

StreamUni: Achieving Streaming Speech Translation with a Unified Large Speech-Language Model

Shoutao Guo^{1,3}, **Xiang Li**⁴, **MengGe Liu**^{4†}, **Wei Chen**⁴, **Yang Feng**^{1,2,3*}

¹Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)

² Key Laboratory of AI Safety, Chinese Academy of Sciences

³ University of Chinese Academy of Sciences, Beijing, China

⁴ Li Auto

guoshoutao22z@ict.ac.cn, lixiang39@lixiang.com, fengyang@ict.ac.cn

Abstract

Streaming speech translation (StreamST) requires determining appropriate timing, known as policy, to generate translations while continuously receiving source speech inputs, balancing low latency with high translation quality. However, existing StreamST methods typically operate on sentence-level speech segments, referred to as simultaneous speech translation (SimulST). In practice, they require collaboration with segmentation models to accomplish StreamST, where the truncated speech segments constrain SimulST models to make policy decisions and generate translations based on limited contextual information. Moreover, SimulST models struggle to learn effective policies due to the complexity of speech inputs and cross-lingual generation. To address these challenges, we propose StreamUni, which achieves StreamST through a unified Large Speech-Language Model (LSLM). Specifically, StreamUni incorporates speech Chain-of-Thought (CoT) in guiding the LSLM to generate multi-stage outputs. Leveraging these multi-stage outputs, StreamUni simultaneously accomplishes speech segmentation, policy decision, and translation generation, completing StreamST without requiring massive policy-specific training. Additionally, we propose a streaming CoT training method that enhances low-latency policy decisions and generation capabilities using limited CoT data. Experiments demonstrate that our approach achieves state-of-the-art performance on StreamST tasks.

Code — <https://github.com/ictnlp/StreamUni>

Datasets —

<https://huggingface.co/datasets/ICTNLP/StreamUni>

Model — <https://huggingface.co/ICTNLP/StreamUni-Phi4>

Introduction

Streaming speech translation (StreamST) (Ma et al. 2019; Ma, Pino, and Koehn 2020a; Dong et al. 2022), known as simultaneous interpretation, generates corresponding translations while continuously receiving incoming source speech

inputs. Given its real-time nature, StreamST is commonly employed in various cross-lingual communication scenarios such as international conferences and real-time subtitles.

Compared to traditional offline speech translation (Gangi, Negri, and Turchi 2019; Alinejad and Sarkar 2020; Lee et al. 2022), StreamST must not only ensure translation quality but also minimize the latency between receiving speech inputs and generating translations (Zhang et al. 2024a). To this end, StreamST requires a generation policy to determine the appropriate timing for outputting each translated word. Additionally, considering that StreamST is often deployed in scenarios lasting tens of minutes to several hours (Ma et al. 2019), and that the relevant content attended to by StreamST is primarily concentrated around real-time inputs (Papi et al. 2024), it becomes necessary to implement a truncation policy that can truncate historical speech inputs and translations. This enables the model to focus on recent speech inputs while preventing information overload that could compromise efficiency. Therefore, an ideal StreamST model requires both an effective generation policy and truncation policy to achieve low latency and high translation quality.

Existing methods primarily belong to simultaneous speech translation (SimulST) rather than StreamST, as they cannot be directly applied to speech streams lasting tens of minutes, but are instead limited to speech clips of only tens of seconds (Tang et al. 2023), which are segmented by upstream modules such as Voice Activity Detection (VAD) (Team 2024). Due to the short duration of speech clips, current SimulST methods focus on the generation policy, which can be broadly categorized into fixed policy and adaptive policy. Fixed policy (Ma et al. 2019; Ma, Pino, and Koehn 2020a) guides the model to alternately read fixed-duration speech chunks and output a predetermined number of words. This approach, which disregards the actual textual content within the speech, typically leads to redundant latency or poor translation quality. Moreover, adaptive policy employs integrate-and-fire (Dong et al. 2022), CTC (Zhang et al. 2024a), and Transducer (Tang et al. 2023) to determine generation policy based on the text density of the input speech, achieving better performance. However, these methods still deliver suboptimal translation quality due to small-scale Transformer (Vaswani et al. 2017) architectures.

[†]This work is done when MengGe Liu works in Li Auto.

*Corresponding author: Yang Feng.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

More recent work attempts to leverage the powerful generation capabilities of Large Speech-Language Models (LSLMs) for SimulST, delivering superior performance. These methods either adopt fixed policy (Agostinelli et al. 2024) or adaptive policy achieved by fine-tuning LSLMs with extensively constructed policy-specific data to enable autoregressive policy prediction (Wang et al. 2024; Cheng et al. 2024). However, such fine-tuning methods not only compromise the inherent generation capabilities of LSLMs but also present difficulties in efficiently transferring to newly advanced LSLMs. Therefore, existing SimulST methods face substantial challenges in enabling LSLMs to conduct effective generation policy learning. Moreover, current methods have limited exploration of truncation policy and must be combined with upstream segmentation models to achieve StreamST. This cascaded strategy constrains both the scope of policy-decision and translation quality for SimulST models, making it difficult to achieve holistic StreamST optimization. Consequently, exploring the adoption of a unified LSLM to implement StreamST has emerged as a highly promising research paradigm.

Despite its advantages, implementing StreamST using a unified LSLM remains challenging, as it requires LSLM to simultaneously handle truncation and generation policies while achieving real-time translation. To determine generation policy, LSLMs need to detect valid content in real-time speech stream and decide on the optimal generation timing and output translations (Dong et al. 2022). As the speech stream grows, LSLMs require the truncation policy to discard historical speech segments and translations, ensuring the model focuses on recent inputs while avoiding excessive computational overhead (Papi et al. 2024). Truncation policy must ensure that discarded speech segment is fully translated and that discarded translations accurately correspond to the discarded speech segments, thereby maintaining truncation integrity. Beyond policy decisions, StreamST also needs to accomplish high-quality translation for continuously incoming speech input streams. However, conventional approaches that separately optimize these three subtasks require constructing substantial amounts of corresponding training data (Wang et al. 2024), which is not only resource-intensive but also present significant difficulties in transferring to newly advanced LSLMs. Therefore, investigating how to enable LSLMs to efficiently accomplish all subtasks in a unified manner for effective StreamST is of paramount importance.

