Understanding and Bridging the Modality Gap for Speech Translation

Qingkai Fang[1,2], **Yang Feng**[1,2*]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2]University of Chinese Academy of Sciences, Beijing, China
{fangqingkai21b, fengyang}@ict.ac.cn

Abstract

How to achieve better end-to-end speech translation (ST) by leveraging (text) machine translation (MT) data? Among various existing techniques, multi-task learning is one of the effective ways to share knowledge between ST and MT in which additional MT data can help to learn source-to-target mapping. However, due to the differences between speech and text, there is always a gap between ST and MT. In this paper, we first aim to understand this *modality gap* from the target-side representation differences, and link the modality gap to another well-known problem in neural machine translation: exposure bias. We find that the modality gap is relatively small during training except for some difficult cases, but keeps increasing during inference due to the cascading effect. To address these problems, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method. Specifically, we regularize the output predictions of ST and MT, whose target-side contexts are derived by sampling between ground truth words and self-generated words with a varying probability. Furthermore, we introduce token-level adaptive training which assigns different training weights to target tokens to handle difficult cases with large modality gaps. Experiments and analysis show that our approach effectively bridges the modality gap, and achieves promising results in all eight directions of the MuST-C dataset.[1]

1 Introduction

End-to-end speech translation (ST) aims to translate speech signals to text in another language directly. Compared to traditional cascaded methods, which combine automatic speech recognition (ASR) and machine translation (MT) models in a pipeline manner, end-to-end ST could avoid error propagation and high latency (Sperber and Paulik, 2020). Recently, end-to-end ST models have achieved comparable or even better results than cascaded ST models (Bentivogli et al., 2021; Anastasopoulos et al., 2021, 2022).

However, due to the scarcity of ST data, it is difficult to directly learn a mapping from source speech to the target text. Previous works often leverage MT data to help the training with multi-task learning (Ye et al., 2022; Tang et al., 2021a). By sharing encoder and decoder between ST and MT, the model tends to learn similar representations from different modalities. In this way, the auxiliary MT task can help build the source-to-target mapping. However, there remains a gap between ST and MT due to the differences between speech and text. In this paper, we measure the *modality gap* with representation differences of the last decoder layer between ST and MT, because the representation of this layer will be mapped into the embedding space to obtain the final translation. A significant modality gap potentially causes different predictions, which makes ST lag behind MT.

Thanks to multi-task learning, we observe that when training with teacher forcing, where both ST and MT use ground truth words as target-side contexts, the modality gap is relatively small except for some difficult cases. However, the exposure bias problem can make things worse. During inference, both ST and MT predict the next token conditioned on their previously generated tokens, which may be different due to the modality gap. Moreover, different predictions at the current step may lead to even more different predictions at the next step. As a result, the modality gap will increase step by step due to this cascading effect.

To solve these problems, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method. To reduce the effect of exposure bias, we introduce scheduled sampling during training, where the target-side contexts are sampled be-

* Corresponding author: Yang Feng.

[1]Code is publicly available at https://github.com/ictnlp/CRESS.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 15864–15881
July 9-14, 2023

tween ground truth words and self-generated words with a changing probability. Based on this, we propose to regularize ST and MT in the output space to bridge the modality gap by minimizing the Kullback-Leibler (KL) divergence between their predictions. This will encourage greater consistency between ST and MT predictions based on partial self-generated words, which is closer to the inference mode. Besides, to handle the difficult cases, we introduce token-level adaptive training for **CRESS**, where each target token is given a varying weight during training according to the scale of the modality gap. In this way, those cases with significant modality gaps will be emphasized. We conduct experiments on the ST benchmark dataset MuST-C (Di Gangi et al., 2019a). Results show that our approach significantly outperforms the strong multi-task learning baseline, with 1.8 BLEU improvements in the base setting and 1.3 BLEU improvements in the expanded setting on average. Further analysis shows that our approach effectively bridges the modality gap and improves the translation quality, especially for long sentences.

2 Background

2.1 End-to-end Speech Translation

End-to-end speech translation (ST) directly translates speech in the source language to text in the target language. The corpus of ST is usually composed of triplet data $\mathcal{D} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$. Here $\mathbf{s} = (s_1, ..., s_{|\mathbf{s}|})$ is the sequence of audio wave, $\mathbf{x} = (x_1, ..., x_{|\mathbf{x}|})$ is the transcription and $\mathbf{y} = (y_1, ..., y_{|\mathbf{y}|})$ is the translation. Similar to previous work (Ye et al., 2021; Fang et al., 2022), our ST model is composed of an acoustic encoder and a translation model. The acoustic encoder is a pre-trained HuBERT (Hsu et al., 2021) model followed by two convolutional layers, which are used to reduce the length of the speech sequence. The translation model follows standard Transformer (Vaswani et al., 2017) encoder-decoder architecture, where the encoder contains N Transformer encoder layers, and the decoder contains N Transformer decoder layers. We first pre-train the translation model with MT data, and then optimize the whole model by minimizing a cross-entropy loss:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{s}, \mathbf{y}_{<i}), \tag{1}$$

$$p(y_i|\mathbf{s}, \mathbf{y}_{<i}) \propto \exp(\mathbf{W} \cdot f(\mathbf{s}, \mathbf{y}_{<i})), \tag{2}$$

where f is a mapping from the input speech $\mathbf{s}$ and target prefix $\mathbf{y}_{<i}$ to the representation of the last decoder layer at step i. $\mathbf{W}$ is used to transform the dimension to the size of the target vocabulary.

2.2 Multi-task Learning for ST

Multi-task learning (MTL) has been proven useful for sharing knowledge between text translation and speech translation (Tang et al., 2021a), where an auxiliary MT task is introduced during training:

$$\mathcal{L}_{\mathrm{MT}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{x}, \mathbf{y}_{<i}), \tag{3}$$

$$p(y_i|\mathbf{x}, \mathbf{y}_{<i}) \propto \exp(\mathbf{W} \cdot f(\mathbf{x}, \mathbf{y}_{<i})). \tag{4}$$

Note that both modalities (i.e., speech and text) share all transformer encoder and decoder layers. Finally, the training objective is written as follows:

$$\mathcal{L}_{\mathrm{MTL}} = \mathcal{L}_{\mathrm{ST}} + \mathcal{L}_{\mathrm{MT}}. \tag{5}$$

3 Preliminary Studies on Modality Gap

With multi-task learning, most of the knowledge of MT can be transferred to ST. However, the performance gap between ST and MT still exists. In this section, we first conduct some preliminary studies with our multi-task learning baseline model to understand where this gap comes from.

