CRAF:Cross-Modal Representation Alignment and Fusion for Speech Translation

Zhenbei Guo[1,2],Wenzhou Wu[1,2],Hua Lai[1,2],Yan Xiang[1,2], Yuxin Huang[1,2], Zhengtao Yu[1,2*]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, Yunnan, 650500, China
2.Yunnan Key Laboratory of Artifcial Intelligence, Kunming, Yunnan, 650500, China
zbguo@kust.edu.cn, 3523919342@qq.com
12309028@kust.edu.cn,sharonxiang@126.com
huangyuxin2004@163.com,ztyu@hotmail.com

Abstract

The end-to-end speech translation task involves directly transforming speech into the text of another language, bypassing the generation of an intermediate transcription. However, existing methods may lose key information during cross-modal length alignment and fail to effectively integrate different representations, resulting in low quality of the fused representation. To address these issues, we propose an efficient method named CRAF for effective cross-modal alignment and fusion for speech translation, which reduces information loss and enhances the integration of cross-modal representations. First, CRAF minimizes information loss by improving the cross-modal length alignment, ensuring the alignment process retains more critical information from the speech modality. Second, CRAF strengthens the integration of cross-modal representations by allowing the model to combine complementary features from diverse modalities, enhancing its capacity to concentrate on the most pertinent and critical information. Finally, we evaluate CRAF by conducting extensive experiments on eight language pairs from the MuST-C dataset. Experiments show that the average BLEU score of CRAF achieves 29.0, outperforming other comparison methods. Our code is available at https://github.com/wu-wen-zhou/first/tree/master.

Keywords: Speech Translation , Cross-modal Task , Length Alignment , Representation Fusion.

1 Introduction

The end-to-end speech translation task (Bérard et al., 2016; Duong et al., 2016) directly translates speech from the source language into text in the target language without generating intermediate text. This task encompasses a broad spectrum of application scenarios, including international conferences, real-time translation, and video subtitling. As a task involving multiple modalities, numerous studies (Ye et al., 2021; Ye et al., 2022; Xu et al., 2023a; Han et al., 2021; Fang et al., 2022) focus on the alignment and fusion of speech and text modalities with the aim of enhancing model performance.

However, existing methods may lose critical information during the alignment of speech and text representations across different lengths. Furthermore, after fusing the speech and text representations, these methods fail to guide the model to focus on the more important parts for semantic understanding, thus limiting the model's performance. 1) Critical information loss: Existing methods (Fang et al., 2022; Han et al., 2021; Gangi et al., 2019; Zhang et al., 2024; Ye et al., 2022; Zhou et al., 2023; Ye et al., 2021) typically compress the length of speech representations during alignment. As a result, key information, such as emotional information in speech representations, is often lost during the compression process, which may lead to an incorrect understanding of the speaker's semantics and affect the quality of the speech representation. For example, SpeechFormer (Papi et al., 2021) effectively mitigates such information loss by introducing improvements in the memory management of the attention mechanism. By

* Corresponding Author
©2025 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031-1042, Jinan, China, August 11-14, 2025.

optimizing how attention computations are handled, SpeechFormer is able to retain more critical information during the process of compressing and aligning speech representations. In addition, the work by (Xu et al., 2023b) provides further insights into the root causes of information loss, attributing it to the use of overly coarse-grained representations in the alignment process. To address this, they propose a progressive downsampling solution that gradually reduces the sequence length while preserving essential details, thereby offering a more refined approach to maintaining information integrity during speech-text alignment; 2) Lack of key information attention mechanism: During the fusion of speech and text representations, existing methods (Fang et al., 2022; Zhou et al., 2023; Xu et al., 2023a) do not guide the model to focus on more important information within the representations, resulting in the model's incomplete understanding of the fused representations and preventing the model from achieving its full performance. For example,VATT (Akbari et al., 2021) demonstrates that cross-modal self-attention enables Transformers to automatically achieve semantic alignment across modalities and focus on semantically relevant key information,even in the absence of explicit supervision. Thus, a practical and interesting research question is: *Can we propose a method that minimizes information loss during speech representation compression while enabling the model to better focus on the important features in the fused representations, thereby improving model performance?*

In this paper, we propose a method named CRAF. CRAF incorporates a dynamic weighted length adapter module that minimizes information loss during the alignment of speech and text representations. It also incorporates a representation fusion module that enables the model to focus on more important features in the fused representations. By integrating these two modules, our model not only reduces information loss but also enhances the ability to focus on key features, thereby achieving improved performance.