To address these challenges, we propose StreamUni, a framework that efficiently enables a unified LSLM to accomplish all subtasks of StreamST in a cohesive manner. StreamUni introduces the speech Chain-of-Thought (CoT) (Huang et al. 2023) that guides LSLMs to progressively generate transcriptions and translations based on the speech inputs. Leveraging multi-stage outputs, the model handles generation policy, truncation policy, and streaming translation generation subtasks. For the generation policy, StreamUni detects effective speech chunks in real-time through intermediate transcriptions to determine optimal generation timing, and decides the current output translation based on the coherence between real-time transcription and previously output translations. For truncation policy, StreamUni

maintains transcription queues across different timestamps and determines speech truncation timing by comparing current and historical transcriptions. Once the source truncation point is identified, StreamUni prompts the LSLM to output complete translations for speech segments preceding the truncation point, subsequently discarding the corresponding translations and speech segments to maintain truncation integrity. The real-time translation generation is obtained by selecting appropriate output translation from the speech CoT based on the generation policy. Through this design, StreamUni achieves StreamST via multi-task results across multiple stages of the speech CoT.

To further enhance streaming performance, we propose a Streaming CoT training scheme that optimizes multi-stage CoT outputs by encouraging LSLMs to predict corresponding transcriptions and complete translations based on partial speech inputs. Therefore, StreamUni unifies all subtasks through the speech CoT and achieves holistic optimization via a unified training strategy. Experiments demonstrate that our method efficiently achieves state-of-the-art performance on StreamST tasks across multiple directions.

Background

Streaming Speech Translation Let the complete speech stream be represented as $\mathbf{s} = (s_1, \dots, s_N)$, where s_i denotes a speech chunk of predefined size, typically around 320ms or 640ms. Given the continuously arriving input speech chunks, the StreamST model progressively generates translation $\mathbf{y} = (y_1, \dots, y_I)$ under a generation policy $\mathbf{g} = (g_1, \dots, g_I)$ where g_i represents the number of speech chunks received when generating y_i . Thus, StreamST can be formulated as:

$$p(\mathbf{y} | \mathbf{s}, \mathbf{g}) = \sum_{i=1}^I p(y_i | \mathbf{s}_{\leq g_i}, \mathbf{y}_{<i}). \quad (1)$$

However, when the incoming speech stream becomes excessively long, StreamST models need to truncate historical speech and translations in real-time, thereby focusing on recent inputs while avoiding excessive inference latency (Irango-Sánchez et al. 2024). Consequently, truncation policy is employed to determine truncation timing. Let the truncation policy for the overall speech input and target translation be $\mathbf{a} = (a_1, \dots, a_M)$ and $\mathbf{b} = (b_1, \dots, b_M)$ respectively, where M denotes the desired number of truncated segments, and a_m and b_m represent the ending positions of the m -th segment within the complete input stream and translation. Under the guidance of the segmentation policy, StreamST is reformulated as:

$$\begin{aligned} p(\mathbf{y} | \mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{b_1} p(y_i | \mathbf{s}_{1:g_i}, \mathbf{y}_{1:i-1}) \\ &+ \sum_{m=2}^M \sum_{i=b_{m-1}+1}^{b_m} p(y_i | \mathbf{s}_{a_{m-1}+1:g_i}, \mathbf{y}_{b_{m-1}+1:i-1}), \end{aligned} \quad (2)$$

where the streaming translation generation will be based solely on the input speech segment and output translation segment that remain after truncation. Therefore, StreamST

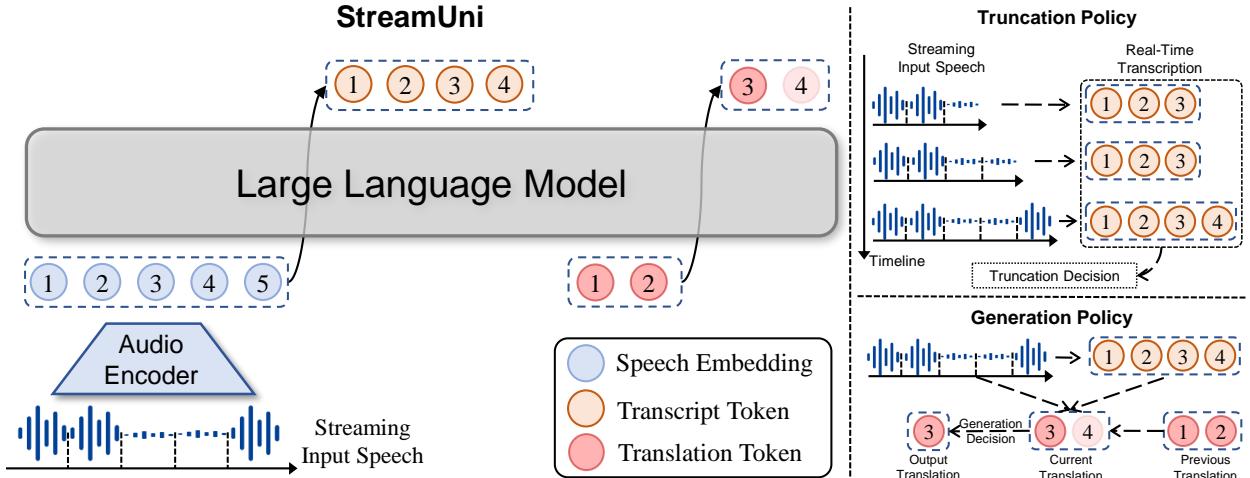


Figure 1: The model framework of StreamUni and illustration of truncation policy and generation policy.

requires determining both truncation and generation policies to guide the model in accomplishing translation generation.

Chain-of-Thought Instruction Chain-of-Thought (CoT) is originally developed for text-based tasks and has been proven to enhance performance on complex tasks by prompting large language models (LLMs) to think step by step before providing final results (Wei et al. 2022; DeepSeek-AI et al. 2025a). For speech inputs, CoT techniques have been widely adopted in speech-to-text cross-modal tasks, where LSLMs first generate transcription and subsequently produce the final outputs (Zhang et al. 2023; Huang et al. 2023). In the context of speech translation, the model first generates transcription $\mathbf{x} = (x_1, \dots, x_J)$, followed by the translation:

$$p(\mathbf{y} | \mathbf{s}) = p(\mathbf{y} | \mathbf{x}, \mathbf{s})p(\mathbf{x} | \mathbf{s}). \quad (3)$$

Method

In this section, we propose StreamUni, a framework that leverages speech CoT to consolidate all subtasks in StreamST. We begin by introducing the architecture of StreamUni and detailing its operational process for achieving StreamST. Subsequently, we present the truncation and generation policies within the StreamUni framework, which governs the management of historical speech and translation and the real-time translation generation. To further enhance the generation capabilities of LSLMs across multiple CoT stages under low-latency conditions, we propose a novel streaming CoT training scheme. The following subsections detail our methodology.

