HW-TSC's Submissions To the IWSLT 2026 Offline Speech Translation Task

Boqi Huang[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], Yu He[1,2], Xiaoqing Lan[1,2]
[1]Huawei Translation Service Center, Beijing, China
[2]School of Software, Northwestern Polytechnical University, Xi'an, China
{huangboqi1, weidaimeng, guojiaxin1, luoyuanchang1, shanghengchao, lizongyao, raozhiqiang, yangjinlong7, wuzhanglin2, heyu97, lanxiaoqing}@huawei.com

Abstract

This paper describes the HW-TSC's submission to the IWSLT 2026 Offline Speech Translation Task, specifically for the English-to-Chinese and English-to-German unconstrained tracks. Our system adopts a robust cascade architecture optimized for long-form, unsegmented audio. To mitigate the hallucination and inconsistency issues common in long-sequence processing, we propose a two-pass transcription strategy: an initial streaming ASR with a 12-second context buffer for sentence-level coherence, followed by Qwen3-ForcedAligner for precise timestamping. Based on these alignments, a second-pass refinement is conducted using Qwen3-Omni on re-segmented 30-second chunks to ensure high-fidelity transcriptions. For the translation module, we employ a context-aware segment merging strategy (up to 150 tokens) to empower the Qwen3 LLM with sufficient semantic context. Experimental results on the tst-2022 benchmark demonstrate the effectiveness of our pipeline, achieving COMET scores of 0.8462 (En-Zh) and 0.7854 (En-De), significantly outperforming the standard cascade baselines.

1 Introduction

The rapid evolution of multimodal large language models (MLLMs) (Yin et al., 2024) and end-to-end speech processing (Bérard et al., 2016) has inaugurated a new era of efficient and robust speech translation. As a pivotal cross-modal task, speech translation bridges linguistic gaps and enables seamless communication in scenarios such as international conferences and media localization. The IWSLT 2026 Offline Unconstrained Track (Adelani et al., 2026) serves as a rigorous benchmark to evaluate systems under realistic conditions. It encourages the use of state-of-the-art pretrained models and advanced engineering to address challenges such as long-form audio and noisy environments.

While end-to-end models such as Whisper (Radford et al., 2023) and SeamlessM4T (Barrault et al., 2023) have made significant progress, they still encounter limitations in data scalability and robustness within unconstrained scenarios, particularly regarding hallucination control in zero-shot settings. In contrast, cascade speech translation systems (Bentivogli et al., 2021) decompose the task into automatic speech recognition (ASR) (Gulati et al., 2020) and machine translation (MT) (Vaswani et al., 2017). This paradigm offers superior flexibility, interpretability, and compatibility with mature pretrained models. Recent studies have further demonstrated that leveraging Large Language Models (LLMs) as translation engines can significantly improve semantic consistency compared to traditional NMT systems (Zhu et al., 2024). Considering the complexity of the IWSLT 2026 test set, we adopt a cascade architecture for our submission, focusing on English-to-Chinese and English-to-German directions.

We leverage the unconstrained track settings to design a robust pipeline centered on the Qwen3 (Yang et al., 2025) model family. First, we apply voice activity detection (VAD) using the Silero VAD framework (Team, 2024) to filter non-speech segments, which reduces computational overhead and mitigates transcription hallucinations. For the ASR module, we implement a streaming inference framework featuring context caching and a two-pass transcription strategy. We utilize Qwen3-Omni (Xu et al., 2025) and Qwen3-4B for initial transcription and employ Qwen3-ForcedAligner (Team, 2026) to generate precise timestamps. Subsequently, we re-segment the audio based on these timestamps for a second-pass refinement to enhance recognition quality. Finally, we merge the refined transcriptions into semantic segments and feed them into the Qwen3-8B for translation, employing a context-aware prompting strategy to maximize translation fidelity (Jean and Cho, 2019).

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 84–90

The remainder of this paper is organized as follows. Section 2 describes the datasets used in our experiments. Section 3 details the system architecture, including audio preprocessing, streaming ASR, and the translation module. Section 4 presents the experimental setup and evaluation results. Finally, Section 5 summarizes our findings and outlines future research directions.

