HW-TSC's submission to the IWSLT 2026 Subtitling track

Xiaoqing Lan[1,2], **Daimeng Wei**[1], **Jiaxin Guo**[1], **Yuanchang Luo**[1], **Hengchao Shang**[1], **Zongyao Li**[1], **Zhiqiang Rao**[1], **Jinlong Yang**[1], **Zhanglin Wu**[1], **Boqi Huang**[1,2],**Yu He**[1,2]
[1]Huawei Translation Service Center, Beijing, China
[2]School of Software, Northwestern Polytechnical University, Xi'an, China
{lanxiaoqing, weidaimeng, guojiaxin1, luoyuanchang1, shanghengchao, lizongyao, raozhiqiang, yangjinlong7, wuzhanglin2, huangboqi1,heyu97}@huawei.com

Abstract

This paper introduces HW-TSC's submission to the IWSLT 2026 Subtitling track. For automatic subtitle generation, we employ a cascaded strategy under unconstrained conditions. First, we construct a large-model-based streaming speech recognition framework, which incorporates VAD voice activity detection, sliding-window context caching, long audio chunking, and the Qwen3 forced alignment model to achieve high-precision transcription and timestamping from English speech to text. Next, we perform text translation using a Qwen3-based translation model. Finally, according to subtitle constraints such as characters per second (CPS) and characters per line (CPL), we identify translation segments that exceed compliance thresholds via quantitative evaluation, and rewrite them using a large language model while preserving core semantic meaning, ultimately producing subtitle files that meet the required standards.

1 Introduction

Faced with the massive volume of audio-visual content produced daily, the automatic subtitling task has attracted widespread attention, driving the demand for high-quality subtitle generation solutions (Álvarez et al., 2016; Vitikainen and Koponen, 2021; Hou et al., 2021; Sun et al., 2025). Current technical approaches can be broadly divided into two categories: the cascaded strategy and the end-to-end strategy. The conventional cascaded strategy adopts a pipelined architecture with distributed multi-module execution, which generally completes subtitle generation sequentially through speech recognition, subtitle segmentation, machine translation and text compression, with each subtask processed independently (Bentivogli et al., 2021). In contrast, the end-to-end strategy aims to generate subtitles directly from audio or audio-visual inputs within a unified framework. Such models enable the joint learning and optimization of multiple subtasks in subtitling, thereby mitigating the error propagation issue inherent in cascaded systems (Liu et al., 2020; Bérard et al., 2016; Papi et al., 2023).

The International Conference on Spoken Language Translation (IWSLT) is a premier annual academic conference dedicated to all research directions in spoken language translation. Since 2023, it has launched the automatic subtitling task, which requires participants to produce high-quality subtitles while satisfying multiple practical constraints, including accurate timeline alignment, reasonable subtitle segmentation, and appropriate reading speed. Unlike previous editions, the 2026 competition has eliminated the separate text compression track, expanded the target language coverage to include Chinese and Japanese for a multilingual setting, and formulated refined limitations on character count and reading speed for different languages (Adelani et al., 2026).

Taking Chinese subtitling as an example, the official constraints are clearly defined as follows: i) each subtitle block (Lines Per Block, LPB) shall contain no more than two lines; ii) the number of Chinese characters per line (Characters Per Line, CPL) is limited to 16; iii) to ensure a comfortable reading experience, the maximum reading speed for Chinese subtitles is restricted to 9 characters per second (Characters Per Second, CPS). Accordingly, qualified subtitles should be organized into text blocks that fully comply with the above specifications, and the degree of rule compliance can be quantified by the proportion of standardized subtitle blocks in all generated results.

In this paper, we adopt a cascaded pipeline for automatic subtitle generation. Considering the outstanding performance of the Qwen3 series models across various domains (Yang et al., 2025; Shi et al., 2026), the proposed cascaded solution is expected to achieve promising results on the subtitling task.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 91–96

Subtitles directly obtained through speech recognition and machine translation often fail to meet practical broadcasting standards. Restricted by visual display limitations, subtitle presentation should be adapted to the video playback rhythm and audience reading speed. Therefore, subtitle compression serves as an essential component within the overall workflow. Non-compliant translated segments are identified via quantitative evaluation metrics. On the premise of preserving the core semantic information of the source text, large language models are employed to compress and rephrase the target texts, with the goal of generating final subtitles that fully meet official formatting and readability constraints.

