European Journal of Engineering and Technologies

Vol. 2 No. 1 2026

Article Open Access

AI-Driven Video Content Optimization Strategies for Immersive Media

Da Xu [1,*]

[1] Video Infra, Meta, Menlo Park, CA, 94025, United States
[*] Correspondence: Da Xu, Video Infra, Meta, Menlo Park, CA, 94025, United States

Received: 19 November 2025
Revised: 29 December 2025
Accepted: 08 January 2026
Published: 14 January 2026

Abstract: With the rapid development of immersive media technology, users' demand for video content interactivity, immersion, and intelligent presentation is constantly increasing. The advantages of AI technology in content perception, creation, integration, and transmission are gradually becoming prominent, and it is an important support for promoting the iterative upgrading of immersive media. This article starts with the basic theory of immersive media and AI driven video content optimization, and constructs an integrated system framework covering perception analysis, multimodal generation, rendering and transmission. It discusses the adaptability of existing technology algorithms, the complexity of polymorphic processing, and the bottleneck of terminal adaptation, and provides solutions such as semantic parsing enhancement, polymorphic fusion optimization, cloud edge collaborative rendering, etc., in order to provide theoretical reference and practical path for promoting the content experience improvement of immersive media and creating intelligent applications.

Keywords: immersive media; artificial intelligence; video content optimization

1. Introduction

Immersive media, as a new generation of information dissemination form, integrates various interactive technologies such as virtual reality (VR), augmented reality (AR), panoramic video, etc., and is widely used in teaching, leisure and entertainment, cultural tourism and other fields. Compared with traditional video content, immersive media puts forward higher requirements for real-time perception, multimodal fusion and high-quality rendering of content. Meanwhile, with the rapid development of artificial intelligence (AI), AI has provided a new technological path for the intelligent generation, semantic understanding, and transmission optimization of immersive video content, which is a major factor affecting user experience [1]. At present, although AI has achieved certain results in immersive content applications, it still faces a series of problems such as algorithm adaptation, multimodal collaboration, and rendering efficiency. Therefore, this article focuses on the theme of "how AI drives the optimization of immersive video content", establishes a multidimensional system architecture, explores problems, and proposes solutions, aiming to provide theoretical and methodological guidance for the systematic and intelligent construction and practical application of immersive media content.

2. Overview of Immersive Media and AI Driven Video Content Theory

Immersive media is a new form of media based on multidimensional sensory interaction technology, which enables users to have a high sense of presence and deep participation in virtual environments. Its core features mainly include spatial perception, situational interaction, and deep immersion, which are widely used in various industries such as education and training, tourism and cultural exhibitions, and digital creativity. Compared to traditional linear videos, the content design of immersive media requires the integration and expression of modal information such as graphics, audio, motion capture, and context, making content analysis, design, operation, and dissemination more difficult [2].

Artificial intelligence technology provides strong support for optimizing immersive media content. The recognition and parsing of video content through deep learning, as well as the generation of new content through generative adversarial networks (GANs), and the optimization of low latency rendering and edge push through intelligent scheduling algorithms, are all manifestations of the intelligent development of immersive content systems driven by AI technology. In recent years, the latest technological developments in AI, such as computer vision, natural language processing, and graphic neural networks, have gradually combined with immersive content platforms, shifting from "manual design" to "automatic understanding and generation". This article focuses on AI driven video content optimization, delving into the characteristics of immersive media content and the principles of intelligent technology, in order to provide support for later architecture design and optimization strategies [3].

3. Construction of AI Driven Video Content System Framework for Immersive Media

3.1. Framework for Intelligent Perception and Semantic Analysis of Video Content

In an immersive media environment, video content presents information not only in a one-way manner, but also in a comprehensive system that includes multiple modalities such as visual, auditory, and spatial dynamics. In order to efficiently understand content, the intelligent perception system and intelligent semantic parsing architecture of artificial intelligence are positioned as important supports. This architecture is based on deep neural networks and integrates various perceptual abilities including image recognition, object detection, action recognition, language processing, etc., to achieve structured extraction of people, places, and objects in videos.

