Scientific and Social Research, 2025, Volume 7, Issue 9
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

A Human-Machine Collaborative Prompt Model for Audio Description of Local Cultural Promotional Videos

Wenyan Shao, Lingqian Zheng*, Xiaoshan Lin, Lirong Yan

College of Foreign Languages, Minjiang University, Fuzhou 350108, Fujian, China

**Author to whom correspondence should be addressed.*

Abstract: This study explores the development of an automated audio description (AD) framework for local cultural promotional videos using a human-machine collaborative approach. The proposed framework integrates a multimodal large language model, Doubao, with human expertise to enhance AD production, particularly for videos featuring culturally rich content. By focusing on the example of the Fujian-based video "Where There Are Dreams, There Is Fu", the study addresses two primary challenges in AD: cross-frame coherence and accurate cultural symbol interpretation. Through iterative human-machine collaboration, the model generates coherent, culturally grounded AD scripts that align with the cognitive patterns of visually impaired audiences. This research highlights the potential of GenAI-driven solutions in creating accessible content for public welfare organizations while maintaining cultural authenticity. The proposed framework offers a scalable, cost-effective approach to improving accessibility and promoting cultural heritage for visually impaired individuals.

Keywords: Audio description; Human-machine collaboration; Multimodal large language models; Cultural heritage

Online publication: September 9, 2025

1. Introduction

The growing demand for accessible media has intensified the need for visually impaired individuals to access information embedded in visual content, including videos, images, and short-form media. Audio description (AD) has consequently emerged as a critical medium for ensuring equitable access to cultural services. Audio description refers to a cultural service that translates key visual elements, such as environmental settings, character actions, facial expressions, and on-screen text, into auditory narratives for visually impaired audiences. This process facilitates comprehension of visual content, enabling equitable information access and cultural participation. Its primary function is to help blind and low-vision viewers auditorily perceive visual content, thereby enhancing their quality of life and promoting social inclusion. It has been found that audio description provides linguistically

formulated visual information that supports film comprehension for visually impaired audiences, noting that AD is more visual and capable of accurately describing on-screen content rather than merely relating pre-scripted narratives [1].

Current AD production faces dual challenges. For one thing, traditional manual production relies on professional collaboration, demonstrating low efficiency, high costs, and an inability to support large-scale accessibility conversion. For another, automated AD exploration primarily focuses on developing large language models specifically for AD generation, requiring substantial funding, time, and ongoing maintenance costs for model iteration, rendering this approach largely infeasible for most public service accessibility organizations. Moreover, automated AD struggles with maintaining scene coherence and character tracking, with particular difficulties in culturally symbolic interpretation, accurate translation, and cognitive adaptation for visually impaired audiences. Recent research has shown that multimodal large models are capable of integrating heterogeneous data from multiple sources, such as text, images, audio, and video, by leveraging cross-modal semantic understanding and knowledge association. This enables the construction of comprehensive and multidimensional cultural heritage knowledge bases, which contextualize previously fragmented cultural elements to support more coherent and enriched interpretation [2]. Moreover, intelligent cross-media content generation and integrated presentation can significantly enrich cultural heritage expressiveness.

This study focuses on integrating local cultural promotional videos with multimodal large language models (LLM), aiming to develop an automated AD generation framework based on human-machine collaboration that utilizes existing multimodal large language model products to produce accessible verbal descriptions aligned with visually impaired cognitive patterns.

2. Human-machine collaborative audio description framework

2.1. Operational principles of large language models

Large language models constitute a fundamental component of GenAI. Based on the Transformer architecture, they capture complex linguistic relationships within text through multi-layered structures that progressively learn textual features from basic syntax to deep semantics. The process involves pre-training on extensive text corpora to acquire fundamental language rules, followed by task-specific fine-tuning for practical application adaptation. It has been found that generative large language models, exemplified by ChatGPT, demonstrate exceptional capabilities in natural language understanding, intent recognition, reasoning, contextual modeling, language generation, and general problem-solving [3]. These strengths, attributable to their massive parameter size and deep network structures, position such models as significant pathways toward artificial general intelligence.

