Generative Error Correction for Emotion-aware Speech-to-text Translation

Zhengdong Yang[1], **Sheng Li**[2], **Chenhui Chu**[1]
[1]Kyoto University [2]Institute of Science Tokyo
zd-yang@nlp.ist.i.kyoto-u.ac.jp li.s.az@m.titech.ac.jp
chu@i.kyoto-u.ac.jp

Abstract

This paper explores emotion-aware speech-to-text translation (ST) using generative error correction (GER) by large language models (LLMs). Despite recent advancements in ST, the impact of the emotional content has been overlooked. First, we enhance the translation of emotional speech by adopting the GER paradigm: Finetuned an LLM to generate the translation based on the decoded N-best hypotheses. Moreover, we combine the emotion and sentiment labels into the LLM finetuning process to enable the model to consider the emotion content. In addition, we project the ST model's latent representation into the LLM embedding space to further improve emotion recognition and translation. Experiments on an English-Chinese dataset show the effectiveness of the combination of GER, emotion and sentiment labels, and the projector for emotion-aware ST. Our code is available at https://github.com/N-Orien/EmoST.

1 Introduction

Speech-to-text translation (ST) is a task where the model takes speech in one language as input and translates it into text in another language. ST performance has greatly improved over the recent years with significant efforts on datasets (Di Gangi et al., 2019; Wang et al., 2021; Jia et al., 2022; Chen et al., 2021; Ye et al., 2023; et al., 2023a) and models (Barrault et al., 2023; et al., 2023b; Radford et al., 2022). However, an essential aspect often overlooked in speech translation is the emotion of speech.

Human speech naturally includes emotions. In real-life conversations, a listener often uses cues from the speaker's voice tone to grasp what is being said. Therefore, emotion can significantly influence the results of translating speech. As the instance shown in Figure 1, the phrase "I can't believe this" can convey a range of emotions, from surprise and shock to awe and excitement, which can alter its translation in another language.

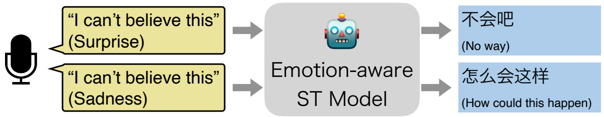

Figure 1: The expectation for an emotion-aware ST model, which can generate appropriate translation based on the emotion of the input speech.

Emotion has been studied in machine translation (or text-to-text translation) studies (Troiano et al., 2020) and other tasks in natural language processing, such as sentiment analysis and recognizing emotions in conversations (Fu et al., 2023). However, there has been little effort focusing on emotion in ST. Seamless Expressive (Barrault et al., 2023) examines the preservation of emotional states in speech-to-speech translation, without addressing the influence of emotions on the semantic aspects of translation. Some datasets (Liang et al., 2021; Chen et al., 2024) are constructed for emotion-aware ST, but further community effort investigating the methodology for this task is required.

Meanwhile, recent advancements in large language models (LLMs) leads to growing interest in leveraging their capabilities in modalities beyond text including speech. Training end-to-end ST models often face challenges due to insufficient speech-text parallel data. However, LLMs are trained on vast amounts of textual data and obtain powerful textual generation abilities, which can enhance the ST performance. This has been proven by recent studies that use LLMs as decoders for ST systems (Wu et al., 2023) or as Generative Error Correction (GER) models to improve ST qualities (Hu et al., 2024).

Speech-text parallel data is scarce, and it is even scarcer when it includes emotion annotations.

Findings of the Association for Computational Linguistics: ACL 2025, pages 20413–20421
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Therefore, leveraging external models like LLMs to help the system understand the correlation between emotion and language can be greatly beneficial. However, to the best of our knowledge, there have not been studies on utilizing LLMs for emotion-aware ST.

