MoLoRA: Boosting LLM-based End-to-end Speech Translation with Mixture of Low-rank Experts

Hao Zhang, Yaqi Chen, Nianwen Si, XuKui Yang*, Wenlin Zhang, Dan Qu*

Information Engineering University, Zhengzhou, China
{haozhang0126,chyaqi,snw1608,gzyangxk,wenlinzzz,qudan_xd}@163.com

Abstract

Recently, End-to-End Speech Translation (E2E-ST) methods leveraging large language models (LLMs) have demonstrated strong generalization capabilities and excellent scalability by integrating pre-trained speech encoders with LLMs, where Low-Rank Adaptation (LoRA) is commonly used for parameter-efficient fine-tuning to reduce training costs. However, LoRA's low-rank assumption often fails in multilingual tasks, as the inherent complexity of cross-lingual semantic relationships and syntactic variations exceeds the representational capacity of low-rank matrices. This leads to parameter conflicts across languages, resulting in suboptimal performance. To address this issue, we propose **M**ixture **o**f **Lo**w-**R**ank **A**daptations (**MoLoRA**), which integrates the Mixture of Experts (MoE) mechanism with LoRA. MoLoRA effectively enhances the model's expressive capacity while maintaining parameter efficiency during training. Specifically, we treat multiple LoRA modules as low-rank experts and introduce a routing mechanism to dynamically activate language-specific experts. Additionally, shared experts are incorporated and consistently activated to model cross-lingual general knowledge. Furthermore, to enhance the robustness and accuracy of speech representations, we propose a **M**ulti-**G**ranularity **R**epresentation **F**usion module (**MGRF**). This module mitigates local distortions in frame-level speech representations caused by noise by fusing frame-level and sentence-level features, thereby providing the LLM with more accurate high-level semantic information. We conduct multilingual experiments on the MuST-C and CoVoST-2 datasets. Our method achieves an average BLEU score of 32.2 across eight language pairs on the MuST-C dataset and an average of 36.3 across three language pairs on the CoVoST-2 dataset, establishing a new state-of-the-art (SOTA) performance.

Introduction

End-to-end speech translation (E2E-ST) (Sperber et al. 2019; Moritz et al. 2025) aims to directly translate speech input into text in the target language without generating an intermediate transcription. This approach offers advantages over traditional cascade systems (Meng and Anastasopoulos 2025; Min et al. 2025)—which consist of automatic speech recognition followed by machine translation—by reducing latency and mitigating error propagation (Liu et al. 2020; Dong et al. 2021). Despite recent advances that have led to promising results, with some E2E-ST models (Zhang et al. 2022) even outperforming cascade systems, training high-performing models remains challenging due to data scarcity and the complexity of cross-modal modeling. To address these challenges, various strategies have been proposed, including pre-training (Wang et al. 2020), multi-task learning (Zhou, Yuan, and Shi 2024; Zhang et al. 2024a), knowledge distillation (Liu et al. 2019), and contrastive learning (Ye, Wang, and Li 2022; Zhang et al. 2023d), all of which have contributed to significant performance gains.

However, many of these methods rely on complex training procedures and carefully designed network architectures, which limit their scalability. On one hand, they often lack data scalability due to the need for complex training process and data preprocessing, making it difficult to extend to new languages with minimal adaptation effort. On the other hand, model scalability is constrained, as most previous E2E-ST systems are either trained from scratch or ini-

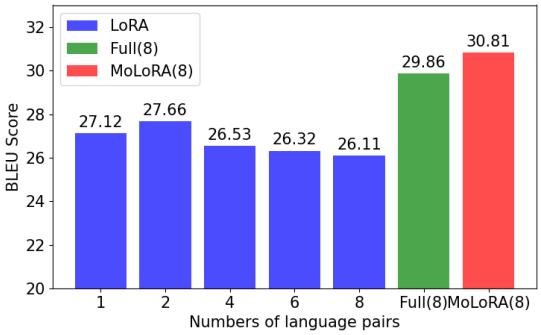

Figure 1: Model performance on the En-De tst-COMMON set with varying numbers of language pairs. We conducted multilingual experiments on the MuST-C dataset. The two rightmost bars represent the performance of full fine-tuning and our proposed MoLoRA method, respectively, when trained on eight language pairs. Standard LoRA degrades significantly due to inter-language interference as languages increase, even underperforming single-pair settings. MoLoRA mitigates this conflict and even outperforms full fine-tuning.

*Corresponding author.

tialized with weak pre-trained parameters, resulting in relatively small model sizes. Recent studies (Chen et al. 2025) have demonstrated a strong scaling law between model capacity and performance, highlighting the need for more scalable and efficient frameworks capable of leveraging larger models.

Recently, E2E-ST systems built upon large language models (LLMs) (Wu et al. 2023; Zhang et al. 2023c; Du et al. 2024; Djanibekov and Aldarmaki 2025; Mundnich et al. 2025; Dou et al. 2025; Liu et al. 2025; Moslem 2025) have gained significant attention. By leveraging powerful pre-trained speech and text representations, these approaches achieve superior translation performance with simplified training pipelines, often surpassing methods that rely on intricate architectures and training schemes. However, the large number of parameters in LLMs entails substantial computational costs for training. As a result, parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al. 2021) are commonly employed to reduce resource requirements.

Despite its computational efficiency, LoRA typically underperforms full fine-tuning in multilingual E2E-ST tasks. As illustrated in Fig. 1, the performance of LoRA-based models degrades noticeably as the number of language pairs increases, sometimes even falling below that of single-language models. LoRA is based on the assumption that weight updates of the pre-trained model during downstream adaptation have low intrinsic dimensionality and can thus be effectively approximated via low-rank matrices. However, multilingual E2E-ST involves complex linguistic structures, cross-lingual transfer, and fine-grained semantic alignment—challenges whose complexity exceeds the representational capacity of low-rank approximations (Nikdan et al. 2024; Biderman et al. 2024). Rather than promoting positive cross-lingual transfer, this shared low-dimensional parameter space leads to parameter competition, where gradients from different languages interfere with one another during training. Consequently, the model struggles to capture language-specific patterns and may suffer from negative transfer, particularly for linguistically distant languages. This limitation poses a critical bottleneck in achieving optimal performance in multilingual E2E-ST.