3.1 Definition of the Modality Gap

The gap between ST and MT is related to the prediction difference at each decoding step, while the prediction depends only on the representation of the last decoder layer. Therefore, we define the *modality gap* at the i-th decoding step as follows:

$$G(\mathbf{s}, \mathbf{y}_{<i} \| \mathbf{x}, \mathbf{y}_{<i}) = 1 - cos(f(\mathbf{s}, \mathbf{y}_{<i}), f(\mathbf{x}, \mathbf{y}_{<i})), \tag{6}$$

where cos is the cosine similarity function $cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / \|\mathbf{a}\| \|\mathbf{b}\|$. A larger cosine similarity indicates a smaller modality gap.

To understand the extent of the modality gap, we count the distribution of $G(\mathbf{s}, \mathbf{y}_{<i} \| \mathbf{x}, \mathbf{y}_{<i})$ based on all triples $(\mathbf{s}, \mathbf{x}, \mathbf{y}_{<i})$ in the MuST-C (Di Gangi et al., 2019a) En→De dev set. As shown in Figure 1, the modality gap is relatively small ($< 10\%$) in most cases, which proves the effectiveness of multi-task learning in sharing knowledge across ST and MT. However, we also observe a long-tail problem: there is a large difference between ST and MT representations in some difficult cases.

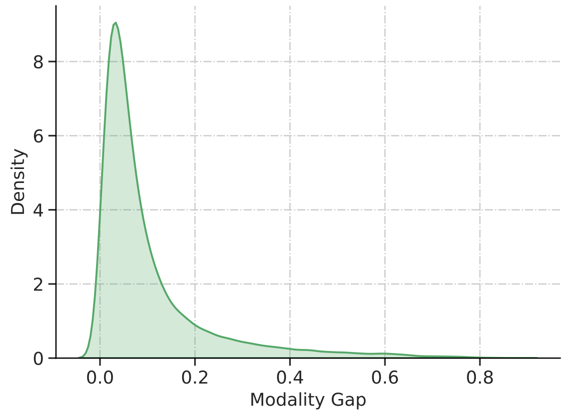

Figure 1: Distribution of the modality gap on MuST-C En→De dev set with kernel density estimation (KDE).

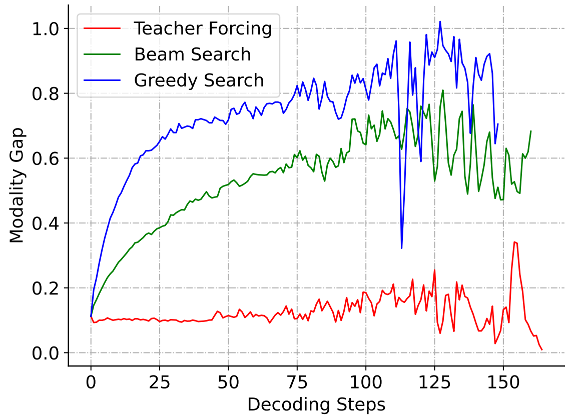

Figure 2: Curves of the average modality gap on MuST-C En→De dev set with decoding steps under *teacher forcing*, *beam search*, and *greedy search* strategies. For *beam search*, we have several candidate translations. The modality gap is calculated with the average representation of all candidates. We set a beam size of 8.

3.2 Connection between Exposure Bias and Modality Gap

Exposure bias, a discrepancy between training and inference, is a well-known problem in neural machine translation (Bengio et al., 2015; Ranzato et al., 2016; Wang and Sennrich, 2020; Arora et al., 2022). During training with *teacher forcing*, both ST and MT predict the next token conditioned on the ground truth target prefix $\mathbf{y}_{<i}$. However, during inference, the predictions of ST and MT depend on their previously generated tokens by the model itself (denoted as $\widehat{\mathbf{y}}^s_{<i}$ and $\widehat{\mathbf{y}}^x_{<i}$ for ST and MT respectively), which might be different due to the modality gap. Furthermore, different predictions at the current decoding step result in different target prefixes for ST and MT, potentially causing even more different predictions at the next step. Such cascading effect will enlarge the modality gap step by step during inference.

To prove our hypothesis, we present the curves of the modality gap with decoding steps under *teacher forcing*, *beam search*, and *greedy search* strategies, respectively. As shown in Figure 2, with *teacher forcing*, there is no significant difference in the modality gap across steps, as both ST and MT depend on the same target prefix at any step. Hence, the modality gap $G(\mathbf{s}, \mathbf{y}_{<i} \| \mathbf{x}, \mathbf{y}_{<i})$ only comes from the difference between input speech $\mathbf{s}$ and text $\mathbf{x}$. However, when decoding with *greedy search*, due to the cascading effect mentioned above, the self-generated target prefix $\widehat{\mathbf{y}}^s_{<i}$ and $\widehat{\mathbf{y}}^x_{<i}$ become increasingly different, making the modality gap $G(\mathbf{s}, \widehat{\mathbf{y}}^s_{<i} \| \mathbf{x}, \widehat{\mathbf{y}}^x_{<i})$ keep increasing with decoding steps. A simple way to alleviate this problem is *beam search*, which considers several candidate tokens rather than a single one at each decoding step. When there is an overlap between candidate tokens of ST and MT, the cascading effect will be reduced, thus slowing down the increase of the modality gap.

4 Method: CRESS

Our preliminary studies in Section 3 show that:

- The modality gap will be enlarged during inference due to exposure bias.
- The modality gap may be significant in some difficult cases.

Inspired by these, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method to bridge the modality gap, especially in inference mode (Section 4.1). Furthermore, we propose a token-level adaptive training method for **CRESS** to handle difficult cases (Section 4.2).

4.1 Cross-modal Regularization with Scheduled Sampling (CRESS)

To bridge the modality gap during inference, we adopt scheduled sampling for both ST and MT to approximate the inference mode at training time. After that, we add a regularization loss between the predictions of ST and MT based on the part of their self-generated words as context. This allows for more consistent predictions between ST and MT during inference, thus reducing the performance gap between ST and MT. Figure 3 illustrates the main framework of our method.