Nevertheless, implementing the CRAF method faces two main challenges. First, during the compression of speech representations, it is highly challenging to preserve as much critical information as possible during the reduction of feature length. Traditional methods, such as employing two CNN layers, can compress speech representation to the desired length yet fail to differentiate the importance of information, resulting in the loss of key details. Second, guiding the model to focus on more important information after fusing the speech and text representations also presents a challenge. Since speech and text belong to different modalities, their representations exist in different distributions and semantic spaces. Although methods like concatenation or weighted fusion can improve performance, it is challenging for the model to fully comprehend the semantic information, limiting performance improvements.

To address the above challenges, we first combine the output representations from different CNN layers using an attention mechanism for fusion. Then, we apply dynamic weighted summation between the output of the first CNN layer and the fused representation to obtain the final compressed representation, which helps reduce information loss. Additionally, after applying Mixup to the speech and text representations, we use an attention mechanism to enable the model to focus on the more important parts of the Mixup representation for semantic understanding.

Our contributions are as follows:

1) We employ a dynamic weighted length adapter, using an attention mechanism and weighted summation strategy to fuse speech representations from different convolutional layers, effectively reducing information loss during the compression of speech representations.

2) We utilize a representation fusion module, where speech and text are randomly sampled and then fused according to the optimal transport matrix. An attention mechanism is then applied to enhance the fusion of speech and text representations, enabling the model to deeply understand the underlying semantic information of the fused representations.

3) We implement a cross-modal representation alignment and fusion model called CRAF and conducted experiments on eight language pairs from the MuST-C dataset. The experimental results show that the model achieves an average BLEU score of 29.0, which is 0.3 points higher than the baseline methods, with the highest improvement reaching 0.5 points.

The remainder of this paper is organized as follows. Section 2 reviews related work on speech-to-text translation. Section 3 provides an overview of the design. Section 4 discusses the design details. Section

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031-1042, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5 presents the experiments. Section 6 presents the numerical results and analysis. Finally, Section 7 concludes the paper.

2 Related Work

End-to-end ST The traditional cascaded approach (Waibel et al., 1991) links speech recognition models with machine translation models sequentially. However, this method faces challenges such as error propagation and high latency. To overcome these limitations, end-to-end speech translation models (Bérard et al., 2016; Duong et al., 2016) have been introduced, enabling the direct conversion of source language speech into target language text without intermediate transcription. This paradigm has become the dominant framework for speech translation (Vila et al., 2018; Salesky et al., 2019; Gangi et al., 2019; Fang et al., 2022; Ye et al., 2022; Zhou et al., 2023; Wang et al., 2020). Nevertheless, end-to-end speech tra‘nslation, as a cross-modal task, faces the challenge of bridging the modality gap between speech and text. Some works have explored this issue. (Ye et al., 2022) uses contrastive learning to align the semantic representations of speech and text; (Fang et al., 2022) employs Mixup to fuse speech and text representations before feeding them into the encoder; (Zhou et al., 2023) first uses optimal transport (OT) to align speech and text representations, and then fuses the encoder outputs of both modalities. However, these methods overlook the information loss that occurs when compressing speech representations, as well as how to make the model focus on the more important information in the representations after fusion. In our work, we explore how to reduce information loss when compressing speech representations and how to make the model focus on more important information in the speech representation after fusion.

Optimal Transport Optimal Transport (OT) is a mathematical framework for calculating the minimal cost of transforming one probability distribution into another, with its theoretical foundations systematically detailed in (Villani, 2008). In recent years, OT has been widely applied to cross-lingual and cross-modal tasks. For example, (Zhou et al., 2023) leverage OT to compute a transport matrix and use Mixup to align speech and text representations, achieving notable performance gains. Similarly, (Chen et al., 2020) and (Tsiamas et al., 2024) utilize OT for fine-grained alignment between textual, visual, and speech representations.

3 Design of Overview

This section describes our model architecture, as shown in Figure 1. The model consists of a speech encoder, a text embedding layer, a representation fusion module, a translation encoder, and a decoder.

Speech Encoder The main function of the speech encoder is to extract semantic information from the original audio and align the lengths of the speech and text representations. It consists of a pre-trained HuBERT[1] (Hsu et al., 2021) model and a Dynamic Weighted Length-Adapter.

Dynamic Weighted Length-Adapter The primary function of this module is to compress speech representations while preserving as much critical information as possible, thereby enhancing model performance. It consists of two convolutional layers, a cross-attention mechanism, a LayerNorm layer, and a weighted sum module.

