ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

Research on Application Scenarios and Implementation Pathways of Automated Video Editing Integrating CLIP/YOLO and Large Models under Cloud-Edge Collaboration

Jiahao Cao

Zhengzhou University, Zhengzhou, Henan, China

Abstract: *To address critical industry pain points in the field of automated video editing—specifically the limited computing power of purely local modes, the prominent privacy risks of purely cloud-based modes, insufficient technology integration, and a lack of multi-scenario adaptability—this study draws upon theories of technology integration and data security to propose an automated video editing framework. This framework integrates CLIP and YOLOv8n with Large Models within a Cloud-Edge collaborative paradigm. The research constructs a three-tier technical architecture comprising "Edge-side Perception, Cloud-side Decision-making, and Collaborative Scheduling." It designs four core functional modules: intelligent material parsing, automated script generation, interactive rendering, and data security protection, thereby establishing a closed technical loop of "Lightweight Perception + Intelligent Decision-making + Elastic Collaboration." Furthermore, targeting three core scenarios—individual creators, enterprise users, and government media—the study formulates differentiated adaptation paths and implementation strategies, focusing on lightweight embedding, customized deployment, and open cooperation, respectively. This work provides a novel perspective for resolving the core industry conflict of "Compute-Security-Adaptability" and lays a solid theoretical foundation for subsequent engineering implementation and technological optimization.*

Keywords: *Edge-Cloud Collaboration; Automated Video Editing; CLIP Model; YOLOv8n; Large Model Fusion*

1. Introduction

1.1 Research Background

Driven by the explosive growth of the short video industry, content production efficiency has emerged as a critical bottleneck restricting industry development. According to the 54th Statistical Report on Internet Development in China released by CNNIC, as of June 2024, the number of short video users in China reached 1.026 billion, accounting for 94.8% of the total internet population, with an average daily usage time exceeding 120 minutes. Faced with massive volumes of video footage, the conflict between efficiency, cost, and scalable production within traditional manual editing models has become increasingly prominent.

From a technical perspective, the iteration of Generative AI and Computer Vision technologies offers new possibilities for automated video editing. Data from IDC indicates that the global market size for AI content generation exceeded $8 billion in 2024, with the Compound Annual Growth Rate (CAGR) of automated video processing technology reaching 47.2%. However, purely cloud-based modes face risks of data privacy leakage, while purely local modes are constrained by hardware computing power, making it difficult to support real-time inference for complex models. Consequently, the Cloud-Edge collaborative architecture, utilizing a division of labor characterized by "lightweight processing + cloud empowerment," has emerged as a key technical pathway to balance efficiency and security.

1.2 Research Significance

Theoretical Significance: This study pioneers the construction of a fusion framework integrating CLIP/YOLO visual perception models with Large Models. By transcending the technical limitations of traditional automated video editing, which relies solely on rule-based templates, it offers a novel

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

theoretical perspective for multimodal content understanding.

Practical Significance: Targeting scenarios such as short video production, corporate publicity, and government media, this study proposes an implementable Cloud-Edge collaborative solution that enhances video production efficiency and reduces labor costs. The research framework provides a valuable reference for technical selection for small and medium-sized content creators, enterprise-level users, and government media, thereby driving the intelligent transformation of the video industry.

2. Related Concepts and Theoretical Basis

2.1 Definition of Core Concepts

2.1.1 Automated Video Editing

Automated video editing refers to a technical system based on Computer Vision (CV) and Natural Language Processing (NLP) that automatically executes the shot segmentation, content understanding, clip selection, and splicing assembly of video materials. Compared with traditional manual editing, its core advantage lies in achieving the full-process automation of "Understanding-Decision-Generation" through algorithms. Typical applications include intelligent variety show editing and short news video production.

