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

Cihan Xiao[1], Matthew Wiesner[1, 2], Debashish Chakraborty[2], Reno Kriz[2], Keith Cunningham[3], Kenton Murray[1, 2], Kevin Duh[1, 2], Luis Tavarez-Arce [2], Paul McNamee[2], Sanjeev Khudanpur[1, 2]

[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Human Language Technology Center of Excellence, Johns Hopkins University
[3]Georgetown University

Correspondence: cxiao7@jhu.edu

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 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 multi-task models. We highlight the effectiveness of cross-modal and cross-task fine-tuning, which improves performance without requiring 3-way parallel data. Our results underscore the flexibility, efficiency, and general applicability of the proposed framework for multi-modal translation.

1 Introduction

The task of speech-to-text translation (ST) encompasses converting spoken content from one language to another, aiming to overcome language barriers to communication. 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, 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 good 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 (manual or ASR-generated) transcripts. 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 ST performance, while ST training conversely enhances ASR accuracy, suggesting mutual reinforcement between the ASR and ST tasks even without 3-way parallel data;
- *Multi-modal inputs (speech + text) consistently enhance translation quality when fused*, even with imperfect ASR transcripts.

Building on these findings, we propose **Whis-**

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3000–3016

per 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-stage decoding strategy*, where the decoder first generates an ASR transcript from speech, then reuses it as context for translation, perhaps emulating human thought processes, even when a transcript is not provided.

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

Experiments on CoVoST2's (Wang et al., 2020b) French-English (`fr-en`) and German-English (`de-en`) subsets demonstrate strong performance. Extended evaluations on conversational telephone speech (CTS) corpora—Fisher-CallHome Spanish (Post et al., 2013), and BBN Mandarin-English (Wotherspoon et al., 2024) further confirm the robustness of our approach across diverse domains. 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 supervision. 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.

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 audio-augmented text generation conditioning on *both inputs*. Our work extends recent work showing that the Whisper can be adapted via fine-tuning to perform a number of novel tasks including, audio-visual speech recognition (Rouditchenko et al., 2024), target-speaker ASR (Guo et al., 2024; Polok et al., 2024; Ma et al., 2024a), translation to non-English languages (Peng et al., 2023), by showing that Whisper can be extended to enable multi-modal translation, i.e., using either only text or both text and speech inputs simultaneously.

2.2 Multi-modal/-task Speech Systems

Recent developments in multi-modal and multi-task systems, e.g., (Tang et al., 2021), are exploring new ways to combine audio and text to improve various language-related tasks. 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 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, T2T translation, T2S translation, S2T translation and S2S translation, it consists of multiple building blocks to leverage uni-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 for speech, with a vocoder for converting the unit-sequences to waveforms. These systems,

[1]We open source our code at https://github.com/BorrisonXiao/Whisper-UT.

along with most existing methods, primarily seek to simply align the representations of the text and speech modalities, limiting the model to still accept only one input modality at a time during inference, which prevents exploitation of *cross-modal cues*.

More recently, speech-centric large language models such as QWen-Audio (Chu et al., 2024) have shown that a unified decoder can be fine-tuned for a broad spectrum of text-conditioned speech tasks—including contextual ASR (Xiao et al., 2025)—but these approaches rely on massive pre-trained text LLMs and demand extensive data and compute during fine-tuning. This is a gap we aim to fill.

A number of related works (Ma et al., 2024b; Zhang et al., 2023; Liu and Niehues, 2024; Le et al., 2024) have also demonstrated that multi-task learning can greatly improve speech translation performance. Here, we focus on model fine-tuning and demonstrate that training end-to-end models for either ASR or ST alone improves performance on the other task, enabling fine-tuning with data that was not original annotated for the target domain task.

