Whisper-UT: A Unified Translation Framework for Speech and Text

Anonymous ACL submission

Abstract

Encoder-decoder models have achieved remarkable success in speech and text tasks, yet efficiently adapting these models to diverse uni/multi-modal scenarios remains an open challenge. In this paper, we propose Whisper-UT, a unified and efficient framework that leverages lightweight adapters to enable seamless adaptation across tasks, including a multi-modal machine translation (MMT) task that explicitly conditions translation on both speech and source language text inputs. By incorporating ASR hypotheses or ground-truth transcripts as prompts, this approach not only enables the system to process both modalities simultaneously but also significantly enhances speech translation (ST) performance through a 2-stage decoding strategy. We demonstrate our methods using the Whisper model, though in principle they are general and could be applied to similar multitask models. We highlight the effectiveness of cross-modal and cross-task fine-tuning, which improves performance without requiring 3-way parallel data. Our approach even outperforms using *ground-truth* transcripts using an in-domain fine-tuned NLLB model on multiple challenging conversational speech translation corpora.

1 Introduction

The task of speech-to-text translation (ST) encompasses converting spoken language from one language to another, aiming to overcome linguistic barriers. Traditionally, the task involves an automatic speech recognition (ASR) module to transcribe spoken words, followed by a machine translation (MT) module to convert the transcribed text into the target language in a cascaded manner (Ney, 1999). The recent development of end-to-end neural architectures and large pre-trained models have substantially propelled advancements in downstream speech tasks, including speech translation, via either self-supervised learning (SSL) (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) or fully supervised learning. Among the pre-trained acoustic models, Whisper (Radford et al., 2022), a transformer-based encoder-decoder multi-task model trained with large-scale data in a supervised manner, has exhibited strong performance on various ST corpora.

However, in real-world scenarios, input modalities and data conditions vary widely. In offline settings, for instance, translating conversational or dialectal speech—characterized by disfluencies, code-switching, and noisy acoustic environments—poses significant challenges to end-to-end models, often resulting in degraded performance. Conversely, scenarios like business meetings or translated media archives frequently provide both source-language speech and transcripts (manual or ASR-generated), yet existing systems fail to exploit this multi-modal synergy.

To address this, we systematically investigate how multi-task encoder-decoder models—using Whisper as a representative case study—can be efficiently adapted to these heterogeneous scenarios. First, we examine fine-tuning strategies for conventional ST (using 3-way parallel speech-transcript-translation data), speech-to-text tasks (ASR-only data), and MT, while also methods for multi-modal translation where both speech and transcripts are available. Our analysis reveals two key insights:

- *Cross-task training induces synergistic benefits*—fine-tuning on in-domain ASR data improves speech translation (ST) performance, while ST training conversely enhances ASR accuracy, suggesting mutual reinforcement between tasks even without 3-way parallel data; and

- *Multi-modal inputs (speech + text) consistently enhance translation quality when fused*, even with imperfect ASR transcripts.

Building on these findings, we propose **Whisper** for **U**nified **T**ranslation[1], or Whisper-UT, a framework that transforms Whisper's decoder into a unified conditional generation model, capable of dynamically conditioning on speech, text, or both modalities. The framework repurposes Whisper's encoder-decoder architecture as a versatile multi-modal interface through two innovations:

1 *A multi-task learning paradigm* with a stochastic task-selection mechanism to adapt the system across ASR, MT, ST, and multimodal translation tasks using a single set of LoRA parameters;

2 *A two-phase decoding strategy*, where the decoder first generates ASR transcripts from speech, then reuses them as context for translation when the transcript is not provided, emulating human thinking process.

Crucially, Whisper-UT requires no architectural modifications—only fine-tuning—ensuring compatibility with any encoder-decoder model.

Experiments on conversational telephony speech (CTS) corpora—Fisher-CallHome Spanish and BOLT Chinese-English demonstrate state-of-the-art performance. Notably, Whisper-UT outperforms the 1.3B-parameter NLLB model in multi-modal settings (speech + ground-truth text) and achieves superior speech-only translation via hypothesis prompting.