In the perception layer, convolutional neural networks (CNN) and Transformer structures extract features from frame level images and construct temporal models to capture motion changes and spatial layout. At the semantic parsing layer, a semantic graph is constructed based on GNN, combined with natural language processing (NLP) technology, to achieve semantic analysis and automatic annotation of user behavior and content tags. This system not only enhances the understanding ability of content machines towards audiovisual data, but also provides the most basic underlying support for post production, transition, and intelligent push [4]. In summary, the intelligent perception and semantic interpretation framework of video content is a necessary prerequisite for AI to empower immersive videos, ensuring the adaptability, matching, and personalized experience of the entire system (Figure 1).

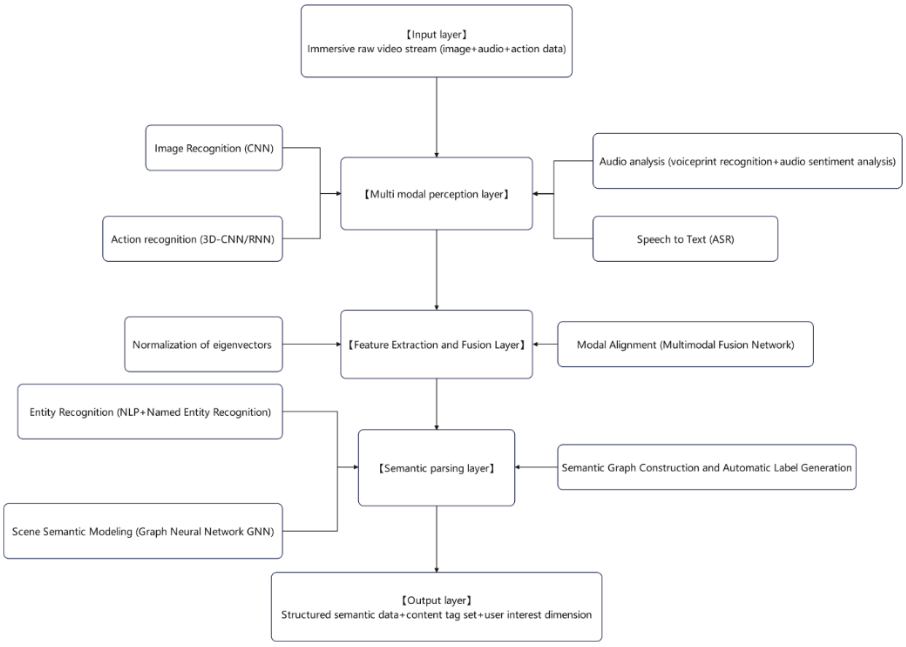
【Input layer】
Immersive raw video stream (image+audio+action data)
Image Recognition (CNN)
Action recognition (3D-CNN/RNN)
【Multi modal perception layer】
Audio analysis (voiceprint recognition+audio sentiment analysis)
Speech to Text (ASR)
Normalization of eigenvectors
【Feature Extraction and Fusion Layer】
Modal Alignment (Multimodal Fusion Network)
Entity Recognition (NLP+Named Entity Recognition)
Scene Semantic Modeling (Graph Neural Network GNN)
【Semantic parsing layer】
Semantic Graph Construction and Automatic Label Generation
【Output layer】
Structured semantic data+content tag set+user interest dimension

Figure 1. Framework diagram for intelligent perception and semantic parsing of video content.

3.2. Multimodal Content Generation and Style Transfer Mechanism

In immersive media systems, the ability to generate multimodal content is the core driver for building immersive experiences. The content generation mechanism empowered by AI not only needs to achieve collaborative construction of multi-source data such as images, audio, and actions, but also has the ability to transfer styles and adapt to contexts, which can meet the demands of different scenarios and users. The multimodal generation system mainly includes three modules: visual content generation, audio environment construction, and semantic consistency control.

In terms of visual generation, high-quality images, dynamic synthesis of scenes and virtual characters can be generated through adversarial networks (GANs) and diffusion models, supporting content style conversion, scene replacement, and special effects generation. By combining acoustic modeling with semantic correlation algorithms, spatialized sound effects and immersive background music are automatically generated to enhance on-site immersion and immersion [5]. In addition, based on cross modal alignment mechanism, CLIP, ALIGN and other models are used to construct a unified semantic space mapping of images, text and speech, improving the semantic consistency and logical coordination of automatically generated content. Meanwhile, the style transfer mechanism is based on pre trained style recognition networks and content decoupling algorithms, which can transfer specified visual styles or cultural element styles to the target material, further achieving customized presence effects. This mechanism is widely used in fields such as virtual reconstruction, teaching simulation, and cultural dissemination, significantly expanding the creative boundaries and user interaction depth of immersive media content.