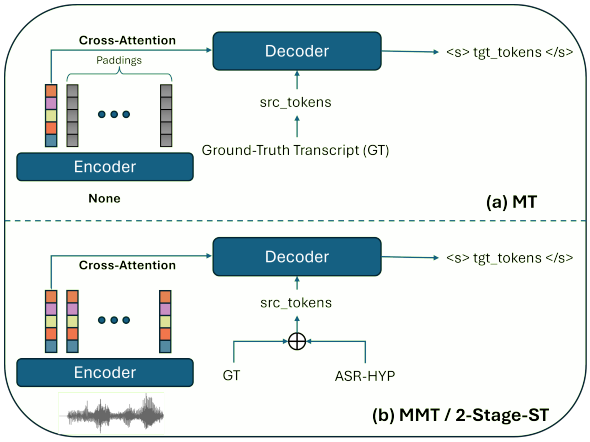

Figure 1: **Overview of our approach.** *ASR-HYP* refers to the ASR hypothesis generated. When GT is used, the task is MMT, otherwise it is referred as 2-Stage-ST. The $\oplus$ symbol refers to the XOR operation. Note that special tokens are omitted to simplify illustration.

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 $P(Z|X, Y)$.

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:

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

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.,

$$\sum_{Y'} P(Z|Y', X)P(Y'|X) \approx \max_{Y'} P(Z|Y', X)P(Y'|X),$$

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

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

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

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

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

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

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

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

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 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, oftentimes the pre-trained encoder is designed specifically for processing speech features, making direct text encoding potentially ineffective. 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-like systems, 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-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.

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.

MT: Translating $Y \rightarrow Z$.

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 Error Simulation in Multi-Modal Translation

For MMT, we introduce an ASR error simulation mechanism to enhance robustness. With probability b, we perturb a batch by replacing the source language tokens, sampled with probability t, with a similar alternative sampled randomly from the top-k nearest neighbors in the embeddings space. To explicitly signal perturbed inputs, we prepend

a special token to the modified sequence, allowing for the model to dynamically re-weight its reliance on the noisy text prefix and the corresponding audio input at inference time. This aims to simulate 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, ASR error simulation is applied with a certain probability to mimic 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 Tasks and Datasets

We test our approach on CoVoST2, a general-domain speech translation benchmark, using its French-English (180 hours) and German-English (119 hours) subsets for training. To assess performance on challenging conversational telephony speech (CTS), we conduct experiments on the Fisher-CallHome Spanish-to-English corpus (186 hours of spontaneous Spanish dialogues) and the BBN Mandarin-to-English corpus (110 hours of Mandarin-English telephony conversations). This setup tests our method's adaptability across both general and domain-specific speech, with CTS posing unique challenges such as disfluencies, code-switching, and informal dialogue structures.

4.2 Evaluation

For both ASR and ST, we normalize the text by lower-casing all characters and removing all punctuations before computing the metrics. For the Fisher Spanish corpus, the BLEU score is computed with multiple references using the Moses (Koehn et al., 2007) toolkit as reported in other work (Weiss et al., 2017a). The evaluation script used is provided in the code.

4.3 Training

To demonstrate our proposed approach, we adopt the LARGE-V2 version of Whisper with 1.6 billion parameters as the base model and fine-tune it for our unified translation modeling. To enable joint training of speech-to-text and text-to-text translation within a single framework, we repurpose the 3-way parallel dataset by strategically replicating its text pairs. Specifically, we create a duplicate of the original dataset where the audio signals are removed, retaining only the source-target text pairs. This allows us to simulate text-only data without introducing external resources, ensuring parity in training scale across objectives.

4.4 Experimental Results

Table 1: Direct Whisper fine-tuning results on the Fisher-Spanish and BBN-Mandarin datasets. The **Objective** column specifies under which training objective the model system is fine-tuned. *None* refers to the original model. Underline highlights the cross-task synergy.

	Dataset	Objective	Task	
			ASR (WER↓)	E2E-ST (BLEU↑)
1	**Fisher**	None	26.7	51.6
2		ASR	19.1	54.9
3		ST	20.3	61.2
4	**BBN**	None	32.2	13.0
5		ASR	18.9	16.2
6		ST	23.1	16.8

4.4.1 Overview

Table 1 presents results from directly fine-tuning Whisper, which reveals a cross-task synergy phenomenon: optimizing for one task (e.g., ASR) not only preserves but often enhances performance on another (e.g., ST), as indicated by underlined improvements across both datasets. Table 2 reports Whisper-UT results on three corpora: CoVoST2 (French → English, German → English), Fisher-Spanish, and BBN-Mandarin. Across all settings, our proposed Whisper-UT variants demonstrate consistent improvements in transcription accuracy (WER↓) and translation quality (BLEU↑).