2 Dataset

2.1 ASR Data and Adapter Tuning

We utilize six large-scale speech datasets to train the audio adapter and fine-tune the decoding capabilities of our ASR module: MuST-C V2 (Cattoni et al., 2021), LibriSpeech (Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), CoVoST 2 (Wang et al., 2020), VoxPopuli (Wang et al., 2021), and Europarl-ST (Iranzo-Sánchez et al., 2020) (Table 1).

Dataset	Duration (Hours)
LibriSpeech	960
MuST-C V2	590
CoVoST 2	1,802
TED-LIUM 3	453
Europarl-ST	161
VoxPopuli	1,270
Total	**5,236**

Table 1: Statistics of the speech corpora used for ASR training.

Unlike standard ASR training, our ASR module connects a Qwen3-Omni speech encoder with a Qwen3-4B language model through a trainable audio adapter. The adapter is not an off-the-shelf component. It is implemented as a lightweight two-layer feed-forward multimodal projector, consisting of a linear layer, a ReLU activation, dropout, and a second linear layer. It maps the frame-level acoustic representations produced by the Qwen3-Omni speech encoder into the hidden embedding space of the Qwen3-4B language model.

We train the adapter together with the ASR decoding module on the speech-transcription data listed in Table 1. The model is optimized with the standard autoregressive cross-entropy loss over the target transcription tokens. Specifically, we supervise the model to predict transcriptions in which the newline character `\n` is explicitly inserted at the end of complete semantic units. This encourages the joint model to function as a sentence-aware transcriber and provides semantically coherent segments for the subsequent timestamp alignment and translation stages.

To enhance robustness for long-form speech, we also apply the data augmentation method, concatenating adjacent audio clips to simulate continuous, unsegmented input during training.

2.2 MT Data and Quality Control

For the machine translation module, we trained two separate bilingual MT models, one for English-to-Chinese and one for English-to-German. Both models were initialized from Qwen3-8B and independently fine-tuned on language-pair-specific parallel corpora. This design avoids mixing the two target languages in a single multilingual model and allows each model to specialize in its corresponding translation direction.

To provide robust linguistic coverage for the unconstrained track, we constructed large-scale bilingual training corpora for both language pairs. For English-to-Chinese, the supervised fine-tuning (SFT) stage used approximately 4.0 million parallel sentence pairs, corresponding to about 5.54B training tokens. In addition, we constructed 128K preference pairs for contrastive preference optimization (CPO), which was used to further improve translation adequacy and output quality. For English-to-German, the SFT stage used approximately 0.9B training tokens. The final MT training data consisted of a mixture of large-scale web-crawled bitext and officially provided task-specific data.

To ensure the reliability of the training data, we implemented a multi-stage quality control pipeline. We first applied a series of heuristic filters to remove noisy samples, including language identification, length-ratio constraints, abnormal character filtering, and sentence-level deduplication. Subsequently, we employed a semantic alignment filtering strategy to estimate the cross-lingual similarity of each sentence pair. Sentence pairs with low semantic alignment scores were discarded, which helped remove low-confidence and misaligned bitext. After this filtering process, we obtained two high-quality language-pair-specific bilingual datasets for the English-to-Chinese and English-to-German MT models, respectively.

Finally, both MT models were further fine-tuned on the officially provided task-specific data. This additional adaptation stage helped the models better match the domain characteristics and discourse style of the IWSLT 2026 test set.

3 Offline Speech Translation System

Our offline speech translation system follows a robust cascade architecture, as illustrated in Figure 1. The pipeline integrates audio preprocessing, streaming ASR with context caching, forced alignment, two-pass transcription optimization, and segment-level machine translation. This modular design is specifically engineered to handle the complexities of long-form, unsegmented audio by ensuring both acoustic precision and semantic coherence.

3.1 Audio Preprocessing

To improve inference efficiency and mitigate transcription hallucinations, we first perform VAD on the input waveforms. We filter out silent segments and non-speech parts, passing only valid speech segments to the subsequent modules. This step effectively reduces redundant computation and prevents unnecessary generation on meaningless audio fragments, establishing a foundation for stable and efficient transcription.