Representation Fusion Module This module first aligns the semantic spaces of speech and text representations using Optimal Transport (OT). Then, it applies a Mixup strategy to randomly sample speech and text representations with a probability P, generating a mixed representation. Subsequently, a self-attention mechanism is employed to further enhance the expressive power of the mixed representation. The module primarily consists of Mixup and a self-attention mechanism.

Text Embedding The text embedding module operates in parallel with the Speech Encoder and is designed to convert text into a sequence of embeddings.

Translation Encoder The translation encoder enhances semantic understanding by processing outputs from both the Speech Encoder and Text Embedding module. It consists of Transformer Encoder layers (Vaswani et al., 2017).

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/hubert

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031-1042, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

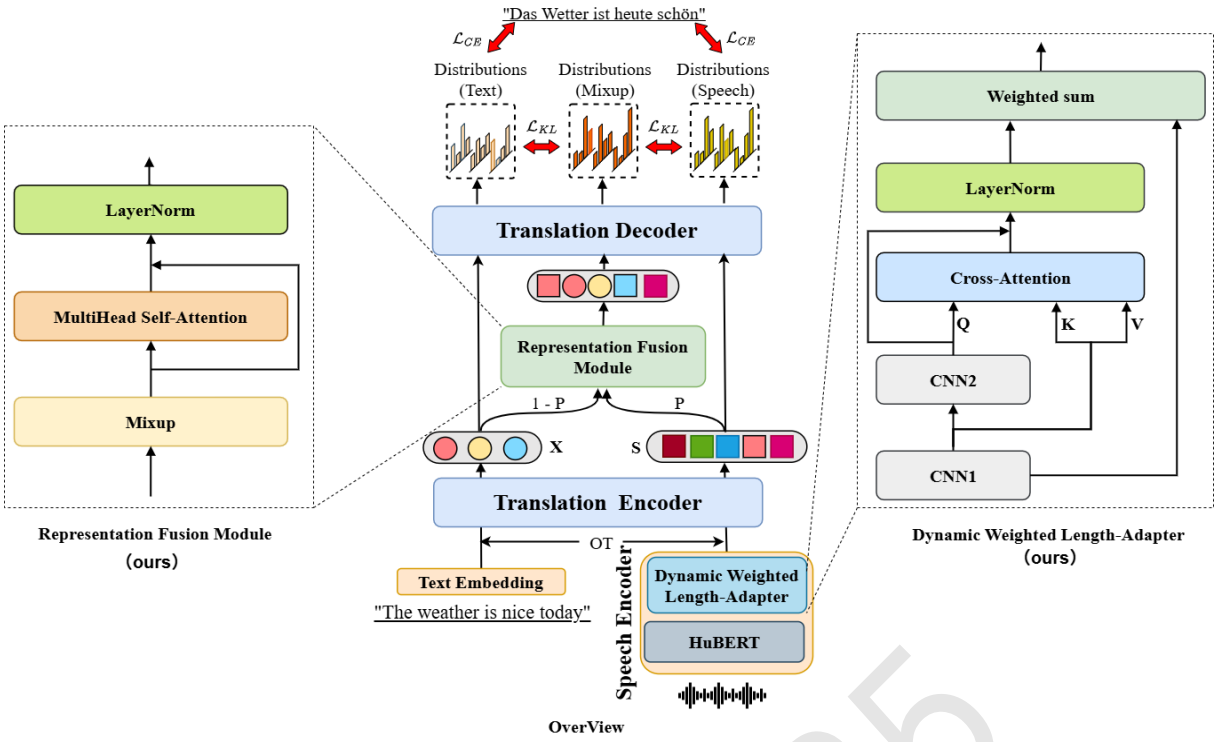

Figure 1: Overview of the proposed framework with CRAF.

Translation Decoder The translation decoder processes the output from both the translation encoder and the Representation Fusion module to deliver the final translation. It incorporates transformer decoder layers (Vaswani et al., 2017).

4 Design Details

In this section, we first introduce the problem definition of speech translation and present **C**ross-Modal **R**epresentation **A**lignment and **F**usion (**CRAF**). Then, we will introduce our training strategy. By using the Dynamic Weighted Length-Adapter and Representation Fusion Module, we can effectively align and fuse cross-modal representations, significantly improving the performance and robustness of speech translation.

4.1 Problem Formulation

A speech-to-text translation dataset is typically represented as a triplet $\mathcal{D} = \{(\mathbf{S}, \mathbf{X}, \mathbf{Y})\}$, where $\mathbf{S}$ is the input audio, $\mathbf{X}$ the source language transcription, and $\mathbf{Y}$ the target language translation. The goal of end-to-end speech translation is to directly convert $\mathbf{S}$ into $\mathbf{Y}$, bypassing the intermediate transcription $\mathbf{X}$.