Table 2: Results on the test sets. *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 with ASR hypothesis.

	Task	Dataset	Model	Task Metrics	Task Results
1	ASR	**CoVoST2** *fr-en* | *de-en*	Baseline (Wang et al., 2020b)	WER↓	18.3 | 21.4
2			Whisper-Large-V2		13.4 | 7.0
3			Whisper-UT		**8.3** | **5.8**
4		**Fisher-Spanish**	SeamlessM4T-Large	WER↓	76.3
5			Whisper-Large-V2		26.7
6			Seq2seq (Weiss et al., 2017b)		23.2
7			Multi-ASR (Inaguma et al., 2019)		22.9
8			STAC-ST (Zuluaga-Gomez et al., 2023)		18.8
9			Whisper-UT		**16.3**
10		**BBN-Mandarin**	SeamlessM4T-Large	WER↓	52.6
11			Whisper-Large-V2		32.2
12			Whisper-UT		**17.4**
13	MT	**CoVoST2** *fr-en* | *de-en*	Baseline (Wang et al., 2020b)	BLEU↑	37.9 | 28.2
14			NLLB-1.3B		**42.3** | **31.0**
15			Whisper-UT		36.5 | 26.9
16		**Fisher-Spanish**	NLLB-1.3B	BLEU↑	48.3
17			Bi-NMT (Inaguma et al., 2019)		**59.6**
18			Whisper-UT		55.9
19		**BBN-Mandarin**	NLLB-1.3B	BLEU↑	8.7
20			Whisper-UT		**15.7**
21	MMT	**CoVoST2** *fr-en* | *de-en*	Whisper-UT	BLEU↑	**46.2** | **40.1**
22		**Fisher-Spanish**	Whisper-UT	BLEU↑	**70.4**
23		**BBN-Mandarin**	Whisper-UT	BLEU↑	**26.0**
24	ST	**CoVoST2** *fr-en* | *de-en*	Baseline (Wang et al., 2020b)	BLEU↑	27.6 | 21.0
25			SeamlessM4T-Large		33.1 | 35.8
26			Whisper-Large-V2		36.7 | 36.8
27			QWen2-Audio (Chu et al., 2024)		38.5 | 35.2
28			Whisper-UT		40.8 | 37.7
29			Whisper-UT-*2-Stage*		**41.4** | **38.1**
30		**Fisher-Spanish**	SeamlessM4T-Large	BLEU↑	14.7
31			Multi-ST (Inaguma et al., 2019)		45.2
32			Multi-task ST/ASR (Weiss et al., 2017b)		48.7
33			Whisper-Large-V2		51.6
34			STAC-ST (Zuluaga-Gomez et al., 2023)		52.6
35			Whisper-UT		62.0
36			Whisper-UT-*2-Stage*		**62.1**
37		**BBN-Mandarin**	SeamlessM4T-Large	BLEU↑	7.0
38			Whisper-Large-V2		13.0
39			Whisper-UT		19.8
40			Whisper-UT-*2-Stage*		**21.6**

4.4.2 Cross-task Synergy

Table 1 reveals 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 (51.6 to 54.9 on Fisher and 13.0 to 16.2 on BBN), and ST fine-tuning reciprocally benefits ASR (26.7 to 20.3 on Fisher and 32.2 to 23.1 on BBN). This suggests that cross-task fine-tuning may mutually reinforce capabilities without architectural changes, inspiring Whisper-UT's unified speech-text framework.

4.4.3 ASR

As shown in Table 2, on CoVoST2, Whisper-UT reduces WER from 13.4/7.0 (Whisper) to 8.3/5.8. Similar gains appear on Fisher (from 18.8 to 16.3) and BBN (from 32.2 to 17.4). These improvements suggest that our stochastic task-interleaving mechanism effectively mitigates catastrophic forgetting, despite the addition of MT and MMT as new tasks. This stability preserves modality-specific expertise while introducing new tasks and enabling cross-task synergy.