Model Framework

The model framework of our approach is illustrated in Figure 1. Given that the previous truncation timing of the input speech stream is a_m , and the current timing is n ($n > a_m$), the currently received speech segment fed into the model can be represented as $\mathbf{s}_{a_m+1:n}$. For segment $\mathbf{s}_{a_m+1:n}$, the LSLM

first utilizes an audio encoder to encode it into speech embeddings. Following the speech CoT instruction, LSLM subsequently generates real-time transcription $\mathbf{x}^{(n)}$ of $\mathbf{s}_{a_m+1:n}$ based on the speech embeddings. StreamUni then determines the truncation policy by comparing $\mathbf{x}^{(n)}$ with maintained historical transcription queue, specifically deciding whether the current timing n should trigger truncation. If it is determined that the current timing n should trigger truncation, StreamUni disregards the generation policy, and continues generating and outputting all subsequent translation based on the input segment $\mathbf{s}_{a_m+1:n}$ and real-time transcription $\mathbf{x}^{(n)}$, building upon the already output translation segment $\mathbf{y}_{b_m+1:i-1}$, where b_m is the translation truncation index corresponding to a_m . Otherwise, StreamUni determines the generation policy based on the real-time transcript $\mathbf{x}^{(n)}$ and uses it to determine the number of output words at current timing. We then elaborate on the generation policy and truncation policy in detail.

Truncation Policy StreamST employs a truncation policy to remove historical speech and translation segments no longer required for subsequent generation. To ensure truncation integrity, each truncated speech segment must maintain semantic alignment with its corresponding translation segment (Iranzo-Sánchez et al. 2024). The above truncation constraints serve dual purposes: (1) preventing the eliminated speech segment containing untranslated content, which would compromise generation quality, and (2) avoiding removal of already-translated content of remaining speech segment, which will result in repetitive translation of remaining speech segment. According to these, we propose the following truncation policy.

For speech stream $\mathbf{s} = (s_1, \dots, s_N)$, StreamUni obtains real-time transcription after receiving each chunk and maintains a historical transcription queue \mathbf{q} . Assuming the end position of the previous truncated input segment is a_m and the chunk to be processed is n ($n > a_m$), \mathbf{q} can be represented as $[\mathbf{x}^{(a_m+1)}, \dots, \mathbf{x}^{(n-1)}]$. StreamUni first obtains

the transcription $\mathbf{x}^{(n)}$ based on $\mathbf{s}_{a_m+1:n}$:

$$\mathbf{x}^{(n)} = \arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{s}_{a_m+1:n}). \quad (4)$$

Subsequently, we compare $\mathbf{x}^{(n)}$ with items in \mathbf{q} to determine the truncation policy. Speech segment truncation occurs if either condition is satisfied:

- If $\mathbf{x}^{(n)}$ remains identical to real-time transcriptions from the previous two chunks ($\mathbf{x}^{(n-1)}$ and $\mathbf{x}^{(n-2)}$), then $a_{m+1} = n$ becomes the speech truncation timing and $\mathbf{s}_{a_m+1:a_{m+1}}$ is discarded. The historical transcription queue is cleared ($\mathbf{q} = []$).
- If $\mathbf{x}^{(l)} (l = n-1, n-2)$ forms a complete sentence terminated by punctuation (?.!;), and $\mathbf{x}^{(n)}$ begins a new sentence following the complete sentence, then truncation timing is $a_{m+1} = l$ and $\mathbf{s}_{a_m+1:a_{m+1}}$ is discarded. The historical transcription queue is cleared, and the newly generated $\mathbf{x}^{(l)} (l = a_{m+1}+1, \dots, n)$ are sequentially added.

After determining the truncation timing, StreamUni generates and outputs the complete translation corresponding to $\mathbf{s}_{a_m+1:a_{m+1}}$ based on previously output translation:

$$\mathbf{y}_{i:b_{m+1}} = \arg \max_{\mathbf{y}} p(\mathbf{y} | \mathbf{s}_{a_m+1:a_{m+1}}, \mathbf{x}^{(a_{m+1})}, y_{b_m+1:i-1}), \quad (5)$$

where b_{m+1} is the index of the last word in the output translation. The translation segment $\mathbf{y}_{b_m+1:b_{m+1}}$ is discarded.

In conclusion, we select truncation timing when users maintain prolonged silence or finish a full sentence, as content prior to this timing is relatively complete and subsequent translations are unlikely to reference earlier inputs. After determining input truncation timing, target truncation timing is decided by outputting complete translation for the truncated input segment, thereby maintaining semantic integrity of the truncation. **More explanation is in Appendix.**

Generation Policy After establishing the truncation policy, we then determine the generation policy, which controls model output at all timing except truncation moments. The generation policy follows two key principles. First, the model should continue generating translation upon detecting the text within input speech; otherwise, no generation is required (Dong et al. 2022). Second, translation generation should lag behind the input source text to provide sufficient context for translation (Liu et al. 2021). Leveraging speech CoT, we implement the generation policy in Figure 1.

Assume the previous truncated segment is the m -th segment, and the speech chunk to be processed is s_n . We can obtain the transcription $\mathbf{x}^{(n)}$ using Eq.(4). Let C denote the number of words in $\mathbf{x}^{(n)}$ and $i-1$ represent the position of the last word in the already output translation. The number of translation words allowed to be output is:

$$O = C - k - (i - 1 - b_m), \quad (6)$$

where the second term k is the delay hyperparameter, and the third term represents the number of retained output translation words. This setting ensures that translation generation

consistently lags behind the input text by k words, providing sufficient context for generation. The current translation generation can be represented as:

$$\mathbf{y}_{i:i-1+O} = \arg \max_{\mathbf{y}} p(\mathbf{y} | \mathbf{s}_{a_m+1:n}, \mathbf{x}^{(n)}, y_{b_m+1:i-1}). \quad (7)$$

Then the generated translation $\mathbf{y}_{i:i-1+O}$ will be output.