To overcome these limitations, we propose **M**ixture **o**f **Lo**w-**R**ank **A**daptations (**MoLoRA**), a novel framework that integrates the Mixture-of-Experts (MoE) paradigm (Liu et al. 2024) with LoRA. MoLoRA enhances model expressiveness in multilingual settings while maintaining the parameter efficiency of LoRA. Specifically, we treat individual LoRA modules as low-rank experts within an MoE architecture. A subset of these experts is designated as shared and permanently activated, capturing cross-lingual commonalities and general knowledge. The remaining experts are dynamically activated through a routing mechanism, allowing the model to adapt to language- or task-specific patterns. This design preserves the lightweight nature of LoRA while alleviating parameter competition in the low-rank space. Through functional specialization and expert collaboration, MoLoRA significantly improves performance in multilingual E2E-ST.

Our contributions are summarized as follows:

1. We propose MoLoRA, a novel framework that effectively scales model capacity and significantly enhances multilingual E2E-ST performance, achieving results that in some cases surpass full fine-tuning.
2. On the MuST-C dataset, MoLoRA achieves an average BLEU score of 32.2 across eight target languages in the multilingual setting, outperforming standard LoRA by 4.3 BLEU points and full fine-tuning by 0.3 BLEU points, demonstrating its effectiveness.
3. On CoVoST-2, MoLoRA consistently improves translation quality, achieving an average BLEU score of 36.3 for English-to-Japanese, English-to-Chinese, and English-to-German translation, thereby establishing a new state-of-the-art.

Related Works

E2E-ST: E2E-ST (Sperber et al. 2019; Moritz et al. 2025) offers structural advantages over cascaded systems through reduced latency and error propagation. However, training effective E2E-ST models remains challenging due to the cross-modal complexity of mapping speech to text in another language and the scarcity of parallel speech-translation data. To address these issues, strategies such as pre-training (Wang et al. 2020; Tsiamas et al. 2024), multi-task learning (Ye, Wang, and Li 2021; Zhang et al. 2023f; Zhou, Yuan, and Shi 2024; Zhang et al. 2024a), contrastive learning (Ye, Wang, and Li 2022; Ouyang, Ye, and Li 2022; Zhang et al. 2023d), and optimal transport (Zhou, Fang, and Feng 2023; Le et al. 2023) have been proposed to enhance representation learning. The encoder in E2E-ST models must simultaneously extract acoustic and semantic features, motivating decoupling approaches (Dong et al. 2021) with CTC-based compression (Liu et al. 2020; Xu et al. 2021) for improved temporal alignment. Data scarcity has driven techniques such as data augmentation (Fang and Feng 2023a; Zhang et al. 2023b), knowledge distillation (Liu et al. 2019), and mixup (Fang et al. 2022; Cheng et al. 2023) to improve efficiency and generalization.

Nevertheless, despite notable improvements, many approaches still rely on complex training pipelines and intricate architectures, and some—trained from scratch or with weak initialization—fail to fully exploit the strong inductive biases and rich representations of pre-trained models, thereby missing opportunities to accelerate convergence and enhance performance.

Speech Large Models: Driven by the rapid advancement of LLMs, there has been growing interest in developing Speech Large Models (SLMs) that integrate speech processing within LLM frameworks to enable multimodal understanding. Notable examples include SpeechGPT (Zhang et al. 2023a), which discretizes speech representations and extends the LLaMA architecture to follow multimodal instructions. Open-source initiatives such as LLaSM (Shu et al. 2023) and SLM (Wang et al. 2023b) further demonstrate the feasibility of adapting LLMs to process speech inputs, paving the way for open, multimodal foundation models. Models like SALMONN (Tang et al. 2023) and

Qwen2-Audio (Chu et al. 2024) have introduced specialized architectures and training strategies to build general-purpose SLMs capable of handling diverse speech tasks. Additionally, recent efforts have focused on developing end-to-end speech-to-speech dialogue systems (Abouelenin et al. 2025), aiming to enable low-latency, emotionally expressive human-machine interactions.

E2E-ST based on LLMs: While general-purpose SLMs aim for broad capabilities, a growing body of work focuses on specialized models tailored for E2E-ST. SpeechLLaMA (Wu et al. 2023) was among the first to demonstrate the feasibility of using a decoder-only LLaMA-based architecture for direct speech-to-text translation. LST (Zhang et al. 2023c) introduced a multi-stage training strategy that substantially improved performance across target languages. LLM-ST (Huang et al. 2023) achieved state-of-the-art results on multiple language pairs by leveraging large-scale multimodal data. LLaST (Chen et al. 2024) proposed a dual LoRA approach to improve parameter efficiency and training stability. Moslem (2025) reduced computational costs via model compression and knowledge distillation, enhancing inference efficiency. Karel et al. (2025) investigated zero-shot speech translation using LLMs. CoT-ST (Du et al. 2024) incorporated a Chain-of-Thought (CoT) mechanism during training, yielding significant quality improvements. Liu et al. (2025) enhanced speech-text alignment by introducing transcribed text as an auxiliary input, improving translation fluency and accuracy.

Despite these advances, most existing approaches emphasize novel training pipelines while largely overlooking the issue of task interference in multilingual settings—particularly when PEFT methods such as LoRA are employed. Sharing a limited low-rank update space across multiple languages can lead to parameter competition, hindering the model's ability to capture language-specific patterns and resulting in suboptimal performance. To address this limitation, we propose a novel framework that integrates the MoE mechanism with LoRA, effectively mitigating parameter conflicts across languages during multilingual training and enhancing overall model performance.

Methodology

The overall architecture of our proposed model is illustrated in Fig. 2.