Therefore, this study pioneers the exploration of the effectiveness of emotion-aware ST by: (a) adopting the LLM GER paradigm, (b) adding emotion and sentiment labels into the GER finetuning process, (c) injecting acoustic representation from the ST model into GER finetuning with a projector.

2 Method

2.1 Generation Error Correction

As illustrated in Figure 2, the GER framework consists of two main components: a pre-trained ST model that produces N-best hypotheses, and a fine-tuned LLM that re-generates the final translation.

2.1.1 N-best Hypotheses Generation

To supply inputs for the GER model, we use a pre-trained ST model to decode N-best hypotheses via beam search. Specifically, given an input speech S in the source language, decoding with beam size M yields $\mathcal{T}_N = \{T_1, T_2, ..., T_N\}$ $(N \leq M)$. In practice, we set $N = M$. These hypotheses serve as preliminary predictions and part of the LLM's input.

2.1.2 GER Finetuning

Inspired by (Hu et al., 2024), we fine-tune an LLM to generate the final translation from N-best hypotheses. Formally,

$$T = M_{EST}(\mathcal{T}_N, I) \tag{1}$$

where I is an instruction prompt (examples shown in Appendix A). The model learns a mapping M_{EST} from $\mathcal{T}_N$ to the true translation T^*. Following a sequence-to-sequence approach, we use T^* as supervision and optimize via cross-entropy:

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(t_l^* \mid t_{l-1}^*, \ldots, t_1^*; \mathcal{T}_N, I) \tag{2}$$

where t_l^* is the l-th token, L is the sequence length, and θ denotes learnable parameters. Considering the large model size of LLMs, we adopt Llama Adapter (Zhang et al., 2023), which inserts learnable prompts into the top L of H Transformer layers (Vaswani, 2017) to capture high-level semantics.

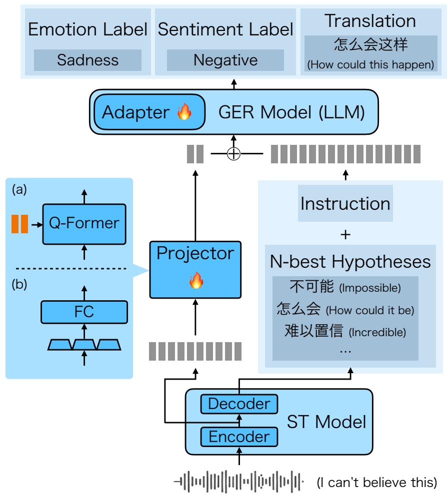

Figure 2: Overview architecture of our proposed model.

2.2 Integration of the Emotion and Sentiment Labels

We incorporate emotion and sentiment labels into the GER fine-tuning process to investigate how emotional content influences translation outcomes. We propose using the GER model to directly predict these labels, which can be considered as a type of multitask learning for the GER model. Based on the hypotheses, the model first generates emotion and sentiment labels and then the translation, as illustrated in Fig 2. In this case, the paradigm and training loss can be defined as:

$$O_{E,T} = M_{EST}(\mathcal{T}_N, I) \tag{3}$$

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(o_l^* | o_{l-1}^*, ..., o_1^*; \mathcal{T}_N, I) \tag{4}$$

where $O_{E,T}$ is the concatenated sequence of E and T, and o_l^* is the l-th token of the ground truth of $O_{E,T}$.

2.3 Injection of Acoustic Representation

Relying solely on textual N-best hypotheses can lose key acoustic cues for emotion prediction. To address this, we inject the ST encoder's acoustic representation into the GER model so that it can leverage both acoustic and textual information. Specifically, we use the projector in Fig 2 to map the encoder's last-layer output $\mathrm{Enc}(S)$ into acoustic embeddings (with the same dimension as the GER model's embeddings):

$$\mathcal{A} = \mathrm{Projector}\big(\mathrm{Enc}(S)\big) \tag{5}$$

We then concatenate $\mathcal{A}$ with textual embeddings (formed by the N-best hypotheses $\mathcal{T}_N$ and instruction I):

$$\mathcal{X} = [\mathcal{A}; \text{Embed}(\mathcal{T}_N, I)] \tag{6}$$

The GER model processes $\mathcal{X}$ in a unified manner, enabling it to process both acoustic and textual inputs for emotion prediction and final translation. During training, projector and adapter parameters are jointly updated.