Natural language processing advancements provide a foundation for human-machine collaboration in AD. For instance, in human-machine collaborative automated advertising, multimodal large language models are employed to automatically generate initial ad drafts by leveraging their capabilities in natural language generation and video-language semantic matching. This process involves intelligent analysis of video scenes and emotional tones, followed by professional linguistic refinement and narrative optimization performed by human experts. The drafts are further perfected through multi-dimensional validation tools to enhance the overall expressiveness and quality of the advertisement. Previous research has argued that although system-generated scripts require human modification, they provide valuable references for professional narrators, enhancing AD production efficiency [4]. This indicates that machines can efficiently bridge visual-textual associations and generate standardized drafts,

thereby forming efficient complementarity and collaborative cooperation with humans in enhancing AD script creation efficiency and quality.

2.2. Challenges in manual AD production and LLM limitations

Manual AD production requires frame-by-frame video analysis, scriptwriting, and voice recording, leading to limited efficiency, extended production cycles, high reliance on specialized professionals, and significant labor costs. Moreover, variations in cognitive and expressive styles among different creators often result in inconsistent interpretations of cultural symbols and linguistic representations when describing the same video content. Such inconsistencies may adversely affect the experience of visually impaired audiences. Furthermore, due to constraints in time and human resources, it remains challenging to develop customized AD content tailored to the diverse cognitive needs of different visually impaired groups.

Conversely, large language models face challenges in accurately interpreting implicit cultural symbols and in capturing nuanced emotional expressions. These limitations often lead to a loss of the humanistic essence within the content. They may also misjudge character relationships or overlook subtle contextual details. Furthermore, current large language models lack the capacity to actively perceive feedback from visually impaired audiences, adapt flexibly in linguistic style, or provide personalized adaptations. It is argued that although technological advances, including machine translation and speech recognition, have enhanced operational efficiency, the integration of multimodal information and cultural adaptation remains dependent on human input [5]. Effective resolution of these issues thus requires collaborative efforts between translators and technology developers to explore integrative solutions. In addition, traditional audio description (AD) production requires substantial professional human resources, while existing automated methods necessitate extensive training to effectively integrate multimodal inputs and adapt output styles [6].

2.3. Key research questions

Significant limitations remain in automated AD for local cultural promotional videos. First, a lack of cross-frame coherence undermines entity tracking. Automated systems often produce fragmented descriptions with weak continuity across segments [7]. Second, cultural symbol interpretation is insufficient. While large language models can detect visual elements such as the Fu (福) character, they often overlook context-dependent meanings, stylistic variations, and the cultural specificity embedded in regional practices.

These shortcomings underscore the necessity of designing human-machine collaborative frameworks that can generate coherent, character-aware, and culturally nuanced AD scripts. Local cultural promotional videos, particularly those highlighting the Fu culture of Fujian province, present a distinctive challenge because they combine dynamic visual storytelling with layered cultural symbolism. Without effective cross-frame tracking and symbol interpretation, the resulting AD risks fragmenting narrative logic and weakening cultural heritage transmission.

Accordingly, this study focuses on two central research questions:

How can human-machine collaboration improve cross-frame coherence in AD scripts, particularly in maintaining character continuity and narrative integrity?

In what ways can human-machine collaborative prompting enhance the interpretation of culturally significant symbols, such as the red Fu character, beyond surface-level recognition?

By addressing these questions, the study aims to contribute a framework that leverages the efficiency of

multimodal AI while incorporating human expertise to achieve accessible and culturally grounded AD for local promotional videos.

3. Research methodology

3.1. Criteria for video selection

The video titled "Where There Are Dreams, There Is Fu", produced in 2017 by Haixia Television under the Fujian Radio and Television Group, was chosen based on two principal considerations. First, its high production quality and meticulously composed visual frames supply ample material for interpretation, offering perceptible and analyzable information for visually impaired audiences. Second, the video thoroughly embodies distinctive elements of Fujian culture, vividly portraying local traditions and everyday atmosphere, which supports an immersive cultural experience through auditory means.

3.2. Procedure of human-machine collaboration

This study employs the large language model of Doubao, selected for its free accessibility and multimodal capabilities suitable for audio description tasks. The procedure comprised two phases. The first phase focused on human-machine collaboration, including video selection, key frame identification, and cultural symbol queries to support machine recognition and interpretation. Key frames were then processed for frame-by-frame description, followed by machine extraction of cross-frame elements, such as characters, scenes, and narrative logic, leading to preliminary AD script generation through draft prompt input. The second phase is interactive validation. LLM-generated scripts were reviewed by experts and visually impaired participants, with draft prompts refined accordingly. Revised prompts were subsequently used by the large language model to regenerate AD scripts through iterative optimization cycles.