LLM-Based E2E-ST

The E2E-ST training corpus typically consists of speech-transcription-translation triples, denoted as $D_{\text{ST}} = \{(s, x, y)\}$, where s, x, and y represent the input speech signal, the source-language transcription, and the target-language translation, respectively. The goal of E2E-ST is to directly generate the translation y from the raw audio signal s, bypassing the intermediate transcription step.

An LLM-based E2E-ST system follows a cascaded architecture composed of three main components: a pre-trained speech encoder (frontend), an adapter network, and an LLM as the backend. Given an input speech signal s, the speech encoder first extracts high-level acoustic representations. The adapter then maps these representations into the embedding space of the LLM. Finally, conditioned on instruction prompts, the LLM autoregressively generates the target translation. During training, the objective is defined solely on the translation output, ignoring speech features and instructional prefixes. The loss function is formulated as the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{(s,y)\in D_{\text{ST}}} \log p_\theta(y \mid s, \text{instruction}) \quad (1)$$

where D_{ST} is the training dataset, θ denotes the trainable parameters of the model, and $p_\theta(y \mid s, \text{instruction})$ is the conditional probability of generating the translation y given the speech input s and a task-specific instruction.

Multi-Granularity Representation Fusion (MGRF)

Frame-level speech representations, although providing fine temporal resolution, are vulnerable to noise and local perturbations that hinder downstream semantic modeling (Zhang et al. 2024b). To this end, we propose the MGRF module, which fuses complementary frame-level and sentence-level features to produce more robust and semantically enriched inputs for the LLM (Fig. 2,left).

Let $h_{\text{frame}} \in \mathbb{R}^{T\times d}$ denote the frame-level features extracted from a pre-trained speech model, where T is the number of time steps and d is the feature dimension. These serve as queries in the fusion process. Concurrently, we extract a sentence-level representation $h_{\text{sentence}} \in \mathbb{R}^{1\times d'}$ from a pre-trained utterance encoder. When $d' \neq d$, a learnable projection $W_p \in \mathbb{R}^{d'\times d}$ is applied to align the dimensions.

The projected vector is replicated T times to form $h'_{\text{sentence}} \in \mathbb{R}^{T\times d}$, which acts as both keys and values in cross-attention:

$$h_{\text{fusion}} = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (2)$$

where $Q = h_{\text{frame}}$, $K = V = h'_{\text{sentence}}$. This mechanism enables frame-level features to attend to global context, yielding a context-aware representation h_{fusion}.

To preserve temporal fidelity while integrating global semantics, we apply:

$$h_{\text{final}} = \text{LayerNorm}(h_{\text{frame}} + h_{\text{fusion}}) \quad (3)$$

The resulting $h_{\text{final}} \in \mathbb{R}^{T\times d}$ combines temporal precision with global semantic regularization, effectively mitigating local distortions and improving representation accuracy.

Mixture of Low-Rank Experts

As previously discussed, the low-rank assumption of LoRA becomes problematic in multilingual settings due to the high complexity and diversity of language-specific adaptations (Nikdan et al. 2024; Biderman et al. 2024). To address this issue, we integrate a MoE mechanism that expands model capacity while maintaining computational efficiency, effectively mitigating task conflicts across languages.

In our framework, each LoRA module functions as an expert. Given an input $\mathbf{h}$, it is processed by n experts

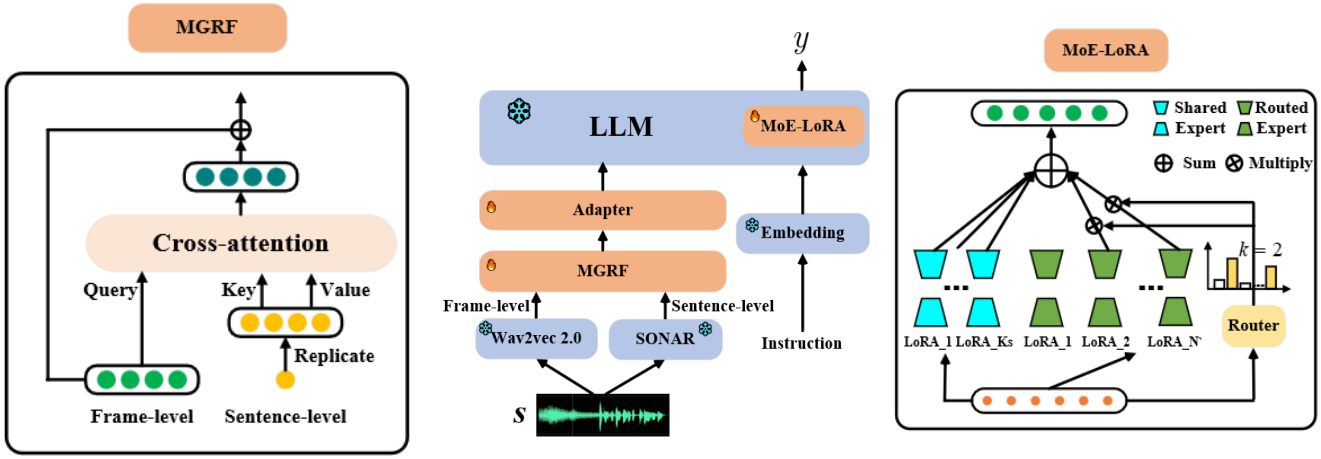

Figure 2: Overview of the MoLoRA architecture. The left side illustrates the MGRF module for fusing frame-level and sentence-level speech representations; LayerNorm is omitted for simplicity. On the right, the LoRA module serves as the expert within the MoE structure, comprising routed experts and shared experts. The routed experts are dynamically activated, while the shared expert is activated consistently.

