AN END-TO-END MULTIMODAL SYSTEM FOR SUBTITLE RECOGNITION AND CHINESE-JAPANESE TRANSLATION IN SHORT DRAMAS

Jing An [1,†], Haofei Chang [2,†], Rui-Yang Ju [3,†], Jinhua Su [4,5†], Yanbing Bai [4,], Xin Qu [1]*

[1] School of AI and Language Sciences, Beijing International Studies University, Beijing, 100024, China
[2] School of Information, Renmin University of China, Beijing, 100872, China
[3] Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan
[4] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China
[5] Simashuhui Ltd., Beijing, 100086, China

ABSTRACT

In recent years, short dramas have achieved impressive success in China, with high-quality subtitle translation serving as a crucial foundation for their expansion into international markets. However, subtitle translation for short dramas presents unique challenges, including frequent use of character names, missing punctuation, fragmented segments, colloquial expressions, and limited context. To address these issues, we propose a novel end-to-end multimodal system that integrates both visual and audio channels. Specifically, we employ Qwen2-VL for subtitle extraction from video frames via optical character recognition (OCR), and Whisper for automatic speech recognition (ASR) of the audio track. The outputs of both are then compared, and the superior result is selected using a fusion strategy that integrates temporal alignment and textual similarity, thereby improving subtitle recognition accuracy. The finalized subtitles are subsequently translated by our translation module. To support Chinese-Japanese translation, we construct a Chinese–Japanese translation dataset based on short dramas and fine-tune the Qwen2.5 on it using Low-Rank Adaptation (LoRA). Experimental results show that our proposed system outperforms the direct application of Qwen2-VL and Whisper. For Chinese–Japanese translation, our fine-tuned model achieves improvements over Qwen2.5, raising chrF++ from 27.8855 to 29.9883 and COMET from 0.6160 to 0.6437.

***Index Terms*—** Subtitle Recognition, Subtitle Translation, Chinese Short Dramas, Multimodal System

1. INTRODUCTION

The growing popularity of short dramas in China has increased the demand for high-quality subtitle translation. Subtitles in these works pose several challenges. First, many subtitle files (e.g., SRT) lack consistent punctuation, complicating sentence boundary detection and syntactic analysis [1, 2]. Second, because subtitles must fit limited on-screen time, they tend to be brief and fragmented, which can result in incomplete information or ambiguity [3]. Third, subtitles often contain incomplete or colloquial utterances that omit subjects or discourse markers, further complicating translation [4, 5]. Fourth, each subtitle segment is often presented in isolation without explicit reference to preceding dialogue or visual context, yet correct interpretation of pronouns and conversational flow requires such context [6].

In addition, subtitle recognition also faces challenges: subtitles can be obscured by complex backgrounds or low-quality frames, and automatic speech recognition (ASR) must deal with background noise, overlapping speech, and accent variation common in short dramas [7, 8].

To address these challenges [9], we propose an end-to-end multimodal system. For subtitle recognition, we apply Qwen2-VL [10] to extract subtitles from video frames via optical character recognition (OCR), and employ Whisper [11] for ASR. While Whisper performs excellently in temporal alignment and producing fluent transcriptions, it struggles with recognizing proper nouns and character names. Conversely, Qwen2-VL presents high accuracy for technical terms but may misclassify non-subtitle textual elements. To leverage their complementary strengths, we propose a multimodal fusion strategy that compares OCR-based frame text and ASR outputs and selects the most accurate subtitles. For subtitle translation, we process the entire video sequence as a unified translation task rather than translating individual clips [12]. We further fine-tune Qwen2.5 [13] on our constructed Chinese–Japanese dataset using Low-Rank Adaptation (LoRA) [14], improving performance on this task [15].

The contributions of this work are as follows: **(1)** We propose a novel end-to-end multimodal system that jointly recognizes subtitles from visual and audio channels of short dramas and performs context-aware subtitle translation. Our system outperforms baselines in both subtitle recognition and Chi-

This work was supported by the Postdoctoral Fellowship Program for Overseas Talent Introduction, Ministry of Education of China.

† These authors contributed equally to this work.