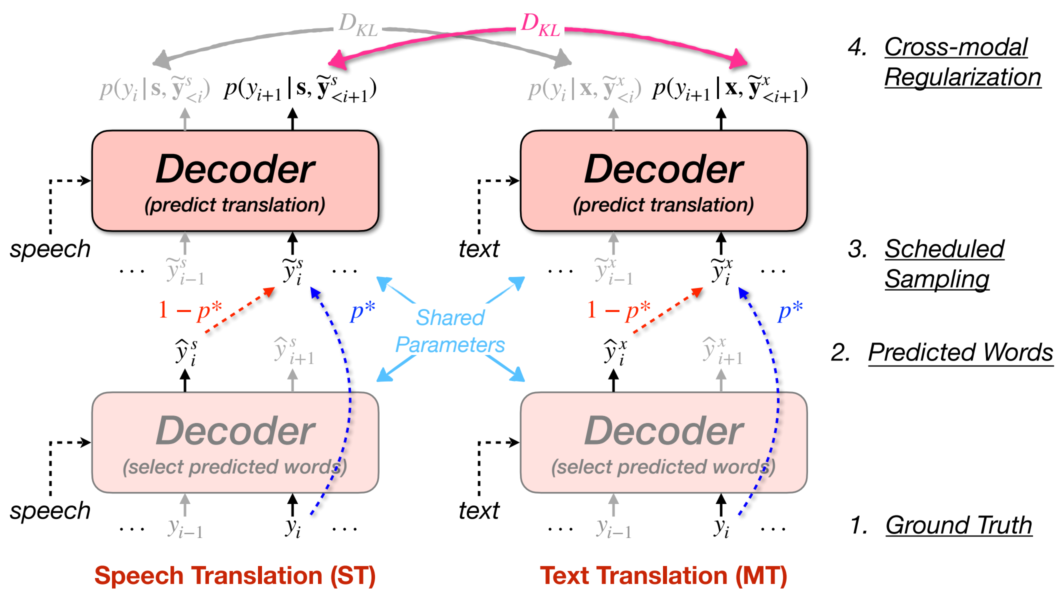

Figure 3: Overview of our proposed **Cress**. Note that the step of selecting predicted words has no gradient calculation and is fully parallelized.

Scheduled Sampling *Scheduled sampling* (Bengio et al., 2015), which samples between ground truth words and self-generated words, i.e., *predicted words*, with a certain probability as target-side context, has proven helpful in alleviating exposure bias. In general, the input at the $\{i+1\}$-th decoding step should be the ground truth word y_i during training. With scheduled sampling, it can also be substituted by a predicted word. Next, we describe how to select the predicted word $\widehat{y}_i^s$ for ST and $\widehat{y}_i^x$ for MT. For ST, we follow Zhang et al. (2019) to select the predicted word $\widehat{y}_i^s$ by sampling from the word distribution $p(y_i|\mathbf{s}, \mathbf{y}_{<i})$ in Equation (2) with *Gumbel-Max* technique (Gumbel, 1954; Maddison et al., 2014), a method to draw a sample from a categorical distribution:

$$\eta = -\log(-\log u), \quad (7)$$

$$\widehat{y}_i^s = \arg\max\left(\mathbf{W} \cdot f(\mathbf{s}, \mathbf{y}_{<i}) + \eta\right), \quad (8)$$

where η is the Gumbel noise calculated from the uniform noise $u \sim \mathcal{U}(0,1)$. Similarly, for MT, there is:

$$\widehat{y}_i^x = \arg\max\left(\mathbf{W} \cdot f(\mathbf{x}, \mathbf{y}_{<i}) + \eta\right). \quad (9)$$

Note that we may omit the superscript and denote the predicted word for both ST and MT by $\widehat{y}_i$ in the following.

How to select between the ground truth word y_i and the predicted word $\widehat{y}_i$? Similar to Bengio et al. (2015); Zhang et al. (2019), we randomly sample from both with a varying probability. We denote the probability of selecting from the ground truth word as p^*. At the beginning of training, since the model is not yet well trained, we select more from the ground truth words (with larger p^*) to help the model converge. In the later stages of training, we select more from the predicted words (with smaller p^*), which is closer to the situation during inference. To achieve this, we decrease p^* with a function of the index of training epochs e:

$$p^* = \frac{\mu}{\mu + \exp(e/\mu)}, \quad (10)$$

where μ is a hyper-parameter. With scheduled sampling, the target-side context becomes $\widetilde{\mathbf{y}} = (\widetilde{y}_1, ..., \widetilde{y}_{|\mathbf{y}|})$, where

$$\widetilde{y}_i = \begin{cases} y_i, & p \le p^* \\ \widehat{y}_i, & p > p^* \end{cases}, \quad (11)$$

where p is sampled from the uniform distribution $\mathcal{U}(0,1)$. Using $\widetilde{\mathbf{y}}^s$ and $\widetilde{\mathbf{y}}^x$ to denote the target-side context of ST and MT respectively, the loss functions of ST and MT become:

$$\mathcal{L}_{\text{ST}}^{\text{CRESS}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}^s_{<i}), \quad (12)$$

$$\mathcal{L}_{\text{MT}}^{\text{CRESS}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}^x_{<i}), \quad (13)$$

Cross-modal Regularization To bridge the modality gap in inference mode, we expect the predictions of ST and MT with scheduled sampling to be consistent. Inspired by recent works of consistency training (Liang et al., 2021; Guo et al., 2022), we regularize ST and MT in the output space. Specifically, we minimize the bidirectional Kullback-Leibler (KL) divergence between the output distributions of ST and MT at each step:

$$\mathcal{L}_{\text{Reg}}^{\text{CRESS}} = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2}(\mathcal{D}_{\text{KL}}(p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s) \| p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x)) + \mathcal{D}_{\text{KL}}(p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x) \| p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s))). \tag{14}$$

With the translation loss in Equation (12) and (13), the final training objective is:

$$\mathcal{L}^{\text{CRESS}} = \mathcal{L}_{\text{ST}}^{\text{CRESS}} + \mathcal{L}_{\text{MT}}^{\text{CRESS}} + \lambda \mathcal{L}_{\text{Reg}}^{\text{CRESS}}, \tag{15}$$

where λ is the hyper-parameter to control the weight of $\mathcal{L}_{\text{Reg}}^{\text{CRESS}}$.