4.2 Cross-Modal Representation Alignment and Fusion (CRAF)

Dynamic Weighted Length-Adapter The dynamic weighted length adapter fuses hierarchical features from different convolutional layers using cross-attention and residual connections. Specifically, the second convolutional layer's output serves as the query, and the first layer's output as the key and value, enabling inter-layer dependency modeling. A weighted summation module then adaptively combines low- and high-level features, generating output features for the Translation Encoder. Let $\mathbf{F}_1$ and $\mathbf{F}_2$ denote the outputs of CNN1 and CNN2, respectively. The process is as follows (Bahdanau et al., 2014):

$$\mathbf{Q} = \mathbf{W_q F_2}, \quad \mathbf{K} = \mathbf{W_k F_1}, \quad \mathbf{V} = \mathbf{W_v F_1}.$$

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031–1042, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Using these, the cross-attention output is computed as:

$$\mathbf{F_{cross}} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}. \tag{1}$$

We obtain the weight μ through the dynamic weighted module:

$$\mu = \text{softmax}\left(\text{MLP}(\mathbf{LayerNorm}(\mathbf{F_{cross}} + \mathbf{F_2}))\right). \tag{2}$$

Finally, a weighted sum is calculated to produce the fused output:

$$\mathbf{F_{out}} = \mu * (F_{cross} + F_2) + (1 - \mu) * F_1. \tag{3}$$

Representation Fusion Module The representation fusion module first aligns speech and text features via OT alignment, then applies random Mixup sampling with probability P to the outputs of the Translation Encoder. A self-attention mechanism further refines the fused sequence, which is then fed into the Translation Decoder. Here, $\mathbf{H^s}, \mathbf{H^x} \in \mathbb{R}^{T\times d}$ are the speech and text feature sequences, and the OT alignment matrix $\mathbf{A} \in \mathbb{R}^{T\times T}$ is computed by minimizing the alignment cost (Peyré and Cuturi, 2019).

$$\mathbf{A} = \arg\min_{\mathbf{A}\in\mathcal{U}(\mathbf{r},\mathbf{c})} \sum_{i=1}^{T}\sum_{j=1}^{T} a_{ij} \cdot c(h_i^s, h_j^x). \tag{4}$$

where a_{ij} denotes the alignment weight, $\mathcal{U}(\mathbf{r}, \mathbf{c})$ represents the set of valid distributions that satisfy the marginal constraints, and $c(h_i^s, h_j^x)$ is a cost function.

Using the alignment matrix $\mathbf{A}$, Mixup is applied to generate a fused representation $\mathbf{M} = (m_1, \ldots, m_T)$ as follows:

$$m_i = \begin{cases} h_i^s, & \text{if } p < P, \\ h_{a_i}^x, & \text{if } p \geq P. \end{cases} \tag{5}$$

where p is a probability parameter that controls the blending ratio between speech and text features.

The fused representation $\mathbf{M}$ is subsequently refined through a multi-head self-attention mechanism:

$$\mathbf{M}' = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{6}$$

where $\mathbf{Q} = \mathbf{W}_q\mathbf{M}$, $\mathbf{K} = \mathbf{W}_k\mathbf{M}$, and $\mathbf{V} = \mathbf{W}_v\mathbf{M}$ are the query, key, and value projections, respectively.

This process ensures the effective integration of speech and text features through representation alignment, while the self-attention mechanism enhances contextual dependencies by assigning greater weight to more important semantic information, enabling the model to focus on key elements and improving their effectiveness for downstream tasks.

4.3 Training Strategy

Our training follows a pretraining-finetuning paradigm.

Pre-training: We first pretrain the model using text pairs $(\mathbf{x}, \mathbf{y})$ from the speech translation corpus as a machine translation task:

$$\mathcal{L}_{MT} = -\mathbb{E}_{(x,y)} \log P(\mathbf{y} \mid \mathbf{x}). \tag{7}$$

Multitask Finetuning: We then perform multitask finetuning with triplets $(\mathbf{s}, \mathbf{x}, \mathbf{y})$, jointly optimizing speech translation and text translation:

$$\mathcal{L}_{CE} = -\mathbb{E}_{(s,x,y)}[\log P(\mathbf{y} \mid \mathbf{s}) + \log P(\mathbf{y} \mid \mathbf{x})]. \tag{8}$$

CCL 2025

To encourage distribution alignment among modalities, we minimize the KL divergence between the fused representation ($\mathbf{M}''$), text ($\mathbf{T}$), and speech ($\mathbf{S}$) distributions:

$$\mathcal{L}_{KL} = \frac{1}{2}\big(\mathrm{KL}(\mathbf{M}'' \parallel \mathbf{S}) + \mathrm{KL}(\mathbf{S} \parallel \mathbf{M}'') + \mathrm{KL}(\mathbf{M}'' \parallel \mathbf{T}) + \mathrm{KL}(\mathbf{T} \parallel \mathbf{M}'')\big). \tag{9}$$

The final training objective is:

$$\mathcal{L}_{ST} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{KL}, \tag{10}$$

where α is the KL divergence loss weight.

