

Original Research Article

An AI-based automatic translation and adaptation system for English comic dramas

Genjuan Ma, Hongxiu Duan, Xuekai Cao, Yuechun Sun

Communication University of China, Nanjing, Nanjing, Jiangsu, 211172, China

Abstract: This paper addresses challenges in AI-driven automatic translation and adaptation of English anime series, including insufficient text extraction accuracy, lack of cultural adaptation, and inadequate synchronization between subtitle layout and audio-video playback. A comprehensive system encompassing "visual analysis – translation adaptation – layout and dubbing – rendering detection" is developed. By employing joint model training, dynamic algorithm optimization, and multimodal collaboration techniques, the system achieves precise optimization across key stages such as text recognition, translation adaptation, and audio-video synchronization, supports output in multiple formats and resolutions, and enhances both translation efficiency and quality.

Keywords: English anime series; AI-powered automatic dubbing; multimodal adaptation; audio-video synchronization; subtitle layout

1. Introduction

With the deep integration of AI technology and the digital content industry, English-language animated series have seen continuously growing market demand as a vital medium for cross-cultural communication. Traditional manual dubbing approaches face challenges such as lengthy production cycles, high costs, and inadequate standardization, making them ill-suited for large-scale distribution needs. Features inherent to these series—including handwritten fonts, cultural references, and vertical-screen layouts—further complicate translation processes. Against this backdrop, developing efficient and precise AI-powered automatic dubbing and adaptation systems has become an industry imperative, providing crucial support for overcoming technical bottlenecks and driving industrial upgrading.

In recent years, there has been strong demand for the global expansion of animated series, and the scale of China's digital content exports has continued to grow, surpassing 300 billion yuan in 2024, with a promising market outlook. However, traditional animation and drama localization methods suffer from significant issues such as low efficiency in human translation, high costs, complex format handling, and difficulties in adapting to cultural contexts, which severely hinder the efficiency of global expansion and the quality of content dissemination. Therefore, we have launched the research project "Animation and Drama Global Expansion · Smart Translation Workshop—An AI-Powered Automatic Localization and Adaptation System for the English Market."

2. Technical challenges in AI-driven automatic translation and adaptation for English anime dramas

2.1. Insufficient accuracy in English manga text extraction

Text extraction from English comic animations is affected by handwriting styles, artistic fonts, background textures, and stroke distortions, resulting in a character recognition error rate of 4%–6% for dedicated comic OCR systems. Regions such as dialogue bubbles and onomatopoeic words are particularly susceptible to irregular shapes and dynamic scaling, leading to detection failures or positioning inaccuracies. After hard subtitle removal, high-density text areas often retain residual traces or exhibit blurring and ghosting effects, impairing subsequent subtitle rendering. Frame-level timeline annotation errors exceeding 0.3 seconds make precise synchronization between text and visual actions challenging^[1].

2.2. Translation and cultural adaptability: Lack of compatibility

The translation process struggles to accurately handle slang, subcultural references, and character-specific tones, often resulting in text with a forced "translation tone." The translated text frequently fails to match the original audio's duration, leading to abnormal speech rates or excessive pauses. Insufficient recognition of cultural norms, sensitive content, and localized expressions across different markets hinders audience comprehension and emotional resonance. Furthermore, the absence of a dynamic terminology database for anime-related terms compromises translation consistency.

2.3. Insufficient synchronization between subtitle layout and audio-video playback

In vertical-screen comics, irregular bubbles and complex line intersections complicate subtitle layout; fixed font sizes and line spacing often lead to overflow, obstruction, or reduced readability. Subtitle timing discrepancies, mismatched dwell durations with speech speed, and inadequate emotional reproduction in AI voiceovers—All characterized by mechanical delivery, temporal misalignment in lip movements, and audio-discrepancy from mouth actions—Significantly diminish viewing immersion.

