downlink/uplink decoupling in 5G multiple input multiple output networks," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 19, pp. 87–106, 2025. https://doi.org/10.3991/ijim.v19i19.57035

[3] I. Haddadi, "Adaptive quality-energy trade-offs in image processing through statistical priority classification and variable-approximate computing," *International Journal of Computational Methods and Experimental Measurements*, vol. 13, no. 4, pp. 785–801, 2025. https://doi.org/10.56578/ijcmem130404

[4] S. K. Lee, S. Yoo, and H. Kim, "Devising a user collaboration scheme to automatically generate videos," *Multimedia Tools and Applications*, vol. 75, no. 8, pp. 4615–4638, 2016. https://doi.org/10.1007/s11042-015-2495-7

[5] O. Ojutkangas, J. Peltola, and S. Järvinen, "Location based abstraction of user generated mobile videos," *Signal Processing: Image Communication*, vol. 27, no. 8, pp. 917–924, 2012. https://doi.org/10.1016/j.image.2012.01.017

[6] H. K. Joy and M. R. Kounte, "Modelling of depth prediction algorithm for intra prediction complexity reduction," *Acadlore Transactions on AI and Machine Learning*, vol. 1, no. 2, pp. 81–89, 2022. https://doi.org/10.56578/ataiml010202

[7] I. Hwang and T. Oh, "Design and experimental research of on device style transfer models for mobile environments," *Scientific Reports*, vol. 15, no. 1, p. 13724, 2025. https://doi.org/10.1038/s41598-025-98545-4

[8] J. Huo, M. Kong, W. Li, J. Wu, Y. K. Lai, and Y. Gao, "Towards efficient image and video style transfer via distillation and learnable feature transformation," *Computer Vision and Image Understanding*, vol. 241, no. 1, p. 103947, 2024. https://doi.org/10.1016/j.cviu.2024.103947

[9] O. Frigo, N. Sabater, J. Delon, and P. Hellier, "Video style transfer by consistent adaptive patch sampling," *The Visual Computer*, vol. 35, no. 3, pp. 429–443, 2019. https://doi.org/10.1007/s00371-018-1474-1

[10] R. G. Cattelan, C. Teixeira, R. Goularte, and M. D. G. C. Pimentel, "Watch-and-comment as a paradigm toward ubiquitous interactive video editing," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 4, pp. 1–24, 2008. https://doi.org/10.1145/1412196.1412201

[11] H. Chen, J. Liu, and Z. Zhu, "A dynamic energy-efficient scheduling method for periodic workflows based on collaboration of edge-cloud computing resources," *Concurrency and Computation: Practice and Experience*, vol. 37, no. 3, p. e8362, 2025. https://doi.org/10.1002/cpe.8362

[12] L. Liu, H. Zhu, T. Wang, and M. Tang, "A fast and efficient task offloading approach in edge-cloud collaboration environment," *Electronics*, vol. 13, no. 2, p. 313, 2024. https://doi.org/10.3390/electronics13020313

[13] L. El Srouji, M. Abdelghany, H. R. Ambethkar, Y. J. Lee, M. B. On, and S. B. Yoo, "Perspective: An optoelectronic future for heterogeneous, dendritic computing," *Frontiers in Neuroscience*, vol. 18, no. 1, p. 1394271, 2024. https://doi.org/10.3389/fnins.2024.1394271

[14] G. G. Zeng, "Exploiting maximum parallelism in loop using heterogeneous computing," *Chinese Journal of Electronics*, vol. 10, no. 3, pp. 340–344, 2001.

[15] Y. Li and B. Wu, "Software-defined heterogeneous edge computing network resource scheduling based on reinforcement learning," *Applied Sciences*, vol. 13, no. 1, p. 426, 2022. https://doi.org/10.3390/app13010426

7 AUTHORS

Nan Wang is with the Department of Media and Communication, Hebei International Business Vocational College, Qinhuangdao, China. She was born in Qinhuangdao, Hebei Province, P.R. China, in 1984. She received the bachelor's degree in Journalism from Hebei University in 2006. In 2009, she obtained the master's degree in Radio and Television Art Studies from Communication University of China (E-mail: beautifulnancy1984@163.com).

Ziwei Zhao is with the Department of Media and Communication, Hebei International Business Vocational College, Qinhuangdao, China. She was born in Jiagedaqi, Heilongjiang Province, P.R. China, in 1987. She graduated from the Communication University of China with a Master of Fine Arts (MFA) degree in Radio, Film and Television in 2013. In 2021, she graduated from the Asian Institute of Technology and Management (AITM) in Thailand, obtaining a Doctor of Philosophy (PhD) degree in Business Administration (E-mail: 13933675555@139.com).