3.2 Streaming Inference and Context Caching

Since the IWSLT 2026 test set does not provide pre-defined segmentation, we must partition the long-form audio ourselves. Given that long-form speech exhibits strong contextual dependencies, maintaining semantic integrity and cross-sentence coherence is crucial for accurate transcription and translation.

To balance recognition consistency with processing efficiency, we implement a streaming inference framework with a sliding window mechanism. We process input audio sequentially in 2-second units. To maintain continuity across blocks, we maintain a dynamic buffer that stores historical audio and corresponding text prefixes, with a maximum context length of approximately 12 seconds.

To ensure sentence-level integrity and avoid mid-sentence truncation, we employ a sentence-aware segmentation strategy. We trigger segmentation and output only when the model generates a newline character or reaches the end of an audio block. This design ensures that each output segment constitutes a complete and semantically coherent sentence.

3.3 First-Pass Transcription

In the first-pass transcription stage, we adopt a joint decoding architecture that integrates speech representations with LLM inference. We utilize a speech encoder derived from Qwen3-Omni to extract high-level acoustic features, and use Qwen3-4B as the text backbone for autoregressive transcription generation. As described in Section 2.1, the speech encoder and the Qwen3-4B language model are connected by the same trainable audio adapter trained on large-scale speech-transcription data.

During inference, the speech encoder first produces frame-level acoustic representations from the input audio. The audio adapter maps these acoustic representations into the token embedding space of the Qwen3-4B language model. The projected acoustic embeddings are injected into the language model input through the `prompt_embeds` mechanism, replacing the embeddings of special audio placeholder tokens. This enables direct fusion of speech representations and LLM decoding.

The model is specifically trained to output newline characters at appropriate semantic pauses. Therefore, the first-pass transcription produces semantically complete initial segments rather than arbitrary fixed-length chunks, providing a reliable textual basis for subsequent timestamp alignment.

3.4 Forced Alignment

To obtain precise temporal boundaries for the transcribed sentences, we apply the Qwen3-ForcedAligner-0.6B model. We perform fine-grained time alignment between the audio waveform and the first-pass transcription output. This model generates accurate sentence-level timestamps, which facilitate reliable audio re-segmentation for the second-pass transcription phase.

3.5 Second-Pass Transcription

We employ a two-pass transcription strategy to further enhance recognition accuracy and robustness. Based on the timestamps obtained from forced alignment, we merge and re-segment the audio into continuous fragments not exceeding 30 seconds. Each fragment is then fed into Qwen3-Omni for a second round of end-to-end speech recognition. This two-pass process effectively corrects recognition errors—particularly in segments involving heavy accents or fast speech—ultimately producing high-quality transcriptions.

3.6 Segment Merging and Translation

Before translation, we merge the refined transcription segments into semantically complete blocks,