We explore using the following two architectures for the projector to obtain $\mathcal{A}$:

Q-Former Q-Former (Li et al., 2023) is a module designed to convert variable-length encoder outputs into a fixed-length representation. A set of learnable queries attends to $\text{Enc}(S)$, producing a compact embedding:

$$\mathbf{Q}^{(0)} = \text{InitQueries} \tag{7}$$

$$\mathbf{Q}^{(l)} = \text{TL}_l(\mathbf{Q}^{(l-1)}, \text{Enc}(S)) \tag{8}$$

$$\mathcal{A} = \text{Linear}(\mathbf{Q}^{(L_q)}) \tag{9}$$

where $l = 1, \dots, L_q$ and TL_l denotes l-th Transformer layers.

1-D Convolution Downsampling Alternatively, we adopt a network with a 1-D convolutional layer followed by two fully-connected layers. Mathematically,

$$\mathcal{A} = \text{Linear}(\text{FC}_2(\text{FC}_1(\text{Conv1D}(\text{Enc}(S))))) \tag{10}$$

3 Experiments

3.1 Dataset

In this study, we use the BMELD dataset (Liang et al., 2021), an emotion-aware English-Chinese ST dataset. It is based on the multimodal emotion dialogue dataset MELD (Poria et al., 2018). The Chinese translations are obtained from available subtitles and then manually post-edited according to the dialogue history by native Chinese speakers, who are post-graduate students majoring in English. As in MELD, the utterances are labeled with 7 different emotions and 3 different sentiments. We added both types of labels into the LLM instructions in our experiments. The dataset statistics are in Appendix B.

3.2 Settings

For the ST model, we use the state-of-the-art SeamlessM4T-Large (Barrault et al., 2023), a Transformer-based model that supports speech-to-text translation for up to 100 languages. For the GER model, we adopt the popular Llama-2-7B (Touvron et al., 2023). For the adapter, we follow the default settings of Llama Adapter (Zhang et al., 2023). For the projector, we use 2 learnable queries and a 2-layer architecture for Q-Former, and a downsample rate of 5 for 1-D convolution downsampling. More hyperparameter details are in Appendix C.

Besides integrating emotion and sentiment labels as GER outputs for multitask learning, we also conducted experiments where ground-truth labels were added into GER inputs to represent the performance upper bound.

3.3 Results

The results are presented in Table 1.[1] We evaluate the quality of translations based on metrics including SacreBLEU (Post, 2018), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020). We also report the accuracy of emotion and sentiment prediction.

Results clearly demonstrate that GER outperforms SeamlessM4T by a notable margin, validating the effectiveness of leveraging LLM capabilities for emotional translation refinement. Additionally, incorporating emotion and sentiment labels further enhances this improvement. Using emotion and sentiment labels as inputs to the GER model provides a performance upper bound that significantly outperforms GER without emotion and sentiment labels, which confirms that adding emotional information is beneficial for ST. However, predicting these labels with the GER model only results in marginal performance gains.

Introducing the projector to inject acoustic representations from the encoder leads to a performance closer to the upper bound. When comparing different projectors, 1-D convolution downsampling is slightly more effective than Q-Former considering all the metrics. In addition, 1-D convolution downsampling also shows a modest improvement in emotion recognition accuracy for both emotion and sentiment labels, highlighting a positive correlation between accurate ST and emotion recognition. Nevertheless, it remains unclear how much portion of ST improvement comes from enhanced emotion recognition.

[1] Results on the SeamlessM4T-Medium model and the MELD-ST dataset (Chen et al., 2024) are in Appendix D and E, respectively.