5 Experiments

In this section, we present the dataset used in our experiments, the experimental setup during training, the inference method, and the baseline methods.

5.1 Dataset

We conducted experiments exclusively on the MuST-C[2] (Di Gangi et al., 2019) dataset, without using additional datasets. The MuST-C dataset, a widely used benchmark for speech translation, is primarily derived from TED Talks. Detailed statistics for the MuST-C dataset are provided in Table 1.

Table 1: Detailed data statistics of the MuST-C dataset.

language	En-De	En-Es	En-Fr	En-Nl	En-Pt	En-Ro	En-Ru	En-It
hours	408	504	492	442	385	432	489	465
sents	234K	270K	280K	253K	211K	240K	270K	258K

5.2 Experimental Setups

Model Configurations Our model, based on CMOT (Zhou et al., 2023) and implemented with Fairseq[3] (Wang et al., 2020), consists of five key components: a speech encoder, a text embedding module, a translation encoder, a translation decoder, and a representation fusion module. The speech encoder uses HuBERT, a pre-trained acoustic model, along with a dynamic weighted length-adapter module. HuBERT serves as a feature extractor for speech representations and requires no additional fine-tuning for downstream tasks.

The dynamic weighted length-adapter module, placed after HuBERT, includes two 1D convolutional layers, a multi-head cross-attention mechanism, a residual connection, and a weighted sum module. The convolutional layers have a kernel size of 5, stride 2, and padding 2, with a hidden layer dimension of 1024. The multi-head cross-attention mechanism has a 512-dimensional input and output, with 4 attention heads.

Both the Translation Encoder and Translation Decoder use a Transformer architecture, each consisting of 6 layers, with a hidden dimension of 512, 8 attention heads, and a feedforward layer of size 2048.

The representation fusion module combines Mixup (sampling rate 0.8) and a multi-head self-attention mechanism with 8 attention heads, processing inputs and outputs of dimensionality 512.

Pre-processing We employ 16-bit, 16 kHz single-channel audio as the speech input for our model. For text preprocessing, we use unigram-based SentencePiece[4] (Kudo and Richardson, 2018) for tokenization and construct a shared vocabulary comprising 10,000 tokens, applicable to both the source and target languages.

Training We employ a pretraining-finetuning framework. Initially, the text embedding module, translation encoder, and decoder are pre-trained on machine translation (MT) tasks,using transcription-target

[2]https://ict.fbk.eu/must-c/
[3] https://github.com/facebookresearch/fairseq/
[4] https://github.com/google/sentencepiece/

Table 2: The BLEU scores on the MuST-C tst-COMMON dataset. "Speech Pre-training" indicates whether a pre-trained speech model was used, and the bold font indicates an improvement in the BLEU score for the corresponding language.

Models	**Speech**	**BLEU**								**Avg.**
	Pretraining	En-De	En-Fr	En-Ru	En-Es	En-It	En-Pt	En-Nl	En-Ro	
Fairseq ST(Wang et al., 2020)	×	22.7	32.9	15.3	27.2	22.7	28.1	27.3	21.9	24.8
XSTNet(Ye et al., 2021)	✓	25.5	36.0	16.9	29.6	25.5	31.3	30.0	25.1	27.5
STEMM(Fang et al., 2022)	✓	25.6	36.1	17.1	30.3	25.6	31.0	30.1	24.3	27.5
MCTN(Zhou and Yuan, 2024)	✓	25.9	36.1	17.1	30.3	25.7	-	-	-	-
ConST(Ye et al., 2022)	✓	25.7	36.8	17.3	30.4	26.3	32.0	30.6	24.8	28.0
$FCCL^m$(Zhang et al., 2023)	✓	25.9	36.8	17.6	30.7	26.4	31.8	30.5	25.0	28.1
ZEROSWOT(Tsiamas et al., 2024)	×	27.3	35.8	17.8	31.7	26.8	31.6	30.9	25.3	28.4
CMOT(Zhou et al., 2023)	✓	27.0	37.3	17.9	31.1	26.9	32.7	31.2	25.3	28.7
CRAF	✓	**27.5**	**37.6**	**18.0**	**31.6**	**27.0**	**33.0**	**31.4**	**25.6**	**29.0**

pairs from the speech translation dataset. Subsequently, these components are fine-tuned for the speech translation task. During pretraining, the learning rate is set to 2e-3 with 8K steps for warm-up. In the fine-tuning phase, the learning rate is reduced to 1e-4 with a warm-up phase of 10K steps.