Our work highlights the untapped potential of multi-task models in adaptive translation systems. By unifying modality handling and enabling efficient task specialization, Whisper-UT bridges the gap between rigid single-modality systems and the dynamic needs of real-world applications.

2 Related Work

2.1 Whisper

Whisper is an end-to-end multi-task speech model that adopts a transformer-like encoder-decoder architecture. Its LARGE-V2 version is pre-trained on 680,000 hours of speech data with multiple supervisions. The model consists of several convolutional blocks, a series of position-encoded transformer-encoder blocks and transformer-decoder blocks. As with the original transformer model (Vaswani et al., 2023), the loss function Whisper used at its pre-training time is the cross-entropy objective for all tasks, including language identification (LID), voice activity detection (VAD), multilingual ASR, ST and speech alignment.

Whisper's decoder supports a prompting mechanism, originally designed for better capturing long-range dependencies of the transcripts/translations to resolve local audio ambiguities. Particularly, long utterances are segmented into chunks and the decoder generates its hypothesis for the current segment conditioning on the previous segment's transcripts. Inspired by the effectiveness of GPT-like decoder-only models in machine translation, we hypothesize that Whisper's decoder, which may be viewed as an audio-conditional language model, is also capable of performing machine translation conditioned on audio inputs.

2.2 Multi-modal speech translation systems

Recent developments in multi-modal speech translation systems are exploring new ways to combine audio and text to improve language translation. mSLAM (Bapna et al., 2022), a multilingual speech and language model, has emerged as a pioneering approach. It aims to construct a shared representation space for both speech and text through joint pre-training on both self-supervised and supervised tasks with various loss objectives, including the translation language modeling (TLM) loss for ST.

SeamlessM4T (Communication et al., 2023) is another innovative model that further refines the integration of multi-modal inputs for speech and text translation tasks. As a single model designed for ASR, text-to-text translation, text-to-speech translation, speech-to-text translation and speech-to-speech translation, it consists of multiple building blocks to leverage mono-modal data, including a w2v-BERT (Chung et al., 2021) as the speech encoder, a 1.3B NLLB model (Team et al., 2022) as the text encoder and decoder, a transformer-based text-to-unit encoder-decoder model, and a vocoder for converting the units to speech. They adopted a staged training strategy to make use of mono-modal data for initializing each building block, followed by several fine-tuning stages to fuse the components for the target tasks.

These systems, along with most existing methods, primarily seek to utilize text and speech modalities by aligning their representations. However, such a design restricts the model to accept inputs in only one modality during inference, preventing it from attending to signals in another modality when available, which is a gap we aim to fill.

[1]We open source our code at [link-hidden-for-review]

3 Methodology

Traditional translation systems treat ST, MT, and ASR as distinct tasks, each requiring separate models or specialized architectures. In this work, we propose a **unified translation** framework that unifies these tasks under a single encoder-decoder paradigm, treating all forms of language conversion—including audio-to-text, text-to-text, and multi-modal translation—as conditional generation tasks. Our approach enables seamless adaptation to various input modalities and data conditions without requiring fundamental architectural changes.

At the core of our method is the insight that ASR can be reformulated as a source-language transcription task, ST as a direct speech-to-text translation task, and MT as a standard text-to-text translation task—all of which can be expressed as instances of sequence-to-sequence learning. Extending this idea, we introduce a **multi-modal translation** task, for which the model conditions on both speech and its corresponding transcript (either human-annotated or ASR-generated) to improve translation quality. This formulation generalizes the conventional ST and MT paradigms, leveraging available transcripts to enhance translation in scenarios where speech alone may be ambiguous or error-prone.

3.1 Translation with Multi-modal Inputs

We first provide a formal definition of the multi-modal translation (MMT) task, or more precisely, the task of speech-and-text-conditioned translation. Let $X = (x_1, x_2, \cdots, x_T)$ denote the speech signal of an utterance, $Y = (y_1, y_2, \cdots, y_M)$ denote the ground-truth transcript of the utterance, and $Z = (z_1, z_2, \cdots, z_N)$ denote its corresponding text translation. The goal of the task is then to find the conditional distribution $P(Z|X, Y)$. We hypothesize that often $H(Z|X, Y) < H(Z|Y)$ in practice, where H denotes the information entropy. In other words, the speech signal may contain additional information for a more accurate translation of the utterance, as it may be able to aid resolving ambiguities such as homographs, tonal variations, and omitted content—such as repetitions and filler words—that may be present in human-annotated transcripts.