4.4.4 MT

In text-only translation, Whisper-UT—trained without architectural modifications—narrowly trails the 1.3B-parameter NLLB model on general-domain CoVoST2 (36.5/26.9 vs. 42.3/31.0 BLEU) but surpasses it by +7.6 and +7.0 BLEU on domain-specific Fisher-Spanish (55.9 vs. 48.3) and BBN-Mandarin (15.7 vs. 8.7) benchmarks, despite using fewer parameters and no dedicated MT pretraining. This divergence highlights two key insights: (1) Whisper's decoder inherently functions as a multilingual language model, capable of text-to-text translation with light-touch adaptation, and (2) its cross-lingual transfer capabilities, honed during speech-centric pretraining, generalize robustly to textual MT in low-resource, domain-specific scenarios. Critically, these results validate our hypothesis that minimal modifications—enabling joint training on speech and text—can unlock Whisper's latent capacity for unified cross-modal translation, bridging the gap between speech and text without sacrificing architectural simplicity.

4.4.5 MMT

When translating with access to both speech and ground-truth transcripts, Whisper-UT achieves 46.2/40.1 BLEU on CoVoST2, 70.4 BLEU on Fisher-Spanish, and 26.0 BLEU on BBN-Mandarin—surpassing all MT baselines. This substantial improvement underscores the complementary nature of audio and text modalities: acoustic cues (e.g., prosody, emotion, pauses, repetitions) resolve ambiguities in noisy transcripts, while lexical context sharpens alignment of speech-derived semantics. By explicitly modeling these mutually compensatory signals, our unified architecture fuses audio and text modalities, yielding more robust translations when multi-modal information is available.

4.4.6 ST

In the ST setting, Whisper-UT achieves competitive performance with single-pass end-to-end decoding: 40.8/37.7 BLEU on CoVoST2 (`fr-en`/`de-en`), 62.0 BLEU on Fisher-Spanish, and 19.8 BLEU on BBN-Mandarin, surpassing QWen2-Audio, SeamlessM4T, and STAC-ST by margins of 2–8 BLEU points. Crucially, the 2-Stage inference variant yields systematic improvements over promptless decoding: +0.6/+0.4 BLEU on CoVoST2 (41.4/38.1 vs. 40.8/37.7), +0.1 BLEU on Fisher-Spanish (62.1 vs. 62.0), and +1.8 BLEU on BBN-Mandarin (21.6 vs. 19.8). These improvements are amplified in error-prone conditions, reflecting successful mitigation of ASR error propagation—a key challenge in cascaded systems. By prepending the special token during training (with simulated ASR noise) and inference (for 2-Stage decoding), the model learns to conditionally distrust imperfect transcripts while retaining their partial utility, rebalancing reliance on audio signals to correct latent errors. These consistent incremental gains validate the effectiveness of our two-stage modeling, demonstrating that even imperfect intermediate transcripts enhance translation fidelity through explicit cross-modal grounding when combined with learned distrust mechanisms.

4.4.7 Summary

The unified Whisper-UT framework achieves robust performance across three key tasks: monolingual ASR, text-only machine translation, and speech translation. Improvements are most pronounced in conversational Mandarin and Spanish settings. Moreover, the 2-Stage decoding strategy provides a reliable way to enhance translation in fully end-to-end deployments. Overall, these results highlight Whisper-UT's ability to unify cross-modal and cross-lingual speech-text tasks within a single architecture, offering a versatile solution for scenarios requiring joint speech-text modeling.

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 error simulation, 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 mitigate 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.

Alexandra Canavan and George Zipperlen. 1996. CALLHOME Mandarin Chinese Speech LDC96S34. Web Download.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

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.

Pascale Fung, Shudong Huang, and David Graff. 2005. HKUST Mandarin Telephone Speech, Part 1 LDC2005S15. Web Download.

