NAIST Offline Speech Translation System for IWSLT 2025

Ruhiyah Faradishi Widiaputri[1], Haotian Tan[1], Jan Meyer Saragih[1], Yuka Ko[1], Katsuhito Sudoh[1,2], Satoshi Nakamura[1,3], Sakriani Sakti[1],
[1]Nara Institute of Science and Technology, Japan,
[2]Nara Women's University, Japan,
[3]Chinese University of Hong Kong, Shenzhen, China,
Correspondence: ruhiyah.faradishi.rc2@naist.ac.jp, ssakti@is.naist.jp

Abstract

This paper presents NAIST's submission to the offline speech translation task of the IWSLT 2025 evaluation campaign, focusing on English-to-German and English-to-Chinese translation. We implemented both cascade and end-to-end frameworks using various components. For the cascade approach, we used Whisper and SALMONN as automatic speech recognition systems, each paired with Qwen2.5 large language model (LLM) for translation. In the end-to-end setting, we used SALMONN as speech translation and also built a custom model combining the Whisper encoder, DeCo projector, and Qwen2.5 LLM. To further leverage the large language model capabilities, we experimented with different prompting strategies. Additionally, since long speech inputs are segmented for processing, we applied hypothesis combination techniques to generate the final translation output. Our results show that combining Whisper and LLMs can yield strong translation performance, even without further fine-tuning in the cascade setup. Moreover, our proposed end-to-end architecture achieved competitive results, despite being trained on significantly less data compared to SALMONN. Finally, we decided to use both SALMONN as an end-to-end speech translation model and our proposed end-to-end model for our IWSLT 2025 submission for both language pairs.

1 Introduction

Spoken Language Translation (SLT) refers to the process of automatically converting spoken audio into written text in another language. Within the IWSLT Shared Task, the Offline Speech Translation Task stands out as one of the longest-running tracks. Its goal is to offer a consistent evaluation setting for speech translation, without the timing and structural limitations typically associated with other tasks—such as real-time constraints in simultaneous interpretation, space restrictions in subtitling, duration matching in dubbing, or the challenges posed by limited data in low-resource language scenarios.

In the 2025 edition of the Offline Speech Translation Task (Abdulmumin et al., 2025), three translation directions are included: English to German, Chinese, and Arabic. This year's challenge places particular emphasis on tackling more practical translation scenarios, such as content from TV shows, academic talks, business news, and speech with diverse accents. Our team at NAIST is participating in the English-German and English-Chinese tracks. Unfortunately, due to limited preparation time, we were not able to take part in the English-Arabic track.

To address these translation tasks, we explore two widely used SLT frameworks: the cascade and end-to-end approaches. The cascade method separates the process into two stages—first transcribing speech using automatic speech recognition (ASR), followed by translating the transcription with a machine translation (MT) system. In contrast, the end-to-end approach generates translations directly from the speech input, integrating both steps into a single model. While the cascade framework benefits from modularity and reuse of existing ASR and MT models, it is susceptible to error propagation. End-to-end systems can mitigate such issues, but they often struggle with data scarcity, as large-scale parallel speech-to-text corpora remain limited.

In particular, we implemented both frameworks using a range of components. For the cascade approach, we explored two ASR systems — Whisper[1] (Radford et al., 2023) and SALMONN[2] (Tang et al., 2024) — each paired with the Qwen2.5[3] (Yang et al., 2024) LLM for machine translation. In the end-to-end setting, we treated SALMONN as a unified speech translation system. Addition-

[1]https://github.com/openai/whisper
[2]https://github.com/bytedance/SALMONN
[3]https://github.com/QwenLM/Qwen2.5

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 360–368

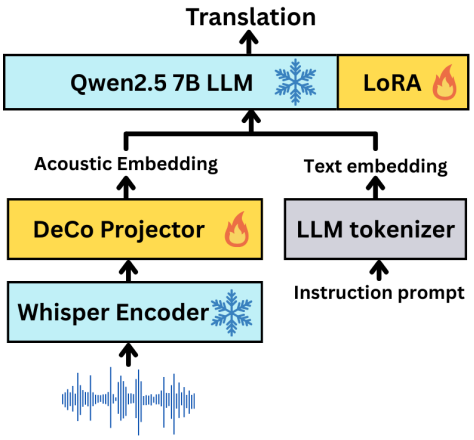

Figure 1: Our proposed end-to-end ST.