2 Automatic Subtitling

We propose a Qwen3-based cascaded automatic subtitling strategy. The overall architecture is illustrated in Figure.1, and the details are as follows:

2.1 Streaming Speech Recognition

We design and implement a multi-module collaborative system for streaming speech recognition and timestamp alignment to generate subtitle texts for long-form audio. Built upon the Qwen3 series as the foundational models, the system integrates key functional modules including voice activity detection, streaming chunked inference, forced alignment, and hallucination filtering.

The system first divides the input audio into segments of fixed duration. The Silero VAD model is adopted to detect voice activity in real time and filter silent segments. Subsequently, a 2-second sliding window is used for streaming audio segmentation and incremental processing. By dynamically maintaining cached contextual audio and textual information, the system guarantees the coherence of recognition outputs. During the recognition and inference stage, audio features extracted by the encoder are fused with textual prompts. The vLLM (Kwon et al., 2023) inference framework is utilized to achieve efficient decoding of large language models, and length-based threshold filtering is adopted to suppress model hallucinations and enhance output reliability. To satisfy the temporal accuracy requirements of subtitling tasks, the forced alignment module of Qwen3 is introduced to realize token-level temporal synchronization between transcribed texts and corresponding audio. Following this pipeline, the system generates standardized subtitles with accurate timestamps, which serve as reliable support for subsequent translation and subtitle compression.

2.2 Text Preprocessing

After completing streaming speech recognition, we process ASR transcripts through text merging and length constraint adjustment to generate English text blocks that meet subtitling specifications. Consecutive ASR segments are merged at the semantic level according to sentence-ending punctuation. Short fragmented outputs from the same sentence are integrated into complete utterances, while the timestamp information of each segment is fully preserved. On this basis, we set an upper limit on word count based on the prior length ratio between English and Chinese. For long text blocks that exceed the threshold, a greedy strategy is applied to reorganize internal subsegments, ensuring that the word quantity of each subtitle block satisfies relevant constraints. This workflow maintains semantic integrity and timestamp accuracy. Meanwhile, it enables subtitle texts to adapt to display space limitations and standard reading speed, and provides high-quality input for the subsequent machine translation stage.

2.3 Machine Translation

Considering the high temporal accuracy of timestamps provided by the ASR system, we only translate English source sentences into the target language during subtitle generation, while keeping all temporal information unchanged. In the IWSLT 2026 Subtitling task, we input structurally standardized English text units into the Qwen3-based translation model for processing, and finally obtain translated results that are strictly aligned with the source content.

3 Subtitle Compression

Subtitle quality after translation requires comprehensive consideration. The official competition rules clearly require all participating teams to balance translation quality with standardized subtitle formatting constraints. To address this requirement, we aim to compress and restructure raw translated texts rationally. By fully retaining valid semantic information from the original content, our method maintains high translation fidelity while enabling the generated subtitles to comply with official CPS, CPL and LPB limitations.

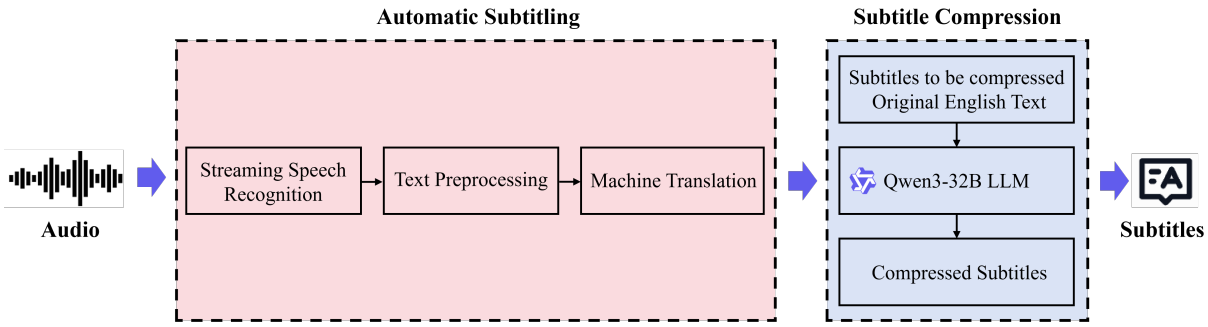

Figure 1: Overall architecture of the proposed cascading subtitle system.

The subtitle compression task requires sentence-level rewriting of subtitle texts while preserving the timestamp information of the original subtitle files. This means that the compression process should preserve the original semantic information, while avoiding fusion compression of multiple sentences.

To adapt translated subtitles to the display constraints of video playback, we propose a progressive subtitle compression scheme based on large language models. Taking Qwen3-32B as the core model, our approach conducts constrained concise rewriting on translated subtitle segments. It ensures complete semantics and coherent expression, and enables the generated subtitles to simultaneously satisfy the dual specifications of CPL and CPS.