We use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a dropout rate of 0.1 and label smoothing of 0.1. The KL divergence weight, α, is set to 2, and the mixup probability is set to 0.8. All experiments are conducted on a single Nvidia GeForce RTX 4090 GPU.

Inference To derive the final model for inference, we average the weights of the last 10 checkpoints. During the decoding phase, a beam search algorithm is employed with a beam size of 13. Evaluation is performed on the MUST-C $\mathbf{tst-COMMON}$ dataset using the $\mathbf{case-sensitive}$ sacreBLEU[5] (Post, 2018) metric, and the reported sacreBLEU scores on the test set represent the final results.

Baselines We evaluate the performance of our proposed method against several established baselines, including fairseq-ST (Wang et al., 2020), XST-Net (Ye et al., 2021), STEMM (Fang et al., 2022), ConST (Ye et al., 2022), CMOT (Zhou et al., 2023), $FCCL^m$ (Zhang et al., 2023), ZEROSWOT (Tsiamas et al., 2024), and MCTN (Zhou and Yuan, 2024). XST-Net employs a progressive multi-task learning framework, simultaneously optimizing speech translation, automatic speech recognition, and machine translation tasks. STEMM improves representation alignment by projecting speech and text features into a shared semantic space, facilitating efficient shared memory learning. ConST integrates a contrastive learning mechanism to enhance the quality of extracted representations. CMOT combines Mixup and Optimal Transport in a unified framework to boost model performance. MCTN is a multi-task collaborative training network designed for joint modeling of ST, MT, and ASR tasks. FCCL enhances the performance of speech translation models by applying contrastive learning at different granularities. ZEROSWOT improves model performance through the use of CTC compression and OT techniques.

6 Numerical Results and Analysis

We will discuss the proposed method through experimental results, visualization analysis, and ablation studies. First, the quantitative results show the performance of our experiments. Second, visualization analysis, using mutual information, confirms the method's effectiveness in reducing information loss. Finally, ablation studies validate the contributions of each model module.

6.1 Experimental Results

The experimental results in Table 2 show significant improvements across eight language pairs in the MuST-C dataset, with the greatest gains in en-de and en-es. Our method achieves an average BLEU

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

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031-1042, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

score of 29.0, demonstrating its effectiveness and robustness. These results highlight both the superior performance and consistent reliability of our approach across various language pairs.

6.2 Analysis

Effect of Dynamic Weighted Length-Adapter In this study, we investigate the informational relationship between speech and text representations through the concept of mutual information (MI). MI quantifies the degree of dependency between two variables by measuring how much the uncertainty of one variable is reduced when the other is known. For continuous representations, MI is mathematically expressed as follows:

$$I(S;X) = \int_X \int_S q(s,x) \log \left(\frac{q(s,x)}{q(s)q(x)} \right) dsdx. \tag{11}$$

where S denotes the speech representation, X represents the text representation, q(s,x) is the joint probability density function of S and X, and q(s) and q(x) are the marginal probability density functions of S and X, respectively.

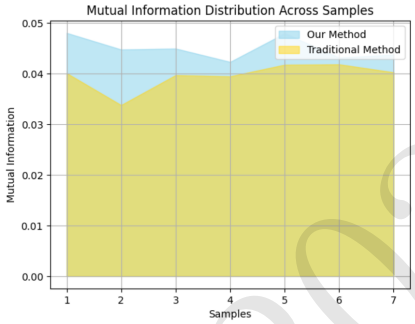

Figure 2: Effect of dynamic weighted length-adapter.

As depicted in Figure 2, the measured samples were divided into seven groups, each consisting of 10 samples, and the average value of each group was calculated to represent the final mutual information value. The yellow region in the figure corresponds to the mutual information between the speech and text representations obtained using the traditional method(which uses only two convolutional layers for length alignment), while the green region represents the mutual information calculated from the representations obtained using our method. Notably, the mutual information values achieved by our method are consistently higher than those obtained by the traditional method , indicating that the features extracted by our approach retain more comprehensive information. This further validates the effectiveness and robustness of our method.