3.3. Immersive Content Rendering and Intelligent Transmission Structure

In immersive media systems, content rendering and transmission efficiency are important factors that affect the smoothness and authenticity of user experience. Traditional image rendering is no longer able to meet the real-time operation needs of large capacity high-definition and multi action forms, and there is an urgent need to use AI technology to optimize image rendering and achieve intelligent image transmission structure design. The core of its architecture mainly includes three components: edge rendering scheduling, low latency encoding compression, and intelligent transmission path optimization.

First, a distributed rendering model based on deep learning is introduced in the rendering layer, which mainly relies on the partial pre-processing and parallel rendering of 3D scenes and high-definition video frames by edge computing nodes (MEC), thus greatly reducing the load and delay of the central server. Secondly, adopting artificial intelligence based video encoding algorithms, such as using convolutional neural networks (CNN) to drive compression strategies to compress frame rates and bit rates according to time-domain requirements, to meet high frame rates and low bit rates, compressed images and occupied bandwidth [6]. Thirdly, a reinforcement learning driven content scheduling mechanism is constructed at the transport layer, mainly combining network situation prediction and user behavior modeling to obtain optimal path optimization, thereby optimizing data transmission speed and stability. The above architecture fully demonstrates that artificial intelligence truly implements rendering layer and transport layer technologies into image rendering and transmission, effectively addressing issues such as device compatibility, network scene changes, and computational pressure, providing effective empowerment support for high-quality and low latency.

4. Issues Faced by AI Driven Video Content for Immersive Media

4.1. Insufficient Adaptability of Content Perception Algorithm

In an immersive media environment, video content is complex, diverse, and unstable, involving a large number of high dynamic images, unstructured objects, and diverse information interactions. Traditional content perception algorithms find it difficult to cope with adaptability. Firstly, the current mainstream image recognition technology and motion detection algorithms usually establish models based on two-dimensional plane data, which makes it difficult to accurately understand the depth relationships and spatial positioning in three-dimensional dimensions; Secondly, due to the inability to implement a comprehensive management mechanism for multimodal information, it can result in scattered and delayed responses to video content. In addition, most of the current algorithm training data comes from standard datasets, which lack training that adapts to the complex and edge scenes of immersive media, as well as the behavioral characteristics of real users, resulting in limited generalization effects. Especially when the interaction scene is high and there are many changes, there are certain limitations on the accuracy and speed of identifying key content. A comparative analysis of the adaptation performance of representative content perception algorithms in immersive media environments is presented in Table 1.

Table 1. Comparison of Adaptation Performance of Common Content Perception Algorithms in Immersive Media Environments.

Algorithm type	Application ability	There are problems
CNN image recognition algorithm	Analysis of Static Image Structure	Difficult to handle dynamic perspectives and spatial depth
Temporal action recognition models (such as RNN)	Action behavior modeling	Low recognition rate for complex interactive behaviors
Multimodal Fusion Perception Algorithm	Joint recognition of image and audio	Unstable fusion strategy and poor semantic consistency

4.2. High Complexity of Multimodal Data Processing

In immersive media, content is presented as a combination of different types of patterns such as graphics, sound, text, and motion trajectories, which increases the demand for computing resources for collaborative computing and real-time fusion in content processing systems. The heterogeneity of multimodal data in terms of dimension, temporal structure, and semantic expression makes it difficult to construct a unified modeling and fusion process. Currently, most existing multimodal processing techniques rely on deep neural networks to independently encode individual modalities and perform modal matching processing, which can easily lead to errors in semantic abstraction levels, resulting in illogical or semantically deviated generated content. Furthermore, the synchronous collection and processing of multimodal data in real interactive scenarios imposes a huge load on the system, especially when the computing power and bandwidth of the edge system are limited, which often leads to response delays or display effect delays, greatly affecting the user experience. A comparative evaluation of the complexity of multimodal data processing tasks is summarized in Table 2.