$(E_1, \ldots, E_n)$, which include both *routed* and *shared* types. To capture common cross-lingual knowledge, K_s shared experts process all tokens deterministically, with their outputs summed using fixed weights of 1. The remaining $N' = N - K_s$ routed experts employ a sparse routing mechanism that selects the top-K experts via:

$$R(\mathbf{h}) = \text{softmax}\left(\text{TopK}(W_r^T \cdot \mathbf{h}, K)\right) \tag{4}$$

where $W_r \in \mathbb{R}^{d \times N'}$ is the trainable routing matrix. This ensures sparse activation while preserving parameter efficiency. Each expert E_i is implemented as a LoRA module: $E_i = B_i A_i$, with $B_i \in \mathbb{R}^{d \times r}$, $A_i \in \mathbb{R}^{r \times k}$, and rank constraint $r \ll \min(d, k)$. The final output is computed as:

$$\tilde{\mathbf{h}} = \underbrace{\sum_{i=1}^{K_s} E_i(\mathbf{h})}_{\text{Shared Experts}} + \underbrace{\sum_{j=1}^{N'} [R(\mathbf{h})]_j \cdot E_j(\mathbf{h})}_{\text{Routed Experts}} \tag{5}$$

This hybrid design combines shared experts for general knowledge representation with routed experts for language-specific specialization, effectively resolving parameter competition and improving multilingual E2E-ST performance.

Expert Balanced Loss

A common challenge in MoE training is uneven expert utilization, where the routing mechanism often favors a few experts while leaving most underused, resulting in inefficient training and suboptimal performance. To address this, we incorporate a load balancing loss (Fedus et al. 2022) that promotes uniform activation of *routed experts* during routing.

For a batch with N' distinct routed experts, the load balancing loss is defined as:

$$\mathcal{L}_{\text{Balance}} = \sum_{i=1}^{N'} f_i P_i \tag{6}$$

where:

$$f_i = \frac{N'}{KT} \sum_{t=1}^{T} \mathbb{I}(h_t \in E_i) \tag{7}$$

$$P_i = \frac{1}{T} \sum_{t=1}^{T} p_{i,t} \tag{8}$$

Here, f_i denotes the normalized empirical activation frequency of expert E_i, calculated by counting tokens assigned to it across the batch (T total tokens, K experts per token). P_i represents the average routing probability assigned to E_i, computed from normalized weights $p_{i,t}$ of each token h_t. The indicator function $\mathbb{I}(\cdot)$ returns 1 when token h_t is assigned to E_i, and 0 otherwise.

The loss achieves its minimum when $f_i = P_i = \frac{1}{N'}$ for all activated experts, ensuring balanced usage across the expert pool. This condition guarantees equal distribution of both routing probabilities and actual activations, maximizing training efficiency.

Training Procedure

The LLM-based E2E-ST system follows a two-stage training approach: *modality alignment* and *downstream task fine-tuning*.

Modality Alignment This stage aims to reduce speech input length and align speech representations with the LLM's embedding space. To avoid language bias, we use Automatic Speech Recognition (ASR) datasets, updating only the adapter and MGRF module parameters while freezing all other components. The alignment is guided by the instruction *"Transcribe the above speech for me"* paired with ASR data. The objective function is:

$$\mathcal{L}(\theta)_{\text{Stage1}} = -\sum_{(s,x)\in \text{ASR}} \log p_\theta(x \mid s, \text{instruction}) \tag{9}$$

where $\theta = \{\text{adapter}, \text{MGRF}\}$ denotes the trainable parameters.

Downstream Task Fine-Tuning This stage focuses on optimizing performance for specific tasks. We freeze the pre-trained speech models and LLM, while training a mixture of low-rank experts alongside the adapter and MGRF modules. The core objective is:

$$\mathcal{L}(\theta)_{\text{Stage2}} = -\sum_{(s,y)\in D_{\text{ST}}} \log p_\theta(y \mid s, \text{instruction}) \quad (10)$$

with $\theta = \{\text{adapter}, \text{MGRF}, \text{router}, \text{LoRA module}\}$. To ensure balanced expert utilization, we incorporate the load balancing loss in this stage.

$$\mathcal{L} = \mathcal{L}(\theta)_{\text{Stage2}} + \alpha \mathcal{L}_{\text{Balance}} \quad (11)$$

where α controls the balancing loss weight in the composite objective.

Experiment Setting

Dataset and Processing

Experiments were conducted on the MuST-C v1.0 (Di Gangi et al. 2019) and CoVoST-2 (Wang, Wu, and Pino 2020) datasets.

MuST-C is a multilingual TED-talks-based dataset containing speech-to-text translation pairs from English into eight languages: German (De), French (Fr), Russian (Ru), Spanish (Es), Italian (It), Romanian (Ro), Portuguese (Pt), and Dutch (Nl). We used the dev set for validation and the tst-COMMON set for testing. CoVoST-2 is a large-scale corpus built on Common Voice, containing 15 languages for English-to-target translation and 21 languages for target-to-English translation. This study focused on the En→Ja, En→Zh, and En→De directions.

For speech input processing, 16-bit, 16 kHz mono-channel audio waveforms were used (Fang et al. 2022). Translation texts were processed with true casing and tokenized using the LLM tokenizer without additional preprocessing.

Model Architecture

The proposed model architecture comprises three main components: the speech frontend, the adapter module, and the LLM backend. Regarding the *speech frontend*, two types of pre-trained models were considered. In the frame-level setting, the CTC-finetuned Wav2Vec 2.0 Large model, which was pre-trained on 53.2k hours of untranscribed speech from LibriVox (Kearns 2014), and further fine-tuned on 960 hours of labeled speech from LibriSpeech combined with pseudo-labels, was employed. The CTC projection head was removed, and the final encoder output was used as the speech representation, resulting in a feature sequence with a dimensionality of 1024. For sentence-level modeling, the SONAR model (Duquenne, Schwenk, and Sagot 2023)—a multilingual and multimodal fixed-size sentence embedding space supporting text representations for 200 languages and speech representations for 37 languages—was adopted.