4.2 Token-level Adaptive Training for CRESS

As mentioned above, the modality gap might be significant in some difficult cases. Inspired by the idea of token-level adaptive training (Gu et al., 2020; Xu et al., 2021b; Zhang et al., 2022a), we propose to treat each token adaptively according to the scale of the modality gap. The training objectives in Equation (12), (13), and (14) are modified as follows:

$$\mathcal{L}_{\text{ST}}^{\text{CRESS}} = -\sum_{i=1}^{|\mathbf{y}|} w_i \cdot \log p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s), \tag{16}$$

$$\mathcal{L}_{\text{MT}}^{\text{CRESS}} = -\sum_{i=1}^{|\mathbf{y}|} w_i \cdot \log p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x), \tag{17}$$

$$\mathcal{L}_{\text{Reg}}^{\text{CRESS}} = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2} w_i (\mathcal{D}_{\text{KL}}(p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s) \| p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x)) + \mathcal{D}_{\text{KL}}(p(y_i|\mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x) \| p(y_i|\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s))), \tag{18}$$

where w_i is the token-level weight defined by a linear function of the modality gap:

$$w_i = B + S \cdot G(\mathbf{s}, \widetilde{\mathbf{y}}_{<i}^s \| \mathbf{x}, \widetilde{\mathbf{y}}_{<i}^x), \tag{19}$$

where B (base) and S (scale) are hyper-parameters to control the lower bound and magnitude of change of w_i. In this way, cases with a large modality gap will be assigned a larger weight and thus emphasized during training. Note that the modality gap is computed on-the-fly during training.

5 Experiments

5.1 Datasets

ST Datasets We conduct experiments on MuST-C (Di Gangi et al., 2019a) dataset, a multilingual ST dataset containing 8 translation directions: English (En) to German (De), French (Fr), Spanish (Es), Romanian (Ro), Russian (Ru), Italian (It), Portuguese (Pt) and Dutch (Nl). It contains at least 385 hours of TED talks with transcriptions and translations for each direction. We use `dev` set for validation and `tst-COMMON` set for evaluation.

External MT Datasets We also introduce external MT datasets to pre-train our translation model in the expanded setting. For En→De/Fr/Es/Ro/Ru directions, we introduce data from WMT (Buck and Koehn, 2016). For En→It/Pt/Nl, we introduce data from OPUS100[2] (Zhang et al., 2020). Table 4 in Appendix A lists the statistics of all datasets.

5.2 Experimental Setups

Pre-processing For *speech* input, we use the raw 16-bit 16kHz mono-channel audio wave. For *text* input, all sentences in ST and external MT datasets are tokenized and segmented into subwords using SentencePiece[3]. For each translation direction, the vocabulary is learned from the source and target texts from the ST dataset, with a size of 10K. For the external MT datasets, we filter out parallel sentence pairs whose length ratio exceeds 1.5.

Model Setting We use the pre-trained HuBERT model[4] to encode the input audio. Two 1-dimensional convolutional layers after HuBERT are set to kernel size 5, stride size 2, and padding 2. For the translation model, we employ Transformer architecture with the base configuration, which contains 6 encoder layers and 6 decoder layers, with 512 hidden states, 8 attention heads, and 2048 feed-forward hidden states for each layer. The translation model is first pre-trained with MT task using *transcription-translation* pairs from the ST dataset (**base setting**), and also sentence pairs from the external MT dataset (**expanded setting**).

During MT pre-training, each batch has up to 33k text tokens. The maximum learning rate is set to 7e-4. During fine-tuning, each batch contains up to 16M audio frames. The maximum learning rate

[2] `http://opus.nlpl.eu/opus-100.php`
[3] `https://github.com/google/sentencepiece`
[4] `https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt`

Models	BLEU								
	En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.
Base setting (w/o external MT data)									
XSTNet (Ye et al., 2021)	25.5	36.0	29.6	25.1	16.9	25.5	31.3	30.0	27.5
STEMM (Fang et al., 2022)	25.6	36.1	30.3	24.3	17.1	25.6	31.0	30.1	27.5
ConST (Ye et al., 2022)	25.7	36.8	30.4	24.8	17.3	26.3	32.0	30.6	28.0
MTL	25.3	35.7	30.5	23.8	17.2	26.0	31.3	29.5	27.4
CRESS	**27.2****	**37.8****	**31.9****	**25.9****	**18.7****	**27.3****	**33.0****	**31.6****	**29.2**
Expanded setting (w/ external MT data)									
Chimera (Han et al., 2021)	27.1	35.6	30.6	24.0	17.4	25.0	30.2	29.2	27.4
XSTNet (Ye et al., 2021)	27.1	38.0	30.8	25.7	18.5	26.4	32.4	31.2	28.8
STEMM (Fang et al., 2022)	28.7	37.4	31.0	24.5	17.8	25.8	31.7	30.5	28.4
ConST (Ye et al., 2022)	28.3	38.3	32.0	25.6	18.9	27.2	33.1	31.7	29.4
†STPT (Tang et al., 2022)	-	39.7	33.1	-	-	-	-	-	-
†SpeechUT (Zhang et al., 2022b)	**30.1**	**41.4**	**33.6**	-	-	-	-	-	-
MTL	27.7	38.5	32.8	24.9	19.0	26.5	32.0	30.8	29.0
CRESS	29.4**	40.1**	33.2*	**26.4****	**19.7****	**27.6****	**33.6****	**32.3****	**30.3**

Table 1: BLEU scores on MuST-C `tst-COMMON` set. The external MT datasets are only used in the expanded setting. * and ** mean the improvements over **MTL** baseline are statistically significant ($p < 0.05$ and $p < 0.01$, respectively). †: speech-text jointly pre-trained models whose training costs are much higher than our models.

is set to 1e-4. We use Adam optimizer (Kingma and Ba, 2015) with 4k warm-up steps. We set dropout to 0.1 and label smoothing to 0.1. The training will early stop if the BLEU score on the `dev` set did not increase for 10 epochs. During inference, we average the checkpoints of the last 10 epochs for evaluation. We use beam search with a beam size of 8. The length penalty is set to 1.2, 1.8, 0.6, 1.4, 0.8, 1.0, 1.4, and 1.0 for En→De, Fr, Es, Ro, Ru, It, Pt and Nl, respectively. We use scareBLEU[5] (Post, 2018) to compute case-sensitive detokenized BLEU (Papineni et al., 2002) scores and the statistical significance of translation results with paired bootstrap resampling[6] (Koehn, 2004). We implement our model with *fairseq*[7] (Ott et al., 2019). All models are trained on 4 Nvidia RTX 3090 GPUs.