2.1.2 Edge-Cloud Collaboration and Core Supporting Technologies

Cloud-Edge Collaboration refers to an architectural pattern where computing tasks are dynamically allocated between local terminals and cloud servers[1]. In this study:YOLOv8n (Lightweight Version) serves as the core detection model on the edge side, capable of achieving real-time object recognition at over 30 frames per second (FPS)[2].The CLIP Model is responsible for cross-modal semantic understanding, establishing a mapping relationship between video content and user requirements by aligning image and text feature vectors. Cloud-based Large Models (such as GPT-4o and ERNIE Bot) undertake complex logical reasoning and creative generation tasks. Together, these three components constitute a technical closed loop of "Perception-Understanding-Decision."

2.2 Core Theoretical Support

2.2.1 Technology Fusion Theory

According to Technology Integration Theory, the cross-fertilization of different technological fields catalyzes new innovation paradigms. In this study, the integration of Computer Vision technologies (YOLO/CLIP) with Generative Large Models overcomes the bottlenecks of single technologies in semantic understanding or real-time performance[3]. For instance, YOLO's object detection results serve as semantic anchors for CLIP, enhancing the precision of cross-modal retrieval; meanwhile, Large Models generate editing scripts based on visual features that align with narrative logic, achieving an exponential amplification of technical capabilities[4].

2.2.2 Data Security and Content Production Efficiency Theory

The "Cannikin Law" (or Buckets Effect) in data security posits that a system's security level is determined by its weakest link. The Cloud-Edge collaborative architecture effectively mitigates data leakage risks by performing sensitive data preprocessing (such as face blurring) locally and transmitting only feature vectors rather than raw footage to the cloud. Simultaneously, according to Metcalfe's Law, the value of a network grows exponentially with the number of nodes. The computing power sharing and model iteration under the Cloud-Edge collaborative mode will drive a scalable increase in content production efficiency.

3. Current Status and Core Issues of the Automated Editing Industry

3.1 Industry Development Status and Core Demands

3.1.1 Mainstream Scenarios and Technical Modes

Currently, automated video editing is primarily categorized into three scenarios:

Consumer-side (C-end) Individual Creation: Represented by CapCut's "Auto Cut" function, which

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

relies on template matching technology to achieve rapid 10-second video production.

Business-side (B-end) Corporate Publicity: Represented by Wondershare Filmora's batch editing tools, which support the mass generation of product promotional videos from multiple materials.

Government-side (G-end) Media Production: Represented by CCTV's AI news editing system, which completes sports highlights production through speech recognition and shot matching.

In terms of technical modes, purely cloud-based solutions are represented by Runway ML, emphasizing model complexity but suffering from high latency. Conversely, purely local solutions are represented by CapCut's mobile app, focusing on real-time performance but constrained by limited computing power[5].

3.1.2 Core Demands

According to the 2024 Ecological Report of Short Video Creators in China, creators have three core demands:

Production Efficiency: 78% of creators hope to shorten the production time for a single video from 2 hours to under 30 minutes.

Content Quality: 65% of users believe current AI editing works suffer from issues such as confused narrative logic and rigid shot transitions.

Privacy and Security: 52% of enterprise users and government media refuse to use purely cloud-based editing tools due to concerns over material leakage.

3.2 Key Issues in Existing Modes

3.2.1 Limitations of Pure Local and Pure Cloud Modes

The purely local mode is constrained by terminal computing power and is unable to run complex algorithms like Large Models. This results in editing effects that are highly dependent on preset templates and lack creative flexibility. For instance, CapCut's "Auto Cut" function only supports fixed combinations of transitions and background music, making it difficult to meet personalized needs.

The purely cloud-based mode faces three major pain points:

Latency Issues: The average time to upload 4K video materials to the cloud exceeds 5 minutes.

Cost Issues: Cloud inference fees are approximately 3-5 times higher than local processing.

Compliance Risks: Regulations such as the EU GDPR require video data storage to meet regional restrictions, increasing the complexity of cross-border data transmission.