The adapter comprises a *length adapter*—two 1D convolutions (kernel 5, stride 2, padding 2, hidden dim 1024)—to downsample speech features temporally, and a *modality adapter*, a linear layer projecting features into the LLM's embedding space (4096-dim for LLaMA2-7B (Touvron et al. 2023)). LoRA (rank 64, $\alpha = 128$) is applied to all linear layers except the input embedding and output head. In the MoE setup, we use 1 shared expert ($K_s = 1$), 4 routed experts ($N' = 4$), top-2 routing ($K = 2$), and a balancing loss coefficient of 0.001 (Eq. (10)) to encourage uniform expert utilization.

Training and Inference

We trained the model using AdamW (Loshchilov and Hutter 2017) with cosine learning rate decay. In Stage 1, the model was trained on LibriSpeech for 6 epochs (batch size 128, warm-up ratio 0.03, initial learning rate 2×10^{-3}); checkpoints were saved every 1,000 steps. In Stage 2, we fine-tuned for 1 epoch with learning rate 2×10^{-4}, no warm-up, and checkpointing every 100 steps. Preliminary studies showed robustness to hyperparameter choices. Training was performed on 16 NVIDIA Tesla V100 GPUs using the ZeRO parallelization strategy (Rajbhandari et al. 2020).

During inference, only the final checkpoint of each training run was evaluated. It was observed that averaging multiple checkpoints, a practice commonly adopted in prior work (Ye, Wang, and Li 2022), did not lead to improved performance in our setup. Beam search with a beam size of 4 was used for sequence generation. To ensure fair and reproducible comparisons with existing methods, case-sensitive, detokenized BLEU scores were computed using sacreBLEU.

Results and Analysis

Results on MuST-C Dataset

The results in Table 1 demonstrate that MoLoRA achieves state-of-the-art performance on the MuST-C benchmark, attaining an average BLEU score of 32.2 across eight language directions—outperforming the previous best (LST-7B) by 0.8 BLEU points. This improvement is particularly significant, considering that MoLoRA surpasses the single LoRA baseline by 4.3 BLEU points (from 27.9 to 32.2) with comparable parameter counts, highlighting its effectiveness in mitigating task interference through the MoE architecture's language-specific expert activation, which optimizes modeling capacity allocation and reduces inter-task competition. Moreover, MoLoRA outperforms full fine-tuning by 0.3 BLEU points on average, demonstrating that the MoE-induced inductive bias toward modularization and specialization enables more effective representations than dense parameter updates—even with fewer trainable parameters. These results underscore the critical importance of architectural innovation for addressing task conflict in multilingual training.

Comparison with Cascaded Baselines

We implement a robust cascaded baseline system by leveraging only the ASR portion of the ST dataset to fine-

Models	En-De	En-Fr	En-Es	En-Ru	En-It	En-Ro	En-Pt	En-Nl	Avg
SATE (Xu et al. 2021)	28.1	-	-	-	-	-	-	-	-
TDA (Du et al. 2022)	27.1	37.4	-	-	-	-	-	-	-
Chimera (Han et al. 2021)	26.3	35.6	30.6	17.4	25.0	24.0	30.2	29.2	27.3
ConST (Ye, Wang, and Li 2022)	28.3	38.3	32.0	18.9	27.2	25.6	33.1	31.7	29.4
FCCL^m (Zhang et al. 2023d)	29.0	38.3	31.9	19.7	27.3	26.8	32.7	31.6	29.7
STEMM (Fang et al. 2022)	28.7	37.4	31.0	17.8	25.8	24.5	31.7	30.5	28.4
M^{3}ST (Cheng et al. 2023)	29.3	38.5	32.4	19.3	27.5	25.9	33.4	32.5	29.9
CRESS (Fang and Feng 2023b)	29.4	40.1	33.2	19.7	27.6	26.4	33.6	32.3	30.3
MSP (Zhang et al. 2023e)	30.2	-	-	-	-	-	-	-	-
Siamese-PT (Le et al. 2023)	26.2	36.9	29.8	16.8	25.9	24.8	32.1	29.8	27.8
STPT† (Tang et al. 2022)	29.2	39.7	33.1	-	-	-	-	-	-
SpeechUT† (Zhang et al. 2022)	30.1	**41.4**	33.6	-	-	-	-	-	-
SRPSE † (Zhang et al. 2024a)	26.9	37.4	31.4	18.3	27.0	25.5	32.8	31.4	28.8
LST-7B † (Zhang et al. 2023c)	29.4	40.9	33.5	20.3	29.9	27.4	36.2	33.7	31.4
Ours									
Full	29.9	41.0	34.0	20.6	30.2	27.9	37.2	34.1	31.9
LoRA	26.1	36.7	31.1	17.5	26.3	23.1	32.2	29.8	27.9
MoLoRA	**30.8**	41.3	**34.4**	**21.2**	**30.2**	**28.0**	**37.4**	**34.3**	**32.2**

Table 1: BLEU scores on the MuST-C `tst-COMMON` set under multilingual training settings. LoRA employs a single shared adapter module (rank $r = 320$) across all language pairs, while MoLoRA utilizes a MoE architecture with LoRA modules as experts (each $r = 64$), ensuring comparable total numbers of trainable parameters. Best results are highlighted in **bold**. "Full": $\sim$ 7B trainable parameters; LoRA and MoLoRA: $\sim$ 7% of the full model's trainable parameters.

Models	En-De	En-Fr	En-Ru
Cascaded			
XSTNet (Ye, Wang, and Li 2021)	25.2	34.9	17.0
STEMM (Fang et al. 2022)	27.5	-	-
SATE (Xu et al. 2021)	28.1	-	-
Ours*	29.5	38.7	19.4
End-to-end			
MoLoRA	**30.8**	**41.3**	**21.2**

Table 2: MoLoRA versus the cascaded ST systems on MuST-C En-De/Fr/Ru test sets. We report the performance of the cascaded model implemented in our previous work. * denotes our newly built cascade system.