In light of the remarkable performance observed with decoder-only language models in machine translation, we presume that encoder-decoder models' audio-conditioned decoder possesses the potential for undertaking the audio-conditioned text translation task. In particular, one may prompt the decoder with source language text, generated either by human annotators or any ASR system, in the translation process, as shown in Figure 1(b). Consequently, the resulting model is trained to learn the distribution $P(Z|X, Y)$.

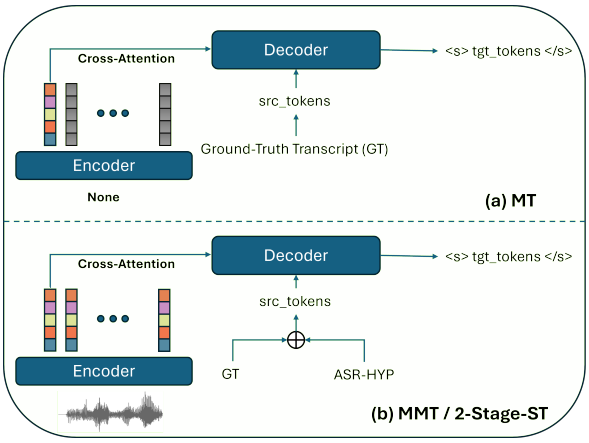

Figure 1: **Overview of our approach.** Note that special tokens are abbreviated in the figure for simplicity. *ASR-HYP* refers to the ASR hypothesis generated. When GT is used, the task is MMT, otherwise it is referred as 2-Stage-ST. The $\oplus$ symbol refers to the XOR operation.

3.2 Translation with Speech-only Inputs

The problem of speech translation can be directly modeled as $P(Z|X)$ or modeled by marginalizing over an underlying latent variable, Y', representing valid transcripts of the audio X:

$$\begin{aligned} P(Z|X) &= \sum_{Y'} P(Z, Y'|X) \\ &= \sum_{Y'} P(Z|Y', X)P(Y'|X) \end{aligned} \quad (1)$$

However, the summation over Y' is generally intractable. One common solution, also adopted by cascaded approaches to speech translation, is to approximate the summation with the single highest weight term in the summation, i.e.,

$$\begin{aligned} &\sum_{Y'} P(Z|Y', X)P(Y'|X) \\ &\quad \approx \max_{Y'} P(Z|Y', X)P(Y'|X), \end{aligned}$$

and furthermore to assume that the best transcript is the most likely one:

$$\begin{aligned} \hat{Y} &= \arg\max_{Y'} P(Z|Y', X)P(Y'|X) \\ &= \arg\max_{Y'} P(Y'|X). \end{aligned} \quad (2)$$

However, cascaded speech translation further assumes that the translation is conditionally independent of the audio given the transcript,

$$P(Z|\hat{Y}, X) = P(Z|\hat{Y}), \tag{3}$$

which is practical in that it enables modular training of components, i.e.,

$$P(Z|X) = P(Z|\hat{Y})P(\hat{Y}|X), \tag{4}$$

where $P(Z|Y)$ and $P(Y|X)$ can be trained separately, but it at the cost of a possible uneeded additional approximation.

End-to-end systems such as Whisper, however, model the problem without explicitly conditioning on the ASR transcripts, Y'. Its single-decoder multi-task paradigm presumably captures a higher-level abstract semantics of the speech signals, such that the ST decoding process is implicitly entangled with the model's ASR ability.