3.3. Preliminary question design

Key frames were extracted from a 36–54 second segment at 2-second intervals because this segment contains continuous plotlines for character tracking and key cultural elements of Fu, with 2-second intervals ensuring plot coherence. Preliminary questions targeted cultural symbols (e.g., red Fu character) are designed to establish cultural background understanding for subsequent accurate interpretation. For example, questions may include: What are the cultural connotations of the Fu character? What symbolic meaning is conveyed by presenting the Fu character to individuals upon departure?

4. Data analysis

4.1. Overview of key frames

Figure 1 shows key frames of 36–54 seconds. These segments require coherent tracking of the same man across four frames and the continuous event of "coloring the Fu character" followed by "delivering the Fu character." Specifically, 36–40 seconds depict the man coloring a red Fu character at a table, while 41–46 seconds show the same man presenting the Fu character at a dock. The 46–48 seconds show the sea under the sunlight, with the waves gently shimmering. Subsequently, during 48–54 seconds, the frame first presents a window with colorful railings adorned with a vivid red Fu character, followed by a wall of a traditional white-wall-and-black-tile building featuring a red Fu character. Associating character features across these appearances was essential to

maintain the integrity of the event chain. A primary challenge, however, was the LLM Doubao's initial failure to track the same man across frames through cross-frame recognition, resulting in narrative discontinuity.

Figure 1. Key frames from "Where There Are Dreams, There Is Fu"; The video of "Where There Are Dreams, There Is Fu" was published on August 26, 2017, accessed on August 15, 2025, https://m.news.cntv.cn/2017/08/26/VIDEm9BGRleRKCHYpeuo7FOR170826.shtml

The red Fu character further illustrates the difficulty of cultural interpretation. For Chinese audiences, red symbolizes auspiciousness and festivity, while Fu denotes life blessings. Beyond general well-wishing, it reflects the heritage of local Fu culture and embodies familial concerns, as exemplified in practices of gifting Fu. However, the LLM Doubao remained limited to surface-level visual recognition, lacking the capacity to interpret contextual meanings of blessings or familial concerns, and unable to differentiate the unique symbolic values embedded in localized Fu character styles.

4.2. Draft prompt analyses

The initial draft prompt instructed: You are a professional video and audio description expert. Your task is to write accurate audio descriptions for the provided video clip, targeting visually impaired audiences. The workflow is as follows. First, context review: recall preliminary questions. Second, video review: examine images in chronological order. Third, description: detail each frame's content while noting continuous events across frames. Fourth, summary: use frame descriptions and visual cues from 10 images to generate a sequential audio description script.

Testing revealed two main limitations. First, inadequate cross-frame coherence impeded entity tracking, as the system failed to associate the man coloring a Fu character with the same man delivering it at the dock, disrupting the continuity of the "coloring Fu to dock delivery" sequence. This weakness in temporal coherence directly undermines narrative integrity and diminishes the audience's ability to follow character-centered events. Second, the interpretation of cultural symbols was superficial, confined to visual recognition without regional grounding. For example, the system identified a "red object" but failed to connect it to the cultural significance of red in Chinese traditions or to the heritage of Fujian Fu culture. The inability to capture these layered meanings limits cultural expressiveness and risks reducing complex heritage symbols to simplistic visual markers, thereby weakening the communicative and inclusive function of AD for visually impaired audiences.

4.3. Revised prompt analyses

The revised prompt specified: You are a professional video and audio description expert. Your task is to write accurate audio descriptions for the provided video clip, targeting visually impaired audiences. The workflow is as follows. First, context review: recall preliminary questions for cultural understanding. Second, video review:

examine images in 2-second intervals, identifying persons by appearance. Third, description: detail frame-by-frame occurrences while noting persistent events. Fourth, summary: integrate cultural background and visual cues from 10 images to generate a 16-second description within 48 words (40–54 second segment). Output format: xx-xx seconds (X characters), textual description.