Pengcheng Guo, Xuankai Chang, Hang Lv, Shinji Watanabe, and Lei Xie. 2024. Sq-whisper: Speaker-querying based whisper model for target-speaker asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Chenyang Le, Yao Qian, Long Zhou, Shujie Liu, Yanmin Qian, Michael Zeng, and Xuedong Huang. 2024. Comsl: A composite speech-language model for end-to-end speech-to-text translation. *Advances in Neural Information Processing Systems*, 36.

Danni Liu and Jan Niehues. 2024. Recent highlights in multilingual and multimodal speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 235–253.

Hao Ma, Zhiyuan Peng, Mingjie Shao, Jing Li, and Ju Liu. 2024a. Extending whisper with prompt tuning to target-speaker asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12516–12520. IEEE.

Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2024b. Cross-lingual transfer learning for speech translation. *arXiv preprint arXiv:2407.01130*.

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.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. *arXiv preprint arXiv:2305.11095*.

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*.

Alexander Polok, Dominik Klement, Matthew Wiesner, Sanjeev Khudanpur, Jan Černockỳ, and Lukáš Burget. 2024. Target speaker asr with whisper. *arXiv preprint arXiv:2409.09543*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

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.

Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. *arXiv preprint arXiv:2406.10082*.

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.

Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. 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 5301–5311, Online. Association for Computational Linguistics.

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. 2017a. Sequence-to-sequence models can directly translate foreign speech. *Preprint*, arXiv:1703.08581.

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

Shannon Wotherspoon, William Hartmann, and Matthew Snover. 2024. Advancing speech translation: A corpus of mandarin-english conversational telephone speech. *Preprint*, arXiv:2404.11619.

Cihan Xiao, Zejiang Hou, Daniel Garcia-Romero, and Kyu J Han. 2025. Contextual asr with retrieval augmented large language model. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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.

Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Rethinking and improving multi-task learning for end-to-end speech translation. *arXiv preprint arXiv:2311.03810*.

Juan Pablo Zuluaga-Gomez, Zhaocheng Huang, Xing Niu, Rohit Paturi, Sundararajan Srinivasan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2023. End-to-end single-channel speaker-turn aware conversational speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7255–7274, Singapore. Association for Computational Linguistics.

A Training Detail

A.1 Parameter Efficient Fine-tuning

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

In order to fit the base model into our hardware, we adopt 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.

A.2 Hyperparameter Settings

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

Hyperparameter	Value
LoRA Rank	200
LoRA Alpha	400
LoRA Dropout	0.1
Max Training Steps	10000
Batch Size	64
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 A.1 render the use of 8 GPUs redundant, yet they are deployed to accelerate the training process.

A.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.

B CTS Data Detail

B.1 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.

B.2 BBN-Mandarin Data Specification

The BBN Mandarin-English conversational telephony speech (CTS) corpus used in our experiments comprises two primary components:

- **HKUST Mandarin ASR Dataset** (90.1 hours): Mandarin conversational speech from telephony interactions, originally designed for ASR research (Fung et al., 2005).
- **CallHome Mandarin ASR Dataset** (20.5 hours): Informal Mandarin dialogues curated for ASR study (Canavan and Zipperlen, 1996).

The BBN team (Wotherspoon et al., 2024) translated these into English to create parallel speech-to-text translation pairs. While our experiments utilized a pre-publication version provided directly by the BBN authors, minor discrepancies (e.g., data splits, preprocessing, or translation refinements)

Table 4: Code-switching example with system outputs.

REF-ASR:	电脑的 MASTER 应该是很 POPULAR 就对了很应该很
HYP-ASR:	电脑的 master 应该是很 popular 就对了很应该很
REF-MT:	MASTER degree of computer science it should be very POPULAR it should be
HYP-E2E-ST:	The computer should be very popular, should be very
HYP-2-Stage-ST:	The computer's master should be very popular that's right very should be very
HYP-MMT:	The computer's MASTER should be very popular that's right very should be very

may exist compared to the final published version. Nevertheless, the corpus retains its core characteristics: conversational telephony domain focus, code-switching prevalence, and disfluency patterns.