We seek to combine the modeling advantages of the cascaded and end-to-end systems and generalize the multi-modal translation setting to reformulate the system's speech-only translation process for approximating Equation 1. Specifically, we relax the conditional independence assumption of cascade approaches, by endowing end-to-end speech translation models with the capacity to also condition on either a ground-truth or hypothesized transcript defined by Equation 2, i.e.:

$$P(Z|X) = P(Z|\hat{Y}, X)P(\hat{Y}|X) \tag{5}$$

In our implementation, we carry out a **two-stage decoding** process. In the first stage, the model is used to produce the ASR hypotheses, and subsequently, in the second stage, the model conditions on them to generate the translations.

An alternative perspective on this modeling is that it fully leverages the system's source-language modeling capability. In end-to-end multi-task models, the decoder can be viewed as implicitly "partitioned" into two roles: source-language modeling and target-language generation. While these functions share parameters and benefit from joint optimization, they may still develop distinct competencies. By conditioning translation on both speech and textual transcripts, this approach explicitly harnesses a well-trained source-language model—potentially even from an external ASR system—allowing the decoder to generate more accurate and fluent translations. This perspective highlights how multi-modal conditioning can serve as a mechanism to refine and reinforce the system's understanding of the source language, ultimately improving translation quality.

3.3 Translation with Text-only Inputs

Integrating MT functionality into a multi-modal encoder-decoder model presents unique challenges. In conventional encoder-decoder MT systems, the source language text is processed through the encoder, which generates contextual representations for the decoder to cross-attend to. However, in our case, the encoder is designed specifically for processing speech features, making direct text encoding potentially ineffective without substantial adaptation. Training the encoder to handle text inputs would require a significant amount of additional data and could lead to catastrophic forgetting, where the model loses its ability to process speech effectively.

Inspired by the success of decoder-only MT models such as GPT-based architectures, we adopt an alternative strategy: instead of modifying the encoder to accommodate text, we encode the source text directly within the decoder, as illustrated in Figure 1(a). Specifically, we prepend the source text as a prefix to the decoder input, leveraging the self-attention mechanism to implicitly model source-to-target dependencies. However, implementing this method within an encoder-decoder framework requires careful handling of the cross-attention mechanism. Since the decoder in our system is designed to attend to encoded speech representations, directly bypassing the encoder would disrupt the model's expected structure. To address this, we introduce a single learnable vector in the encoder, serving as an indicator that informs the decoder that text input is being processed. The remaining encoder output is padded with zeros, and we modify the cross-attention mask such that the decoder attends only to this learnable embedding. This design ensures that the model's architecture remains structurally intact while effectively repurposing the decoder for text-based translation. By adopting this strategy, we enable seamless integration of MT capabilities into our unified translation system without requiring extensive modifications to the speech encoder.

3.4 Whisper-UT: Unified Translation System

To achieve a unified translation framework that encompasses multiple translation paradigms, we

propose Whisper-UT, a system designed to handle ASR, ST, MT, and MMT within a single model. Our approach is built on multi-task learning, leveraging 3-way parallel data and text-only MT data to optimize multiple objectives in a stochastic fashion.

3.4.1 3-way Parallel Data Objectives

We formulate the learning process with six distinct training objectives, categorized based on the availability of parallel data.

For the 3-way dataset that provide speech, transcripts, and translations $\{X, Y, Z\}$, we define three primary objectives:

ASR Objective. Learning the mapping $X \rightarrow Y$, i.e., predicting the source language transcript from speech.

E2E-ST Objective. Directly predicting the target language text Z from speech X.

MMT Objective. Predicting Z while attending to both X (speech) and Y (source transcript).

3.4.2 Text-Only Data Objectives

Since 3-way parallel datasets are scarce in reality, we incorporate text-only MT data $\{Y, Z\}$ and define additional objectives:

Source Language Modeling (SLM): Predicting the next source token in Y, acting as an ASR surrogate for text-only samples.

Target Language Modeling (TLM): Predicting the next token in Z, improving the decoder's target language modeling ability.

Machine Translation (MT): Translating $Y \rightarrow Z$ using the decoder.

For MMT and MT objectives, we allow gradients to propagate back through the source language tokens, implicitly enhancing the model's source language modeling ability.