3. Comprehensive system design strategy for AI-driven automatic translation and adaptation of English anime series

3.1. Optimized design of the visual analysis module

By employing a joint training architecture combining MangaOCR and YOLOv8, we expanded the English handwritten and artistic font dataset to include 500,000 samples, optimized the feature extraction network, and enhanced text recognition accuracy. Frame-level image segmentation technology was utilized to identify dialogue bubbles and onomatopoeic word regions, while multi-scale detection algorithms were integrated to accommodate irregular and dynamically scaling bubble scenarios. An improved LaMa model was introduced with a two-stage restoration strategy: First using a mask generation network to precisely locate subtitle regions, then applying texture transfer algorithms for image restoration. A frame-level timeline annotation system was developed that, through correlation analysis between audio waveforms and motion frames, achieves precise alignment between text and visual frames with an annotation error within 0.1 seconds.

3.2. Design of the multimodal translation and cultural adaptation engine

A three-tier translation architecture comprising a "base engine + specialized corpus + duration constraints" was established, with GPT-4o serving as the core engine and integrating the Sakura corpus containing 300,000 anime-specific terms, supported by a dynamic update mechanism. A duration-controlled translation algorithm was implemented that analyzes syllable counts and pause intervals in the source audio to build a translation length prediction model, ensuring the discrepancy between translated text and original audio duration remains within 0.8 seconds^[2]. A regional cultural adaptation module was designed, featuring a sensitive content lexicon categorized by target markets—Covering eight major categories including religious taboos and cultural customs with 120,000 entries—And employing semantic analysis for automatic filtering and replacement of sensitive content. A term consistency verification system was developed, utilizing cosine similarity algorithms to compare translation results across similar concepts, thereby ensuring consistency throughout series translations.

3.3. Design of the subtitle adaptive layout system

Develop a dynamic typography algorithm based on bubble characteristics to dynamically detect bubble size, shape, and position parameters in real time, automatically adjusting subtitle font size, line spacing, and line break rules while limiting each line to no more than 32 characters. Utilize layered rendering technology by placing subtitles on separate layers with a 1.5 pt black border and 80% opacity, reserving a 10% bottom safety margin to prevent obstruction of platform UI elements and key visual components. Implement a dual-track processing module for onomatopoeic words that employs speech recognition to identify word types and apply either phonetic or semantic translation methods, synchronized with corresponding typographic styles^[3]. Integrate a timeline synchronization algorithm that adjusts subtitle duration based on speaker speed analysis, achieving precise synchronization with a timestamp accuracy of 5 ms.

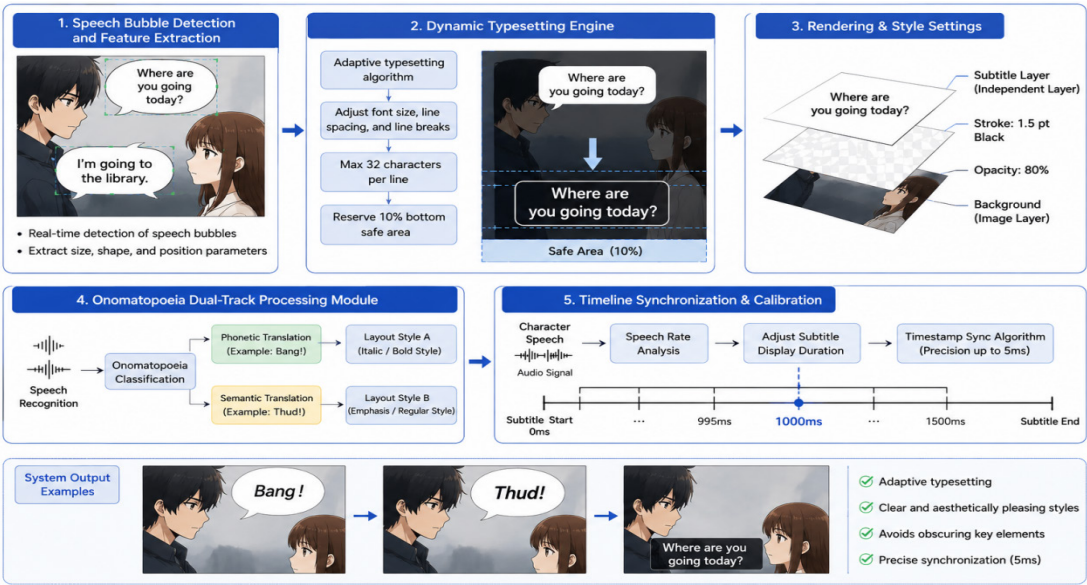


Figure 1. Design of the subtitle adaptive layout system.