During the processing procedure, we first parse the start and end timestamps of each subtitle and calculate its actual playback duration, so as to automatically identify unqualified segments that exceed length or rate constraints. For these non-compliant segments, our approach is to adopt a two-stage constrained optimization strategy. In the first stage, greedy decoding with a temperature of 0 is applied. Under deterministic generation without random sampling, the model condenses translated texts by only removing redundant auxiliaries and conjunctions, thereby preserving complete semantic meaning and sentence structure. If the constraints remain unsatisfied after initial simplification, the second-stage deep compression is activated with the temperature adjusted to 0.3.

The entire compression process refers strictly to the original English text. With instructional constraints, we ensure the retention of proper nouns such as personal names, geographical locations and institutional entities, and avoid any modification or omission of core semantic information. The final compressed subtitles can fully conform to screen display limits and audience reading speed requirements. This approach effectively improves constraint compliance while maintaining the accuracy, readability and viewing experience of news-style subtitles.

We propose a subtitle rewriting scheme based on Qwen3-32B. This large language model possesses powerful capabilities in semantic understanding and text generation, and can be deployed to accomplish the simplification and rewriting of translated subtitles. Specifically, we design professional role prompts and constraint rules for the model, which clearly define key requirements including character length, characters per second, and the reservation of proper nouns. In each inference step, we input both the original English text and its Chinese translation into the model. The model then performs text compression and concise rewriting under strict compliance with the predefined rules. The detailed prompts are presented as follows:

Role: Professional News Subtitle Refinement Expert

Task: Refer to the original English text to condense the Chinese translation, and strictly comply with the following constraints:

1. Total character count $\leq$ 32 characters.

2. Characters displayed per second $\leq$ 9 characters.

3. Fully retain all personal names, place names, country names, institutional names and proper nouns.

4. Only remove redundant auxiliaries and conjunctions; do not modify, abbreviate or omit core nouns.

5. Ensure coherent expression and complete semantics, with priority given to news accuracy.

4 Experiments

4.1 Dataset

Participants are required to automatically generate subtitles for three categories of audiovisual documents, with English as the sole spoken language. Audio and video files of the development and evaluation sets are provided in MP4 format for Asharq-Bloomberg and ITV materials, while YODAS data adopts the WAV format. The competition requires participants to use only the audio tracks contained in the provided videos. The video tracks are delivered with low visual quality, and are mainly adopted to verify temporal synchronization and other display-related characteristics of on-screen subtitles.

1) ITV: The ITV dataset is provided by ITV Studios, a subsidiary of the largest commercial broadcasting organization in the United Kingdom. This institution develops and produces diverse content including TV dramas, entertainment programs and documentaries across 13 countries worldwide, and distributes such works globally with high-quality official subtitles. The data adopted in this competition consists of MP4-format video materials featuring colloquial language and naturally variable speaking rates. Meanwhile, the standardized subtitle specifications of this dataset enable effective evaluation of model performance in translating colloquial and scenario-based audiovisual content.

2) Asharq-Bloomberg: The Asharq-Bloomberg dataset is derived from financial news programs produced by SRMG, the largest integrated media group in the Middle East and North Africa (MENA) region. All data is delivered in MP4 format, with content focusing on global finance, business and economic information under a rigorous and authoritative news production framework. This dataset features a relatively formal and standardized linguistic style with high information density, and contains a wide range of domain-specific terminology in the economic and financial fields.

3) YODAS: The YODAS (YouTube-Oriented Dataset for Audio and Speech) originates from diverse original video content on the YouTube platform. This dataset covers large-scale real-world speech data across multiple languages, with official audio resources provided in WAV format for experimental evaluation. It features a highly colloquial linguistic style and flexible, informal expressions. Meanwhile, the acoustic conditions are complex and diversified, frequently involving background noise, background music, and overlapping speech among multiple speakers.

4.2 Evaluation Metrics

Subtitle quality evaluation needs to take both translation performance and compliance with subtitle production constraints into account. The evaluation metrics specified by the competition are detailed as follows:

1) SubER (Subtitle Edit Rate): SubER (Wilken et al., 2022) serves as the primary metric to measure the overall quality of automatically generated subtitles, which comprehensively reflects translation accuracy, temporal consistency and subtitle standardization.

2) BLEU/BLEURT: These metrics are adopted to evaluate the quality of translated texts. BLEU (Papineni et al., 2002; Post, 2018) measures the matching degree between machine translations and reference texts based on n-gram precision, while BLEURT (Sellam et al., 2020) is a deep neural metric. Together, they reflect the accuracy and fluency of translation. Following official guidelines, appropriate tokenization strategies are applied to different languages in this competition.