	GER	E/S Labels	Projector	BLEU	BLEURT	COMET	Acc. (E)	Acc. (S)
SeamlessM4T		-	-	11.87	43.34	67.10	-	-
Ours	✓	-	-	15.54†	51.57†	69.90†	-	-
	✓	GER Outputs	-	15.61†	51.81†	69.82†	49.79	53.06
	✓	GER Outputs	Q-former	15.91†‡	51.86†	**70.09**†	48.90	53.44
	✓	GER Outputs	Conv1D	**15.97**†‡	**52.07**†‡	69.97†	**50.17**	**53.52**
Ours (Upper-bound)	✓	GER Inputs	-	16.28†‡	52.50†‡	70.64†‡	-	-

Table 1: ST results on the BMELD dataset. The ST model is SeamlessM4T-large, and the GER model is Llama-2-7B. † and ‡ indicate that the results are significantly better than "SeamlessM4T" and "Ours with GER only" at $p < 0.05$, respectively.

E/S Labels	Projector	BLEU	BLEURT	Acc. (E)	Acc. (S)
-	-	15.54	51.57	-	-
-	Conv1D	15.22	51.85	-	-
E Only	Conv1D	**16.24**†‡	51.96†	**50.90**	-
S Only	Conv1D	16.01†‡	**52.09**†	-	**54.33**
Both	Conv1D	15.97†‡	52.07†	50.17	53.52

Table 2: Ablation studies on emotion (E) and sentiment (S) labels. The model architecture is the same as the one with Conv1D as the projector in Table 1. † and ‡ indicate that the results are significantly better than "- -" and "- Conv1D" at $p < 0.05$, respectively.

Source	Don't worry Phoebe, you're gonna meet someone. If I can meet a great guy, so can you.
Reference	Neu#Neu#别着急，菲比，你会遇到梦中情人的。
SeamlessM4T	别担心 , 如果我能见到一个好人 , 你也能见到 (Don't worry, if I can meet a good person, you can too)
GER	我会找到一个好男人，你也会找到一个好男人。 (I will find a good man and you will find a good man.)
+ E/S Labels	Neu#Neu#不用担心，如果我能找到一个好男人，你也能找到。 (Do not worry, if I can find a good man, you can too.)
+ Conv1D	Neu#Neu#别担心，如果我能找到一个好男人，你也能找到。 (Don't worry, if I can find a good man, you can too.)
Source	Those contracts absolutely had to go out today!
Reference	Ang#Neg#合约必须今天发出去！
SeamlessM4T	这些合同今天就要结束了 (These contracts end today)
GER	这些合同今天就要结束了。 (These contracts end today.)
+ E/S Labels	Neu#Neu#这些合同今天就要结束了。 (These contracts end today.)
+ Conv1D	Ang#Neg#这些合同今天就要结束了！ (These contracts end today!)

Table 3: Translation examples of different methods. The combination of emotion labels, sentiment labels, and translations is presented with the separator "#".

To disentangle the contributions of emotion information and projector-based acoustic injection, we conduct some ablation studies focusing on emotion and sentiment labels, as shown in Table 2. When emotion and sentiment labels are omitted, introducing the projector alone yields relatively poor performance, indicating that the projector alone does not substantially elevate ST quality. However, once emotion and sentiment labels are included, the synergy between acoustic features and emotional content becomes evident, resulting in considerable improvements.

Additionally, we examine the effect of using only emotion labels and only sentiment labels. The results show that using a single type of label can yield comparable, if not slightly better, performance than using both types. This suggests that either emotion or sentiment labels alone may capture sufficient information for emotion-aware ST.

4 Analysis

4.1 Case Study

Table 3 presents example translations comparing different system configurations. In the first example, the baseline SeamlessM4T model mistranslates the phrase "great guy" as "good person," losing the intended nuance of "good male partner." Incorporating the GER method addresses this misinterpretation but omits some details and logic in the resulting sentence. The translation becomes more contextually accurate when emotion and sentiment labels are further integrated.