3.2.2 Insufficient Technical Fusion and Demand Adaptation

Currently, most solutions merely implement a simple superposition of technologies rather than deep integration. For example, while some tools integrate YOLO object detection, they fail to combine it with CLIP's semantic understanding capabilities, resulting in an inability to recognize complex scenes such as a "smiling person." Furthermore, existing products are mostly oriented towards C-end users, with insufficient customization capabilities for enterprise-level users and government media. Deficiencies such as a lack of multi-material batch processing and brand style unification make it difficult to meet the professional demands of the B-end and G-end markets.

4. Design of Edge-Cloud Collaborative Automated Editing Scheme

4.1 Design Philosophy and Core Objectives

Guided by the core design philosophy of "Data Stays Local, Knowledge Migrates to the Cloud," this proposed solution addresses the inherent conflict between the computational bottlenecks of purely local infrastructure and the privacy security risks of purely cloud-based models. It constructs an automated video editing technical framework based on Cloud-Edge Heterogeneous Computing. The specific core objectives are as follows:

4.1.1 Computation Optimization Objective

To achieve a computational allocation strategy characterized by "Edge-Centric Perception and Cloud-

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

Centric Decision-Making," lightweight models such as YOLO are deployed on terminal GPUs/NPUs. These models perform computationally intensive tasks—specifically real-time object detection and scene recognition—locally, thereby reducing bandwidth pressure and latency associated with raw video transmission at the source. Simultaneously, complex inference tasks, including CLIP-based cross-modal feature extraction, are migrated to cloud-based elastic computing pools. This establishes a technical closed loop of "Terminal Preprocessing + Cloud Deep Computation," which ensures real-time editing performance while effectively mitigating the high costs associated with full cloud-side inference.

4.1.2 Intelligent Decision-Making Objective

To enhance the semantic understanding capabilities of automated video editing, this study introduces a Large Language Model (LLM)-driven narrative orchestration mechanism, transcending the logical limitations inherent in traditional template-based editing[6]. By leveraging CLIP to extract cross-modal features from video frames and text, and integrating the LLM's profound comprehension of narrative structure and emotional pacing, a decision-making chain of "Visual Perception – Semantic Parsing – Narrative Generation" is constructed. The LLM automatically generates editing scripts based on user requirements, intelligently matching key shots, transitions, and background music. This facilitates a paradigm shift in output from "mechanical assembly" to "logical storytelling," thereby achieving a higher level of creative intelligence.

4.1.3 Security Compliance Objective

To mitigate privacy risks during material transmission, a security mechanism based on metadata interaction is established. In the Cloud-Edge collaborative workflow, the terminal transmits only metadata—such as scene tags, character features, and timestamps—to the cloud, rather than the complete raw video footage. Upon receiving editing instructions from the cloud, the terminal performs the final rendering locally. This interaction mode, adhering to the principle of "Data Stays Local," not only leverages the decision-making advantages of cloud-based Large Models but also blocks leakage paths for sensitive materials at the source, thereby meeting both individual privacy needs and industry compliance requirements.

4.1.4 Scenario Adaptation Objective

Adopting a modular architecture design, this solution flexibly addresses the core demands of diverse user groups. For individual users, lightweight mobile tools are provided to support rapid short-video editing. For enterprise users, a Cloud-Edge collaborative batch production pipeline is constructed to meet the scalable requirements of professional content, such as advertising. For government scenarios, a metadata-based security mechanism is utilized to enable the compliant processing of sensitive content. This ensures that a unified technical framework can be dynamically configured based on the specific computing constraints, privacy levels, and creative objectives of different scenarios.

4.2 Overall Technical Architecture and Division Logic

4.2.1 Edge Perception Layer: Lightweight Visual Processing Module

Serving as the foundational perception unit of the technical architecture, the Edge-Side Perception Layer primarily deploys lightweight YOLO series models and CLIP cross-modal models to undertake preliminary processing tasks for raw video. Its core functions encompass video frame-level feature extraction, object recognition, and semantic annotation. Through lightweight model optimization, the layer adapts to terminal hardware computing power, thereby preventing complex calculations from creating excessive performance overhead on local devices. Crucially, this layer does not transmit complete raw video to the cloud; instead, it exclusively synchronizes processed feature information and semantic tags. This mechanism ensures data privacy security at the source while simultaneously alleviating network bandwidth pressure.