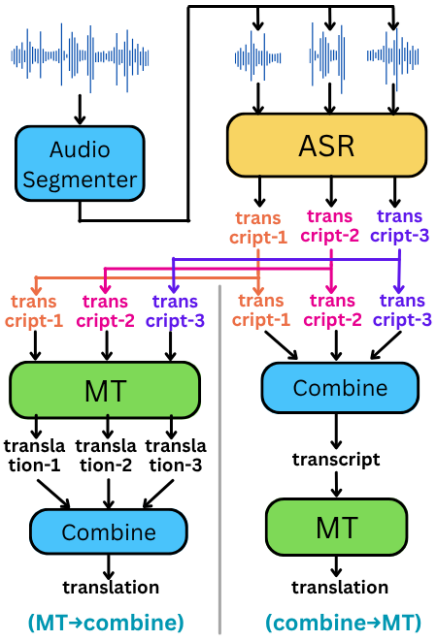

Figure 2: Two strategies for combining outputs of short segments in cascaded speech translation: (left) each short ASR outputs is translated individually, and the translations are merged afterward; (right) ASR outputs are first merged into a single text before translation.

ally, we developed a custom end-to-end model that integrates the Whisper encoder, a DeCo (Yao et al., 2024) projection module, and Qwen2.5 as the decoder. To further investigate the capabilities of large language models, we conducted experiments with different prompting strategies and analyzed their impact on translation performance. Since long speech inputs are segmented and processed separately, we also explore hypothesis combination strategies to produce the final translation output.

2 System Description

As outlined earlier, this work explores both cascade and end-to-end approaches to speech translation, utilizing a range of components including Whisper, SALMONN, Qwen2.5, and others. In the following sections, we first describe the model architectures of these components in Section 2.1. We then explain our methods for applying zero-shot, few-shot learning, or fine-tuning to both the cascaded and end-to-end speech translation settings in Sections 2.2 and 2.3, respectively.

2.1 Model Architecture

Whisper is an encoder-decoder Transformer model (Vaswani et al., 2017), trained on a wide range of speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection, using up to 680,000 hours of weakly-supervised labeled audio data. Whisper is highly robust across diverse environments and performs well in zero-shot settings without the need for fine-tuning. The model is available in various sizes, from `tiny` to `large`. Additionally, improved versions of the `large` model have been released, known as `large-v2` and `large-v3`. In this work, we experimented with both Whisper `large-v2` and `large-v3` as ASR.

SALMONN is a multimodal LLM that can perceive and understand general audio inputs including speech, audio events, and music. The model integrates two auditory encoders: a Whisper `large-v2` speech encoder and a fine-tuned BEATs (Chen et al., 2022) encoder for non-speech audio. These are connected to a Vicuna 13B LLM (Chiang et al., 2023) via a window-level Q-Former module (Li et al., 2023). SALMONN is pre-trained through a three-stage cross-modal learning process on diverse datasets. In this work, we use SALMONN for both ASR and end-to-end ST.

Qwen2.5 is the latest series of large language models from the Qwen LLM family, which has demonstrated top-tier performance across various benchmarks. The open-sourced Qwen2.5 models are dense, Transformer-based decoder architectures. Two types of Qwen2.5 models have been released: base models and instruction-tuned models, available in sizes ranging from 0.5B to 72B parameters. For the cascade system, we used the instruction-tuned version of Qwen2.5 with 7B parameters as MT.

In addition to existing models, we propose a

Split	en-de		en-zh	
	Dataset	Size	Dataset	Size
Train	CoVoST + Europarl	261526	CoVoST	229347
Dev	CoVoST + Europarl	16530	CoVoST	15233
	tst2022	2045		
Test	tst2021	2025	tst2022	2130

Table 1: Data statistics.

novel end-to-end ST model that integrates the Whisper `large-v3` encoder, a DeCo projector, and the Qwen2.5 LLM. As illustrated in Figure 1, the Whisper encoder first extracts acoustic features from the input speech. The DeCo projector, consisting of a 2D adaptive average pooling layer for downsampling followed by two linear projection layers, bridges the speech-text modality gap by mapping the acoustic features from the Whisper encoder into the LLM embedding space as acoustic embeddings. Qwen2.5 then performs the translation based on prompt instructions. In this system, we use the base Qwen2.5 7B model.