Streaming CoT Training

After introducing the overall model framework, StreamUni can now perform StreamST using existing LSLMs (Microsoft et al. 2025; Xu et al. 2025). However, existing LSLMs are trained on multi-task datasets containing complete speech inputs paired with corresponding responses. In streaming scenarios with continuously growing speech stream, LSLMs must handle speech inputs of different lengths, which we refer to as streaming generation capability. Furthermore, our approach unifies policy decisions and streaming translation generation through speech CoT, which requires enhanced streaming generation capability across multiple stages of speech CoT. Therefore, we propose the Streaming CoT training scheme, which improves the capabilities of policy decision and streaming translation generation by augmenting streaming speech CoT data.

Our method constructs streaming CoT data using existing non-streaming CoT triplets of speech, transcription, and translation. Given the input speech stream $\mathbf{s} = (s_1, \dots, s_N)$, our approach randomly truncates the stream through uniform sampling to obtain $\mathbf{s}_{\leq i}$. We then employ timestamp alignment tools to extract the corresponding transcription $\mathbf{x}^{(i)}$ for $\mathbf{s}_{\leq i}$ from the complete transcription \mathbf{x} . Our Streaming CoT training encourages the LSLM to predict full translation based on partial speech and transcription:

$$\mathcal{L} = - \sum_{\mathbf{s}_{\leq i} \sim \mathcal{U}(\mathbf{S})} \log p(\mathbf{y} | \mathbf{x}^{(i)}, \mathbf{s}_{\leq i}) p(\mathbf{x}^{(i)} | \mathbf{s}_{\leq i}), \quad (8)$$

where \mathbf{S} is $\{\mathbf{s}_{\leq 1}, \dots, \mathbf{s}_{\leq N}\}$ and $\mathbf{s}_{\leq i} \sim \mathcal{U}(\mathbf{S})$ represents uniform sampling from set \mathbf{S} . This formulation trains accurate transcription prediction for policy decisions while requiring complete translation prediction to enhance generation capability and prevent premature termination. For efficiency, we employ sampling rather than training on all possible speech inputs for a instance. Through this training approach, our method efficiently enhance streaming CoT generation capability, thereby improving the capabilities of policy decision and streaming translation generation in low latency. In experiments, our training method requires integration with traditional non-streaming training approaches to achieve greater performance gains.

Experiments

Datasets

We mainly conduct experiments on streaming speech translation (StreamST) and simultaneous machine translation (SimulST) tasks.

MuST-C English⇒German (En⇒De) This dataset (Di Gangi et al. 2019) is collected from TED talks. The dataset contains both document-level and human-annotated

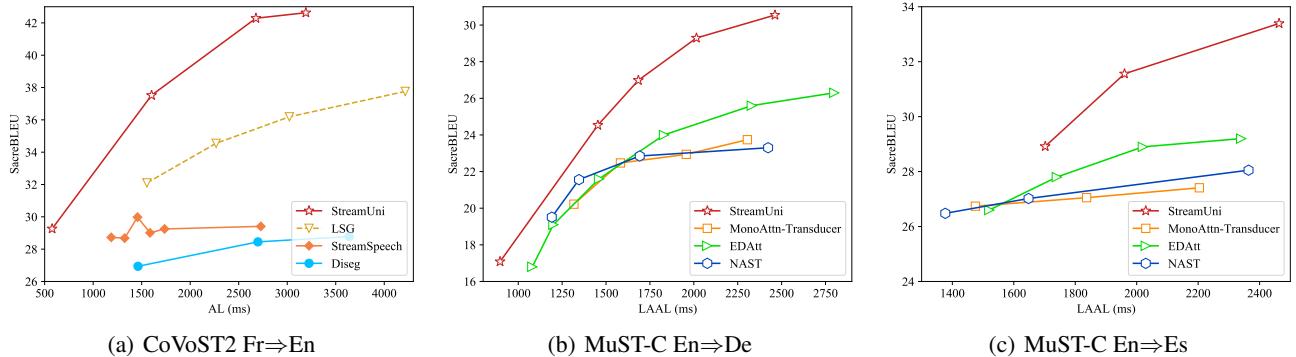


Figure 2: Performance of different methods on SimulST task.

sentence-level speech translation data, enabling evaluation of both SimulST and StreamST tasks.

MuST-C English⇒Spanish (En⇒Es) The dataset is constructed following the same approach as MuST-C En-De and serves as an evaluation benchmark for both StreamST and SimulST tasks.

CoVoST2 English⇒Chinese (En⇒Zh) This dataset only contains sentence-level speech translation data and is used to evaluate SimulST tasks (Wang, Wu, and Pino 2020).

CoVoST2 French⇒English (Fr⇒En) This dataset is also used to evaluate SimulST tasks.

System Settings

In this subsection, we delineate the settings of our StreamUni method and then present the comparative methods for each task separately.

For our approach, we adopt Phi-4-Multimodal (Microsoft et al. 2025) as the primary backbone LSLM and fine-tune it using the speech CoT data across four language directions. Specifically, the En⇒Zh direction contains 50 hours of streaming CoT data and 50 hours of non-streaming CoT data, while the other three directions each comprise 100 hours of non-streaming CoT data. The CoT instruction used for LSLM inference is: ‘Transcribe the audio to text, and then translate the audio to {target.lang}. Use <sep> as a separator between the original transcript and the translation’. During inference, the chunk size is set to 320ms for the En-Zh direction and 640ms for the other directions. To control inference latency, we configure k as {1, 3, 5, 7, 9}. When applied to the SimulST task, StreamUni executes only the generation policy. **Additional training hyperparameters are provided in the Appendix.** Beyond Phi-4-Multimodal, we also experiment with Qwen2.5-Omni (Xu et al. 2025) as the base LSLM to validate the generalizability of our method, leveraging its thinker for policy-decision and translation generation.

For SimulST task, we compare our method with **DiSeg** (Zhang and Feng 2023), **NAST** (Ma et al. 2023), **EDAtt** (Papi, Negri, and Turchi 2023), **StreamSpeech** (Zhang et al. 2024b), **LSG** (Guo et al. 2025) and **MonoAttn-Transducer** (Ma, Feng, and Zhang 2025). We also design a baseline called **Phi4-Wait- k** , which also uses fine-tuned Phi-4

Multimodal as our StreamUni but employs a generation policy that waits for $k-1$ chunks and then outputs one word for each subsequently received chunk.