C Qualitative Analysis of Code-Switching

The code-switching example presented in Table 4 demonstrates two critical insights:

- **ASR Preservation of Linguistic Salience:** The 2-Stage decoding system successfully retains the code-switched terms "master" and "popular" (WER $\approx 0\%$ for these tokens), while E2E-ST completely omits "master". This suggests that: 1) direct audio-to-translation mapping struggles with lexical disambiguation of homophones ("master" vs. contextually expected "computer"), and 2) explicit intermediate ASR provides discrete textual anchors that guide translation decisions.
- **Cross-Modal Faithfulness:** While the reference MT (REF-MT) omits the final "很" (translated as "very") from the source utterance "很应该很", our ASR transcript preserves all repetitions. This discrepancy highlights how audio-derived prosodic cues (e.g., emphatic stress on the final "很") enable 2Stage-ST and MMT to retain pragmatic emphasis ("...that's right very should be very") where text-only MT truncates for conciseness. By aligning acoustic signals (stress patterns) with textual redundancy, our framework distinguishes intentional repetition—a discourse marker of conviction in Mandarin—from superficial noise, demonstrating superior faithfulness to both linguistic content and pragmatic intent compared to E2E ST pipelines.

The example validates our hypothesis that two-stage processing particularly benefits scenarios where: 1) ASR can reliably capture linguistically salient content (code-switches, proper nouns), and 2) Audio signals contain complementary paralinguistic information (prosodic boundaries, emphasis) that each modality alone cannot convey. This dual-modality advantage explains 2-Stage-ST's performance gain over E2E-ST on BBN-Mandarin despite identical model parameters.

D Ablation Study

We conduct ablation experiments presented in Table 5 on the two CTS datasets (Fisher-Spanish and BBN-Mandarin), as their domain-specific challenges—disfluencies, code-switching, and spontaneous dialogue—diverged significantly from Whisper's pretraining data. This allows us to isolate our framework's adaptability beyond pretraining biases and quantify its efficacy in resource-constrained, real-world scenarios.

D.1 Text-only MT Training and Its Effects

Rows 7 and 17 show the results of the MT-only fine-tuning experiment, demonstrating that the model achieves strong text translation performance even with limited in-domain data—BLEU 63.4 on Fisher-Spanish and 16.0 on BBN-Mandarin. This outperforms the original NLLB-1.3B model, though it remains modestly behind its fine-tuned counterpart. 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 61.1/20.4 (Fisher/BBN) BLEU score, despite MMT being a novel objective that the model was not explicitly trained on. In fact, on the BBN corpus, the MT-trained model exhibits MMT capabilities that surpass its original training objective, achieving a BLEU score of 20.4 (MMT) compared to 16.0 (MT). This finding reinforces our earlier observation of cross-task synergy.

Table 5: Ablation studies on the CTS test sets. 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.

	Dataset	Model	Objective	Task (num_beams = 1)				
				ASR (WER↓)	E2E-ST (BLEU↑)	MT (BLEU↑)	MMT (BLEU↑)	2-Stage-ST (BLEU↑)
1	Fisher	NLLB-1.3B	None	–	–	48.3	–	–
2			MT	–	–	**67.3**	–	–
3		Whisper	None	26.7	51.6	–	–	–
4			ASR	19.1	54.9	–	–	–
5			ST	20.3	61.2	–	–	–
6			ASR + ST	16.3	**62.2**	–	–	–
7			MT	60.3	51.0	63.4	61.1	52.4
8			ASR + ST + MT + LM	**16.0**	61.7	55.2	–	–
9			MMT	16.4	57.4	1.4	67.5	58.6
10			UT-OOD	**16.0**	61.5	44.2	70.0	61.6
11			UT-CTS	16.3	62.0	55.9	**70.4**	**62.1**
12	BBN	NLLB-1.3B	None	–	–	8.7	–	–
13			MT	–	–	**22.7**	–	–
14		Whisper	None	32.2	13.0	–	–	–
15			ASR	18.9	16.2	–	–	–
16			ST	23.1	16.8	–	–	–
17			ASR + ST	18.5	20.2	–	–	–
18			MT	37.7	12.7	16.0	20.4	15.5
19			ASR + ST + MT + LM	17.7	20.4	14.8	–	–
20			MMT	17.5	19.5	1.0	25.2	20.6
21			UT-OOD	17.5	**20.6**	11.1	25.3	21.5
22			UT-CTS	**17.4**	19.8	15.7	**26.0**	**21.6**