2.2 Zero-shot and Few-shot Learning for Cascaded ST

For ASR, we used Whisper `large-v2`, Whisper `large-v3`, and SALMONN in a zero-shot setting (without fine-tuning). After we segmented long audio into shorter clips (see Subsection 3.2) and generated the transcription for each segment individually, we experimented with two hypothesis combination methods, as illustrated in Figure 2: (1: MT→combine) translating each transcription segment individually and then combining the translations, or (2: combine→MT) combining the transcriptions and translating the merged text using MT. The combination was performed by simply concatenating the transcriptions or translations of each speech segment in their original order.

For LLM-based MT, we initially experimented with the instruction-tuned version of Qwen2.5 7B using seven zero-shot prompts, ranging from simple to detailed instructions, as listed in Appendix A. For English-to-German, we selected the two best-performing prompts based on average BLEU and COMET scores, and then applied few-shot learning with $k = 1, 3, 5, 7$, and 10, using examples derived from transcriptions generated by the best ASR on the development set (en-de `tst2022`).

ASR	en-de tst2021	en-zh tst2022
SALMONN	16.33	15.9
Whisper `large-v2`	9.79	**10.33**
Whisper `large-v3`	**8.89**	10.77

Table 2: WER scores of Whisper and SALMONN ASRs.

2.3 Zero-shot and Fine-tuning for End-to-end ST

For end-to-end ST, we used SALMONN and our proposed end-to-end ST model. SALMONN was evaluated under two settings: zero-shot and fine-tuning, using the datasets described in Section 3.1. Inference and fine-tuning of SALMONN followed the default settings and hyperparameters provided in the official SALMONN source code, except that we used a maximum of 22 and 30 epochs for fine-tuning with CoVoST + Europarl en-de as the validation set, and 22 epochs for fine-tuning with en-zh data (see Table 4).

Our proposed ST model was also fine-tuned. During the fine-tuning phase, we fully trained the projector while fine-tuning the LLM using LoRA (Hu et al., 2022), with the parameters of both the Whisper encoder and the LLM kept frozen. To improve translation performance and simplify training, we incorporated ASR as an auxiliary task. Specifically, we used a single prompt that instructed the LLM to output both the transcription and its corresponding translation, separated by the `<end>` symbol. This symbol then served as the stopping criterion during inference.

3 Experiment Setup

3.1 Dataset

The training and development datasets used in this work consist of speech-to-text parallel data listed under the IWSLT 2025 constrained setup, namely CoVoST v2 (Wang et al., 2020) and Europarl v1.1 (Iranzo-Sánchez et al., 2020). For development and test sets, we used the most recent past development sets provided by IWSLT 2025, tst2022 and tst2021. Specifically, for end-to-end English-to-German speech translation, we used the en-de CoVoST v2 and en-de Europarl v1.1 train sets for training, either the en-de CoVoST v2 dev set or the en-de tst2022 set for validation, and en-de tst2021 for testing. For end-to-end English-to-Chinese speech

en-de tst2021				
Prompt	**MT→combine**		**combine→MT**	
	BLEU	**COMET**	**BLEU**	**COMET**
1	24.90	78.12	27.72	**85.56**
2	25.59	77.15	**29.18**	84.22
3	26.04	79.43	28.54	83.58
4	21.53	73.71	27.84	82.90
5	**28.36**	**81.11**	26.80	83.46
6	26.91	78.92	27.86	83.62
7	26.13	77.79	28.87	82.59
en-zh tst2022				
Prompt	**MT→combine**		**combine→MT**	
	BLEU	**COMET**	**BLEU**	**COMET**
1	41.41	86.17	44.66	86.10
2	43.48	85.44	46.37	**86.81**
3	45.86	84.36	46.83	86.75
4	41.20	82.99	45.60	86.24
5	44.45	**86.24**	46.01	86.66
6	**46.10**	83.55	**47.15**	86.71
7	44.92	83.87	45.71	86.39