Optimizations included standardized output with time and word count labels, explicit word limits, appearance-based character identification, and omission of the first 4 seconds with firecracker sounds. These adjustments improved the performance of the LLM Doubao by enabling consistent identification of the same man across four frames, presenting the full "coloring to delivering" sequence, and incorporating culturally grounded interpretations of the Fu character as symbolizing auspiciousness, celebration, blessings, and familial concerns. The outputs aligned more closely with the cognitive habits of visually impaired audiences, ensuring both informational accuracy and cultural nuance. Moreover, the revised prompt design demonstrated the feasibility of combining machine efficiency with human cultural guidance, showing that prompt engineering can directly enhance the interpretive depth and narrative coherence of AD. This not only contributes to methodological innovation in accessibility research but also provides a scalable solution for producing culturally sensitive AD across different types of visual media.

5. A human-machine prompt model for audio description

Based on the discussions, we propose a human-machine collaborative prompt model. The model integrates human expertise with the large language model to enhance both the coherence and cultural accuracy of audio description scripts.

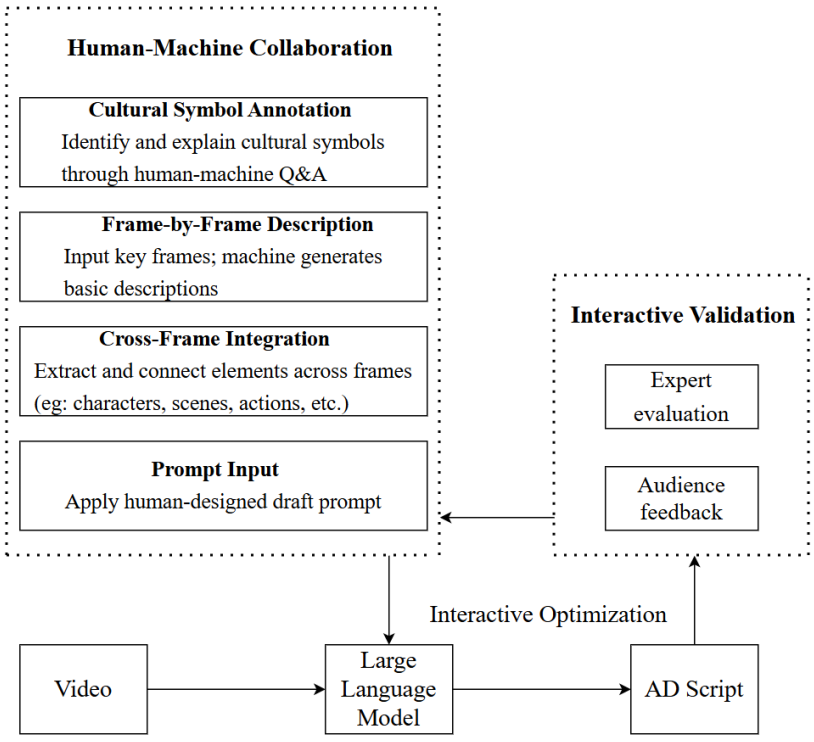

Figure 2. A human-machine collaborative prompt model

Figure 2 presents a Human-Machine Prompt Model integrating implementation steps and prompt optimization logic to visualize human-machine collaborative AD script generation through prompt-language model interaction. The flowchart illustrates LLM-based AD script generation through human-AI collaboration and iterative validation. Using video as source material, human-AI collaboration extracts core elements for LLM queries to enable cultural symbol recognition and interpretation, followed by frame-by-frame description, cross-frame element integration, and draft prompt input for AD script generation. Generated scripts are refined through expert and audience feedback, continuously optimizing prompts and collaboration until final validation.

Results demonstrate that this collaborative solution addresses two core deficiencies of multimodal AI in local cultural promotional video AD: inadequate character tracking and limited cultural symbol interpretation. In response to the first research question, human-machine collaboration improves cross-frame coherence by ensuring entity continuity, for example, linking the man painting a Fu character with the same man later delivering it at the dock, thereby maintaining narrative integrity. In response to the second research question, human-machine collaborative prompting enhances cultural symbol interpretation by guiding the system to connect surface visual recognition of the red Fu character with its deeper connotations of auspiciousness, blessings, and local cultural heritage. Personalized prompting compensates by enabling cross-frame entity association (e.g., identifying the man painting and delivering Fu as the same person) and targeted cultural questioning (e.g., exploring the connotations of Fu) that incorporate meanings beyond surface recognition.