For scheduled sampling, the decay parameter is set to $\mu = 15$. For cross-modal regularization, the weight parameter is set to $\lambda = 1.0$. For token-level adaptive training, we did a grid search for base and scale parameters on MuST-C En→De `dev` set with $B \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ and $S \in \{0.05, 0.10, 0.20, 0.50, 1.00\}$. Finally, we set $B = 0.7$ and $S = 0.05$ for all translation directions. We start applying token-level adaptive training after the 20th epoch during training.

[5]`https://github.com/mjpost/sacrebleu`

[6]sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

[7]`https://github.com/pytorch/fairseq`

Baseline Systems We include several strong end-to-end ST models for comparison: Chimera (Han et al., 2021), XSTNet (Ye et al., 2021), STEMM (Fang et al., 2022), ConST (Ye et al., 2022), STPT (Tang et al., 2022), and SpeechUT (Zhang et al., 2022b). Besides, the multi-task learning baseline in Section 2.2 is also included as a strong baseline, which is denoted as **MTL**. We use **CRESS** to denote our method with token-level adaptive training.

Among these models, Chimera, XSTNet, STEMM, and ConST combine pre-trained Wav2vec 2.0 (Baevski et al., 2020) and pre-trained translation model together, and then fine-tune the whole model on ST datasets. Our implemented **MTL** and **CRESS** follow a similar design, but we use HuBERT instead of Wav2vec 2.0 as we find HuBERT gives a stronger baseline (See Table 5 for details). STPT and SpeechUT jointly pre-train the model on speech and text data from scratch, which achieve better performance but also bring higher training costs[8].

5.3 Main Results on MuST-C Dataset

Table 1 shows the results on MuST-C `tst-COMMON` set in all eight directions. First, we find that our implemented **MTL** is a strong baseline compared with existing approaches. Second, our proposed **CRESS** significantly outperforms **MTL** in both settings, with 1.8 BLEU improvement in the base set-

[8]For example, the pre-training of SpeechUT takes 3 days with 32 V100 GPUs.

#	Adaptive Training	Regularization	Scheduled Sampling	BLEU
1	✓	✓	✓	**29.4**
2	×	✓	✓	29.0
3	×	✓	×	28.4
4	×	×	✓	28.0
5	✓	×	×	27.5
6	×	×	×	27.7

Table 2: BLEU scores on MuST-C En→De `tst-COMMON` set with different combinations of training techniques.

ting and 1.3 BLEU improvement in the expanded setting on average, demonstrating the superiority of our approach. Besides, we report ChrF++ scores on MuST-C in Appendix E, and we also provide results on CoVoST 2 (Wang et al., 2020a) En→De dataset in Appendix C.

6 Analysis and Discussion

Results in Section 5.3 show the superiority of our method. To better understand **CRESS**, we explore several questions in this section. All analysis experiments are conducted on MuST-C En→De dataset in the expanded setting.

(1) Do scheduled sampling, cross-modal regularization, and token-level adaptive training all matter? Scheduled sampling, regularization, and token-level adaptive training are effective techniques to improve the performance of translation models. To understand the role of each, we conduct ablation experiments in Table 2. When only applying token-level adaptive training (#5), we observe a performance decline of 0.2 BLEU since only adaptive training can not bridge the modality gap. When training with scheduled sampling only (#4), we observe a slight improvement of 0.3 BLEU, probably due to the alleviation of exposure bias. When training with cross-modal regularization only (#3), which encourages the consistency between predictions of ST and MT with ground truth target contexts, we observe an improvement of 0.7 BLEU. If we combine both (#2), we obtain a much more significant boost of 1.3 BLEU, proving that both scheduled sampling and cross-modal regularization play a crucial role in our method. Furthermore, with token-level adaptive training (#1), the improvement comes to 1.7 BLEU, which shows the benefit of treating different tokens differently according to the modality gap.

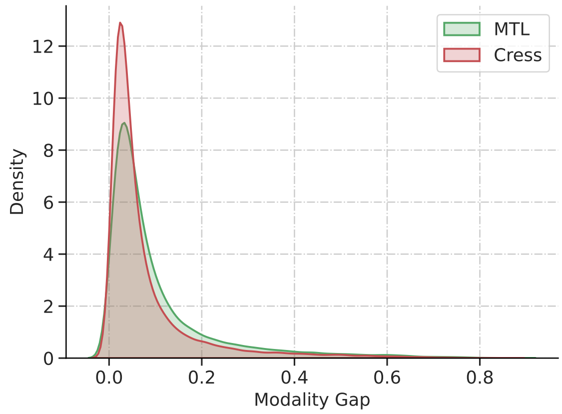

Figure 4: Distributions of the modality gap on MuST-C En→De dev set of **MTL** and **CRESS** with kernel density estimation (KDE).

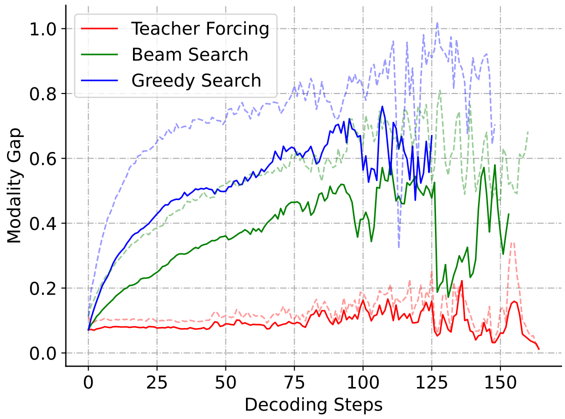

Figure 5: Curves of the average modality gap with decoding steps under three strategies. The dotted line refers to **MTL**, and the solid line refers to **CRESS**.

(2) Does CRESS successfully bridge the modality gap? To validate whether our approach successfully bridges the modality gap between ST and MT, we revisit the experiments in Section 3. Figure 4 shows the distribution of the modality gap with teacher forcing. We observe a general decrease in the modality gap compared with **MTL**. We also plot the curves of the modality gap with decoding steps of **CRESS** under teacher forcing, greedy search, and beam search strategies. As shown in Figure 5, our approach significantly slows down the increase of the modality gap compared with **MTL** baseline, suggesting that the predictions of ST and MT are more consistent during inference, demonstrating the effectiveness of our method in bridging the modality gap.