* Corresponding author: ybbai@ruc.edu.cn

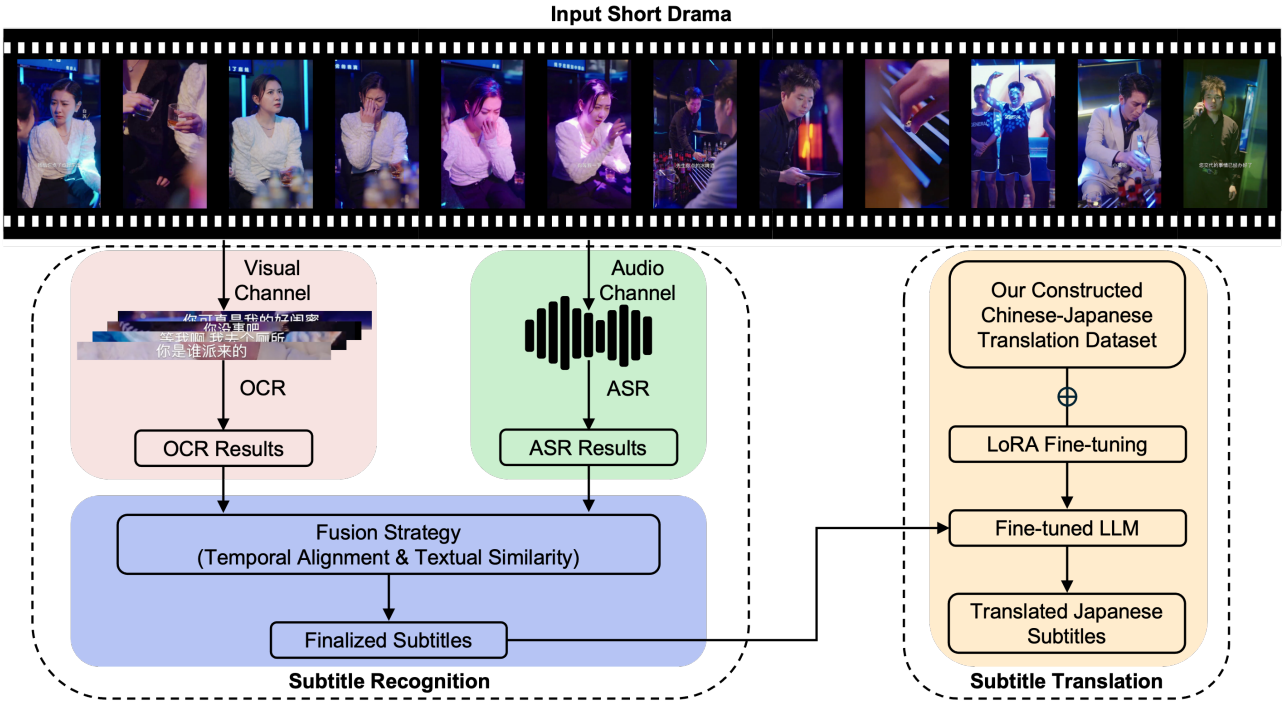

Fig. 1. Architecture of the proposed end-to-end multimodal system, comprising subtitle recognition and translation modules.

nese–Japanese subtitle translation. **(2)** We introduce a multimodal fusion strategy that compares frame-extracted text with ASR transcriptions to select the most accurate subtitles. **(3)** We construct a short-drama–based Chinese–Japanese translation dataset, and provide details on dataset statistics.

2. RELATED WORK

2.1. Subtitle Recognition

The subtitle extraction task primarily relies on OCR technology. In recent years, the rapid development of visual–language models (VLMs) has shown remarkable performance in this task [16, 17, 18]. Among these models, Qwen2-VL [10] stands out as a state-of-the-art model, offering robust perceptual capabilities and the ability to process images at various resolutions. Its architecture integrates visual encoding with large language model (LLM) inference, enabling accurate text recognition even in challenging conditions such as cluttered backgrounds, variable lighting, and diverse text styles. Furthermore, its multimodal fusion mechanism facilitates a deeper contextual understanding of visual content, which is particularly useful when textual information is ambiguous.

In addition, ASR is a key approach for extracting subtitles from video audio tracks [8, 19, 20]. Whisper [11] represents a significant advancement in ASR, developed using large-scale weakly supervised training on 680,000 hours of multilingual audio data. The model employs a Transformer-based encoder–decoder architecture, which allows it to perform well across languages, accents, and acoustic conditions without language-specific fine-tuning. Its robust temporal alignment and multilingual capabilities make it particularly effective for subtitle recognition, especially when processing colloquial speech typical of short dramas.

2.2. Subtitle Translation

With the widespread adoption of LLMs, their performance in subtitle translation has achieved satisfactory results [21, 22, 23]. Qwen2.5 [13], a state-of-the-art LLM, presents strong multilingual capabilities and can efficiently perform Chinese-Japanese translation. Its advanced instruction-following ability, combined with extensive training on multilingual corpora, enables it to tackle complex translation scenarios. Its architecture also supports efficient fine-tuning through methods such as LoRA [14], allowing domain-specific adaptation while maintaining computational efficiency.