4.2.2 Cloud Decision Layer: Large Model-Driven Creative Core

The Cloud-Based Decision-Making Layer serves as the intelligent core of the automated editing system. Leveraging the Natural Language Understanding (NLU) and logical reasoning capabilities of Large Models, this layer achieves the intelligent generation of editing scripts. Upon receiving feature and tag information transmitted from the edge, it integrates user input instructions (such as video style, purpose, and duration) to automatically construct a shot assembly schema that aligns with narrative logic. This includes core content such as shot sequencing, transition matching, and caption generation. Furthermore, the layer supports personalized customization, capable of aligning with specific

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

requirements like enterprise brand guidelines or government communication protocols. This enables the generation of scenario-adapted editing schemes, effectively balancing universality with specificity.

4.2.3 Collaborative Scheduling Layer: Resource Adaptation Hub

Acting as the critical nexus bridging the edge and the cloud, the Collaborative Scheduling Layer functions primarily to facilitate the dynamic allocation of computational tasks and resource adaptation. By employing a flexible scheduling mechanism, this layer dynamically adjusts the division of labor between the cloud and the edge based on real-time factors such as terminal hardware configurations, network transmission conditions, and task complexity. Under conditions of poor network connectivity or sufficient terminal computing power, the system increases the proportion of local processing. Conversely, when tasks require complex semantic understanding or creative generation, the system leverages cloud-based computing power to execute core functions. Simultaneously, model version management and compatibility mechanisms are established to ensure technical adaptability across diverse terminal devices and system environments, thereby laying a solid foundation for the widespread application of the proposed solution.

4.3 Design of Core Functional Modules

4.3.1 Intelligent Material Parsing Module

Supported by the technical foundation of the Edge-Side Perception Layer, this module implements the automated cleaning and structuring of materials. Its core functions include: Blur detection based on Laplacian Variance to automatically reject low-quality shots; Audio climax extraction based on the RMS (Root Mean Square) Energy Envelope to achieve precise musical beat matching; Semantic label construction to provide accurate "Visual-Auditory" feature indexing for subsequent script generation[7].

4.3.2 Automatic Script Generation Engine

As the core of the Cloud-Based Decision-Making Layer, this engine focuses on the precise alignment of "Requirements-Materials-Creativity." It supports adaptation to multiple mainstream styles, utilizing Chain-of-Thought (CoT) technology to guide Large Models in generating targeted scripts based on distinct scenario characteristics (e.g., the "Pain Point-Solution" logic for product promotion, or the rigorous logic required for government publicity)[8].Simultaneously, it incorporates cross-modal semantic alignment algorithms to ensure that shot selection is highly consistent with the emotional tone of the script, thereby enhancing the narrative tension of the content.

4.3.3 Interactive Optimization and Rendering Module

Balancing automation with user autonomy, a visual interactive interface is designed to support users in performing secondary adjustments to AI-generated scripts. This includes operations such as shot replacement, sequence reordering, and transition modification. The rendering module adapts to multi-terminal output requirements, supporting the flexible selection of mainstream video formats and resolutions. This ensures that generated videos can directly meet the publishing requirements of various distribution platforms, thereby reducing subsequent processing costs for users.

4.3.4 Data Security Protection Module

Centered on the core principle of "Data Localization," a full-process security protection mechanism is designed. Raw video materials remain stored on the local terminal throughout the process, with only non-sensitive feature information transmitted via encryption. The cloud retains no complete material data, processing script generation tasks only temporarily. An operation log traceability mechanism is established to ensure the full traceability of data flow. This comprehensively meets the privacy protection and compliance requirements of individual, enterprise, and government scenarios.