The second example highlights how emotional cues enhance punctuation, which can convey emotion. The original utterance is delivered with an angry tone, making an exclamation mark an appropriate ending. Without emotional and sentiment labels, or if they are mispredicted, the model fails to generate the correct punctuation. However, when the approach with a projector correctly predicts the underlying emotion, the model accurately appends the exclamation mark.

4.2 Human Evaluation of Emotion Preservation

We conducted a human evaluation of emotion preservation in translations on a subset of the BMELD test data.

Data Selection We drew 101 utterances from the test set, excluding neutral-labeled samples to ensure clear emotional content (e.g., anger, sadness, joy). We selected these samples based on the distribution of non-neutral emotion labels in the overall test set. For each label, we selected samples with

	GER	E/S Labels	Projector	Average	STD	ICC3k
SeamlessM4T		-	-	3.13	0.97	0.735
Ours	✓	-	-	3.11	0.90	0.748
	✓	GER Outputs	-	3.12	0.99	0.824
	✓	GER Outputs	Q-former	3.09	0.93	0.723
	✓	GER Outputs	Conv1D	**3.15**	0.99	0.737
Ours (Upper-bound)	✓	GER Inputs	-	3.27	0.97	0.732

Table 4: Human evaluation of emotion preservation on the BMELD dataset. The ST model is SeamlessM4T-large, and the GER model is Llama-2-7B.

higher diversity across translations from different systems. The *diversity* is measured by counting the number of unique tokens used across the different systems' outputs (i.e., each token that appears in at least one system's translation but not in all).

Evaluation Protocol Each utterance's translations were randomly shuffled to avoid bias from system ordering. Three evaluators rated each translation from 1 to 5 for how well it matches the emotion and sentiment labels.

Evaluation Results In Table 4, we present the average scores, standard deviations (STD), and intraclass correlation coefficients (ICC) among the three evaluators to assess rating reliability. All the ICCs are above 0.7, which indicates good reliability.

Regarding the average scores, there appears to be little difference among the first five models considering the standard deviation, suggesting that using GER with E/S labels as outputs does not enhance emotion preservation. However, we argue that this is reasonable, as it is challenging for a single sentence of text to clearly convey emotion. Our goal in incorporating emotional content is not only to preserve emotion but also to improve the translation's semantic accuracy (as in the example of Figure 1).

Additionally, using E/S labels as inputs shows a moderate improvement, indicating that integrating these labels can lead to better emotion preservation. However, we need a more effective method to further enhance emotion recognition, which we plan to address in future work.

5 Conclusion

In this paper, we pioneered the investigation of emotion-aware ST using LLMs. We proposed adopting the GER method, integrating emotion and sentiment labels, and injecting acoustic information from the speech into the GER finetuning process. The experimental results showed its effectiveness. Future works include verifying the effectiveness of our method with other LLMs and datasets, as well as increasing the diversity of the N-best list and exploring semi-supervised strategies for the task.

Limitations

- The LLM used in the experiments is Llama-2-7B, which, while powerful, may not capture the full potential of larger or more advanced models. The limited model size may constrain the quality of translations and the handling of complex linguistic nuances, particularly when related to emotion and sentiment.
- Our experiments are only conducted on three language pairs (en-zh, en-ja, en-de), and hence the generalizability of our findings to other languages remains to be validated. In future work, we plan to expand our evaluations to additional datasets that can be used for emotion-aware ST training, such as mExpresso[2] and EmoFilm (Morgenroth et al., 2025).
- We only adopt beam search to generate the N-best hypotheses, which can possibly limit the generation diversity and information presented to the GER model. We will investigate different decoding strategies, such as Diverse Beam Search (Vijayakumar et al., 2016) and temperature sampling (Holtzman et al., 2019), to increase the diversity of the N-best list, enabling more diverse translations for different emotions.
- Our approach relies on fine-tuning the LLM and projectors on a supervised dataset, which can limit scalability when labeled emotional speech resources are scarce. In future work, we plan to address this limitation by exploring semi-supervised strategies. The focus will be on generating and filtering emotion pseudo-labels for ST data, combined with semi-supervised training methods such as bootstrapping (Lee et al., 2013) and the Π-model (Laine and Aila, 2016).