3. METHOD

3.1. System Architecture

We propose a novel end-to-end multimodal system for subtitle extraction and translation in short dramas, as shown in Fig. 1. The system first extracts video frames from the input short drama at fixed intervals, enabling parallel processing of the visual and audio channels. For subtitle recognition, Qwen2-VL [10] and Whisper [11] are employed for OCR and ASR,

respectively. The outputs of these two channels are then integrated using our proposed multimodal fusion strategy, which leverages temporal alignment and textual similarity to select accurate and time-aligned Chinese subtitles. These subtitles are subsequently fed into our fine-tuned Qwen2.5 [13], optimized via LoRA [14] on our constructed Chinese–Japanese subtitle translation dataset, ultimately producing contextually appropriate Japanese subtitles.

3.2. Subtitle Recognition

We employ two parallel channels for subtitle recognition. In the visual channel, frames are sampled from the input short drama at 1.0-second intervals, balancing comprehensive coverage with computational efficiency while ensuring that no subtitle content is missed within the typical display duration. Each frame is processed using Qwen2-VL [10] model for OCR. To further improve efficiency, we implement a caching mechanism for both video frames and Qwen2-VL recognition results, thereby avoiding redundant computations. In the audio channel, Whisper [11] model transcribes the entire audio track, generating text segments with precise start and end timestamps.

Our proposed fusion strategy integrates temporal alignment with text similarity to effectively combine visual and audio information. Whisper's time-stamped text segments serve as primary anchors. For each Whisper segment, a 1.5-second tolerance window is used to identify all temporally overlapping Qwen2-VL OCR results, accounting for potential offsets between audio and visual subtitles during production. The edit distance similarity between each candidate Qwen2-VL text and the corresponding Whisper text is computed using the RapidFuzz library's ratio algorithm. When the highest similarity exceeds a 60% threshold, the result of Qwen2-VL is adopted to replace the outputs of Whisper while retaining the original timestamp, balancing OCR accuracy with ASR temporal consistency. If no Qwen2-VL text meets the threshold, the original Whisper segment is preserved. This strategy effectively leverages OCR precision alongside ASR fluency and timing, producing high-quality Chinese subtitles.

3.3. Subtitle Translation

For subtitle translation, we employ supervised fine-tuning to adapt LLMs to this specific task. To address the contextual limitations discussed in the motivation, we treat the translation of the entire subtitle sequence of a short drama as a holistic task. This method preserves dialogue coherence, pronoun reference, and narrative fluency, all of which are essential for accurate subtitle translation. For model selection, we adopt Qwen2.5 as the baseline and perform efficient fine-tuning using 4-bit quantization in combination with LoRA [14]. Specifically, we set the LoRA rank r to 16 and the scaling factor α to 32. A rank of 16 provides sufficient adaptability while maintaining parameter efficiency. This setting is recommended for 3B-scale models to mitigate overfitting on limited subtitle data. Moreover, the 2:1 α-to-r ratio facilitates effective learning rate adjustment and supports stable convergence.

4. EXPERIMENTS

4.1. Dataset

We construct a Chinese–Japanese subtitle translation dataset for Chinese short dramas, comprising multiple commercially available clips and their corresponding subtitles with speaker IDs. The short drama content is sourced from entertainment companies to ensure authenticity and professional quality. The ground-truth Chinese subtitles are manually annotated by experts to guarantee accuracy and reliability. The target Japanese subtitles are manually translated by native Japanese researchers and further verified to ensure high-quality and culturally appropriate translations. The dataset reflects typical characteristics of subtitle content, including highly colloquial language, minimal punctuation, short segments, and limited contextual information. It provides a robust benchmark for evaluating model performance and dialogue coherence in real-world scenarios.

4.2. Evaluation Metrics and Baseline

For subtitle recognition, we employ four metrics: Character Error Rate (CER), Accuracy, Bilingual Evaluation Understudy (BLEU), and the enhanced character n-gram F-score (chrF++). For subtitle translation, we evaluate models using BLEU, chrF++, and the Crosslingual Optimized Metric for Evaluation of Translation (COMET).

In addition, we adopt zero-shot Qwen2-VL-2B [10] and Whisper-medium [11] as baselines for subtitle recognition, and zero-shot Qwen2.5-3B [13] as the baseline for subtitle translation.