Table 3: BLEU and COMET scores of Qwen2.5 7B Instruct as zero-shot MT with seven prompts. Inputs are Whisper large-v3 ASR outputs.

translation, we used the en-zh CoVoST v2 train and dev sets for training and validation, and en-zh tst2022 as the test set. For the cascade speech translation system, since both the ASR and MT components were evaluated in zero-shot or few-shot settings, no training data was used. Instead, we evaluated directly on the test set. Few-shot examples were selected from the development set randomly.

For CoVoST v2 and Europarl v1.1, we pre-processed the datasets by removing samples with missing audio or target text, samples with audio that was too short or too long, and samples with noisy audio. We also performed basic text cleaning. Table 1 presents the details of the data used in this work.

3.2 Model Setup

Since Whisper was trained on 30-second audio chunks and cannot process longer input directly, we segmented long audio into shorter clips of less than 30 seconds before feeding them into the ASR or the end-to-end ST system. Segmentation for en-de tst2021, en-de tst2022, and en-zh tst2022 was performed using the Gentle forced aligner based

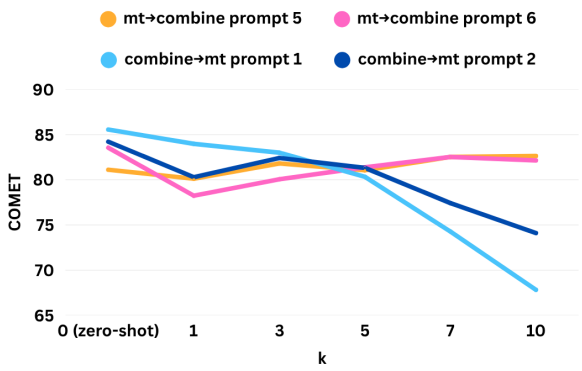

Figure 3: COMET Scores for English-to-German few-shot translation.

on the provided reference transcripts[4]. For audio without reference transcriptions, segmentation was carried out using the Silero Voice Activity Detector[5] (Silero Team, 2024).

For cascaded ST, we first performed zero-shot inference using the ASR models Whisper `large-v2`, Whisper `large-v3`, and SALMONN on the test set of each language pair. We then calculated the WER based on the transcriptions of the long audio inputs, where individual segment transcriptions were first merged using simple string concatenation before computing WER. Next, we selected the better-performing transcriptions between Whisper `large-v3` and SALMONN (Whisper `large-v2` was used solely for comparison with SALMONN, as SALMONN's speech encoder is based on Whisper `large-v2`) and used them as input for the Qwen2.5 7B Instruct MT.

We evaluated seven zero-shot prompts for both hypothesis combination methods (Figure 2). Subsequently, we calculated BLEU and COMET scores (Rei et al., 2020) for the long audio translations. BLEU scores were computed using SacreBLEU (Post, 2018), and COMET scores were obtained using the default COMET model `Unbabel/wmt22-comet-da`. For the English-to-German pair, we further explored few-shot learning with $k = 1, 3, 5, 7$, and 10 using the two prompts that achieved the highest average BLEU and COMET scores.

For the end-to-end model, we first evaluated the SALMONN checkpoint in a zero-shot setting. We then fine-tuned both SALMONN and our proposed end-to-end ST model for each language pair and calculated BLEU and COMET scores on the long

[4] https://github.com/strob/gentle
[5] https://github.com/snakers4/silero-vad

Languages and Data	Model	Detail	BLEU	COMET
en-de train: -CoVoST en-de -Europarl en-de test: tst-2021	SALMONN	Zeroshot	22.33	71.86
		Dev: CoVoST + Europarl en-de (Max epoch: 22)	31.20	86.19
		Dev: CoVoST + Europarl en-de (Max epoch: 30)	**31.76**	86.21
		Dev: tst-2022 en-de	30.81	**86.25**
	OUR-ST	Dev: CoVoST + Europarl en-de	28.84	84.73
		Dev: tst-2022 en-de	29.03	85.05
en-zh train: CoVoST en-zh test: tst-2022	SALMONN	Zeroshot	46.97	76.77
		Dev: CoVoST en-zh (Max epoch: 22)	**48.63**	80.47
	OUR-ST	Dev: CoVoST en-zh	44.78	**83.47**

Table 4: BLEU and COMET scores of SALMONN and ours in end-to-end ST.

audio translations.