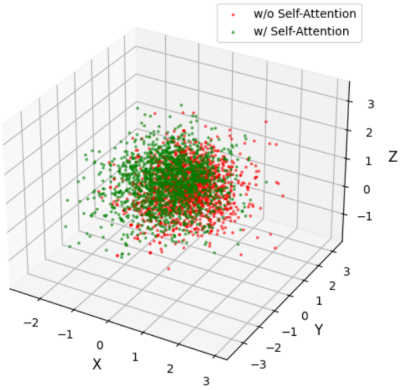

Figure 3: Effect of representation Fusion Module.

The Effect of Representation Fusion Module Figure 3 shows that applying self-attention (green

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

dots) makes feature representations more compact and semantically clustered, enabling the model to better focus on important features. Without self-attention (red dots), the representations are more scattered, indicating less effective semantic capture.

6.3 Ablation Study

Our ablation experiments take the En-De language pair as an example.

The Effect of the Dynamic Weighted Length-Adapter and Representation Fusion Module We conducted experiments to evaluate the effectiveness of combining the dynamic weighted length-adapter and representation fusion modules. As shown in Table 3, when only the dynamic weighted length-adapter module is used, the BLEU score reaches 27.2, indicating reduced information loss during length alignment. When only the representation fusion module is retained, the BLEU score is 27.3. These results demonstrate that integrating both components achieves the optimal performance improvement.

Table 3: BLEU scores with different modules

Module	BLEU
−Self-Attention	27.2
− Cross-Attention	27.3
CRAF	27.5

Dynamic Weighted Sum To evaluate the effectiveness of the Dynamic Weighted Sum proposed in our paper, we compare it with the static weighted sum. The calculation formula for the static weighted sum is as follows:

$$Fused_rep = \mu \cdot cross + (1 - \mu) \cdot CNN1. \tag{12}$$

where **cross** represents the output of **Cross-attention** , **CNN1** represents the first convolution layer, and **Fused_rep** is the final feature representation after length alignment, with μ being the weight.

For μ in equation (12), we tested values of 0.2, 0.4, 0.6, 0.8, and 1.0. As shown in Figure 4, the experimental results revealed that setting μ to 0.4 and 0.8 yields a maximum BLEU score of 27.2. However, this result remains lower than that achieved by our dynamic weighted sum module.

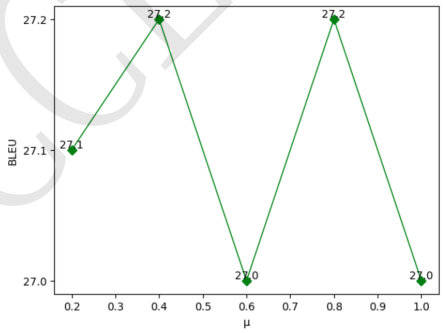

Figure 4: BLEU score with different static weights μ.

Why Use Self-Attention in the Representation Fusion Module? We investigated whether the performance improvement of our method is due to the increased number of parameters. The results in Table 4 show that by replacing self-attention with cross-attention, I and IV fuse the speech representation with M' using the speech representation and M' as queries, respectively; II and III fuse the text representation with M' using the text representation and M' as queries. Comparing I with IV shows that using the speech representation as the query yields better results. Comparing II with III shows that using M' as the query improves performance. However, none of these approaches outperform our method, demonstrating the effectiveness and advantage of our approach under the same parameter settings.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1031-1042, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 4: Effect of Self-attention

Module	BLEU
I	27.3
II	27.1
III	27.3
IV	27.2
CRAF	27.5

The Effect of the Mixup To verify the effectiveness of the Mixup fusion method, we designed three sets of comparative experiments: length-wise concatenation (Concatenation-L), feature-wise concatenation (Concatenation-F), and dynamic weighted fusion. Specifically, Concatenation-L refers to directly concatenating the speech and text representations along the sequence length dimension, while Concatenation-F concatenates them along the feature dimension. In the dynamic weighted fusion strategy, we first compute separate weight coefficients for the speech and text representations, and then fuse them via a weighted sum, ensuring that the weights sum to one.