En→X	De	Ja	Zh	Avg
SALMONN (Tang et al. 2023)	18.6	22.7	33.1	24.8
SeamlessM4T-V2 (Barrault et al. 2023)	-	23.5	34.6	-
BLSP (Wang et al. 2023a)	14.1	-	-	-
Qwen2-Audio (Chu et al. 2024)	29.9	28.6	45.2	34.5
CoT-ST (Du et al. 2024)	28.7	30.8	**47.7**	35.7
AI-STA (Liu et al. 2025)	-	**31.4**	46.0	-
MoLoRA	**30.1**	31.3	47.5	**36.3**

Table 3: BLEU scores for English-to-German (En→De), English-to-Japanese (En→Ja), and English-to-Chinese (En→Zh) translation on the CoVoST-2 datasets. We conduct multilingual training by mixing three languages.

tune the speech encoder, and separately using only the Machine Translation (MT) portion to fine-tune the LLaMA2 7B model. These adapted components are then combined to form a two-stage cascaded system. We further compare MoLoRA against both this newly constructed cascade and previously reported cascaded approaches. As shown in Table 2, MoLoRA achieves superior performance, demonstrating the advantage of end-to-end modeling in avoiding error propagation between stages, and outperforming all cascaded baselines.

Results on CoVoST-2 Dataset

Table 3 presents the overall experimental results on the CoVoST-2 dataset. The proposed method demonstrates a similar trend, achieving an average BLEU score of 36.3 across the En→De, En→Ja, and En→Zh language directions, surpassing previous studies. Notably, this performance was achieved without utilizing transcription information, which further underscores the potential of the proposed method. The ability to achieve such results without leveraging transcriptions indicates that our approach can effectively capture and utilize the inherent linguistic features from raw audio inputs alone. This capability not only enhances the robustness of the model but also broadens its applicability, as it reduces dependency on additional transcriptions that may not always be available or reliable.

Ablation Study

We perform ablation studies on the MuST-C dataset to evaluate the contribution of key components in the proposed MoLoRA framework, focusing on the En→De translation direction. The results are presented in Table 4. The full MoLoRA model achieves a BLEU score of 30.8. When the sentence-level representation module is removed, performance drops to 30.4 BLEU, demonstrating its critical role in correcting local distortions in frame-level speech representations. Replacing the multi-granularity fusion strategy with a simple summation of frame-level and sentence-level representations decreases performance to 30.5 BLEU, con-

Model Variant	En-De
MoLoRA	**30.8**
w/o Sentence-level Representation	30.4**
w/o Multi-granularity Fusion Strategy	30.5**
w/o Shared Experts	30.6**
w/o Load Balancing Loss	30.6**

Table 4: Ablation study on different components of the proposed method. We evaluate the En→De translation performance (BLEU) in a multilingual setting trained on all eight language pairs from MuST-C. ** means MoLoRA's performance advantage over each ablated variant is significant at $p < 0.01$.

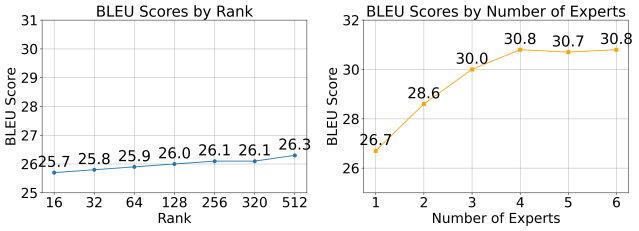

Figure 3: BLEU score on the En→De direction as a function of rank and number of routed experts, under multilingual training on all eight language pairs of the MuST-C dataset. In the left figure, the number of experts is 1, corresponding to the conventional LoRA fine-tuning method. In the right figure, there is one shared expert, and the x-axis indicates the number of routed experts.

firming its effectiveness in integrating fine-grained acoustic details with global semantic information. Disabling the shared experts reduces the BLEU score to 30.6, highlighting their importance in capturing cross-lingual commonalities. Finally, omitting the load balancing loss also leads to a 0.2 BLEU degradation (to 30.6), underscoring its role in ensuring uniform expert utilization and mitigating training instability caused by imbalanced expert loads.

How to Expand the Overall Capacity?

The capacity limitations of LoRA can be addressed by increasing the rank r or expanding the number of routed experts N'. Results in Fig. 3 show that increasing r yields marginal gains due to the inherent insufficiency of low-rank approximations in capturing full-parameter fine-tuning updates, which are characterized by numerous small yet non-zero singular values (Nikdan et al. 2024; Biderman et al. 2024). This limitation is exacerbated in multilingual settings, where a single low-rank matrix fails to model complex cross-lingual patterns, paradoxically worsening language interference. In contrast, increasing N' significantly improves performance by enabling specialized experts to capture distinct input features while maintaining computational efficiency through sparse activation. Our experiments reveal an optimal threshold at four experts, beyond which overfitting occurs due to overly granular data partitioning.

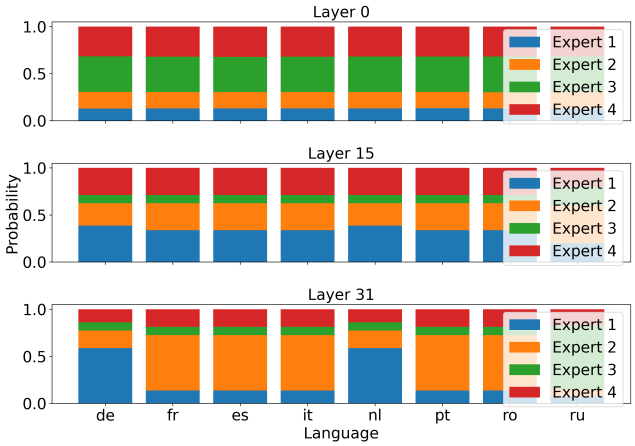

Figure 4: Proportion of tokens assigned to each expert in the eight language pairs of the MuST-C dataset at layers 0, 15, and 31.

Is There a Clear Correspondence Between Experts and Tasks?