The model has two key features. First, it leverages existing machine learning tools and datasets to analyze footage and automatically identify key visual elements, establishing a foundation for narrative transformation. Second, it allows flexible adjustment of outputs through dialogic collaboration, aligning descriptions with the cognitive habits of visually impaired audiences while embedding cultural specificity. Unlike conventional large-model pre-training, this approach provides adaptive, user-centered interaction and cultural sensitivity. The framework employs existing LLM via draft prompt queries and iterative human optimization to produce qualified AD scripts, finalized through expert integration. It offers an efficient, cost-effective solution, particularly suitable for public welfare organizations seeking accessible conversion of cultural videos. Nonetheless, it requires human scaffolding. Individuals must recognize cultural symbols to pose effective questions, while obscure local symbols remain challenging. To enhance adaptability, simple feedback mechanisms could be incorporated, involving visually impaired users as evaluators of coherence, cultural clarity, and character tracking. Feedback on identity confusion or insufficient cultural interpretation would trigger targeted prompt adjustments, continuously improving alignment with user needs. As recent research argues, human-machine collaborative creation positions humans at the ideological core, such as value construction and experiential integration, while machines contribute to expressive functions, such as symbol design and narrative structuring, achieving synergy through a division of labor between human creativity and machine scalability [8].

6. Conclusion

This study developed a model of human-machine collaboration combined with an existing multimodal LLM to provide an automated AD solution for local cultural promotional videos. Using the video case of “Where There Are Dreams, There Is Fu”, the model demonstrated its capacity to generate AD scripts that achieve character tracking and cultural interpretation while aligning with the cognitive habits of visually impaired audiences. The approach requires neither high financial investment nor advanced technical infrastructure, thus enabling public

welfare organizations to produce AD for local cultural promotional videos efficiently and at low cost. In doing so, it supports visually impaired communities in accessing cultural representations of their hometowns, thereby fostering belonging and cultural pride. Moreover, the framework offers a reference for extending accessibility transformation to other video genres such as documentaries and short videos. Future research could incorporate broader feedback from visually impaired individuals across regions and with diverse cognitive profiles to further enhance the adaptability of the scripts.

Funding

This research was supported by the project of the National College Student Innovation Training Program, "Development of an Audio Description Coding Framework for the Visually Impaired Based on Large Language Models" (Project No. 202510395023).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Rohrbach A, Torabi A, Rohrbach M, et al., 2017, Movie Description. International Journal of Computer Vision, 123(1): 94–120.

[2] Wei L, 2025, Narration, Identity, and Immersion: Strategies for Leveraging Multimodal Large Language Models for Enhancing Cultural Heritage Protection and Inheritance in the New Era. Journal of Yunnan Minzu University (Philosophy and Social Sciences Edition), 42(1): 31–39.

[3] Liu XB, Hu BT, Chen KH, et al., 2023, Key Technologies and Future Development Directions of Large Language Models: Insights from ChatGPT. Bulletin of National Natural Science Foundation of China, 37(5): 758–766.

[4] Campos VP, de Araújo TMU, de Souza Filho GL, et al., 2020, CineAD: A System for Automated Audio Description Script Generation for the Visually Impaired. Universal Access in the Information Society, 19(1): 99–111.

[5] Yuan MT, Ye SC, 2025, Starting from "Audiovisual Translation": Understanding the Cross-cultural Auditory Communication of Audio Description in Accessible Filmmaking. Film and Television Industry Research, 2(1): 68–78.

[6] Chu P, Wang J, Abrantes A, 2024, LLM-AD: Large Language Model-based Audio Description System. Arxiv, 2405(983): 1–9.

[7] Braun S, Starr K, Delfani J, et al., 2021, When Worlds Collide: AI-created, Human-mediated Video Description Services and the User Experience. Lecture Notes in Computer Science, 13096(1): 147–167.

[8] Sun BL, Wu L, 2025, Research on the Internal Logic and Evolution of Human-Machine Collaborative Creation. Chinese Editor, 24(8): 26–33.