4.3. Implementation Details

To ensure fair model comparison, all experiments are conducted on an NVIDIA GeForce RTX 3090 GPU on an Ubuntu system. To evaluate our fine-tuned Qwen2.5-3B [13] model for subtitle translation, we employ a five-fold cross-validation strategy on our constructed dataset with stratified random splits to enhance robustness and ensure reliable evaluation. Each fold is trained for 10 epochs with a learning rate of 2×10^{-4}, a batch size of 4, and four gradient accumulation steps. The chosen learning rate balances convergence speed and stability for LoRA fine-tuning. The small batch size accommodates GPU memory constraints, and gradient accumulation effectively increases the batch size for stable training. To prevent overfitting, we apply an early stopping strategy with a patience of three epochs for early stopping.

Table 1. Quantitative comparison of subtitle recognition.

Model	CER↓	Accuracy↑	BLEU↑	chrF++↑
Qwen2-VL [10]	0.2984	**0.9216**	72.3279	70.4881
Whisper [11]	0.2491	0.7819	81.2538	57.5461
Ours	**0.1598**	0.9174	**85.5974**	**77.963**

Table 2. Quantitative comparison of subtitle translation.

Model	BLEU↑	chrF++↑	COMET↑
Qwen2.5 [13]	9.7665	27.8855	0.6160
Ours[a]	**9.8440**	**29.9883**	**0.6437**

[a]Ours are obtained using five-fold cross-validation.

4.4. Quantitative Comparison

4.4.1. Subtitle Recognition

Table 1 presents a quantitative comparison of subtitle recognition performance across different models. Qwen2-VL [10] achieves the highest accuracy of 0.9216 but suffers from a relatively high CER of 0.2984. Whisper [11] obtains a lower CER of 0.2491 and a competitive BLEU score of 81.2538, yet its accuracy drops to 0.7819 and its chrF++ score remains limited at 57.5461. In contrast, our proposed model achieves the best overall performance, with the lowest CER of 0.1598, the second-highest accuracy of 0.9174, and the highest BLEU score of 85.5974 and chrF++ score of 77.963. These results highlight the superior effectiveness of our model in subtitle recognition and demonstrate that the proposed fusion strategy successfully integrates the strengths of both baseline models.

4.4.2. Subtitle Translation

For subtitle translation, we adopt a five-fold cross-validation procedure on our constructed dataset. Table 2 presents the performance of our fine-tuned model compared to the baseline, demonstrating that the fine-tuned model consistently outperforms the zero-shot baseline across all metrics. Specifically, our model achieves improvements of 0.79% in BLEU, 7.54% in chrF++, and 4.50% in COMET compared to the baseline. These results indicate that fine-tuning Qwen2.5 [13] using LoRA on our dataset effectively captures domain-specific language patterns and idiomatic expressions in subtitle translation.

4.5. Characteristics of Our Proposed Dataset

Our proposed dataset is derived from a complete Chinese short drama, provided by a commercial entertainment company. It is organized by episode and comprises 79 pairs of samples, each comprising three components: a video file (.mp4), a Chinese subtitle file (.srt) with speaker IDs, and a corresponding Japanese subtitle file (.txt). Dataset statistics are presented in Table 3. All Chinese and Japanese subtitles were manually annotated by native-speaking researchers to ensure both translation quality and cultural appropriateness. The inclusion of character annotations in the subtitles provides essential context for speaker-specific dialogue analysis, while the distribution of subtitle segments per episode is shown in Fig. 2.

Table 3. Statistical information of the proposed dataset.

Dataset composition		**Subtitle statistics**	
Subtitle files	79	Total duration (min)	130.56
Paired episodes	79	Subtitle segments	3,692

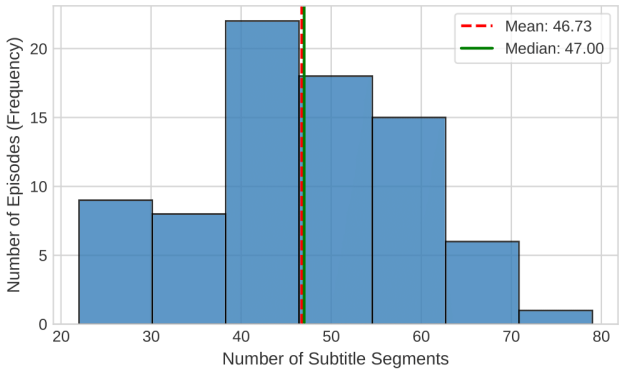

Fig. 2. Distribution of subtitle segments per episode.