We investigate whether MoLoRA experts develop language-specific specializations by analyzing expert selection patterns on the MuST-C dataset with eight language pairs, as shown in Fig. 4. At lower layers (layer 0), expert selection remains random with no language bias, indicating insufficient differentiation between language pairs at this stage. In contrast, higher layers (layer 31) exhibit clear specialization, where Romance languages (French, Spanish, Italian, Portuguese, Romanian) share similar expert allocation patterns due to linguistic similarities. This demonstrates that deeper layers enable experts to capture nuanced language-specific characteristics, establishing explicit expert-task correspondences. By routing tokens to language-specialized experts, MoLoRA effectively mitigates task interference and enhances multilingual performance.

Conclusion and Future Work

In this work, we propose MoLoRA, a hybrid framework integrating MoE and LoRA for multilingual E2E-ST. By treating LoRA modules as low-rank experts within MoE, MoLoRA mitigates task interference across languages. We further introduce the Multimodal Gating and Representation Fusion (MGRF) module, which jointly leverages sentence- and frame-level features to produce semantically enriched representations for LLMs. Experiments on MuST-C and CoVoST-2 show that MoLoRA achieves state-of-the-art performance with computational cost comparable to standard LoRA. Future work will focus on refining routing to reduce redundancy and further improve efficiency.

Acknowledgements

We thank the anonymous reviewers for their constructive suggestions. This work was partially supported by the Henan Provincial Natural Science Foundation (No. 252300420990), the Key Scientific Research Program of

Henan Province (No. 252102211040), and the National Natural Science Foundation of China (No. 62171470).

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.-A.; Ellis, B.; Elsahar, H.; Haaheim, J.; et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*.

Biderman, D.; Portes, J.; Ortiz, J. J. G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

Chen, W.; Tian, J.; Peng, Y.; Yan, B.; Yang, C.-H. H.; and Watanabe, S. 2025. OWLS: Scaling laws for multilingual speech recognition and translation models. *arXiv preprint arXiv:2502.10373*.

Chen, X.; Zhang, S.; Bai, Q.; Chen, K.; and Nakamura, S. 2024. LLaST: Improved end-to-end speech translation system leveraged by large language models. *arXiv preprint arXiv:2407.15415*.

Cheng, X.; Dong, Q.; Yue, F.; Ko, T.; Wang, M.; and Zou, Y. 2023. M 3 st: Mix at three levels for speech translation. In *Proceedings of ICASSP*, 1–5.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Di Gangi, M. A.; Cattoni, R.; Bentivogli, L.; Negri, M.; and Turchi, M. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of NAACL*, 2012–2017.

Djanibekov, A.; and Aldarmaki, H. 2025. Sparqle: Speech queries to text translation through llms. *arXiv preprint arXiv:2502.09284*.

Dong, Q.; Ye, R.; Wang, M.; Zhou, H.; Xu, S.; Xu, B.; and Li, L. 2021. Listen, Understand and Translate: Triple Supervision Decouples End-to-end Speech-to-text Translation. In *Proceedings of AAAI*, 10343–10344.

Dou, H.; Tian, X.; Lyu, X.; Zhu, J.; Li, J.; and Guo, L. 2025. Speech Translation Refinement using Large Language Models. *arXiv preprint arXiv:2501.15090*.

Du, Y.; Ma, Z.; Yang, Y.; Deng, K.; Chen, X.; Yang, B.; Xiang, Y.; Liu, M.; and Qin, B. 2024. CoT-ST: Enhancing LLM-based Speech Translation with Multimodal Chain-of-Thought. *arXiv preprint arXiv:2409.19510*.

Du, Y.; Zhang, Z.; Wang, W.; Chen, B.; Xie, J.; and Xu, T. 2022. Regularizing end-to-end speech translation with triangular decomposition agreement. In *Proceedings of AAAI*, volume 36, 10590–10598.

Duquenne, P.-A.; Schwenk, H.; and Sagot, B. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Fang, Q.; and Feng, Y. 2023a. Back translation for speech-to-text translation without transcripts. *arXiv:2305.08709*.

Fang, Q.; and Feng, Y. 2023b. Understanding and bridging the modality gap for speech translation. *arXiv:2305.08706*.

Fang, Q.; Ye, R.; Li, L.; Feng, Y.; and Wang, M. 2022. STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation. In *Proceedings of ACL*, 7050–7062.

Fedus, W.; Zoph, B.; Shazeer, N.; et al. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Han, C.; Wang, M.; Ji, H.; and Li, L. 2021. Learning Shared Semantic Space for Speech-to-Text Translation. In *Findings of ACL*, 2214–2225.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; Ye, R.; Ko, T.; Dong, Q.; Cheng, S.; Wang, M.; and Li, H. 2023. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

Kearns, J. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1): 7–8.

Le, P.-H.; Gong, H.; Wang, C.; Pino, J.; Lecouteux, B.; and Schwab, D. 2023. Pre-training for Speech Translation: CTC Meets Optimal Transport. *arXiv:2301.11716*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, H.; Chen, A.; Chen, K.; Bai, X.; Zhong, M.; Qiu, Y.; and Zhang, M. 2025. Adaptive Inner Speech-Text Alignment for LLM-based Speech Translation. *arXiv preprint arXiv:2503.10211*.

Liu, Y.; Xiong, H.; Zhang, J.; He, Z.; Wu, H.; Wang, H.; and Zong, C. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proceedings of INTERSPEECH*, 1128–1132.

Liu, Y.; Zhu, J.; Zhang, J.; and Zong, C. 2020. Bridging the Modality Gap for Speech-to-Text Translation. *arXiv:2010.14920*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv:1711.05101*.

Meng, C.; and Anastasopoulos, A. 2025. GMU Systems for the IWSLT 2025 Low-Resource Speech Translation Shared Task. *arXiv preprint arXiv:2505.21781*.