We then compared the best-performing cascade combination, the best SALMONN results, and the best fine-tuned version of our end-to-end model for each language pair with two strong baselines—Whisper `large-v3` ASR + NLLB 3.3B (Costa-Jussà et al., 2022) and SeamlessM4T v2 Large (Barrault et al., 2023). Both baselines were evaluated in zero-shot settings, and the cascaded system used the MT→combine scenario due to the NLLB models' limited maximum input lengths. We also compared them with the top three models from previous IWSLT submissions that used the same test sets: IWSLT 2021 (Anastasopoulos et al., 2021) for English-to-German and IWSLT 2022 (Anastasopoulos et al., 2022) for English-to-Chinese.

It is important to note that, since we assumed the use of pre-trained acoustic models (i.e., Whisper in this case) is allowed under the "constrained with large language models" setting, we submitted the fine-tuned SALMONN results under the "unconstrained" track, and our proposed end-to-end ST model under the "constrained with large language models" track for IWSLT 2025.

4 Experiment Results

4.1 Cascaded ST

Table 2 presents the WER scores of Whisper and SALMONN as ASR systems, evaluated on the en-de tst2021 and en-zh tst2022 sets in a zero-shot setting. As shown in the table, Whisper `large-v2` and Whisper `large-v3` performed comparably, and both models outperformed SALMONN, despite SALMONN incorporating a Whisper `large-v2` encoder and an LLM.

Table 3 presents the BLEU and COMET scores of Qwen2.5 7B Instruct as MT across seven zero-shot prompt variations, using the output of Whisper `large-v3` as input. Interestingly, although the WER for en-de tst2021 is lower than that for en-zh tst2022, Qwen's translation performance for the latter is significantly better. In other words, Qwen translates English to Chinese more effectively than English to German. It can also be observed that combining the ASR transcriptions for a single input audio before translation yields better translation results than combining the translations afterward. Furthermore, the the fourth prompt is shown to give slightly lower performance compared to other prompts. In the MT→combine scenario, prompts that perform well for English-to-German also work well for English-to-Chinese, and vice versa. However, this relationship does not hold in the combine-before-translate scenario (combine→MT).

Figure 3 shows COMET scores for MT with few-shot learning for English to German, using the two best prompts for each hypothesis recombination strategy: prompts 5 and 6 for `MT→combine`, and prompts 1 and 2 for `combine→MT`. As shown in the figure, adding examples in the `MT→combine` strategy improves translation quality. However, the opposite trend is observed for `combine→MT`. This is possibly due to the fact that, in `combine→MT`, the MT input becomes significantly longer, which can affect the model's ability to effectively utilize few-shot examples.

4.2 End-to-end ST

Table 4 shows the BLEU and COMET scores for two end-to-end ST models—SALMONN and our