3) CPS (Characters Per Second): This metric denotes the number of characters displayed per second and is used to evaluate the compliance of subtitle reading speed. The competition sets an upper limit of 9 characters per second for Chinese content; any value exceeding this threshold is deemed inconsistent with audience reading habits.

4) CPL (Characters Per Line): This indicator refers to the maximum number of characters per subtitle line and is designed to restrict the single-line length of subtitles. For Chinese subtitles, the upper limit is set to 16 characters per line, so as to guarantee the visual aesthetics and readability of on-screen subtitles.

5) LPB (Lines Per Block): This metric defines the maximum number of lines contained in each subtitle segment to constrain the line count of individual subtitles. The competition stipulates an upper limit of two lines per block, in compliance with the presentation standards adopted for conventional film and television subtitles.

4.3 Experimental Results

Experiments are conducted on the IWSLT2026 development dataset, and all metric scores are calculated with the results presented in Table1. The case-sensitive and punctuation-aware SubER is

Task	Condense	SubER (↓)	BLEU (↑)	CPS% (↑)	CPL% (↑)	LPB% (↑)
Asharq	×	60.10	28.95	94.33	87.65	100.00
	✓	59.29	29.00	98.81	98.97	100.00
ITV	×	63.50	22.77	70.08	96.25	100.00
	✓	62.94	22.40	87.16	99.29	100.00
YODAS	×	54.86	29.43	75.48	93.14	100.00
	✓	54.24	29.63	87.60	99.62	100.00

Table 1: Subtitle condensation results of three IWSLT tasks for en2zh on the 2026 development set.

adopted as the primary evaluation metric, while the BLEU score is used to measure translation quality. Before BLEU score calculation, mweralign (Post and Hoang, 2025) is employed to realign machine-generated subtitles with reference subtitles, which serves as a variant of the AS-WER algorithm. The compliance metrics are calculated with the script provided by Papi et al. (Papi et al., 2023).

Our subtitle compression strategy leads to an upward trend in both SubER and BLEU scores. Text compression is generally considered to impair translation quality, while the experimental results show a certain discrepancy from this perception. Two major reasons account for this phenomenon. On one hand, large language models effectively preserve core semantics and cut redundant expressions during compression, rendering generated subtitles consistent with references in stylistic features, sentence structures and diction. On the other hand, reference subtitles are standardized and concise, with shorter length compared with verbatim oral translations. Compressed outputs share similar textual characteristics with references, which also contributes to the improvement of evaluation metrics. In terms of compliance metrics, we achieve nearly 100% compliance with the LPB constraint, and simultaneously attain effective improvements in CPS and CPL.

5 Conclusion

This paper presents the system submitted by the HW-TSC team for the unrestricted subtitle track of IWSLT2026. For the automatic subtitle generation task, we propose a Qwen3-based cascaded strategy. By constructing a streaming speech recognition framework and adopting Qwen3-powered translation models, our method produces target translations that are strictly aligned with source utterances. To adapt generated subtitles to the display constraints of video playback, we further develop a large language model empowered progressive subtitle compression scheme. Experimental results demonstrate that the proposed compression strategy effectively improves overall subtitle quality, which validates the effectiveness and application potential of our method. In future work, we will further explore high-quality subtitle generation solutions suitable for multilingual scenarios.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, Tiago Luís, Sérgio Paulo, Nicola Piccinini, Haritz Arzelus, João Neto, Carlo Aliprandi, and Arantza Del Pozo. 2016. Automating live and batch subtitling of multimedia contents for several european languages. *Multimedia Tools and Applications*, 75(18):10823–10853.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Jingyi Hou, Yayun Qi, Xinxiao Wu, and Yunde Jia. 2021. Cross-lingual knowledge distillation for chinese video captioning. *Chinese Journal of Computers*, 44(9):1907–1921.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Danni Liu, Jan Niehues, and Gerasimos Spanakis. 2020. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Matt Post and Hieu Hoang. 2025. Effects of automatic alignment on speech translation metrics. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 84–92.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Haoying Sun, Shuyi Li, Zeyu Xi, and Lifang Wu. 2025. Spatiotemporal enhancement of video captioning integrating a state space model and transformer. *Journal of Signal Processing*, 41(2):279–289.

Kaisa Vitikainen and Maarit Koponen. 2021. Automation in the intralingual subtitling process: Exploring productivity and user experience. *Journal of Audiovisual Translation*, 4(3):44–65.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. Suber-a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.