For the StreamST task, we compare our method with **StreamAttFW** and **StreamAttP** (Papi et al. 2024). Furthermore, we implement a baseline, called **Phi-4-VAD**, that replaces our truncation policy with VAD (Team 2024) while keeping all other components consistent with our approach.

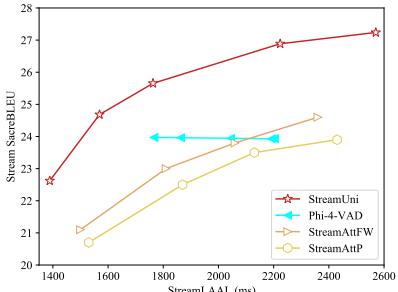
Evaluation

In evaluating streaming generation systems, we need to assess two critical aspects: latency and generation quality. To quantify latency, we utilize the Average Lagging (AL) (Ma et al. 2019) and Length-Adaptive Average Lagging (LAAL) metrics (Papi et al. 2022), which measures the delay between input reception and output generation. For translation quality, we use the SacreBLEU (Post 2018) and COMET (Rei et al. 2022) metrics. For the SimulST task, we employ the SimulEval tool (Ma et al. 2020) to evaluate our StreamUni. In the StreamST task, we follow the setup of Papi et al. (2024). We first use mWERSegmenter (Matusov et al. 2005) for aligning document-level translation with references and then convert these alignments into consistent metrics used in the SimulST task. In this task, the latency metric is termed StreamLAAL, and translation quality is assessed using Stream SacreBLEU.

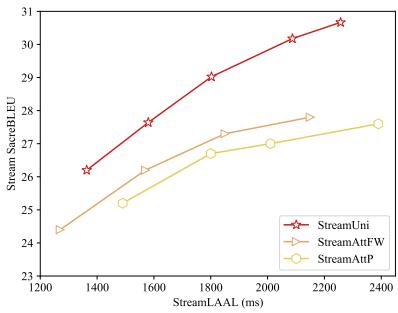
Main Results

We evaluate our methods on SimulST and StreamST tasks.

As illustrated in Figure 2, our method achieves optimal SimulST performance across all datasets. Compared to traditional SimulST approaches employing Encoder-Decoder architectures (e.g., NAST and EDAtt), our method harnesses the comprehension and reasoning capabilities of LSLMs (Microsoft et al. 2025), yielding substantial performance improvements across all latency settings. Although methods like LSG also leverage LSLMs and demonstrate promising results, their policy decisions rely on heuristic rules (Guo et al. 2025), resulting in suboptimal performance. Our method employs intermediate outputs from speech CoT to enable real-time detection of valid user inputs and make generation decisions at optimal timing. Through our superior



(a) MuST-C En⇒De



(b) MuST-C En⇒Es

Figure 3: StreamST performance of different methods.

policy and streaming CoT training scheme, we achieve further performance gains.

Our method also demonstrates superior performance on the StreamST task, as shown in Figure 3. Traditional StreamST approaches, including StreamFW and StreamAttP (Papi et al. 2024), rely on attention interpretability to determine generation and truncation policies for streaming translation. In contrast, our approach utilizes speech CoT for real-time detection of valid speech inputs to inform generation policies, while implementing truncation policies through alignments between speech input and translations. This design enables more effective policy decisions and enhanced performance. Compared to Phi-4-VAD, which employs VAD for truncation policy, our method achieves truncation policy through the semantic alignments between speech inputs and translation, resulting in more appropriate timing and enhanced performance.

Analysis

To provide deeper understandings into our approach, we conduct comprehensive analyses, with each experiment detailed below.

Ablation Study

We first conduct ablation studies to investigate the impact of different configurations on the performance.

Figure 4 presents a performance comparison of our method under various training methods and generation policies. Unlike Phi4-Wait- k , which employs heuristic rules for generation decisions without considering speech con-

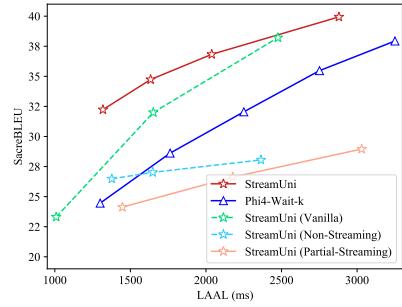


Figure 4: SimulST performance of different generation policies and training methods on CoVoST2 En⇒Zh dataset. The ‘Vanilla’ setting uses no additional training, ‘Non-Streaming’ is trained only on non-streaming data, and ‘Partial-Streaming’ adapts the proposed training method to encourage partial translation prediction rather than full translation. The employed LSM is Phi-4-Multimodal.

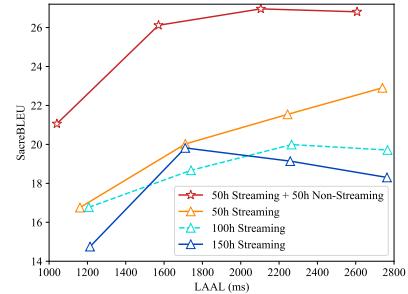


Figure 5: The SimulST performance of different training data recipes on MuST-C En⇒De dataset. Our method utilizes 50 hours each of proposed streaming and non-streaming data, while other methods employ only proposed streaming data of varying durations.

tent (Ma et al. 2019), our method determines generation timing by detecting valid speech inputs, thereby achieving superior performance through more informed generation policies. Beyond generation policy, our proposed streaming CoT training scheme enhances performance across all latency settings, particularly under low-latency. However, the streaming CoT training data must be combined with non-streaming data to achieve maximum performance gains.

To validate this hypothesis, we conduct experiments using training data from a single language direction. As illustrated in Figure 5, simply increasing data volume when using only streaming CoT data fails to yield performance improvements. Superior performance across different latency settings is achieved exclusively when both streaming and non-streaming CoT data are employed simultaneously. We hypothesize that this mixed-data approach effectively stimulates the streaming generation capabilities of model while enabling it to perceive complete speech input boundaries, thereby preventing over-translation and achieving enhanced overall performance.

Furthermore, we investigate the effectiveness of our pro-

Direction	Truncation	COMET	SacreBLEU
En⇒De	Human Model	82.45 83.42	32.51 31.59
	Model	80.83 82.86	35.84 34.97
En⇒Es	Human Model	80.83 82.86	35.84 34.97
	Model	80.83 82.86	35.84 34.97

Table 1: Performance of our method on the MuST-C dataset with document-level speech inputs when the generation policy is disabled and only the truncation policy is employed. ‘Human’ refers to human-annotated truncation timing, while ‘Model’ indicates that the model autonomously determines the truncation policy.