(3) How base and scale hyper-parameters influence token-level adaptive training? B (base) and S (scale) are two important hyper-parameters

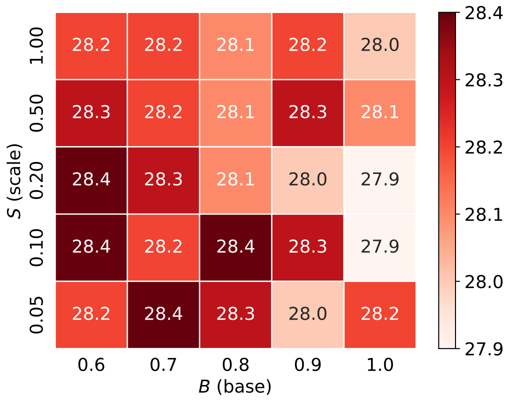

Figure 6: The heat map of BLEU scores on MuST-C En→De dev set with different combinations of B and S. The BLEU score without token-level adaptive training is 28.0.

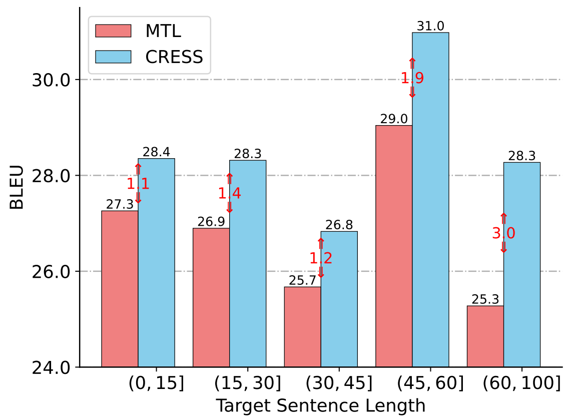

Figure 7: BLEU scores on MuST-C En→De dev set at different target sentence lengths.

in token-level adaptive training. We investigate how different combinations of B and S influence performance. As shown in Figure 6, token-level adaptive training can bring improvements in most cases. In particular, it usually performs better with smaller B and smaller S, leading to a boost of up to 0.4 BLEU. We conclude that treating different tokens too differently is also undesirable. We use $B = 0.7$ and $S = 0.05$ for all translation directions.

(4) Does CRESS successfully reduce the performance gap between ST and MT? As shown in Table 3, our method not only brings improvements to ST, but also gives a slight average boost of 0.3 BLEU to MT. We suggest that this may be due to the effect of regularization. More importantly, we find that the performance gap between ST and MT for **CRESS** is significantly reduced compared to the **MTL** baseline (6.0→5.0), which further demonstrates that the improvement in ST is mainly due to the effective reduction of the modality gap.

(5) Is CRESS more effective for longer sentences? The autoregressive model generates the translation step by step, making the translation of long sentences more challenging. We divide the MuST-C En→De dev set into several groups according to the length of target sentences, and compute the BLEU scores in each group separately, as shown in Figure 7. We observe that **CRESS** achieve significant improvements over the baseline in all groups, especially for sentences longer than 45, which shows the superiority of our method when translating long sentences.

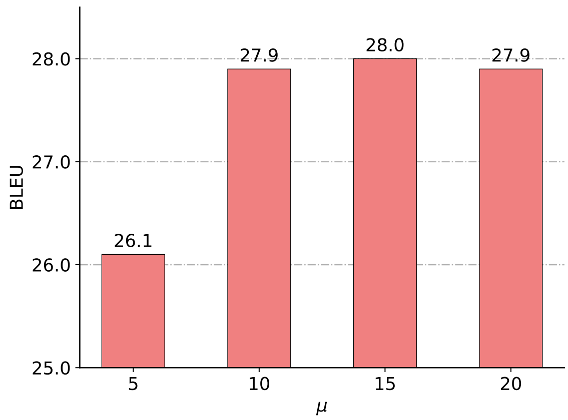

Figure 8: BLEU scores on MuST-C En→De dev set (expanded setting) with different μ. Here token-level adaptive training is not used for training.

(6) How the decay parameter in scheduled sampling influence the performance? In scheduled sampling, the probability of selecting the ground truth word p^* keeps decreasing during training as the function in Equation (10). Here, the hyper-parameter μ is used to control the shape of the function. As μ increases, the probability p^* decreases more slowly, and vice versa. We investigate the impact of μ in Figure 8, and find that (1) the model performs worse when p^* drops too quickly, and (2) when μ is within a reasonable range, there is not much impact on the final BLEU score. We use $\mu = 15$ in our experiments.

7 Related Work

End-to-end Speech Translation End-to-end speech translation (Bérard et al., 2016; Weiss et al., 2017) has shown great potential for overcoming error propagation and reducing latency compared to traditional cascaded ST systems (Salesky et al.,

Models	Task	BLEU									
		En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.↑	Δ ↓
MTL	ST	27.7	38.5	32.8	24.9	19.0	26.5	32.0	30.8	29.0	6.0
	MT	33.5	**46.6**	**38.3**	30.9	22.1	33.0	38.6	36.7	35.0	
CRESS	ST	**29.4**	**40.1**	**33.2**	**26.4**	**19.7**	**27.6**	**33.6**	**32.3**	**30.3**	**5.0**
	MT	**34.1**	**46.6**	38.1	**31.1**	**22.4**	**33.3**	**39.5**	**37.6**	**35.3**	

Table 3: BLEU scores of both ST and MT on MuST-C `tst-COMMON` set (expanded setting). Δ indicates the average gap in BLEU between ST and MT.