D.2 Effectiveness of Multi-task Learning

In rows 6 and 17, we conduct straightforward multi-task fine-tuning experiments by duplicating the speech dataset with both ASR and ST supervision, concatenating the datasets, and employing random sampling within each batch. These experiments confirm that multi-task training is beneficial, as it enhances BLEU score from 61.2 to 62.2 and WER is reduced from 20.3 to 16.3 on the Fisher-Spanish corpus. A similar trend is observed on the BBN set as well. This suggests that jointly optimizing multiple relevant objectives allows the model to better capture linguistic patterns and improve generalization across tasks.

D.3 MMT-Multi-task Training and Its Implications

Rows 9 and 20 evaluate MMT-multi-task fine-tuned models, that is, the model is trained with $q = 0$ and $b = 0$. Notably, the MMT inference results outperform even the strong fine-tuned NLLB-1.3B baseline in MT performance, 70.4 vs. 67.4 on Fisher and 26.0 vs. 22.7 on BBN— demonstrating that MMT provides tangible benefits over traditional cascaded MT approaches.

However, a gap remains between different MMT settings. Specifically, when using the ASR hypothesis as input instead of the ground-truth transcript, i.e., the 2-Stage-ST decoding, performance drops from 67.5 to 58.6 on Fisher and from 25.2 to 20.6 on BBN. While this still exceeds the results from direct ST (52.4 vs. 51.0 on Fisher and 20.6 vs. 19.5 on BBN), the model tends to over-rely on the transcript in the absence of explicit modeling. Specifically, without the special tag to signal potential errors, the model treats the input transcript as fully reliable ground truth—an assumption that breaks down when using ASR outputs, which may contain recognition errors. These highlight both the effectiveness of explicit modeling and the limitations introduced by ASR errors.

D.4 Unified Translation (UT) Training

D.4.1 Overview

Finally, the UT-trained system (rows 11 and 22) achieves the best MMT and 2-Stage-ST results, with MMT reaching 70.4/26.0 BLEU and 62.1/21.6 BLEU, respectively, on the Fisher-Spanish and

BBN-Mandarin corpora, proving the method's effectiveness. Applying the error simulation strategy in this training scheme improves the robustness of the two-stage approach, narrowing the performance gap between MMT and 2-Stage-ST decoding. Specifically, on Fisher, the gap decreases from 8.9 to 8.3 BLEU (row 9 vs. 11), and on BBN, from 4.6 to 4.4 BLEU (row 20 vs. 22), indicating more stable performance under ASR-transcript input.

D.4.2 Analysis of Transcript-Conditioning

On the Fisher test set, the 2-Stage-ST decoding strategy of the Whisper-UT model actually falls slightly behind the simpler ASR+ST multi-task E2E-ST model. Direct multi-task training of ASR and ST (row 6) achieves a BLEU of 62.2, whereas conditioning on ASR hypotheses under the unified-translation objective (row 11, 2-Stage-ST) yields 62.1—a 0.1 BLEU drop. Through manual inspection, we found this gap is driven largely by inconsistent translation of filler words: the same Spanish filler (e.g., "eh," "um") in ASR transcripts is rendered inconsistently in output, magnifying ASR transcript "errors" during translation. Moreover, because Whisper's ASR and ST performance on Fisher Spanish are both strong already (WER $\approx$ 16, BLEU $\approx$ 60), there is little mismatch for transcript conditioning to resolve, so the transcript signal offers marginal benefit.

In contrast, on the BBN corpus, the UT model demonstrates a clear advantage. The ASR+ST multi-task E2E-ST model (row 16) scores 20.2 BLEU, while the Whisper-UT 2-Stage-ST decoder (row 22) jumps to 21.6 BLEU—a significant 1.4-point gain. This larger benefit arises because BBN combines relatively low WER ($\approx$ 18) with much lower translation quality (BLEU $\approx$ 20), indicating that the model's ST ability lags behind its ASR competence. In this scenario, explicitly leveraging ASR transcripts helps fill the performance gap, yielding more accurate translations under the unified objective.