3.4.3 Dynamic Loss Weighting

To balance the competing objectives, we employ a **stochastic task selection mechanism** with beta-distributed loss weighting inspired by (Zhang and Patel, 2024):

$$\alpha \sim \text{Beta}(\beta_1, \beta_2), \tag{6}$$

which determines the final multi-task loss:

$$\mathcal{L}_{\text{mtl}} = (1 - \alpha)\mathcal{L}_{\text{asr}}^{CE} + \alpha\mathcal{L}_{\text{st}}^{CE}, \tag{7}$$

where $\mathcal{L}_{\text{asr}}^{CE}$ is the ASR loss (or SLM loss for text-only samples), and $\mathcal{L}_{\text{st}}^{CE}$ is either the ST loss or the MMT loss, selected via stochastic task selection.

The stochastic weighting scheme is motivated by empirical findings that equal task weighting leads to gradient interference, degrading performance across tasks.

3.4.4 Utterance-Level Task Selection

Each batch is sampled from a mixture of the 3-way parallel data and text-only MT data. We define the loss computation as follows:

- **ASR Loss:** Always computed for speech-based samples; replaced with SLM loss for text-only samples (zero-padded input except for a learnable vector).
- **ST vs. MMT Objective:**
  - With probability q, apply standard ST loss; for text-only data, this is equivalent to the TLM loss.
  - With probability $(1 - q)$, apply MMT loss, where the decoder cross-attends to both speech features and source text tokens; for text-only data, this becomes the conventional MT loss.

3.4.5 Masking in Multi-Modal Translation

For MMT, we introduce a masking mechanism to enhance robustness:

- With probability b, the source language tokens are partially masked.
- Each token in the source text is independently masked with probability t.

This simulates real-world noise in transcripts (e.g., ASR errors, omissions), encouraging the model to rely on both modalities for translation.

3.5 Unified Training Framework

In summary, our unified training framework integrates ASR, ST, MMT, and MT into a single multi-task learning process. To achieve this, we first concatenate both speech-text and text-only datasets, allowing for random sampling within each batch. For every batch, we compute the ASR loss, which corresponds to the source language modeling loss when dealing with text-only samples. The ASR and ST loss weights are dynamically balanced by sampling a weight α from a Beta distribution. Next, we stochastically determine whether the batch follows the ST/TLM objective or the MMT/MT objective.

If the batch is selected for MMT training, masking is applied with a certain probability to simulate transcription imperfections and enhance robustness. By combining these components, Whisper-UT serves as a unified model for ASR, ST, MT, and MMT, leveraging both textual and speech inputs efficiently.

4 Experiments

4.1 Experimental Setup

4.1.1 Tasks and Datasets

In this work, we focus primarily on the translation of conversational telephony speech (CTS). We fine-tune and evaluate our approach on two datasets, including the Fisher-CallHome Spanish-to-English corpus with 186 hours, and the BOLT Chinese-to-English corpus with 110 hours, of conversational telephony speech in Spanish and Chinese in their training partition, respectively.

4.1.2 Pre-processing

CTS corpora usually consist of short utterances segmented from a full recording, reflecting the alternating speech of participants during conversations. However, we found empirically that fine-tuning on such segments, presumably due to a mismatch in sample lengths compared to Whisper's pre-training data, leads to significant performance degradation. The resulting model tends to repetitively produce frequent filler words in the training corpus at inference time regardless of the input. Therefore, we re-segmented the utterances by merging them chronologically, with durations (in seconds) sampled from a Gaussian distribution, e.g. $\mathcal{N}(15, 5^2)$. As Whisper's feature extractor automatically pads the features up to 30 seconds, such re-segmentation also significantly reduced the training cost in terms of memory and time.

4.1.3 Data Augmentation

We apply the conventional speed perturbation (Ko et al., 2015) with parameters 0.9, 1.0, 1.1 to the speech prior to the training stage. Additionally, we adopt SpecAug (Park et al., 2019) to randomly mask extracted speech features during training.

4.1.4 Parameter Efficient Fine-tuning

To efficiently adapt the model to these conversational scenarios without overfitting or incurring excessive computational cost, we leveraged several parameter-efficient fine-tuning (PEFT) techniques.