Min, A.; Hu, C.; Ren, Y.; and Zhao, H. 2025. When End-to-End is Overkill: Rethinking Cascaded Speech-to-Text Translation. *arXiv preprint arXiv:2502.00377*.

Moritz, N.; Xie, R.; Gaur, Y.; Li, K.; Ahmed, S. M. Z.; Seide, F.; and Fuegen, C. 2025. Transcribing and Translating, Fast and Slow: Joint Speech Translation and Recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Moslem, Y. 2025. Efficient Speech Translation through Model Compression and Knowledge Distillation. *arXiv preprint arXiv:2505.20237*.

Mundnich, K.; Niu, X.; Mathur, P.; Ronanki, S.; Houston, B.; Elluru, V. R.; Das, N.; Hou, Z.; Huybrechts, G.; Bhatia, A.; et al. 2025. Zero-resource speech translation and recognition with LLMs. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Nikdan, M.; Tabesh, S.; Crnčević, E.; and Alistarh, D. 2024. Rosa: Accurate parameter-efficient fine-tuning via robust adaptation. *arXiv preprint arXiv:2401.04679*.

Ouyang, S.; Ye, R.; and Li, L. 2022. WACO: word-aligned contrastive learning for speech translation. *arXiv:2212.09359*.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16.

Shu, Y.; Dong, S.; Chen, G.; Huang, W.; Zhang, R.; Shi, D.; Xiang, Q.; and Shi, Y. 2023. Llasm: Large language and speech model. *arXiv:2308.15930*.

Sperber, M.; Neubig, G.; Niehues, J.; and Waibel, A. 2019. Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation. *TACL*, 7: 313–325.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Tang, Y.; Gong, H.; Dong, N.; Wang, C.; Hsu, W.-N.; Gu, J.; Baevski, A.; Li, X.; Mohamed, A.; Auli, M.; et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv:2204.05409*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Tsiamas, I.; Gállego, G. I.; Fonollosa, J. A.; and Costa-jussà, M. R. 2024. Pushing the limits of zero-shot end-to-end speech translation. *arXiv preprint arXiv:2402.10422*.

Wang, C.; Liao, M.; Huang, Z.; Lu, J.; Wu, J.; Liu, Y.; Zong, C.; and Zhang, J. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Wang, C.; Wu, A.; and Pino, J. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Wang, C.; Wu, Y.; Liu, S.; Zhou, M.; and Yang, Z. 2020. Curriculum Pre-training for End-to-End Speech Translation. In *Proceedings of ACL*, 3728–3738.

Wang, M.; Han, W.; Shafran, I.; Wu, Z.; Chiu, C.-C.; Cao, Y.; Wang, Y.; Chen, N.; Zhang, Y.; Soltau, H.; et al. 2023b. SLM: Bridge the thin gap between speech and text foundation models. *arXiv:2310.00230*.

Wu, J.; Gaur, Y.; Chen, Z.; Zhou, L.; Zhu, Y.; Wang, T.; Li, J.; Liu, S.; Ren, B.; Liu, L.; et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. *arXiv:2307.03917*.

Xu, C.; Hu, B.; Li, Y.; Zhang, Y.; Huang, S.; Ju, Q.; Xiao, T.; and Zhu, J. 2021. Stacked Acoustic-and-Textual Encoding: Integrating the Pre-trained Models into Speech Translation Encoders. In *Proceedings of ACL*, 2619–2630.

Ye, R.; Wang, M.; and Li, L. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv:2104.10380*.

Ye, R.; Wang, M.; and Li, L. 2022. Cross-modal Contrastive Learning for Speech Translation. In *Proceedings of NAACL*.

Zhang, C.; Zhou, Y.; Zhao, R.; Chen, Y.; and Shi, X. 2024a. Representation Purification for End-to-End Speech Translation. *arXiv preprint arXiv:2412.04266*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv:2305.11000*.

Zhang, D.; Ye, R.; Ko, T.; Wang, M.; and Zhou, Y. 2023b. DUB: Discrete Unit Back-translation for Speech Translation. *arXiv:2305.11411*.

Zhang, H.; Si, N.; Chen, Y.; Zhang, W.; Yang, X.; Qu, D.; and Jiao, X. 2023c. Tuning Large language model for End-to-end Speech Translation. *arXiv preprint arXiv:2310.02050*.

Zhang, H.; Si, N.; Chen, Y.; Zhang, W.; Yang, X.; Qu, D.; and Zhang, W. 2023d. Improving Speech Translation by Cross-Modal Multi-Grained Contrastive Learning. *TASLP*, 31: 1075–1086.

Zhang, H.; Si, N.; Zhang, W.; Yang, X.; and Qu, D. 2024b. Improving Speech Translation by Understanding the Speech From Latent Code. *IEEE Signal Processing Letters*, 31: 1259–1263.

Zhang, Y.; Xu, C.; Hu, B.; Zhang, C.; Xiao, T.; and Zhu, J. 2023e. Improving end-to-end speech translation by leveraging auxiliary speech and text data. In *Proceedings of AAAI*, volume 37, 13984–13992.

Zhang, Y.; Xu, C.; Li, B.; Chen, H.; Xiao, T.; Zhang, C.; and Zhu, J. 2023f. Rethinking and Improving Multi-task Learning for End-to-end Speech Translation. *arXiv:2311.03810*.

Zhang, Z.; Zhou, L.; Ao, J.; Liu, S.; Dai, L.; Li, J.; and Wei, F. 2022. SpeechUT: Bridging Speech and Text with Hidden-Unit for Encoder-Decoder Based Speech-Text Pre-training. In *Proceedings of EMNLP*, 1663–1676.

Zhou, Y.; Fang, Q.; and Feng, Y. 2023. CMOT: Cross-modal Mixup via Optimal Transport for Speech Translation. *arXiv:2305.14635*.

Zhou, Y.; Yuan, Y.; and Shi, X. 2024. A multitask co-training framework for improving speech translation by leveraging speech recognition and machine translation tasks. *Neural Computing and Applications*, 36(15): 8641–8656.