D.5 Impact of Out-of-Domain Text Data

D.5.1 Dataset Setup

To evaluate the robustness of our unified framework to domain shifts in text data, we replace the in-domain machine translation (MT) pairs (derived from CTS audio transcripts, as described in Section 4.3) with out-of-domain (OOD) text pairs. Specifically:

Spanish: We use 197 hours of text pairs from three sources:

- CoVoST 2 (Wang et al., 2020b) (diverse web-mined speech),
- mTEDx (Salesky et al., 2021) (TED talk subtitles), and
- Europarl-ST (Koehn, 2005) (parliamentary proceedings).

Mandarin: We include 130 hours from:

- CoVoST (Wang et al., 2020a) (multilingual web content),
- GALE (Song et al., 2016) (broadcast news and interviews), and
- proprietary in-house datasets (mixed genres).

The OOD sets contrast sharply with CTS data in domain (e.g., formal talks vs. casual dialogues) and lexical style. To isolate the effect of data domain (not scale), we match the total training steps to our baseline CTS experiments, ensuring comparable optimization cycles. This setup tests whether cross-modal alignment generalizes to heterogeneous text distributions.

D.5.2 Analysis of OOD Text Data Injection

Injecting out-of-domain text under the unified objective appears to have limited benefit and in some cases even disrupted established behaviors. On Fisher, UT-OOD (row 10) lags behind UT-CTS across every translation metric—most notably MT accuracy, which jumps from 44.2 BLEU with OOD data to 55.9 BLEU when text is drawn from the CTS domain. This suggests that the linguistic and stylistic mismatch of web-mined, TED talk, and parliamentary text fails to reinforce the speech-to-text alignment learned on conversational telephone speech, and may inject conflicting patterns that the model struggles to reconcile.

A similar story unfolds on BBN. On BBN, the impact of injecting OOD text is most pronounced in the MT task. Under UT-OOD (row 21), the model's MT performance barely improves over the base unified setting and remains far below the CTS-matched variant—rising only to 11.1 BLEU compared with 15.7 BLEU for UT-CTS (row 22). In contrast, UT-CTS consistently lifts MT and MMT performance by several BLEU points and

Model	Pre-training				Fine-tuning (hrs)			
	Speech (hrs)			Text				
	ASR	ST	Total	(token | sentence)	3-Way	ASR-only	ST-only	Total
Whisper-large-v2	555k	126k	680k	–	–	–	–	–
NLLB-1.3B	–	–	–	N/A | $>$ 40B	–	–	–	–
SeamlessM4T-Large	N/A	N/A	$>$ 1M	N/A | $>$ 40B	N/A	N/A	N/A	$>$ 400k
STAC-ST	–	–	–	–	206	–	–	206
Bi-NMT	–	–	–	–	206	–	–	206
Multi-ST	–	–	–	–	472	–	–	472
Multi-ASR	–	–	–	–	269	–	–	269
QWen2-Audio	N/A	N/A	$>$ 5M	2.4T | N/A	N/A	N/A	N/A	520k
Whisper-UT (Ours)	555k	126k	680k	–	110 $\sim$ 180	–	–	110 $\sim$ 180

Table 6: Comparison of pre-training and fine-tuning data scales for Whisper-UT and baseline models. "3-Way Parallel" refers to datasets with aligned speech, transcripts, and translations. Note that "N/A" means some data is used for the specific training, yet the exact amount is not available.

slightly improves ASR quality. Together, these findings imply that substituting in-domain transcripts with heterogeneous text corpora does not generalize well in a cross-modal training regime and can inadvertently weaken the model's ability to leverage the unified translation objective.

E Model Training Data Overview

Here, we also present a rough sketch of the training data amounts for our model and the compared methods, as summarized in Table 6. However, it is important to note that due to differences in training methodologies, stages, and the unavailability of precise details for some systems, this comparison should be interpreted with caution and may contain ambiguities. We encourage readers to consult the original publications for more accurate and comprehensive descriptions of the training data used in each model.