Training Settings	WER (↓)	SacreBLEU (↑)
Streaming CoT + Non-Streaming CoT	20.74	35.22
Non-Streaming CoT	27.83	33.60
Vanilla	31.62	33.34

Table 2: Performance of multiple stages of speech CoT under different training configurations. ‘Streaming CoT + Non-Streaming CoT’ denotes our employed training recipe. ‘Non-Streaming CoT’ only utilizes Non-Streaming CoT training data. ‘Vanilla’ represents the baseline without any further training.

posed truncation policy. Rather than comparing the accuracy of model-determined truncation timing against human-annotated truncation points, we focus on the final generation quality, which represents our ultimate objective. Given document-level speech inputs with an average duration exceeding 10 minutes (Di Gangi et al. 2019), we evaluate the generation quality of our fine-tuned model under different truncation policies. Figure 1 illustrates the results, where we report document-level metrics rather than sentence-level metrics after alignment. Notably, while our proposed truncation strategy performs slightly below the human-annotated policy on SacreBLEU, it surpasses human annotation on COMET score. This demonstrates the effectiveness of our approach and provides valuable insights for future research utilizing semantic alignment models to implement segmentation policies.

Speech CoT Argumentation

StreamUni unifies streaming translation generation and policy decisions through speech CoT. The accuracy of outputs at each CoT stage significantly impacts overall StreamST performance, particularly under low-latency settings. To investigate this, we construct a low-latency speech evaluation dataset based on CoVoST2 En⇒Zh to assess generation capabilities across different CoT stages.

For dataset construction, we randomly truncate each speech clip and obtain transcriptions using WhisperX (Bain et al. 2023), then generate reference translations using DeepSeek-V3-0324 (DeepSeek-AI et al. 2025b). We eval-

Task	Model	LAAL(↓)	SacreBLEU(↑)
ST	Phi-4-MM	N/A	28.55
	Qwen-Omni	N/A	24.21
SimulST	Phi-4-MM	1112.48	22.51
		1448.43	24.27
	Qwen-Omni	949.36	20.64
		1449.83	21.80

Table 3: Performance of various vanilla LSLMs on ST and SimulST tasks. ‘ST’ denotes speech translation that utilizes complete speech inputs for translation, while ‘SimulST’ represents the simultaneous speech translation task that incorporates our proposed generation policy.

uate models trained with different schemes through speech CoT inference. As shown in Table 2, our approach achieves superior performance across all CoT stages, delivering excellent capabilities of policy decision and streaming translation generation.

Extending to Other LSLMs

Beyond the analytical experiments of our method, we further extend our evaluation to Qwen2.5-Omni-7B (Xu et al. 2025) to validate the generalizability of our approach across different LSLMs. The experimental results are presented in Table 3. Phi-4-Multimodal consistently outperforms Qwen-Omni on both ST and SimulST tasks, demonstrating that LSLMs with stronger speech translation capabilities achieve superior SimulST performance. This finding further validates that our StreamUni method can effectively leverage and scale with the enhanced capabilities of LSLMs, thereby demonstrating the generalizability of our approach.

Related Work

Streaming speech translation (StreamST) aims to generate real-time translations for continuously arriving speech stream, requiring the simultaneous completion of generation policy, segmentation policy, and streaming translation generation. Early research focused on sentence-level speech segments and is called simultaneous speech translation (SimulST), predominantly employing encoder-decoder architectures (Vaswani et al. 2017). Initial SimulST methods (Ma, Pino, and Koehn 2020b) determine generation policy based on the number of input chunks. Subsequently, researchers explore content-adaptive generation policy by leveraging auxiliary ASR tasks (Zeng, Li, and Liu 2021; Chen et al. 2021; Zhang et al. 2024b), integrate-and-fire (Dong et al. 2022), monotonic attention (Communication et al. 2023), transducer (Liu et al. 2021; Tang et al. 2023), and CTC (Graves et al. 2006; Ma et al. 2023) to make decisions based on speech content.

With the advancement of Large Speech-Language Models (LSLMs), researchers have begun exploring their application to SimulST tasks (Agostinelli et al. 2024; Guo et al. 2025). However, relying solely on LSLMs for SimulST still requires coordination with multiple auxiliary models

to achieve complete StreamST, introducing cascaded errors and hindering end-to-end optimization (Li et al. 2021). Consequently, researchers have attempted to develop unified methods capable of handling all StreamST tasks within a single model framework. Early attempts utilize attention mechanisms for generation and segmentation decisions (Papi et al. 2024), while subsequent work constructs dedicated policy-specific datasets to enable autoregressive prediction for policy decisions (Cheng et al. 2024). Nevertheless, these approaches suffer from significant challenges in large-scale data construction and advanced model transferability, while facing difficulties in fully leveraging the pre-training capabilities of foundation models.

Conclusion

In this paper, we propose StreamUni, a framework that efficiently enables unified LSLM to accomplish all subtasks of StreamST in a cohesive manner. Experimental results demonstrate that our method efficiently achieves state-of-the-art performance on StreamST tasks across multiple directions.