en-de (test: tst2021)			BLEU	COMET
Baseline	Baseline cascade	Whisper large-v3 ASR + NLLB 3.3B MT	**34.04**	84.62
	Baseline end-to-end	SeamlessM4T v2 Large	31.37	74.45
Existing IWSLT submissions	HW-TSC	Constrained - cascade	20.30	-
	KIT (Nguyen et al., 2021)	Constrained - cascade	19.00	-
	AppTek (Bahar et al., 2021)	Constrained - end-to-end	18.30	-
Our best systems	Our best cascade	Whisper large-v3 ASR + Qwen2.5 7B Inst MT (prompt: 2 - scenario: combine→MT - k: 0)	29.18	84.22
	Our best SALMONN	Dev: CoVoST + Europarl en-de - max epoch=30	31.76	**86.21**
	Our best end-to-end ST	Dev: tst-2022 en-de - max step: 100,000	29.03	85.05
en-zh (test: tst2022)			**BLEU**	**COMET**
Baseline	Baseline cascade	Whisper large-v3 ASR + NLLB 3.3B MT	30.31	76.48
	Baseline end-to-end	SeamlessM4T v2 Large	32.81	68.12
Existing IWSLT submissions	USTC-NELSLIP cascade (Zhang et al., 2022a)	Cascade	35.70	-
	YI cascade (Zhang et al., 2022b)	Cascade	35.00	-
	HW-TSC (Li et al., 2022)	Cascade	33.40	-
Our best systems	Our best cascade	Whisper large-v3 ASR + Qwen2.5 7B Inst MT (prompt 6 - scenario: combine→MT - k: 0)	47.15	**86.81**
	Our best SALMONN	Dev: CoVoST en-zh - max epoch: 22	**48.63**	80.47
	Our best our end-to-end ST	Dev: tst-2022 en-zh - max step: 100,000	44.78	83.47
	Our submitted end-to-end ST	Dev: tst-2022 en-zh - max step: 51,000	40.69	83.03

Table 5: Performance comparison between our best ST systems, baselines, and previous IWSLT submissions. Our submitted systems are shaded in gray.

proposed model. As shown in the table, fine-tuning the publicly released SALMONN checkpoint with additional ST data improves translation performance. Similar to the results observed when using Qwen as the MT component in the cascaded approach, the end-to-end models also achieve significantly higher BLEU scores for English-to-Chinese translation than for English-to-German. However, the COMET scores show the opposite trend. Additionally, for English-to-German translation, the choice of development set had minimal impact on performance. Lastly, despite being fine-tuned on substantially less data than SALMONN, our proposed end-to-end models achieve competitive results, especially compared to the zero-shot SALMONN.

4.3 Comparison with Baselines and Previous Submissions

Table 5 shows the comparison between our best cascaded ST, our best SALMONN end-to-end ST, our best proposed end-to-end ST, with two strong baselines: Whisper ASR + NLLB 3.3B cascade baseline and SeamlessM4T v2 Large end-to-end baseline, as well as the top three previous IWSLT submissions (from IWSLT 2021 for en-de and IWSLT 2022 for en-zh)[6]. As shown in the table, for both en-de and en-zh pairs, our cascaded and end-to-end ST systems performed significantly better than the IWSLT submissions.

For the en-de pair, despite using the same ASR, the cascaded Whisper ASR + NLLB 3.3B system achieved a higher BLEU score than our best cascaded model. This suggests that for English-to-German MT, NLLB 3.3B still outperforms Qwen2.5 7B Instruct. Our end-to-end models, on the other hand, achieved comparable BLEU scores to SeamlessM4T v2 Large and outperformed it in terms of COMET scores. In contrast, for the en-zh pair, both our cascaded and end-to-end ST systems performed significantly better than the baselines, indicating that Qwen2.5 7B Instruct outperforms NLLB 3.3B for English-to-Chinese translation.

We decided to use end-to-end models (SALMONN and our proposed end-to-end ST) for our IWSLT 2025 submission, which are shaded in

[6]The IWSLT submissions were selected based on their `BLEU NewRef` scores as reported in the official findings; however, the scores shown in the table are `BLEU TEDRef` to allow fair comparison with our systems.

gray in the table. However, due to time constraints, the submission for English-to-Chinese using our proposed model did not use the best-performing checkpoint, but rather the best checkpoint at step 51,000.

5 Conclusion

This paper describes NAIST's submission to the IWSLT 2025 offline speech translation task, focusing on English-to-German and English-to-Chinese translation. We found that using Whisper as the ASR combined with Qwen2.5 LLM as the MT in a zero-shot setting was already capable of producing good translations. Furthermore, in the zero-shot setting, translation quality for long audio was better when the transcriptions of individual segments were combined first and then translated together, compared to translating each segment individually and combining the translations afterward. However, few-shot learning yielded better results in the latter case. Fine-tuning the SALMONN model further improved its translation quality. Additionally, our custom end-to-end model demonstrated competitive performance with SALMONN, despite being trained on significantly less data. Finally, we observed that both Qwen2.5 and SALMONN performed better on English-to-Chinese translation than on English-to-German.

