ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202509.00125

Postprint: Research on the Technological Development and Multi-scenario Applications of Digital Humans

Authors: Guo Quanzhong, Gu Kexin

Date: 2025-09-19T15:52:13+00:00

Abstract

Dynamic interaction functionality constitutes the key to "human-like" industry perception and underscores the enormous potential of digital humans as a novel form of content productivity. The system can comprehend user inputs and make real-time decisions by integrating contextual information, host persona, and semantic trajectories, generating personalized responses that are both contextually appropriate and emotionally resonant.

Full Text

Preamble

By Guo Quanzhong and Gu Kexin / Media Outlook / Research on the Technological Development and Multi-Scenario Applications of Digital Humans

On June 15, 2025, the inaugural live stream of Luo Yonghao's digital human attracted over 13 million viewers, with gross merchandise volume (GMV) exceeding 55 million yuan. Sales figures for core product categories such as 3C and food items surpassed those from Luo's in-person debut in May, establishing a new benchmark for digital human live-streaming e-commerce. Throughout the broadcast, the two digital humans exhibited highly realistic capabilities in interactive actions, content generation, and user responsiveness. This milestone not only broke records but also demonstrated the application of multi-dimensional tags—including voice emotion, intonation control, facial movements, and body expression—which provide precise parameters for speech synthesis and video generation. These tags enable natural intonation transitions, synchronized lip movements with speech, and expressive gestures, thereby achieving an organic unity of "voice, form, and meaning" that enhances overall realism and immersion.

Dynamic interaction functionality constitutes the key to "human-like" industry perception and underscores the enormous potential of digital humans as a novel form of content productivity. The system can comprehend user inputs and make real-time decisions by integrating contextual information, host persona, and semantic trajectories, generating personalized responses that are both contextually appropriate and emotionally resonant.

Key Technologies Behind Digital Human Live Streaming

The digital human live-streaming solution encompasses five core innovations: script-driven digital human multimodal coordination, script generation that integrates multimodal planning and deep thinking, real-time interaction with dynamic decision-making, text-controllable speech synthesis, and high-consistency hyper-realistic long video generation. Together, these technologies achieve a high degree of unity in digital humans' "spirit, form, voice, appearance, and speech."

Language Model-Driven Multimodal Coordination Mechanism

In digital human live-streaming systems, script generation forms the foundation for constructing highly realistic interactive experiences, with its core lying in language model-driven multimodal coordination. This process comprises three integrated modules: line generation, multimodal driving, and dynamic interaction.

Line generation must deliver content while aligning with the host's persona and linguistic style. Through style modeling and character modeling, the system ensures personalized yet consistent language expression. In multi-host scenarios, it achieves overall coordination in semantic logic, rhythmic pacing, and emotional style. To enhance authenticity and depth, the system incorporates content planning, knowledge enhancement, and fact-checking mechanisms to mitigate AI hallucination risks. In the case of Luo Yonghao's digital human, leveraging Wenxin Large Model 4.5 Turbo and training on real-person corpora to distill linguistic patterns enabled precise simulation of logical structures and expression habits.

Multimodal driving ensures high coupling between language generation and audio-visual output. As the language model generates lines, it simultaneously outputs content-associated tags for voice emotion, intonation control, facial movements, and body expression. These tags provide precise parameters for speech synthesis and video generation, enabling natural intonation transitions, synchronized lip movements with speech, and expressive gestures that achieve an organic unity of "voice, form, and meaning."

Dynamic interaction represents the key to creating "human-like" experiences. The system understands user input and makes real-time decisions by integrating contextual information, host persona, and semantic trends, generating personalized responses that are both contextually appropriate and emotionally resonant.

Advances in Speech Synthesis Technology

As digital human technology expands, the naturalness of speech synthesis has become crucial for user immersion and emotional connection. In live-streaming contexts, audiences no longer content themselves with hearing perfectly enunciated recitations but expect hosts' voices to exhibit emotional fluctuations and personal styles that enhance interaction authenticity and persuasiveness.