For this study, we concentrate on the LARGE-V2 version of Whisper with 1.6 billion parameters. In order to fit it into our hardware, we adopted a list of strategies:

- **Low-Rank Adaptation (LoRA).** LoRA (Hu et al., 2021) introduces a trainable adapter comprised of rank decomposition matrices on top of the fixed pre-trained model's weight matrices in specified layers so that the number of trainable parameters can be considerably reduced.
- **Gradient checkpointing.** Gradient checkpointing (Chen et al., 2016) stores intermediate activations in the forward pass, and recomputes the remaining activations during back-propagation.
- **Zero Redundancy Optimizer (ZeRO).** ZeRO (Rajbhandari et al., 2020) is an algorithm that partitions data, optimizer states, gradients, and parameters for speeding up the training of large neural models with low communication costs.

4.1.5 Baselines

To evaluate our approach, we compared our results against several strong baselines spanning both end-to-end and cascaded systems. For MT, we employed NLLB as our primary baseline. For ST and ASR, we used Whisper-large-v2 in its pre-trained form as well as after fine-tuning. For each fine-tuned model, we also tune the number of trainable parameters and report results for the best-performing configuration.

4.1.6 Evaluation

For both ASR and ST, we normalize the text by lower-casing all characters and removing all punctuations before computing the metrics. In addition, we also apply the global map scoring (GLM) to standardize digits and abbreviations. For the Fisher Spanish corpus, the BLEU score is computed on the first set of references, in contrast to the multi-reference BLEU reported in other work (Weiss et al., 2017). The evaluation script used is provided in the code.

4.2 Training

To demonstrate the effectiveness of our proposed approach, we adopt Whisper as the base model and fine-tune it for our unified translation modeling. By leveraging its robust speech and language

Table 1: Results on the Fisher Spanish test set. The **Objective** column specifies under which training objective the model system is fine-tuned. The *UT* objective refers to the unified-translation objective described in section 3.5. The **Task** column specifies the target inference task. *E2E-ST* refers to the promptless E2E speech translation setting, *MMT* refers to the translation process that conditions on both the ground-truth transcript and the speech signals, while *2-Stage-ST* refers to the MMT process which conditions on the model's own ASR hypotheses.

	Model	**Objective**	**Task** (num_beams = 1)				
			ASR (WER↓)	E2E-ST (BLEU↑)	MT (BLEU↑)	MMT (BLEU↑)	2-Stage-ST (BLEU↑)
1	NLLB-1.3B	MT	-	-	**43.4**	-	-
2	Whisper	None	26.7	30.9	-	-	-
3		ASR	19.1	36.7	-	-	-
4		ST	20.3	40.0	-	-	-
5		ASR + ST	16.3	**40.6**	-	-	-
6		MT	60.3	30.8	41.3	40.2	-
7		MMT	16.4	36.2	1.1	44.3	38.6
8		UT	**15.6**	40.3	29.8	**44.6**	**44.2**

Table 2: Results on the BOLT Chinese-English test set.

	Model	**Objective**	**Task** (num_beams = 1)				
			ASR (WER↓)	E2E-ST (BLEU↑)	MT (BLEU↑)	MMT (BLEU↑)	2-Stage-ST (BLEU↑)
1	NLLB-1.3B	MT	-	-	**22.7**	-	-
2	Whisper	None	32.2	13.0	-	-	-
3		ASR	18.9	16.2	-	-	-
4		ST	23.1	16.8	-	-	-
5		ASR + ST	18.5	20.2	-	-	-
6		UT	**17.5**	**20.6**	11.1	**25.3**	**25.2**

modeling capabilities, we aim to showcase how a single model can be adapted to perform various speech/text tasks through our proposed multi-task learning framework.