3.4. Design of the AI voice acting and lip sync system

The system features a speech library containing 150 authentic voice samples from the UK, US, and Australia, with support for quantitative adjustment of emotional parameters. Utilizing affective speech synthesis technology, it automatically matches corresponding tone characteristics based on semantic text analysis. A dynamic speech rate calibration module adjusts the speaking speed in real time according to translation length, covering a range of 0.8 to 1.2 times the original speed to ensure smooth dubbing performance^[4]. The lip-sync alignment algorithm analyzes phoneme timing patterns in the audio track and matches them frame-by-frame with the character's mouth movements, synchronously driving the AI lip-generation model to produce dynamically synchronized lip motions. Facial feature constraints are incorporated to ensure generated lip movements align precisely with the character's facial structure and expression states^[5].

3.5. Engineering rendering and quality inspection design

Utilizing an FFmpeg batch compression framework, it supports output in six mainstream formats including MP4 and MOV, is compatible with resolutions of 720p, 1080p, and 4K, and features a dynamic bitrate adjustment range of 1.5–8 Mbps. A multi-dimensional AI quality inspection system has been developed, establishing evaluation metrics across four dimensions — Subtitle recognition accuracy, translation consistency, layout compliance, and audio-video synchronization—And performing automated validation through machine vision and natural language processing technologies. A batch processing scheduling system has been designed to handle parallel processing of up to 100 episodes per batch, with each episode processed within 12 minutes.

Table 1. System output format and adaptation parameters table.

output format	resolution ratio	Bitrate Range (Mbps)	Subtitle Format	Audio Sampling Rate (kHz)
MP4	720p	1.5-3.0	WebVTT	44.1
MP4	1080p	3.0-5.0	WebVTT	48.0
MP4	4K	5.0-8.0	WebVTT	48.0
MOV	720p	1.5-3.0	SRT	44.1
MOV	1080p	3.0-5.0	SRT	48.0
MOV	4K	5.0-8.0	SRT	48.0

4. Epilogue

The AI-powered automatic translation and adaptation system for English animated series employs multi-module technologies to address key challenges such as text extraction, translation adaptation, and

audio-video synchronization. By integrating CV, NLP, and TTS technologies, the system achieves full-process automation and precision while balancing efficiency and quality. Its implementation will advance standardization in the English animated series translation industry, provide technological empowerment for cross-cultural content dissemination, and demonstrate significant practical value for the sector.

References

- [1] Li Jiayi. Three-tier closed-loop human-machine collaboration: Research on AI animated drama technology stack adaptation [J]. Radio and Television Network, 2026, 33(5): 62–64.
- [2] Xu Liang. Generative AI Empowering Cross-Cultural Empathy and the Global Expansion of China's Audiovisual Content — An Investigation Based on iFlytek's Multilingual Intelligent Translation Practices [J]. Guochuan Observation, 2026, (5): 36-41.
- [3] Zhong Fangfang. Central Literature Translation Criticism: Current Status, Challenges, and Model Development [J]. Journal of Chongqing Second Normal University, 2026, 39(2): 34–40 + 127–128.
- [4] Zhou Yuxin. Exploring pathways for integrating film and television dubbing cases into English translation teaching [N]. China Film News, 2026-03-04(011).
- [5] Gong Mingqi. Artificial Intelligence-Assisted Film and Television Translation: New Models and Challenges in Cross-Cultural Communication [J]. New Era Foreign Language Education Series, 2025, (1): 120–129.