2019; Di Gangi et al., 2019c,b; Bahar et al., 2019a). One challenge in training end-to-end ST models is the scarcity of ST data. To address this problem, researchers employed MT data to help training with techniques like pre-training (Bansal et al., 2019; Stoian et al., 2020; Wang et al., 2020b,c; Alinejad and Sarkar, 2020; Le et al., 2021; Dong et al., 2021a; Zheng et al., 2021; Xu et al., 2021a; Tang et al., 2022), multi-task learning (Le et al., 2020; Dong et al., 2021b; Ye et al., 2021; Tang et al., 2021a,b; Indurthi et al., 2021), knowledge distillation (Liu et al., 2019; Inaguma et al., 2021), and data augmentation (Jia et al., 2019; Bahar et al., 2019b; Lam et al., 2022; Fang and Feng, 2023). However, due to the *modality gap* between speech and text, it is still difficult to fully exploit MT data with the above techniques. To overcome the modality gap, Han et al. (2021) projects features of both speech and text into a shared semantic space. Fang et al. (2022); Zhou et al. (2023) mixes up features of speech and text to learn similar representations for them. Ye et al. (2022) brings sentence-level representations closer with contrastive learning. Bapna et al. (2021, 2022); Chen et al. (2022); Tang et al. (2022); Zhang et al. (2022b) jointly train on speech and text and design methods to align two modalities. Different from previous work, in this work, we understand the modality gap from the target-side representation differences, and show its connection to exposure bias. Based on this, we propose the **C**ross-modal **Re**gularization with **S**cheduled **S**ampling (**CRESS**) method to bridge the modality gap.

Exposure Bias Exposure bias indicates the discrepancy between training and inference. Several approaches employ Reinforcement Learning (RL) (Ranzato et al., 2016; Shen et al., 2016; Bahdanau et al., 2017) instead of Maximum Likelihood Estimation (MLE) to avoid this problem. However, Wu et al. (2018) shows that RL-based training is unstable due to the high variance of gradient estimation. An alternative and simpler approach is scheduled sampling (Bengio et al., 2015), which samples between ground truth words and self-generated words with a changing probability. Zhang et al. (2019) extends it with Gumbel noise for more robust training. In this paper, we adopt this approach to approximate the inference mode due to its training stability and low training cost.

Output Regularization for MT Regularization in the output space has proved useful for MT. Liang et al. (2021) proposes to regularize the output predictions of two sub-models sampled by dropout. Guo et al. (2022) regularizes the output predictions of models before and after input perturbation. In this paper, we regularize the output predictions across modalities, which encourages more consistent predictions for ST and MT.

Token-level Adaptive Training Token-level adaptive training for MT is first proposed in Gu et al. (2020), which assigns larger weights to low-frequency words to prevent them from being ignored. Xu et al. (2021b); Zhang et al. (2022a) computes the weight with bilingual mutual information. In this paper, we compute the weights with the modality gap between ST and MT.

8 Conclusion

In this paper, we propose a simple yet effective method CRESS to regularize the model predictions of ST and MT, whose target-side contexts contain both ground truth words and self-generated words with scheduled sampling. Based on this, we further propose a token-level adaptive training method to handle difficult cases. Our method achieves promising results on MuST-C benchmark. Further analysis shows that our method can effectively bridge the modality gap and improve the translation quality, especially for long sentences. In the future, we will explore how to apply our method to other tasks.

Limitations

Although our proposed method achieves promising results and outperforms most baseline systems on the ST benchmark, it still has some limitations: (1) the performance of our method still lags behind a recent work SpeechUT, although our approach has the advantage of consuming less time and resources; (2) we observe that the modality gap is still not eliminated and the effect of exposure bias on the modality gap still exists; (3) the performance of our approach on larger datasets and larger models remains to be explored; (4) how to apply our approach to other tasks also needs to be further investigated.

Ethics Statement

In this paper, we present an effective method **CRESS** for speech translation. While our model achieves superior performance on the widely used ST benchmark MuST-C, applying it directly to real scenarios is still risky. This is due to the fact that our training corpus only contains hundreds of hours of audio recordings from TED talks, which is far from covering all domains of the real world. Besides, the datasets we used in this paper (MuST-C, WMT, and OPUS-100) are all publicly available. We also release the code implemented with a well-known framework fairseq. These guarantee the reproducibility of our work.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments. This work was supported by National Key R&D Program of China (NO. 2018AAA0102502)

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). 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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019a. A comparative study on end-to-end speech to text translation. In *Proc. of ASRU*, pages 792–799. IEEE.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019b. On using specaugment for end-to-end speech translation. In *Proc. of IWSLT*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proc. of NAACL-HLT*, pages 58–68.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *CoRR*, abs/2202.01374.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS workshop on End-to-end Learning for Speech and Audio Processing*.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. In *INTERSPEECH*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. 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)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A. Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 21–31.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting transformer to end-to-end spoken language translation. In *Proc. of INTERSPEECH*, pages 1133–1137. International Speech Communication Association (ISCA).

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *The Thirty-fifth AAAI Conference on Artificial Intelligence, AAAI*.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021b. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qingkai Fang and Yang Feng. 2023. Back translation for speech-to-text translation without transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014*, pages 16–23. ISCA.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online. Association for Computational Linguistics.

Emil Julius Gumbel. 1954. Statistical theory of extreme values and some practical applications: a series of lectures. In *Nat. Bur. Standards Appl. Math. Ser.*, volume 33. US Government Printing Office.

Dengji Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2022. Prediction difference regularization against perturbation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of NAACL*, pages 1872–1881.

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2021. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7723–7727.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proc. of ICASSP*, pages 7180–7184.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. 2020. Libri-light: A benchmark for ASR with limited or no supervision. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7669–7673. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 245–254, Dublin, Ireland. Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech 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 2: Short Papers)*, pages 817–824, Online. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. 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 827–838, Online. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pages 1128–1132.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. 2014. A$*$ sampling. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2757–2761. ISCA.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In *Proc. of NAACL-HLT*, pages 2786–2792.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Mihaela C. Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021a. Improving speech translation by understanding and learning from the auxiliary text translation task. 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 4252–4261, Online. Association for Computational Linguistics.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213.

Jörgen Valk and Tanel Alumäe. 2021. VOXLINGUA107: A dataset for spoken language recognition. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 652–658. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021b. Large-scale self- and semi-supervised learning for speech translation. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2242–2246. ISCA.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020a. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proc. of AAAI*, volume 34, pages 9161–9168.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020c. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Proc. of INTERSPEECH*, pages 2625–2629.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021a. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. 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 2619–2630, Online. Association for Computational Linguistics.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021b. Bilingual mutual information based adaptive training for neural machine 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 2: Short Papers)*, pages 511–516, Online. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022a. Conditional bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2377–2389, Dublin, Ireland. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proc. of ICML*.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

A Statistics of all datasets

En→	ST (MuST-C) hours	#sents	External MT name	#sents
De	408	234K	WMT16	3.9M
Fr	492	280K	WMT14	31.2M
Es	504	270K	WMT13	14.2M
Ro	432	240K	WMT16	0.6M
Ru	489	270K	WMT16	1.9M
It	465	258K	OPUS100	0.7M
Pt	385	211K	OPUS100	0.7M
Nl	442	253K	OPUS100	0.7M

Table 4: Statistics of all datasets. #sents refers to the number of parallel sentence pairs.