For the *Unified Translation (UT)-trained system*, we extend the training data to include additional MT text-only data. Specifically, for Spanish, we incorporate the transcript and translation of 197 hours of audio from CoVoST 2(Wang et al., 2020b), mTEDx(Salesky et al., 2021), and Europarl-ST(Koehn, 2005). For Chinese, we include 130 hours from CoVoST(Wang et al., 2020a), GALE(Song et al., 2016) and a collection of in-house datasets. We refer to these additional MT datasets as the *OOD* set (out-of-domain). To ensure a fair comparison across models, we maintain a consistent number of training steps for all systems, including the UT-trained system, despite the addition of the OOD set.

4.3 Analysis of Experimental Results

The results presented in Table 1 and 2 provide key insights into the effectiveness of our proposed unified translation model. Since the trends observed are consistent across both datasets, the following analysis will primarily focus on the Fisher Spanish data as a representative example.

4.3.1 Cross-task Synergy

Our findings reveal that fine-tuning on one task does not only improve performance on the target task but also benefits other tasks as well. Notably, ASR fine-tuning enhances ST performance, and ST fine-tuning reciprocally benefits ASR. This is likely due to the incorporation of in-domain data, which aids the overall language modeling capability and adaptation to the target acoustic environment. For instance, ASR fine-tuning improves ST performance from 30.9 to 36.7, while ST fine-tuning boosts ASR accuracy, reducing WER from 26.7 to 20.3. This observation motivated us to explore a unified model approach, as real-world scenarios often lack task-specific training data, yet leveraging in-domain data from other modalities can serve as a proxy objective for improving target task performance.

4.3.2 Effectiveness of Multi-task Learning

In Row 5, we conducted a straightforward multi-task fine-tuning experiment by duplicating the speech dataset with both ASR and ST supervision,

concatenating the datasets, and employing random sampling within each batch. This experiment confirms that multi-task training is beneficial, as it enhances BLEU score from 40.0 to 40.6 and WER is reduced from 19.1 to 16.3. This suggests that jointly optimizing multiple relevant objectives allows the model to better capture linguistic patterns and improve generalization across tasks.

4.3.3 Text-only MT Training and Its Effects

Row 6 presents the MT-only fine-tuning experiment, which reveals that even with limited in-domain data, the model achieves strong performance in text translation (BLEU 41.3). This suggests that Whisper's decoder inherently possesses some text translation capabilities or at least has sufficiently strong source and target language modeling abilities such that minimal adaptation enables it to perform the MT task. Interestingly, this MT training also gives the system MMT ability, as suggested by the 40.2 BLEU score, despite MMT being a novel objective that the model was not explicitly trained on. This finding reinforces our earlier observation of cross-task synergy.

However, while MT fine-tuning does not degrade ST performance (compared to the baseline), it significantly worsens ASR performance, increasing WER from 26.7 to 60.3. This suggests that fine-tuning without an ASR objective may lead to catastrophic forgetting in speech modeling. Thus, incorporating an ASR objective is essential to prevent such degradation and to regularize the model.

4.3.4 MMT-Multi-task Training and Its Implications

Row 7 evaluates an MMT-Multi-task fine-tuned model, that is, the model is trained with $q = 1$ and $b = 0$. Notably, the system's MMT inference result outperforms even the strong fine-tuned NLLB-1.3B baseline's MT performance (44.3 vs. 43.4), demonstrating that MMT provides tangible benefits over traditional cascaded MT approaches. Additionally, the training also enhances the E2E-ST results, increasing ST BLEU from 30.9 to 36.2, further supporting the idea of cross-task synergy.

However, a gap remains between different MMT settings. Specifically, when using the ASR hypothesis as input instead of the ground-truth transcript, performance degrades from 44.3 to 38.6. Although this still surpasses the direct ST performance (36.2), it indicates that transcript quality plays a crucial role. This highlights both the effectiveness of explicit modeling and the limitations introduced by ASR errors.

4.3.5 Unified Translation (UT) Training

Finally, the UT-trained system (Row 8) achieves the best MMT and 2-Stage-ST results, with MMT reaching 44.6 BLEU and 44.2 BLEU respectively, proving the method's effectiveness. Applying the masking strategy in this training scheme makes the two-stage approach more robust; only a minor gap between MMT and two-stage ST performance remains.