Acknowledgments

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loc Barrault, Luisa Bentivogli, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, and 1 others. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th international conference on spoken language translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Parnia Bahar, Patrick Wilken, Mattia A Di Gangi, and Evgeny Matusov. 2021. Without further ado: Direct and simultaneous speech translation by apptek in 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 52–63.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. 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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yinglu Li, Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen,

Min Zhang, Shimin Tao, and 1 others. 2022. The hw-tsc's offline speech translation system for iwslt 2022 evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 239–246.

Tuan-Nam Nguyen, Thai-Son Nguyen, Christian Huber, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, and Sebastian Stüker. 2021. Kit's iwslt 2021 offline speech translation system. In *Proceedings Of The 18th International Conference On Spoken Language Translation (IWSLT 2021)*, pages 125–130.

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, Belgium, Brussels. Association for Computational Linguistics.

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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

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.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, and 1 others. 2022a. The ustc-nelslip offline speech translation systems for iwslt 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207.

Ziqiang Zhang, Junyi Ao, Long Zhou, Shujie Liu, Furu Wei, and Jinyu Li. 2022b. The yitrans end-to-end speech translation system for iwslt 2022 offline shared task. *arXiv preprint arXiv:2206.05777*.

A Prompts for LLM MT

Table 6 shows seven zero-shot prompts for LLM MT. `<tgt-lang>` denotes the target language for the translation, which may be either "German" or "Chinese". For few-shot prompting, we add some examples in the format of Prompt 1 right before the input.

For example, using Prompt 3 with few-shot prompting, the full prompt will be:

> You are given a source English sentence. It is a transcription of spontaneous speech, which may include repetitions, fillers, or disfluencies. Translate it into `<tgt-lang>` as it is. Do not change the structure or nuance.
>
> English: `<example-src-text-1>`
> `<tgt-lang>`: `<example-tgt-text-1>`
> ...
> English: `<example-src-text-k>`
> `<tgt-lang>`: `<example-tgt-text-k>`
> English: `<input>`
> `<tgt-lang>`:

Prompt 1	English: {} <tgt-lang>:
Prompt 2	"Translate the sentence from English to <tgt-lang>. English: {} <tgt-lang>: "
Prompt 3	"You are given a source English sentence. It is a transcription of spontaneous speech, which may include repetitions, fillers, or disfluencies. Translate it into <tgt-lang> as it is. Do not change the structure or nuance. English: {} <tgt-lang>: "
Prompt 4	"You are given a source English sentence. It is a transcription of spontaneous speech, which may include repetitions, fillers, or disfluencies. Your task is to: 1. Translate it into <tgt-lang> as it is. Do not change the structure or nuance. 2. Convert any written-out numbers (e.g., one, twenty) into numerical digits (e.g., 1, 20). 3. If you detect any indirect words, enclose them in „“. 4. Add punctuations '.', ',' if necessary. English: {} <tgt-lang>: "
Prompt 5	"Translate from English to <tgt-lang>, using the appropriate tone for the topic. Do not mention the topic. Output only the <tgt-lang> translation. English: {} <tgt-lang>: "
Prompt 6	"You are given a source English sentence. It is a transcription of spontaneous speech, which may include repetitions, fillers, or disfluencies. Translate it into <tgt-lang> as it is. Do not change the structure or nuance. Do not mention the topic. Output only the <tgt-lang> translation. English: {} <tgt-lang>: "
Prompt 7	"You are given a source English sentence. It is a transcription of spontaneous speech, which may include repetitions, fillers, or disfluencies. Your task is to: 1. Translate it into <tgt-lang> as it is. Do not change the structure or nuance. Do not mention the topic. Output only the <tgt-lang> translation. 2. Convert any written-out numbers (e.g., one, twenty) into numerical digits (e.g., 1, 20). 3. If you detect any indirect words, enclose them in „“. 4. Add punctuations '.', ',' if necessary. English: {} <tgt-lang>: "

Table 6: Zero-shot prompts for Qwen2.5 LLM as MT.