[2] https://huggingface.co/facebook/seamless-expressive

Ethical Considerations

This study exclusively uses publicly available datasets (BMELD and MELD-ST) for emotion-aware speech-to-text translation, ensuring compliance with ethical and privacy standards. Our work does not involve any private or sensitive data collection. In addition, we confirm that the dataset and models used in our study were obtained and utilized in full compliance with their respective licenses and intended use guidelines.

Acknowledgment

This work was supported by the SPRING program of Kyoto University and JSPS KAKENHI Grant Number JP23K28144.

References

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

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*.

Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. 2024. MELD-ST: an emotion-aware speech translation dataset. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10118–10126. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Agarwal et al. 2023a. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*.

Paul K. Rubenstein et al. 2023b. AudioPaLM: A Large Language Model That Can Speak and Listen. *Preprint*, arXiv:2306.12925.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. 2024. Gentranslate: Large language models are generative multilingual speech and machine translators. *arXiv preprint arXiv:2402.06894*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS Corpus and Massively Multilingual Speech-to-Speech Translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat translation. 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 5711–5724, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Elenor Morgenroth, Stefano Moia, Laura Vilaclara, Raphael Fournier, Michal Muszynski, Maria Ploumitsakou, Marina Almató-Bellavista, Patrik Vuilleumier, and Dimitri Van De Ville. 2025. Emo-film: a multimodal dataset for affective neuroscience using naturalistic stimuli. *Scientific Data*, 12(1):684.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

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

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models (2016). *arXiv preprint arXiv:1610.02424*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2023. GigaST: A 10,000-hour Pseudo Speech Translation Corpus. In *Proc. INTERSPEECH 2023*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

A Instruction Prompts

The following is the instruction prompt used for GER fine-tuning in our experiments. It includes three variations: without emotion and sentiment labels, with labels as GER outputs, and with labels as GER inputs. The output is highlighted in blue.

GER

Below is the best-hypotheses transcribed from speech translation system. Please try to revise it using the words which are only included into other-hypothesis, and write the response for the true transcription.

Best-hypothesis:
那你为什么不给你的号码呢?.

Other-hypothesis:
所以你为什么不给你的号码呢?. 所以你为什么不给你的号码?. 那你为什么不给自己号码呢?. 那你为什么不给你的号码?.

Response:
给我留个电话嘛。

GER + E/S Labels as Outputs

You will be shown the best-hypotheses transcribed from speech translation system. Please try to predict the emotion and the sentiment of the speech, and try to revise the best-hypothesis using the words which are included in other-hypothesis. Please write the response in the following format:
Emotion
Sentiment
True transcription.

Best-hypothesis:
那你为什么不给你的号码呢?.

Other-hypothesis:
所以你为什么不给你的号码呢?. 所以你为什么不给你的号码?. 那你为什么不给自己号码呢?. 那你为什么不给你的号码?.

Response:
neutral
neutral
给我留个电话嘛。

GER + E/S Labels as Inputs

Below is the best-hypotheses transcribed from speech translation system, as well as the emotion and the sentiment of the speech. Please try to revise the best-hypothesis using the words which are included in other-hypothesis while considering the emotion and sentiment, and write the response for the true transcription.

Best-hypothesis:
那你为什么不给你的号码呢?.

Other-hypothesis:
所以你为什么不给你的号码呢?. 所以你为什么不给你的号码?. 那你为什么不给自己号码呢?. 那你为什么不给你的号码?.