5. Application Scenario Adaptation and Practice Paths

5.1 Core Application Scenario Adaptation Design

5.1.1 C-End Scenario: Individual Short Video Creation

Addressing the dual demands of individual users for creative convenience and content personalization, this study constructs an automated closed-loop system covering "Perception Analysis – Intelligent Matching – Terminal Rendering":

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

Service Architecture Lightweighting: We implement an algorithm containerization strategy, converting visual analysis and non-linear editing engines into independent API interfaces. The mobile end functions solely for instruction interaction. By replacing raw material transmission with metadata flow and seamlessly embedding the system via plugins, the underlying model loading process is effectively shielded from the user.

Processing Flow Automation: A quality control gate based on physical features is embedded. By calculating the Laplacian Variance and Mean Gray Value of video frames, the system automatically filters blurred or overexposed low-quality frames during the pre-rendering stage, achieving zero-intervention material cleaning.

Style Mapping Adaptation: A cross-modal emotional alignment mechanism is introduced, matching audio tonality by analyzing the distribution of "emotion tags" in video frames. Integrating YOLOv8n intelligent ROI cropping and optical flow frame interpolation algorithms, the system automatically optimizes composition and smooths video fluidity, enabling the batch generation of cinematic-quality content.

5.1.2 B-End Scenario: Enterprise Publicity Video Production

Facing the rigid demand for scalable output and unified brand image on the enterprise side, the solution designs an industrial operational pipeline of "Configuration-Oriented – Parallel Computing – Standardized Output":

Visual Standardization: A parametric rendering pipeline is established, supporting the dynamic loading of Enterprise Visual Identity (VI) configuration files to achieve automatic unification of watermarks and color tones. YOLOv8n is integrated for subject reconstruction and cinematic aspect ratio cropping, forcibly locking core visual subjects and solidifying output resolution and composition ratios to guarantee brand asset consistency.

High-Throughput Parallel Computing: Based on a Batch-Ready architecture optimization of the visual engine, terminal-side GPU power is utilized for batch semantic inference on multi-channel materials. Driven by global configuration, the system achieves full-link automated batch processing, spanning from material parsing and parameter configuration to final rendering.

Professional Narrative Customization: Industry-specific narrative templates are introduced. Large Models are leveraged to orchestrate storyboard scripts adapted to specific business logic (such as technical interpretation or product demonstrations). Combined with smooth slow-motion generated by optical flow technology, this significantly enhances the professional quality and visual impact of product close-up shots.

5.1.3 G-End Scenario: Government Media Content Dissemination

In view of the strict standards for content rigor and data sovereignty in the government sector, a dedicated link of "Physical Isolation – Rule Constraint – Agile Response" is constructed:

Content Generation Compliance: A constraint generation system based on a domain knowledge base is established. Government-specific Prompt templates and sensitive word filtering gateways are pre-set in the cloud-based Large Model decision layer, supplemented by a local expert rule engine for dual verification. Once risks or network restrictions are detected, the system immediately downgrades to an offline preset orchestration mode, ensuring the absolute safety and controllability of political narratives.

Data Physical Isolation: A full-link data localization strategy is executed, strictly adhering to the baseline that "raw materials do not leave the domain." Only desensitized JSON metadata is permitted for Edge-Cloud interaction. The rendering process is forcibly completed on the local terminal, eliminating the risk of uploading sensitive materials to the cloud from the underlying architecture, thereby complying with government data security review requirements.

Instant Response to Emergencies: Terminal-side resource scheduling is optimized. Integrating FFmpeg ultrafast presets and multi-threaded parallel rendering technology, the physical rendering time for breaking news is drastically reduced, meeting the minute-level timeliness requirements for policy releases and emergency reporting.

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

5.2 Practice Paths by Subject

5.2.1 Individual Users: Lightweight Embedding Path

For consumer-facing users, a lightweight implementation path of "Host Application + Plugin Container" is established.

Service Encapsulation: Core Python algorithm scripts are encapsulated into mobile plugin containers. By utilizing API hooks to integrate with mainstream short video platforms, users are relieved from the need to configure underlying environments.