Table 5: Effect of Mixup

Module	BLEU
Concatenation-L	27.2
Concatenation-F	27.0
Dynamic Weighted Sum	27.1
Mixup	27.5

As shown in Table 5, the Mixup-based fusion method achieves the best performance. We attribute this to Mixup's use of an optimal transport (OT) matrix, which enables distribution-aware sampling between speech and text representations. This results in a fused representation that is primarily guided by the speech features, supplemented by text features that are distributionally aligned with the speech. Such a fusion strategy effectively mitigates modality mismatch and allows the textual information to complement the speech features, thereby enhancing overall model performance. Furthermore, the superior performance of Concatenation-L over Dynamic Weighted Sum may be due to the fact that the former preserves the full information of both modalities by concatenating speech and text representations along the length dimension. In contrast, the latter performs frame-wise weighted averaging, which may lead to partial information loss and weaken the discriminative power of the fused representation. Compared with Concatenation-F, Concatenation-L also performs better, likely because feature-dimension concatenation requires aligning the sequence lengths of speech and text. This often involves truncating the speech representations or padding the text representations, either of which may introduce redundancy or cause the loss of crucial information, thus interfering with effective semantic modeling.

Beam Size We conducted an ablation study on the choice of beam size during decoding and visualized the results for different values. As shown in Figure 5, the model achieves its best performance when the beam size is set to 13.

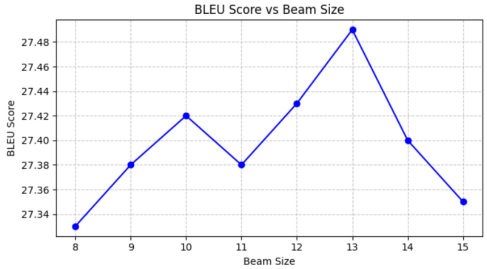

Figure 5: BLEU score with different beam size.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

7 Conclusion

In this study, we present **C**ross-Modal **R**epresentation **A**lignment and **F**usion (**CRAF**), designed to align the lengths of speech and text representations while minimizing information loss and enhancing feature fusion between the two modalities. To validate the effectiveness of **CRAF**, we conducted extensive experiments on the MuST-C dataset. The results demonstrate that our method significantly reduces information loss and strengthens the fusion of cross-modal representations, thereby improving overall model performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants: U24A20334, 62366027, 62466030), Yunnan Provincial Key R&D Program (202502AD080014, 202303AP140008).

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NeurIPS*, Barcelona, Spain, December.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, pages 2012–2017, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *NAACL*, pages 949–959, San Diego, California, June. 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 Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, pages 7050–7062, Dublin, Ireland, May. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting transformer to end-to-end spoken language translation. In *Interspeech*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *ACL-IJCNLP*, pages 2214–2225, 01.

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, October.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Speechformer: Reducing information loss in direct speech translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1706, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

CCL 2025

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11:355–206, 01.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *WMT*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, pages 2786–2792, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ioannis Tsiamas, Gerard Gállego, José Fonollosa, and Marta Costa-jussa. 2024. Pushing the limits of zero-shot end-to-end speech translation. In *ACL*, pages 14245–14267, Bangkok, Thailand, 01.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Laura Vila, Carlos Escolano, José Fonollosa, and Marta Costa-jussa. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*, pages 60–63, 11.

Cédric Villani, 2008. *Optimal transport – Old and new*, volume 338, pages xxii+973. Springer, 01.

A. Waibel, A.N. Jain, A.E. McNair, H. Saito, A.G. Hauptmann, and J. Tebelskis. 1991. Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *ICASSP*, pages 793–796 vol.2.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In Derek Wong and Douwe Kiela, editors, *AACL-IJCNLP*, pages 33–39, Suzhou, China, December. Association for Computational Linguistics.

Chen Xu, Yuhao Zhang, Chengbo Jiao, Xiaoqian Liu, Chi Hu, Xin Zeng, Tong Xiao, Anxiang Ma, Huizhen Wang, and Jingbo Zhu. 2023a. Bridging the granularity gap for acoustic modeling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *ACL*, pages 10816–10833, Toronto, Canada, July. Association for Computational Linguistics.

Chen Xu, Yuhao Zhang, Chengbo Jiao, Xiaoqian Liu, Chi Hu, Xin Zeng, Tong Xiao, Anxiang Ma, Huizhen Wang, and Jingbo Zhu. 2023b. Bridging the granularity gap for acoustic modeling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10816–10833, Toronto, Canada, July. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *ArXiv*, abs/2104.10380.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *NAACL*, pages 5099 – 5113, Seattle,United States. Association for Computational Linguistics.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Wei-Qiang Zhang. 2023. Improving speech translation by cross-modal multi-grained contrastive learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:1075–1086, February.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, and Furu Wei. 2024. Speechlm: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:2177–2187, March.

Yue Zhou and Yuxuan Yuan. 2024. A multitask co-training framework for improving speech translation by leveraging speech recognition and machine translation tasks. *Neural Computing and Applications*, 36:1–16, 02.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *ACL*, pages 7873–7887, Toronto, Canada, July. Association for Computational Linguistics.

CCL 2025