B Impact of Different Acoustic Encoders

Our model is composed of an acoustic encoder and a translation model. To investigate the impact of different acoustic encoders, we also conduct experiments using Wav2vec 2.0[9] (Baevski et al., 2020) as the acoustic encoder. As shown in Table 5, we find that (1) HuBERT performs slightly better than Wav2vec 2.0 with an improvement of 0.5 BLEU, and (2) our proposed **CRESS** achieves consistent improvements with different acoustic encoders. In practice, we use HuBERT to build our systems, since we believe that developing on a strong baseline will make our results more convincing and demonstrate the robustness of our approach.

Acoustic Encoder	MTL	CRESS
HuBERT (Hsu et al., 2021)	27.5	**29.4**
Wav2vec 2.0 (Baevski et al., 2020)	27.0	**28.9**

Table 5: BLEU scores on MuST-C En→De `tst-COMMON` set (expanded setting) with different acoustic encoders.

C Results on CoVoST 2 En→De

We also conduct experiments on CoVoST 2 (Wang et al., 2020a) to examine the performance of our approach on large datasets. CoVoST 2 is a large-scale multilingual speech translation corpus that covers translations from 21 languages into English and from English into 15 languages. It is one of the largest open ST datasets available currently. In this paper, we evaluate our approach on the En→De direction, which contains 430 hours of speech with annotated transcriptions and translations. We use `dev` set for validation and `test` set for evaluation.

We use the same pre-processing, model configuration, and hyper-parameters as MuST-C (see details in Section 5.2). The results are shown in Table 6. First, we find our **CRESS** significantly outperforms the **MTL** baseline, with 1.8 BLEU improvement in the base setting and 1.6 BLEU improvement in the expanded setting, which demonstrates the effectiveness and generalization capability of our method across different datasets, especially on the large-scale dataset. Second, our result is competitive with previous methods, although they use larger audio datasets (≥60K hours) and larger model size (≥300M), while we only use 960 hours of audio data and 155M model parameters.

D Discussion about the Training Speed

During training, our approach requires an additional forward pass to select predicted words compared with the baseline, which will impair the training speed. Practically, we find the training time for 1 epoch of **CRESS** is 1.12 times longer than **MTL**, which is actually negligible. This is because the step of selecting predicted words is fully parallel and has no gradient calculation, which does not incur a significant time overhead.

E ChrF++ Scores on MuST-C Dataset

We also report ChrF++ score (Popović, 2017) using sacreBLEU toolkit[10] on MuST-C dataset in Table 7. We observe that **CRESS** outperforms **MTL** with 1.4 ChrF++ improvement in the base setting and 1.0 ChrF++ improvement in the expanded setting.

[9] `https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt`

[10] sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.0.0

Models	**Audio Datasets**	**#Params**	**BLEU**
wav2vec-2.0 (LS-960) (Wang et al., 2021b)	LS-960	300M	20.5
wav2vec-2.0 (LV-60K) (Wang et al., 2021b)	LV-60K	300M	25.5
wav2vec-2.0 + Self-training (LV-60K) (Wang et al., 2021b)	LV-60K	300M	**27.1**
LNA (Joint Training) (Li et al., 2021)	LV-60K	1.05B	25.8
SLAM-TLM (Bapna et al., 2021)	LV-60K	600M	**27.5**
XLS-R (0.3B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	317M	23.6
XLS-R (1B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	965M	26.2
XLS-R (2B) (Babu et al., 2022)	VP-400K, MLS, CV, VL, BBL	2162M	**28.3**
MTL (base setting)	LS-960	**155M**	21.4
CRESS (base setting)	LS-960	**155M**	23.2 (+1.8)
MTL (expanded setting)	LS-960	**155M**	25.1
CRESS (expanded setting)	LS-960	**155M**	**26.7** (+1.6)

Table 6: BLEU scores on CoVoST 2 En→De `test` set. LS-960: LibriSpeech (Panayotov et al., 2015) (960 hours). LV-60K: Libri-Light (Kahn et al., 2020) (60K hours). VP-400K: VoxPopuli (Wang et al., 2021a) (372K hours). MLS: Multilingual LibriSpeech (Pratap et al., 2020) (50K hours). CV: CommonVoice (Ardila et al., 2020) (7K hours). VL: VoxLingua107 (Valk and Alumäe, 2021) (6.6K hours). BBL: BABEL (Gales et al., 2014) (1K hours).

Models	**ChrF++**								
	En→De	En→Fr	En→Es	En→Ro	En→Ru	En→It	En→Pt	En→Nl	Avg.
Base setting (w/o external MT data)									
MTL	52.4	60.4	56.4	50.9	41.7	52.6	57.3	56.1	53.5
CRESS	**54.0****	**62.0****	**57.6****	**52.4****	**43.1****	**53.8****	**58.5****	**57.6****	**54.9**
Expanded setting (w/ external MT data)									
MTL	54.9	62.6	58.6	51.9	44.2	53.4	57.9	56.9	55.0
CRESS	**56.1****	**63.7****	**58.9***	**53.1****	**44.5***	**54.2****	**59.3****	**58.3****	**56.0**

Table 7: ChrF++ scores on MuST-C `tst-COMMON` set. The external MT datasets are only used in the expanded setting. * and ** mean the improvements over **MTL** baseline are statistically significant ($p < 0.05$ and $p < 0.01$, respectively).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5

☑ B1. Did you cite the creators of artifacts you used?
Section 5

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We just use for research purposes, no commercial use and no derivative works of the original data.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Our use of dataset and pre-trained models is consistent with their intended use, which does not require much discussion.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data we used are publicly available on the website, and widely used in the research community. We cannot change the training/test data in order to be consistent with previous work.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☑ Did you run computational experiments?

Section 5, Section 6, Appendix

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5, Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5, Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5, Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.