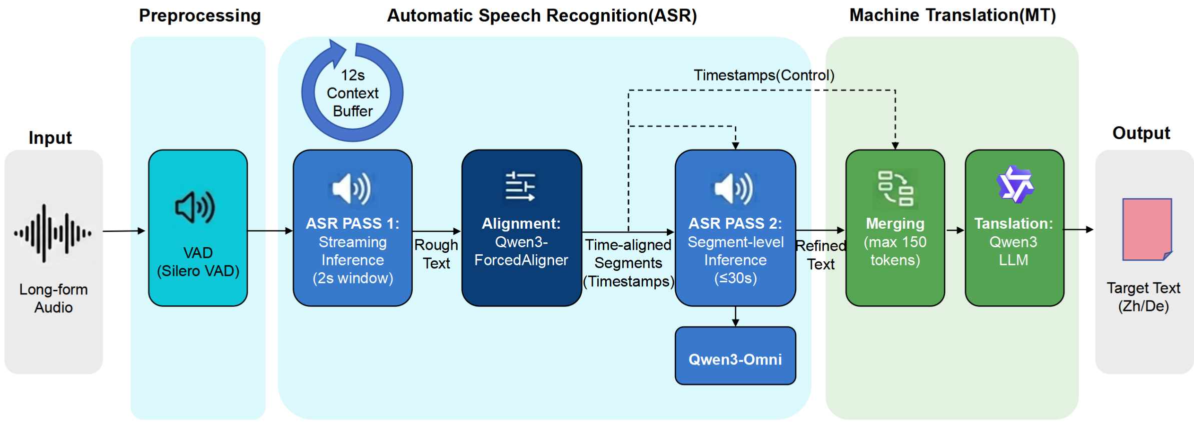

Figure 1: The overall architecture of the HW-TSC offline speech translation system. The pipeline integrates Silero VAD, two-pass ASR (streaming inference and refinement via Qwen3-Omni), Qwen3-ForcedAligner, and context-aware machine translation.

restricted to a maximum length of 150 tokens. This strategy maintains contextual coherence while preventing issues associated with excessively long model inputs. Finally, we feed the merged source text into the Qwen3-8B LLM to generate the final translation in the target language, completing the offline speech translation pipeline.

4 Experiments and Results

4.1 Experimental Setup

We evaluated our system on the tst-2022 benchmark dataset. For the ASR module, we used Word Error Rate (WER) (von Neumann et al., 2025) as the primary metric. For the machine translation and the integrated system, we utilized COMET (Rei et al., 2022) score to measure translation quality. All experiments were conducted on the Qwen3 model family, utilizing Silero VAD for audio preprocessing.

4.2 ASR Performance Analysis

To establish a comprehensive understanding of our system's capabilities, we conduct the ASR evaluation in two stages: model selection under ideal conditions and system validation under realistic unconstrained conditions. All experiments are conducted on the *tst-2022* English-to-Chinese (En-Zh) dataset.

4.2.1 Model Selection and Contextual Capacity

We first evaluate the transcription performance of various models within the Qwen3 family. Since the *tst-2022* dataset provides official segmentation, we use these ground-truth boundaries to establish a performance upper bound. To investigate the models' ability to leverage long-range dependencies, we further test a strategy where official segments are merged into longer audio chunks not exceeding 30 seconds. The results are summarized in Table 2.

ASR Model	Segmentation Method	WER (%)
Qwen3-ASR-0.6B	Official Seg.	7.31
Qwen3-ASR-0.6B	Official Seg. + 30s	5.79
Qwen3-ASR-1.7B	Official Seg.	5.40
Qwen3-ASR-1.7B	Official Seg. + 30s	3.93
Qwen3-Omni	Official Seg.	5.55
Qwen3-Omni	**Official Seg. + 30s**	**3.14**

Table 2: ASR WER (%) on the *tst-2022* test set with official segmentation. Results show that larger models and longer semantic context (up to 30s) significantly enhance recognition accuracy.

The results indicate that **Qwen3-Omni** possesses the strongest acoustic modeling and linguistic reasoning capabilities. Notably, increasing the fragment length to 30 seconds yields a substantial reduction in WER across all models, proving that maintaining semantic integrity via longer segments is crucial for MLLM-based ASR systems.

4.2.2 Evaluation of the Two-pass Strategy

In the realistic IWSLT offline track, official segmentation is unavailable. To verify the effectiveness of our proposed pipeline, we evaluate the system starting from raw, unsegmented waveforms. We compare a **Single-pass** baseline (utilizing streaming ASR with a 12s context buffer but without second-pass refinement) against our full **Two-pass** strategy.

As shown in Table 3, our Two-pass strategy achieves a WER of **3.01%**, representing a significant improvement over the single-pass baseline

System Configuration	WER (%)
Single-pass (Streaming Baseline)	6.31
Two-pass (Proposed Pipeline)	**3.01**

Table 3: ASR performance comparison on *tst-2022* without official segmentation. Our two-pass strategy effectively refines initial transcriptions.

(6.31%). Impressively, this performance even surpasses the best result obtained using official segmentation (3.14% in Table 2). This suggests that our timestamp-based re-segmentation, guided by the Qwen3-ForcedAligner, creates audio boundaries that are more optimized for the model's receptive field than the original manual annotations. This refined transcription serves as a high-fidelity input for the subsequent translation stage.