Emotion:
neutral

Sentiment:
neutral

Response:
给我留个电话嘛。

B Dataset Statistics

In addition to the BMELD dataset used in the main paper, we also conducted experiments on the MELD-ST dataset (Chen et al., 2024), which is constructed in a similar manner but without post-editing, containing both English-Japanese and English-German language pairs. Table 5 presents the dataset statistics for the three language pairs used in our experiments: one from BMELD and two from MELD-ST. It includes the number of samples in the training, validation, and test sets, along with the distribution of emotion and sentiment labels. Both datasets are derived from MELD dataset but differ in data partitioning and translation sources, resulting in slight variations in dataset size.

C Experimental Setup and Hyperparameters

The Q-former layers uses the same hyperparameters as the vanilla Transformer layer. The 1-D convolution downsampling network uses a hidden dimension of $2,048$. For the adapters, the number of tunable Transformer layers L is set to $H-1$, which means all layers except the first one are tunable with inserted prompts. The prompt length U is set to 10. As a result, the total number of trainable parameters is 12.3M when using Q-Former and 17M when using 1-D convolution downsampling.

We conducted our experiments on a single A100 80G GPU, with each experiment being a single run. We train for 2 epochs with the AdamW optimizer (Loshchilov and Hutter, 2019), with the learning rate initialized at $1e^{-2}$ and then linearly decreased to $1e^{-5}$ during training. The batch size is set to 4, with accumulation iterations set to 8 (i.e., the real batch size is 32).

As for evaluation, we used SacreBLEU with its own tokenizer: “zh” for Chinese, “ja-mecab” for Japanese, and “13a” for German. We used the BLEURT-20 model for BLEURT.

D Results with SeamlessM4T Medium

Table 6 presents results using the same settings as Table 1, except that SeamlessM4T-Large is replaced with SeamlessM4T-Medium. The results indicate that the impact of introducing the projector is less significant compared to using SeamlessM4T-Large. This is likely because SeamlessM4T-Large has a higher-dimensional encoder output, providing more information for the projector to utilize. Additionally, the performance upper bound (using emotion and sentiment labels as GER inputs) shows an unexpected low BLEU score while maintaining a high BLEURT score, indicating that BLEURT may be a more reliable evaluation metric than traditional BLEU.

E Results on MELD-ST

We also conducted experiments on MELD-ST using the same settings as for BMELD. Tables 7 and 8 present the results for the en-ja and en-de language pairs, respectively. However, the results show less consistent improvements compared to BMELD, with a noticeable gap between the two evaluation metrics.

The primary reason is likely the lower quality of training data, as translations in MELD-ST training sets were not manually verified. Other possible factors include the weaker performance of Llama-2-7B in Japanese and German than in Chinese.

Additionally, the relatively smaller cultural gap between English and German may reduce the impact of incorporating emotion and sentiment labels, as direct translation already performs well. For example, in MELD-ST, the English phrase “I’m sorry for your loss” spoken with an angry tone is translated into Japanese as “同情してあげる.” Although “同情” simply means “to sympathize,” adding “してあげる” implies that the

Dataset	Split	Total	Neu.	Joy.	Sad.	Fea.	Ang.	Sur.	Dis.	Neu.	Pos.	Neg.
	Train	9,987	4,709	1,743	682	268	1,109	1,205	271	4,709	2,334	2,944
BMELD	Valid	1,084	460	162	109	40	146	146	21	460	231	393
(en-zh)	Test	2,601	1,251	400	208	50	345	279	68	1,251	518	832
	Train	8,069	3,836	1,284	603	209	982	917	238	3,836	1,715	2,518
MELD-ST	Valid	1,008	482	176	84	31	116	97	22	482	229	297
(en-ja)	Test	1,008	479	186	73	25	85	121	39	479	253	276
	Train	9,314	4,402	1,571	656	232	1,096	1,096	261	4,402	2,084	2,828
MELD-ST	Valid	1,164	550	202	99	31	127	130	25	550	271	343
(en-de)	Test	1,164	550	218	92	32	102	131	39	550	288	326