Table 2. Comparison of Complexity of Multimodal Data Processing Tasks.

Processing tasks	technical requirement	Main challenges
Image audio alignment	Cross modal time synchronization and matching	Different time scales and high difficulty in semantic coupling
Text Image Generation	Integration of natural language and visual coding	Significant differences in representation and poor consistency in generated content
Action Speech Matching	Dynamic Trajectory and Audio Event Recognition	The timing sequence varies in length and lacks a common mode characteristic benchmark

4.3. Rendering and Transmission Adaptation Performance Is Limited

Immersive media requires extremely high real-time performance and visual experience, while existing rendering and transmission systems have significant performance bottlenecks in terminal adaptation, multi-resolution output, and network dynamic response. During the process of rendering high frame rate and high-definition videos, users may encounter issues such as rendering delay, image tearing, or unsynchronized audio and video, which greatly affect their immersive experience. In terms of rendering, traditional GPU rendering cannot effectively achieve high-quality output on lightweight terminal devices such as VR headsets and mobile devices, and the complexity of algorithms cannot meet their resource requirements. In terms of transmission, slow or even paused content downloads due to network instability or bandwidth changes are not addressed by any intelligent allocation and storage mechanisms. A comparative analysis of adaptation issues in immersive content rendering and transmission solutions is presented in Table 3.

Table 3. Comparison of Adaptation Issues for Immersive Content Rendering and Transmission Solutions.

Module category	Technical Proposal	Adaptation problem performance
Rendering end	Local GPU rendering	Insufficient computing power and unstable frame rate of the device
Rendering end	Cloud based unified rendering output	Strong network dependency, easily affected by latency and bandwidth
Transmission end	Real time streaming protocol (such as RTSP)	Network fluctuations are sensitive and can easily lead to data loss and lag
Transmission end	Segmented content caching mechanism	Inconsistent loading, affecting the immersive coherence experience

5. AI Video Content Optimization Strategy for Immersive Media

5.1. Building a Multi level Semantic Analysis and Generation Framework

In order to further enhance the intelligent generation quality of immersive media content, it is necessary to build an immersive media multi-dimensional semantic analysis and creation system based on the three levels of "perception understanding generation" to address semantic ambiguity and logical omission issues. The core idea of this system is to combine graph neural networks (GNNs), pre trained language models (such as BERT, GPT), and visual language matching strategies to achieve full process control from raw multimodal data to semantic driven creation processes.

In the semantic parsing stage, the system maps video content to a structured knowledge graph through entity recognition, event extraction, and semantic association mapping, extracting key semantic nodes and their logical relationships. In the generation stage, using generative AI to guide semantic consistency, style coherence, and user interest matching in the content production process ensures that the generated results have interactivity and adaptability. In addition, by introducing a hierarchical attention mechanism, multiple modalities and granularity information semantics can be dynamically fused and selectively expressed, further enhancing the level of intelligence. This framework not only enhances the system's perception and construction ability of immersive scenes, but also lays the foundation for subsequent multimodal and personalized service push.

5.2. Optimizing Multimodal Fusion and Enhancement Processing Models

In immersive media, multimodal fusion plays a core role in content understanding and output coordination in the immersive media system, and its processing effect directly affects the authenticity and consistency of video content. To address the limitations of current models in semantic alignment and perceptual unity, a multimodal fusion processing model based on attention mechanism and residual enhancement is proposed. This model uses a Cross modal Attention Fusion Network to jointly model multiple modal information.

Taking the immersive teaching system as an example, the system needs to recognize the teacher's gesture (image), explanation voice (audio) and teaching text (text) at the same time. The model encodes the three modal features separately, and then represents them uniformly through the fusion module. The fusion method adopts residual weighting strategy:

$$H_{fused} = \alpha \cdot H_{image} + \beta \cdot H_{audio} + \gamma \cdot H_{text} + \varepsilon \quad (1)$$

Among them, α, β, and γ are trainable weights and ε is the residual term used to preserve low-level modal information and enhance the fidelity of the system to details and context. This optimization model enables the fused content to have higher semantic consistency, interactive response speed, and visual performance, making it well suited for immersive scenes with abundant interaction and information content.