However, traditional speech synthesis systems suffer from mechanical expression and limited emotional range, struggling to adapt to live-streaming's dynamic contexts and lacking emotional tension, which degrades user experience. To address this, Baidu introduced a "text-controllable speech synthesis" solution. Based on large language models, this approach deeply integrates speech synthesis with script content, host persona, and semantic-intonation tags, optimizing both "what to say" and "how to say it." The system incorporates semantic understanding, host style modeling, and fine-grained prosody control during voice generation, enabling coordinated generation of speech content and emotional expression. For instance, during product introductions, digital hosts can naturally adjust intonation from calm narration to passionate excitement, enhancing linguistic appeal and persuasive power.

Hyper-Realistic Long Video Generation Technology

In digital human live streaming, image generation and driving represent the most technically challenging aspects. Unlike audio or text, video generation requires not only image modeling and motion control but also maintaining high consistency over extended durations to ensure precise synchronization of digital human images, actions, and voice.

To address these challenges, Baidu constructed a "high-consistency hyper-realistic digital human long video generation" technical system. This solution takes scripts, voice, historical video data, and skeletal movements as multimodal inputs. Through multimodal video analysis and understanding, it generates highly expressive clips, complex "human-object-scene" interaction segments, and large-motion, large-expression segments, which are then uniformly scheduled across long time sequences. This ensures that voice, lip movements, expressions, and actions remain highly synchronized, achieving true consistency in "voice, appearance, and speech." In Luo Yonghao's digital human live-streaming practice, the system's independent modeling of characters and products effectively ensured stable persona, accurate actions, and semantic synchronization during prolonged interactions, delivering a highly coordinated and realistic live-streaming experience.[1]

Digital Humans Gradually Moving Towards Large-Scale Application

As artificial intelligence and virtual modeling technologies mature, digital humans are transitioning from laboratories to practical applications, with significantly accelerated commercialization and scene implementation. Today, digital humans are being deployed across multiple communication and service contexts with enhanced practicality and adaptability. From content industries to public services, brand marketing to cultural communication, digital humans are embedding themselves into multiple dimensions of social operation with the advantages of being "all-weather, controllable, and high-efficiency," releasing enormous potential.

Accelerating Penetration into Multiple Industry Scenarios

As AI technology evolves, digital humans are no longer limited to live-streaming e-commerce but are accelerating penetration into diverse fields, demonstrating blossoming development across multiple fronts. Their flexible image presentation, continuously optimized interaction capabilities, and persistent online availability make them a new vehicle for digital transformation across industries.

In customer service, digital humans serve as virtual agents in intelligent customer service systems, providing 24/7 consultation and business processing services that significantly improve efficiency while reducing labor costs. In government services and urban management, digital humans are deployed in smart government halls, policy dissemination platforms, and navigation systems, providing policy explanations, business guidance, and process prompts to the public through visual interfaces. This effectively alleviates pressure on offline service windows and enhances the intelligence level of government services.

In education and training, digital humans are gradually replacing traditional recorded course instructors, undertaking online teaching, Q&A, and situational interaction tasks. Their vivid images and interactivity add interest and immersion to educational content, demonstrating clear advantages particularly in language teaching and vocational training that require high-frequency interaction.

In cultural tourism, digital humans are created as virtual tour guides, providing interactive services such as route planning and scenic spot explanations. They can design optimal tour routes based on visitors' interests and time constraints, and vividly introduce the historical culture and characteristics of attractions, enhancing tourist destinations' visibility and appeal.

With their highly realistic image presentation and multimodal interaction capabilities, digital humans are quietly transforming connections between industries and users. In these application scenarios, digital humans function not only as communicators undertaking information transmission, service guidance, and knowledge popularization but are also evolving into an intelligent interface that

carries brand image, optimizes user experience, and enhances service effectiveness. This deep integration not only drives innovation in industry service models but also opens broader imaginative space for the integration of AI and social life.

Mainstream Media Embracing Virtual Digital Human Technology

In 2021, the National Radio and Television Administration issued the "14th Five-Year Plan for Science and Technology Development in Radio, Television, and Online Audio-Visual Services," proposing vigorous promotion of virtual anchors in news, weather, variety shows, and science education, and exploring the introduction of virtual anchors into program interactions to enhance personalization and engagement.[2]

Against the backdrop of deepening media convergence, digital transformation has become a critical imperative for mainstream media development. As a product of AI and communication practice integration, virtual digital humans are becoming a powerful tool for media to enhance communication power and interactivity. In recent years, mainstream media have actively explored digital human applications in news broadcasting, government services, live-streaming e-commerce, and cultural tourism promotion.