References

- Agostinelli, V.; Wild, M.; Raffel, M.; Fuad, K. A. A.; and Chen, L. 2024. Simul-LLM: A Framework for Exploring High-Quality Simultaneous Translation with Large Language Models. *arXiv preprint arXiv:2312.04691*.
- Alinejad, A.; and Sarkar, A. 2020. Effectively pretraining a speech translation decoder with Machine Translation data. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8014–8020. Online: Association for Computational Linguistics.
- Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.
- Chen, J.; Ma, M.; Zheng, R.; and Huang, L. 2021. Direct Simultaneous Speech-to-Text Translation Assisted by Synchronized Streaming ASR. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.
- Cheng, S.; Huang, Z.; Ko, T.; Li, H.; Peng, N.; Xu, L.; and Zhang, Q. 2024. Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent. *arXiv:2407.21646*.
- Communication, S.; Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.-A.; Ellis, B.; Elsayhar, H.; Haheim, J.; Hoffman, J.; Hwang, M.-J.; Inaguma, H.; Klaiber, C.; Kulikov, I.; Li, P.; Licht, D.; Maillard, J.; Mavlyutov, R.; Rakotoarison, A.; Sadagopan, K. R.; Ramakrishnan, A.; Tran, T.; Wenzek, G.; Yang, Y.; Ye, E.; Evtimov, I.; Fernandez, P.; Gao, C.; Hansanti, P.; Kalbassi, E.; Kallet, A.; Kozhevnikov, A.; Gonzalez, G. M.; Roman, R. S.; Touret, C.; Wong, C.; Wood, C.; Yu, B.; Andrews, P.; Balioglu, C.; Chen, P.-J.; Costa-jussà, M. R.; El-bayad, M.; Gong, H.; Guzmán, F.; Heffernan, K.; Jain, S.; Kao, J.; Lee, A.; Ma, X.; Mourachko, A.; Peloquin, B.; Pino, J.; Popuri, S.; Ropers, C.; Saleem, S.; Schwenk, H.; Sun, A.; Tomasello, P.; Wang, C.; Wang, J.; Wang, S.; and Williamson, M. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv:2312.05187*.
- DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*.
- DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Guo, D.; Yang, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Zhang, H.; Ding, H.; Xin, H.; Gao, H.; Li, H.; Qu, H.; Cai, J. L.; Liang, J.; Guo, J.; Ni, J.; Li, J.; Wang, J.; Chen, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Song, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhao, L.; Wang, L.; Zhang, L.; Li, M.; Wang, M.; Zhang, M.; Tang, M.; Li, M.; Tian, N.; Huang, P.; Wang, P.; Zhang, P.; Wang, Q.; Zhu, Q.; Chen, Q.; Du, Q.; Chen, R. J.; Jin, R. L.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Xu, R.; Zhang, R.; Chen, R.; Li, S. S.; Lu, S.; Zhou, S.; Chen, S.; Wu, S.; Ye, S.; Ye, S.; Ma, S.; Wang, S.; Zhou, S.; Yu, S.; Zhou, S.; Pan, S.; Wang, T.; Yun, T.; Pei, T.; Sun, T.; Xiao, W. L.; Zeng, W.; Zhao, W.; An, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Li, X. Q.; Jin, X.; Wang, X.; Bi, X.; Liu, X.; Wang, X.; Shen, X.; Chen, X.; Zhang, X.; Chen, X.; Nie, X.; Sun, X.; Wang, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yu, X.; Song, X.; Shan, X.; Zhou, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, Y. K.; Wang, Y. Q.;

- Wei, Y. X.; Zhu, Y. X.; Zhang, Y.; Xu, Y.; Xu, Y.; Huang, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Li, Y.; Wang, Y.; Yu, Y.; Zheng, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Tang, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Wu, Y.; Ou, Y.; Zhu, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Zha, Y.; Xiong, Y.; Ma, Y.; Yan, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Wu, Z. F.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Huang, Z.; Zhang, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Gou, Z.; Ma, Z.; Yan, Z.; Shao, Z.; Xu, Z.; Wu, Z.; Zhang, Z.; Li, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Gao, Z.; and Pan, Z. 2025b. DeepSeek-V3 Technical Report. arXiv:2412.19437.
- Di Gangi, M. A.; Cattoni, R.; Bentivogli, L.; Negri, M.; and Turchi, M. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Dong, Q.; Zhu, Y.; Wang, M.; and Li, L. 2022. Learning When to Translate for Streaming Speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Gangi, M. A. D.; Negri, M.; and Turchi, M. 2019. Adapting Transformer to End-to-End Spoken Language Translation. In *Interspeech 2019*, 1133–1137.
- Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ACM International Conference Proceeding Series*, 369–376.
- Guo, S.; Zhang, S.; Ma, Z.; and Feng, Y. 2025. Large Language Models Are Read/Write Policy-Makers for Simultaneous Generation. arXiv:2501.00868.
- Huang, Z.; Ye, R.; Ko, T.; Dong, Q.; Cheng, S.; Wang, M.; and Li, H. 2023. Speech Translation with Large Language Models: An Industrial Practice. arXiv:2312.13585.
- Iranzo-Sánchez, J.; Iranzo-Sánchez, J.; Giménez, A.; Civera, J.; and Juan, A. 2024. Segmentation-Free Streaming Machine Translation. arXiv:2309.14823.
- Lee, A.; Chen, P.-J.; Wang, C.; Gu, J.; Popuri, S.; Ma, X.; Polyak, A.; Adi, Y.; He, Q.; Tang, Y.; Pino, J.; and Hsu, W.-N. 2022. Direct Speech-to-Speech Translation With Discrete Units. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3327–3339. Dublin, Ireland: Association for Computational Linguistics.
- Li, D.; I, T.; Arivazhagan, N.; Cherry, C.; and Padfield, D. 2021. Sentence Boundary Augmentation for Neural Machine Translation Robustness. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7553–7557.
- Liu, D.; Du, M.; Li, X.; Li, Y.; and Chen, E. 2021. Cross Attention Augmented Transducer Networks for Simultaneous Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Ma, M.; Huang, L.; Xiong, H.; Zheng, R.; Liu, K.; Zheng, B.; Zhang, C.; He, Z.; Liu, H.; Li, X.; Wu, H.; and Wang, H. 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*.
- Ma, X.; Dousti, M. J.; Wang, C.; Gu, J.; and Pino, J. 2020. SIMULEVAL: An Evaluation Toolkit for Simultaneous Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
- Ma, X.; Pino, J.; and Koehn, P. 2020a. SimulMT to SimulST: Adapting Simultaneous Text Translation to End-to-End Simultaneous Speech Translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 582–587. Suzhou, China: Association for Computational Linguistics.
- Ma, X.; Pino, J.; and Koehn, P. 2020b. SimulMT to SimulST: Adapting Simultaneous Text Translation to End-to-End Simultaneous Speech Translation. *arXiv preprint arXiv:2011.02048*.
- Ma, Z.; Feng, Y.; and Zhang, M. 2025. Overcoming Non-monotonicity in Transducer-based Streaming Generation. arXiv:2411.17170.
- Ma, Z.; Zhang, S.; Guo, S.; Shao, C.; Zhang, M.; and Feng, Y. 2023. Non-autoregressive Streaming Transformer for Simultaneous Translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Matusov, E.; Leusch, G.; Bender, O.; and Ney, H. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*. Pittsburgh, Pennsylvania, USA.
- Microsoft; Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; Chen, D.; Chen, D.; Chen, J.; Chen, W.; Chen, Y.-C.; ling Chen, Y.; Dai, Q.; Dai, X.; Fan, R.; Gao, M.; Gao, M.; Garg, A.; Goswami, A.; Hao, J.; Hendy, A.; Hu, Y.; Jin, X.; Khademi, M.; Kim, D.; Kim, Y. J.; Lee, G.; Li, J.; Li, Y.; Liang, C.; Lin, X.; Lin, Z.; Liu, M.; Liu, Y.; Lopez, G.; Luo, C.; Madan, P.; Mazalov, V.; Mitra, A.; Mousavi, A.; Nguyen, A.; Pan, J.; Perez-Becker, D.; Platin, J.; Portet, T.; Qiu, K.; Ren, B.; Ren, L.; Roy, S.; Shang, N.; Shen, Y.; Singhal, S.; Som, S.; Song, X.; Sych, T.; Vaddamanu, P.; Wang, S.; Wang, Y.; Wang, Z.; Wu, H.; Xu, H.; Xu, W.; Yang, Y.; Yang, Z.; Yu, D.; Zabir, I.; Zhang, J.; Zhang, L. L.; Zhang, Y.; and Zhou, X. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743.
- Papi, S.; Gaido, M.; Negri, M.; and Bentivogli, L. 2024. StreamAtt: Direct Streaming Speech-to-Text Translation with Attention-based Audio History Selection. In Ku, L.-W.; Martins, A.; and Srikanth, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3692–3707. Bangkok, Thailand: Association for Computational Linguistics.
- Papi, S.; Gaido, M.; Negri, M.; and Turchi, M. 2022. Over-Generation Cannot Be Rewarded: Length-Adaptive Average