Automatic Fault Tolerance: Anomaly monitoring and service degradation mechanisms are deployed on the client side. In the event that audio rhythm analysis or visual feature extraction is obstructed, the system automatically switches to a default fixed-step slicing mode, ensuring stability and output rate with zero manual intervention.

5.2.2 Enterprise Users: Customized Deployment Path

For business-facing users, a customized implementation path of "Docker Containerization + Configuration-Driven Deployment" is implemented.

Environment Isolation: Independent computing units containing FFmpeg and PyTorch environments are delivered via Docker images, supporting GPU virtualization deployment within enterprise private cloud environments.

Parameter Injection: Business logic is managed based on a global CONFIG dictionary. This allows enterprises to dynamically set brand watermarks, rendering bitrates, and encoding formats by modifying configuration files, while seamlessly integrating with existing CMS (Content Management Systems) via RESTful interfaces.

5.2.3 Industry Ecosystem: Open Cooperation Path

For the industry ecosystem, a collaborative implementation path of "Semantic Metadata Standardization" is promoted.

Protocol Unification: A standardized JSON semantic description protocol is formulated to regulate the field definitions for "emotion," "category," and "temporal tags" of video frames, thereby breaking down data barriers between different editing tools.

Algorithm Co-construction: Low-level CLIP model fine-tuning interfaces are opened to support industry developers. This enables the optimization of visual encoder weights using vertical domain data (e.g., medical, education), facilitating the construction of a ubiquitous automated editing algorithm ecosystem.

5.3 Implementation Guarantee Design

5.3.1 Technical Adaptation Guarantee

Building a robust system for complex hardware environments.

Heterogeneous Computing Abstraction: An Edge-side Hardware Abstraction Layer (HAL) is constructed to automatically detect computing units upon startup. The system dynamically loads FP16 quantized models or degrades to pure OpenCV algorithms, ensuring availability even on low-computing-power terminals.

Network Circuit Breaking Mechanism: A timeout circuit breaking strategy is designed within the collaborative scheduling layer. Once the cloud API response latency exceeds a set threshold, the system immediately assumes control, switching to a local preset expert rule library to execute fallback logic, thereby guaranteeing service continuity in weak network environments.

5.3.2 Security Compliance Guarantee

Constructing an active defense system based on the principle of "data sovereignty".

Non-intrusive Interaction: Metadata-level information flow is implemented, permitting only the upload of desensitized semantic feature vectors for inference. Raw video materials are forcibly locked in local storage, eradicating cloud leakage risks at the root.

Full-link Auditing: An operation log audit module is embedded in the local rendering engine. It fully

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

records the decision basis and material sources of the automated editing process, meeting the data compliance and traceability requirements of government and enterprise scenarios.

5.3.3 Iterative Optimization Guarantee

Designing a data-driven mechanism for continuous algorithmic evolution.

Parameter Hot Update: The system supports the cloud delivery of weight configuration files, enabling the dynamic adjustment of threshold ratios for "positive/negative" semantics in the visual scoring model. This allows for rapid responses to changes in aesthetic trends without the need for version releases.

Adaptive Calibration: A/B testing and negative feedback loop mechanisms are established. By analyzing user "undo/modify" behaviors on generated results, the system reverse-corrects the prompt engineering and ranking logic of the CLIP model, achieving continuous adaptive optimization of algorithm precision.

6. Conclusion

6.1 Research Conclusion

Addressing critical industry pain points in the field of automated video editing—specifically the limited computing power of purely local modes, the high privacy security risks of purely cloud-based modes, insufficient technological integration, and a lack of multi-scenario adaptability—this study, grounded in theories of technology integration and data security, has completed the design of an automated video editing solution. This solution integrates CLIP, YOLOv8n, and Large Models within a Cloud-Edge collaborative paradigm.