During the 2025 Spring Festival, the digital human news anchor created by Hangzhou Cultural Broadcasting Group achieved zero-error broadcast of news programs, attracting widespread attention. Hangzhou Daily Newspaper Group has built a digital human matrix covering diverse scenarios including news, live streaming, and cultural tourism promotion, demonstrating strong systematic application. Beijing Radio and Television Station's "Time Xiaoni" has also received positive user feedback through its accurate broadcasting and natural expression in news and government services. Digital human personas such as "Shen Ya," "Gu Xiaoyu," "Xiaoyang," and "Cheng Shuangshuang," created by broadcasting systems in Shanghai, Zhejiang, Hunan, and other provinces, demonstrate strong personalization in appearance and voice design while closely aligning with young users' preferences in interaction mechanisms. They have become important tools for local mainstream media to expand young audiences and enhance platform appeal.

Shaanxi Station's "Weiyang" and Henan Station's "Princess Jinfeng" integrate regional cultural elements into image design, combining communication and cultural inheritance functions and opening an innovative path of "virtual + culture."

Driving Content Value Reconstruction and Communication Method Upgrading

First, liberating productivity and enhancing content production efficiency. Traditional media content production is often constrained by human resources, scheduling, and on-site conditions, while digital humans greatly alleviate these

limitations. Virtual anchors can achieve 7$×$24-hour uninterrupted broadcasting, significantly improving information update frequency and timeliness. Simultaneously, through integration with news writing and editing workflows, digital humans can achieve automated news generation and broadcasting, substantially reducing content production costs.

Second, expanding expression methods and building immersive communication experiences. Digital humans possess visualizable and interactive communication characteristics that break traditional media's one-way output limitations. With diverse virtual image designs and realistic voice synthesis capabilities, digital humans can flexibly switch identities across scenarios, presenting more affable and engaging content expressions. In applications such as news broadcasting, thematic explanations, and live interaction, digital humans achieve immersive expression that blends virtual and real elements, enhancing users' viewing experience and emotional resonance. For young audiences, this novel communication method is more attractive and increases their attention and stickiness to media content.

Third, driving intelligent system upgrading in the media industry. Virtual digital human applications extend beyond front-end content output to compel comprehensive intelligent upgrading of mainstream media's entire chain processes. From content planning, corpus management, voice and image synthesis to distribution scheduling and user feedback analysis, digital human system implementation requires media organizations to comprehensively optimize data collection, algorithm training, and technology deployment. This promotes the formation of more intelligent and platform-based production and operation systems. By introducing AI intelligent middle platforms, content knowledge graphs, and large model interfaces, media organizations are gradually building sustainable "human-machine collaborative" content production systems, laying the foundation for future reconstruction of the information communication ecology.[3]

Digital humans are deeply embedding themselves into multiple communication and service scenarios, demonstrating broad application potential and development prospects. Whether in content generation, situational interaction, image construction, or system integration, digital humans have injected new momentum and possibilities into the industry. Future research and practice will focus on how to achieve continuous value release in broader industry implementations—a core issue of common concern for academia and industry. Against the backdrop of continuously advancing new quality productive forces, digital humans will become an important medium driving the leap-forward development of social informatization and intelligence.

---

[1] Machine Heart. Behind Lao Luo's Digital Human Going Viral, AI Directors Are Quietly Rewriting Live-Stream Scripts [EB/OL]. (2025-06-20) [2025-07-19]. https://mp.weixin.qq.com/s/sOLFAEFWU4fh_TGI84cB6Q. [2] Lu Jing. Application and Development of Virtual Digital Human Technology in Media Con-

vergence [J]. News Culture Construction, 2024(5): 8-10. [3] Xie Xinzhou and Zeng Ni. Opportunities and Challenges: Application and Future Development of Virtual Digital Humans in the Media Industry [J]. Youth Journalist, 2024(4): 73-77, 93.

Authors: Guo Quanzhong, Professor at the School of Journalism and Communication, Minzu University of China, Doctoral Supervisor, Director of the Research Center for Internet Platform Enterprise Development and Governance; Gu Kexin, Master's Student at the School of Journalism and Communication, Minzu University of China.

(Editor: Li Jing)

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.