5. CONCLUSION

This work presents a novel end-to-end multimodal system for subtitle recognition and Chinese–Japanese translation, specifically designed to address the unique challenges of short drama subtitles. For subtitle recognition, we propose a dual-channel fusion strategy that integrates OCR and ASR techniques, leveraging the complementary strengths of Qwen2-VL for visual recognition and Whisper for audio transcription. Our proposed method yields superior performance compared to existing baselines. For subtitle translation, we fine-tune Qwen2.5 using LoRA. The fine-tuned model consistently outperforms zero-shot baselines across all evaluation metrics. In addition, we construct a dataset for Chinese-to-Japanese subtitle translation in short dramas, which captures the distinctive characteristics of subtitle content, including colloquial expressions, fragmented segments, and strict temporal constraints, providing a valuable resource for this task.

For future work, we plan to expand the dataset to cover a wider range of domains and language pairs, enhancing the generalizability of our method. We also intend to explore more advanced fusion strategies, such as attention-based mechanisms, to further improve recognition accuracy. Moreover, we aim to incorporate real-time constraints and optimize deployment efficiency, facilitating practical applications in production environments.

6. REFERENCES

[1] Ines Rehbein, Josef Ruppenhofer, and Thomas Schmidt, "Improving sentence boundary detection for spoken language transcripts," in *LREC*, 2020, pp. 7102–7111.

[2] Gregor Donabauer, Udo Kruschwitz, and David Corney, "Making sense of subtitles: Sentence boundary detection and speaker change detection in unpunctuated texts," in *WWW Companion*, 2021, pp. 357–362.

[3] Daniel Li, I Te, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield, "Sentence boundary augmentation for neural machine translation robustness," in *ICASSP*, 2021, pp. 7553–7557.

[4] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow, "Evaluating discourse phenomena in neural machine translation," in *NAACL*, 2017, pp. 1304–1313.

[5] Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang, "Modeling coherence for discourse neural machine translation," in *AAAI*, 2019, vol. 33, pp. 7338–7345.

[6] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov, "Context-aware neural machine translation learns anaphora resolution," in *ACL*, 2018, pp. 1264–1274.

[7] Hongyu Yan and Xin Xu, "End-to-end video subtitle recognition via a deep residual neural network," *Pattern Recognition Letters*, vol. 131, pp. 368–375, 2020.

[8] Danni Liu, Jan Niehues, and Gerasimos Spanakis, "Adapting end-to-end speech recognition for readable subtitles," in *IWSLT*, 2020, pp. 247–256.

[9] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," *NeurIPS*, vol. 36, pp. 72842–72866, 2023.

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, et al., "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023, pp. 28492–28518.

[12] Colin Cherry, Naveen Arivazhagan, Dirk Padfield, and Maxim Krikun, "Subtitle translation as markup translation," in *Interspeech*, 2021, pp. 2237–2241.

[13] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, et al., "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, pp. 3, 2022.

[15] Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen, "Eliciting the translation ability of large language models via multilingual finetuning with translation instructions," *TACL*, vol. 12, pp. 576–592, 2024.

[16] Haiyang Yu, Jinghui Lu, Yanjie Wang, Yang Li, Han Wang, Can Huang, and Bin Li, "Eve: Towards end-to-end video subtitle extraction with vision-language models," *arXiv preprint arXiv:2503.04058*, 2025.

[17] Sankalp Nagaonkar, Augustya Sharma, Ashish Choithani, and Ashutosh Trivedi, "Benchmarking vision-language models on optical character recognition in dynamic video environments," *arXiv preprint arXiv:2502.06445*, 2025.

[18] Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yi-Fan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, et al., "Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios," *arXiv preprint arXiv:2505.21333*, 2025.

[19] Qi Liu, Zhehuai Chen, Hao Li, Mingkun Huang, Yizhou Lu, and Kai Yu, "Modular end-to-end automatic speech recognition framework for acoustic-to-word model," *IEEE/ACM TASLP*, vol. 28, pp. 2174–2183, 2020.

[20] Haoyuan Yang, Yue Zhang, and Liqiang Jing, "Speech recognition on tv series with video-guided post-correction," *arXiv preprint arXiv:2506.07323*, 2025.

[21] Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe, "Translating movie subtitles by large language models using movie-meta information," in *ACL*, 2025, pp. 315–330.

[22] Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Zongyao Li, Yuanchang Luo, Jinlong Yang, Zhiqiang Rao, and Hao Yang, "Combining the best of both worlds: A method for hybrid nmt and llm translation," in *Findings of ACL*, 2025, pp. 5140–5148.

[23] Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang, "Multilingual machine translation with open large language models at practical scale: An empirical study," in *NAACL*, 2025, p. 5420–5443.