- Lagging for Simultaneous Speech Translation. In Ive, J.; and Zhang, R., eds., *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, 12–17. Online: Association for Computational Linguistics.
- Papi, S.; Negri, M.; and Turchi, M. 2023. Attention as a Guide for Simultaneous Speech Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.
- Rei, R.; C. de Souza, J. G.; Alves, D.; Zerva, C.; Farinha, A. C.; Glushkova, T.; Lavie, A.; Coheur, L.; and Martins, A. F. T. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In Koehn, P.; Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Fraser, A.; Freitag, M.; Graham, Y.; Grundkiewicz, R.; Guzman, P.; Haddow, B.; Huck, M.; Jimeno Yepes, A.; Kocmi, T.; Martins, A.; Morishita, M.; Monz, C.; Nagata, M.; Nakazawa, T.; Negri, M.; Névéol, A.; Neves, M.; Popel, M.; Turchi, M.; and Zampieri, M., eds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 578–585. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.
- Tang, Y.; Sun, A.; Inaguma, H.; Chen, X.; Dong, N.; Ma, X.; Tomasello, P.; and Pino, J. 2023. Hybrid Transducer and Attention based Encoder-Decoder Modeling for Speech-to-Text Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Team, S. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. <https://github.com/snakers4/silero-vad>.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.
- Wang, C.; Wu, A.; and Pino, J. 2020. CoVoST 2 and Massively Multilingual Speech-to-Text Translation. *arXiv preprint 2007.10310*.
- Wang, M.; Vu, T.-T.; Wang, Y.; Shareghi, E.; and Haffari, G. 2024. Conversational SimulMT: Efficient Simultaneous Translation with Large Language Models. *arXiv:2402.10552*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. *arXiv:2503.20215*.
- Zeng, X.; Li, L.; and Liu, Q. 2021. RealTranS: End-to-End Simultaneous Speech Translation with Convolutional Weighted-Shrinking Transformer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.
- Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Zhang, S.; Fang, Q.; Guo, S.; Ma, Z.; Zhang, M.; and Feng, Y. 2024a. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8964–8986. Bangkok, Thailand: Association for Computational Linguistics.
- Zhang, S.; Fang, Q.; Guo, S.; Ma, Z.; Zhang, M.; and Feng, Y. 2024b. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- Zhang, S.; and Feng, Y. 2023. End-to-End Simultaneous Speech Translation with Differentiable Segmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Truncation Policy in StreamUni

Our truncation policy is designed to truncate historical speech inputs and translations in real-time, enabling the model to focus on recent speech inputs while avoiding additional inference costs. To this end, we establish two key principles: (1) preventing elimination of speech segments containing untranslated content, which would compromise generation quality, and (2) avoiding removal of already-translated content from remaining speech segments, which would result in repetitive translations. Our approach first identifies the truncation timing for source speech inputs, then uses this as an anchor to determine the corresponding truncation point for output translations.

For speech inputs, we consider appropriate truncation timing to be when users pause speaking or complete a sentence. Therefore, we design two triggering rules for speech truncation. The first rule targets prolonged user silence, while the second targets moments when users finish speaking a complete sentence. When neither condition is met for an extended period, causing the processed speech duration to exceed a predefined threshold (30 chunks), our method designates this moment as the truncation point.

After determining the speech truncation timing, we identify the corresponding truncation point for target translations to ensure semantic consistency between discarded content on both source and target sides. To achieve this alignment, we instruct the model to output all translations preceding the source truncation point and subsequently discard them.

This approach implements an effective truncation policy that maintains translation quality while ensuring computational efficiency.

Hyperparameters		Settings	
	Base_model	Base_model	Phi-4-Multimodal
LSLM	Training Details	batch_size	32
		learning_rate	4e-5
		weight_decay	0.01
		lr_scheduler	WarmupLR
		betas	(0.9, 0.95)
		optimizer	AdamW
		zero_optimization	stage_2

Table 4: Settings of StreamUni.

Training and Evaluation Details

We provide comprehensive details of our training methodology. For training data construction, we focus on building streaming CoT data for the En \Rightarrow Zh direction and incorporate an equal duration of non-streaming CoT data. For other language pairs, we directly utilize non-streaming data. The dataset released in this work is intended for academic research purposes only. Any commercial use is strictly prohibited. Our training details are detailed in Table 4.

For SimulST evaluation, we employ SimulEval (Ma et al. 2020) as the standard assessment framework. For StreamST evaluation, we first utilize mWERSegmenter (Matusov et al. 2005) alignment tools to map the generated document-level translations to sentence-level references. Subsequently, we compute latency metrics and translation quality on the aligned sentences. We refer to these metrics as Stream LAAL (Papi et al. 2024) and Stream SacreBLEU, respectively.