The core contributions of this research are threefold:

Architectural Innovation: It constructs a three-tier architecture comprising "Edge-side Perception, Cloud-side Decision-making, and Collaborative Scheduling." By leveraging lightweight edge processing, intelligent cloud decision-making, and dynamic resource allocation, this architecture circumvents the inherent defects of single-mode systems, offering a referenceable architectural paradigm for similar technologies.

Functional Closed-Loop: It designs four core functional modules, forming a closed loop of "Material Parsing – Script Generation – Interactive Rendering – Security Protection," thereby breaking through the limitations of traditional template-based editing.

Scenario Adaptation: Addressing the differentiated demands of Consumer (C), Business (B), and Government (G) scenarios, it formulates adaptation paths focusing on lightweight embedding, customized deployment, and open cooperation, respectively, enhancing the solution's industrial adaptability value.

The core significance of this study lies in providing a systematic design framework and innovative ideas. It offers a reference for the theoretical exploration and subsequent engineering implementation of automated video editing technology, providing a novel perspective for resolving the core industry conflict of "Compute-Security-Adaptability."

6.2 Discussion and Future Prospects

As this research focuses on solution design without full-scale implementation verification, certain limitations remain: the feasibility of the architecture requires validation through engineering practice, functional details need optimization in real-world scenarios, and the depth of scenario adaptation awaits expansion. From an industry perspective, "balancing intelligence with security compliance" and "extending general solutions to vertical domains" remain critical issues requiring further exploration.

Future research can be deepened in three aspects:

Engineering Optimization: Optimize edge-side model lightweighting and cloud-side scheduling algorithms in conjunction with hardware characteristics to enhance the feasibility of architectural implementation.

Functional Refinement: Focus on adapting to complex narrative logic and upgrading interactive experiences to strengthen the competitiveness of core functions.

ISSN 2616-5775 Vol. 9, Issue 3: 63-71, DOI: 10.25236/AJCIS.2026.090308

Scenario Verticalization: Customize specialized functions for fields such as education and healthcare, and collaborate with the industry to establish technical standards. This will drive automated video editing technology toward meeting specific industry needs, providing support for the intelligent upgrade of the short video industry.

References

[1] Zhou Junlong, Hou Xiangpeng, Lan Lan, et al. Cloud-Edge-End Collaborative Computing and Intelligence[J]. Embedded Technology and Intelligent Systems, 2025, 2(4): 261-267.
[2] Lv Kun, Zhang Weixu, Jing Jipeng. Research on Disruptive Technology Topic Identification Based on CLIP-LDAGV Multimodal Information Fusion: A Case Study of the New Energy Field[J]. Journal of the China Society for Scientific and Technical Information, 2025, 44(3): 353-368.
[3] Hou Yonghong, Zheng Haochun, Gao Jiajun, et al. Zero-Shot Action Recognition Based on CLIP Model and Knowledge Database[J]. Journal of Tianjin University (Science and Technology), 2025, 58(1): 91-100.
[4] Liu Jie, Qiao Wensheng, Zhu Peipei, et al. Zero-Shot Referring Image Segmentation Based on Fine-Tuning of Image-Text Large Model CLIP[J]. Application Research of Computers, 2025, 42(4): 1248-1254.
[5] Vardhan S H, Tanya S, R Indra. Automating YouTube Video Uploads Using Cloud-Native Technologies[J]. International Journal of Science and Research Archive, 2025, 15(02): 900-907.
[6] Li Zeyu, Chen Yang, Zhao Wentao. Optimization of Lightweight YOLOv8n Object Detection Algorithm Based on Cloud-Edge Collaboration[J]. Computer Engineering and Applications, 2025, 61(8): 112-119.
[7] Wang Chen, Li Xue, Liu Chang. Research on Intelligent Video Editing Script Generation Driven by Large Models[J]. Journal of Frontiers of Computer Science and Technology, 2025, 19(6): 1215-1224.
[8] Zhang L, Chen W, Li Y. Rapid Adaptation in Photovoltaic Defect Detection: Integrating CLIP with YOLOv8n for Efficient Learning[J]. Energy Reports, 2025, 1(2): 45-52.