4.3 Machine Translation Refinement

To evaluate the effectiveness of our context-aware translation module, we utilize the high-fidelity transcriptions generated by our best-performing **Two-pass ASR system** (3.01% WER) as the input for the translation task. We specifically investigate the impact of our **Segment Merging** strategy by comparing two configurations: a sentence-by-sentence translation baseline and our proposed approach, which merges refined segments into semantic blocks not exceeding 150 tokens.

Configuration	En-Zh	En-De
Qwen3-8B (Sentence-level)	0.8397	0.7628
Qwen3-8B + Merging	**0.8462**	**0.7854**

Table 4: Machine translation performance (COMET) on *tst-2022* using Two-pass ASR transcriptions. Segment merging significantly enhances contextual consistency.

As shown in Table 4, the segment merging strategy yields a measurable improvement in COMET scores across both language directions. While the Two-pass ASR provides accurate textual input, translating isolated sentences often leads to a loss of discourse-level information. By grouping refined transcriptions into larger semantic blocks, the **Qwen3-8B** model better captures long-range dependencies and cross-sentence logic. This strategy effectively mitigates the risk of semantic fragmentation, ensuring that the final output maintains the high stylistic and contextual fidelity required for the IWSLT offline track.

4.4 Overall System Evaluation

Finally, we evaluate the end-to-end performance of the integrated HW-TSC system across different language directions. Since the audio subsets for the English-to-Chinese (En-Zh) and English-to-German (En-De) tracks in the *tst-2022* dataset are not identical, we report the ASR Word Error Rate (WER) and translation COMET scores separately for each track. We compare our full cascaded pipeline against a representative baseline system (Single-pass ASR + standard MT).

As shown in Table 5, the integrated HW-TSC system consistently outperforms the baseline across all metrics and subsets. For the En-Zh track, our pipeline reduces the WER from 6.31% to 3.01%, resulting in a significant COMET gain of +0.0538. A similar trend is observed in the En-De track, where the improved transcription fidelity directly contributes to superior translation quality. These results confirm that the synergistic integration of two-pass refinement and context-aware segment merging effectively addresses the challenges of unconstrained offline speech translation, regardless of the target language.

5 Conclusion

In this paper, we presented the HW-TSC team's submission to the IWSLT 2026 Offline Speech Translation Task. We implemented a robust cascade system centered on the **Qwen3 model family**, specifically optimized for the English-to-Chinese and English-to-German unconstrained tracks. Our core contribution lies in the integration of a **two-pass transcription strategy**—combining streaming inference with context caching and fine-grained forced alignment—which effectively addresses semantic fragmentation and hallucinations in long-form audio. **Notably, our pipeline demonstrates the ability to surpass official segmentation baselines by optimizing audio boundaries through model-based alignment.**

Experimental results on the *tst-2022* benchmark demonstrate that our pipeline significantly outperforms standard cascade baselines across all metrics. Specifically, the combination of context-aware segment merging and the multimodal capabilities of Qwen3-Omni ensures both high ASR accuracy and superior translation consistency. These findings confirm that a carefully optimized cascade architecture, leveraged with state-of-the-art **Multimodal Large Language Models (MLLMs)**, remains a

System	ASR WER ↓		Translation COMET ↑	
	En-Zh	En-De	En-Zh	En-De
Baseline (Single-pass + Std. MT)	6.31	6.54	0.7924	0.7409
HW-TSC (Full Pipeline)	**3.01**	**3.25**	**0.8462**	**0.7854**

Table 5: Overall system performance on *tst-2022*. The table reflects separate ASR metrics for each language track due to differences in the corresponding audio subsets.

powerful and flexible paradigm for complex, real-world speech translation scenarios. For future work, we aim to explore **knowledge distillation techniques** to integrate our two-pass insights into more efficient end-to-end frameworks, further reducing inference latency while maintaining the robust performance of our current pipeline.

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.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

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*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer speech & language*, 66:101155.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Qwen Team. 2026. Qwen3-asr technical report.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thilo von Neumann, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2025. Word error rate definitions and algorithms for long-form multi-talker speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

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*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, arXiv:2306.13549. Version 4.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.