Table 5: Statistics for the datasets we used (BMELD, MELD-ST). There are 7 types of emotion labels: Neutral (Neu.), Joy (Joy.), Sadness (Sad.), Fear (Fea.), Anger (Ang.), Surprise (Sur.), Disgust (Dis.); and 3 types of sentiment labels: Neutral (Neu.), Positive (Pos.), Negative (Neg.)

	GER	E/S Labels	Projector	BLEU	BLEURT	Acc. (E)	Acc. (S)
SeamlessM4T	-	-	-	11.50	41.52	-	-
Ours	✓	-	-	12.80†	49.97†	-	-
	✓	GER Outputs	-	13.67†‡	50.25†	50.25	52.21
	✓	GER Outputs	Q-former	13.71†‡	**50.30**†‡	**50.40**	**53.02**
	✓	GER Outputs	Conv1D	**13.74**†‡	50.12†	49.63	51.98
Ours (Upper-bound)	✓	GER Inputs	-	13.08†	50.47†‡	-	-

Table 6: ST results on the BMELD dataset. The ST model is SeamlessM4T-medium, and the GER model is Llama-2-7B. † and ‡ indicate that the results are significantly better than "SeamlessM4T" and "Ours with GER only" at $p < 0.05$, respectively.

	GER	E/S Labels	Projector	BLEU	BLEURT	Acc. (E)	Acc. (S)
SeamlessM4T	-	-	-	2.20	27.57	-	-
Ours	✓	-	-	3.02†	**26.40**	-	-
	✓	GER Outputs	-	**3.49**†	25.69	**51.09**	**54.56**
	✓	GER Outputs	Q-former	3.16†	24.96	50.10	53.47
	✓	GER Outputs	Conv1D	2.92†	25.59	50.00	53.87
Ours (Upper-bound)	✓	GER Inputs	-	3.58†‡	26.25	-	-

Table 7: ST results on the MELD-ST dataset for the en-ja language pair. The ST model is SeamlessM4T-large, and the GER model is Llama-2-7B. † and ‡ indicate that the results are significantly better than "SeamlessM4T" and "Ours with GER only" at $p < 0.05$, respectively.

	GER	E/S Labels	Projector	BLEU	BLEURT	Acc. (E)	Acc. (S)
SeamlessM4T	-	-	-	**11.74**	52.68	-	-
Ours	✓	-	-	10.96	**54.04**†	-	-
	✓	GER Outputs	-	11.07	53.53†	51.55	54.38
	✓	GER Outputs	Q-former	11.14	54.01†	**51.89**	**57.13**
	✓	GER Outputs	Conv1D	11.19	53.42†	50.95	54.55
Ours (Upper-bound)	✓	GER Inputs	-	11.28	54.29†	-	-

Table 8: ST results on the MELD-ST dataset for the en-de language pair. The ST model is SeamlessM4T-large, and the GER model is Llama-2-7B. † indicates that the results are significantly better than "SeamlessM4T" at $p < 0.05$.

speaker is doing the act of sympathizing as a favor, which carries a nuance of superiority that aligns with the angry emotion. If the phrase were spoken with another emotion, it would likely be translated as "同情するよ." This difference is due to unique Japanese expressions that cannot be directly translated back into English. In contrast, the German translation is "Es tut mir leid für deinen Lieblingspulli," which literally means "I'm sorry about your favorite sweater." Here, the semantic change occurs because MELD-ST is based on TV series subtitles in different languages, which do not necessarily correspond directly. Except for this change, the German translation uses expressions very similar to the English, without any special adjustment for the angry tone. Even if the emotion were different, the translation would likely remain the same.