5.3. Building Cloud Edge Collaborative Rendering and Push Architecture

To alleviate the system bottleneck of immersive content in real-time rendering and high load transmission, a cloud edge collaborative architecture is constructed to effectively optimize content delivery speed and device applicability expansion. Combining cloud computing capabilities with real-time response capabilities of edge nodes to achieve centralized processing of content rendering, intelligent scheduling, and personalized push functions.

In terms of system design, the cloud mainly undertakes preprocessing and overall model construction tasks, such as complex 3D scene rendering, deep semantic information generation, etc; Edge nodes dynamically receive, decode, and optimize content performance based on user location and device performance, supporting low latency feedback and personalized rendering strategies. At the same time, based on content popularity prediction and user behavior modeling, cache preloading of edge nodes for high-frequency access to content is implemented to reduce network traffic usage and loading time. The performance benefits of this cloud-edge collaborative architecture compared to a fully centralized rendering system are summarized in Table 4.

Table 4. Performance Comparison between Centralized Rendering and Cloud Edge Collaborative Architecture.

performance index	Centralized rendering architecture	Cloud edge collaborative architecture
network delay	High, affected by remote transmission	Low, timely edge response
Terminal adaptability	Poor, requiring high-performance equipment	Well, adapt to multiple terminal types
Content push efficiency	Low, slow unified scheduling	High, supporting intelligent scheduling and pre caching
User experience consistency	Vulnerable to network fluctuations	Stable, flexible scene response

6. Conclusion

With the popularity of immersive media applications, the quality of video content presentation and the level of system big data intelligence have become rigid indicators for improving user experience. This article focuses on video content optimization methods driven by artificial intelligence, and establishes a complete technical system for intelligent perception, multimodal generation, rendering and transmission. It analyzes and discusses key issues such as poor algorithm adaptability, complex data processing, and prominent transmission bottlenecks that often exist in existing methods. At the same time, relevant optimization strategies are proposed, such as multi-level semantic parsing, cross modal enhanced fusion, cloud edge collaborative push, etc. This study provides a theoretical basis and practical reference for the intelligent construction of immersive media systems. In the future, further exploration of real-time content generation and personalized interaction modes driven by large models will drive the development of immersive experiences to a higher level.

References

1. Y. Yang, S. N. Sannusi, and A. R. Ahmad Rizal, "Exploring the role of immersive media technologies in environmental communication: A case study of underwater earth projects," *Environmental Communication*, vol. 19, no. 3, pp. 432-448, 2025. doi: 10.1080/17524032.2024.2420785
2. H. K. Ravuri, J. Struye, J. van der Hooft, T. Wauters, F. De Turck, J. Famaey, and M. Torres Vega, "Context-Aware and Reliable Transport Layer Framework for Interactive Immersive Media Delivery Over Millimeter Wave," *Journal of Network and Systems Management*, vol. 32, no. 4, p. 78, 2024. doi: 10.1007/s10922-024-09845-5
3. P. H. Gueldner, K. E. Kerr, N. Liang, T. K. Chung, T. Tallarita, J. Wildenberg, and I. Sen, "Artificial intelligence-based machine learning protocols enable quicker assessment of aortic biomechanics: A case study," *Journal of Vascular Surgery Cases, Innovations and Techniques*, vol. 11, no. 4, p. 101806, 2025. doi: 10.1016/j.jvscit.2025.101806
4. M. M. Hasan, M. Pramanik, I. Alam, A. Kumar, R. Avtar, and M. Zhran, "Assessing the efficacy of artificial intelligence based city-scale blue green infrastructure mapping using Google Earth Engine in the Bangkok metropolitan region," *Journal of Urban Management*, 2024. doi: 10.1016/j.jum.2024.11.009
5. X. Hui, S. H. Raza, S. W. Khan, U. Zaman, and E. C. Ogadimma, "Exploring regenerative tourism using media richness theory: emerging role of immersive journalism, metaverse-based promotion, eco-literacy, and pro-environmental behavior," *Sustainability*, vol. 15, no. 6, p. 5046, 2023. doi: 10.3390/su15065046
6. A. Sanatizadeh, Y. Lu, K. Zhao, and Y. Hu, "Engagement or entanglement? The dual impact of generative artificial intelligence in online knowledge exchange platforms," *Information & Management*, 2025. doi: 10.1016/j.im.2025.104178

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.