This system enhances ST performance, indicating that in the absence of a transcript, the two-stage approach generates high-quality translations. However, MT performance remains limited, likely due to two factors: (1) it's trained only on the OOD data, and (2) the decision to maintain fair training conditions by not increasing training steps. This suggests that to leverage the full potential of the UT framework may require further optimization and additional training.

4.3.6 Summary

Our results highlight the benefits of cross-task learning, multi-task training, and explicit modeling techniques. The strong synergy between tasks suggests that a unified modeling approach is an effective way to enhance performance across ASR, ST, and MT tasks, particularly in low-resource or cross-modality scenarios.

5 Conclusion

In this paper, we introduced Whisper-UT, a unified translation framework that integrates ASR, ST, MT, and MMT within a single multi-task learning paradigm. In addition to this unified framework, we propose an explicit modeling approach for speech translation that conditions on both speech signals and textual prompts, effectively leveraging ASR hypotheses or ground-truth transcripts. Our training strategy, incorporating stochastic task selection and modality-aware masking, ensures effective multi-task learning while mitigating catastrophic forgetting. Experimental results show that Whisper-UT achieves strong performance across various translation tasks, demonstrating the benefits of cross-task synergy. Future work will explore scaling to more languages and extending to broader multi-modal scenarios.

6 Limitations and Ethical Considerations

While our approach demonstrates strong improvements, several limitations remain. To ensure fair comparisons, we kept training steps consistent across models, meaning our best-performing system may not have reached its full potential with extended training.

Due to resource constraints, we fine-tuned Whisper rather than training from scratch, which might limit the full integration of the objectives. Ideally, to demonstrate cross-task fine-tuning, we would start from a pretrained model that natively support each of our tasks, (MT, MMT, ST, ASR), but building state-of-the-art, or close to state-of-the-art systems requires building from existing models, such as Whisper, and adapting to Whisper to additionally perform these tasks, while a contribution in its own right, ultimately requires a two-stage fine-tuning approach that complicates analysis of the effectiveness of cross-task fine-tuning. Furthermore, while we believe our method to be general, i.e., it could be applied to similar models such as the OWSM model (Peng et al., 2024), we have only demonstrated our results using the Whisper model.

Training of machine learning models is a costly, energy-intensive process, so our method, which introduces a novel means of efficiently adapting existing large pre-trained models to new tasks, may limit the ethical concerns about the costs, financial, environmental, or other, associated with training ML models. Furthermore, the success of our approach, specifically cross-task fine-tuning, implies that speech translation systems can be more easily trained for new domains, including languages with limited training resources.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

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. *Preprint*, arXiv:2202.01374.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *Preprint*, arXiv:1604.06174.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

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. *Preprint*, arXiv:2106.07447.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

H. Ney. 1999. Speech translation: coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 1, pages 517–520 vol.1.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *Preprint*, arXiv:2102.01757.

Zhiyi Song, Gary Krug, and Stephanie Strassel. 2016. Gale phase 3 and 4 chinese newswire parallel text.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. *Preprint*, arXiv:2002.01320.

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

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *Preprint*, arXiv:1703.08581.

Ke Zhang and Vishal M. Patel. 2024. Modelmix: A new model-mixup strategy to minimize vicinal risk across tasks for few-scribble based cardiac segmentation. *Preprint*, arXiv:2406.13237.

A Hyperparameter Settings

Table 3 presents the hyperparameter configurations used for training our Whisper-UT model.

Hyperparameter	Value
LoRA Rank	128
LoRA Alpha	256
LoRA Dropout	0.1
Max Training Steps	8000
Batch Size	8
Gradient Accumulation Steps	1
Warmup Steps	500
Learning Rate	$1e^{-5}$
Weight Decay	$5e^{-4}$
SpecAug Mask Feature Probability	0.1
SpecAug Mask Time Probability	0.05

Table 3: Hyperparameter configurations used for training.

Experiments in this work are conducted with 8 V100-32GB GPUs. However, PEFT methods outlined in Section 4.1.4 render the use of 8 GPUs redundant, yet they are deployed to accelerate the training process.