Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent

Cross Language Agent Team
ByteDance Research

July 20, 2024

Abstract

In this paper, we present **C**ross **L**anguage **A**gent - **S**imultaneous **I**nterpretation, CLASI, a high-quality and human-like Simultaneous Speech Translation (SiST)[1] System. Inspired by professional human interpreters, we utilize a novel data-driven read-write strategy to balance the translation quality and latency. To address the challenge of translating in-domain terminologies, CLASI employs a multi-modal retrieving module to obtain relevant information to augment the translation. Supported by LLMs, our approach can generate error-tolerated translation by considering the input audio, historical context, and retrieved information. Experimental results show that our system outperforms other systems by significant margins. Aligned with professional human interpreters, we evaluate CLASI with a better human evaluation metric, valid information proportion (VIP), which measures the amount of information that can be successfully conveyed to the listeners. In the real-world scenarios, where the speeches are often disfluent, informal, and unclear, CLASI achieves VIP of 81.3% and 78.0% for Chinese-to-English and English-to-Chinese translation directions, respectively. In contrast, state-of-the-art commercial or open-source systems only achieve 35.4% and 41.6%. On the extremely hard dataset, where other systems achieve under 13% VIP, CLASI can still achieve 70% VIP. Demonstrations and human-annotated test sets are available at https://byteresearchcla.github.io/clasi.

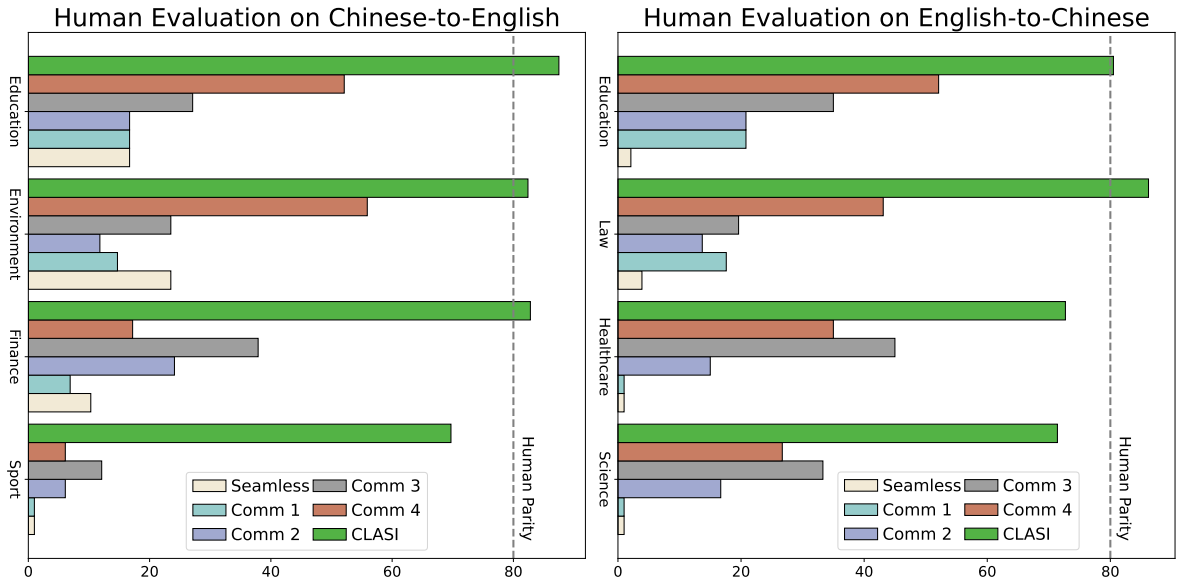

Figure 1: Performance evaluation. CLASI significantly outperforms the leading commercial and open-source systems using a more reliable VIP metric, achieving human interpreter parity.

[1] In this paper, we use Simultaneous Interpretation and Simultaneous Speech Translation interchangeably.

1 Introduction

Simultaneous speech translation (SiST) is recognized as one of the most challenging tasks in the translation domain [33]. Machine-assisted automatic interpretation has been receiving much attention in the natural language processing (NLP) community [18, 19, 66, 65]. Traditional simultaneous translation approaches [11, 24, 89] usually employs a cascaded system, involving a streaming Automatic Speech Recognition (ASR) model, a punctuation model and a Machine Translation (MT) model. However, such cascaded systems often suffer error propagation and latency from the ASR module. Despite these advancements in both academic SiST models [7, 22, 41, 51, 63, 84, 87] and commercial SiST engines, the translation quality is still far from satisfactory. As shown in Figure 1, we conduct a human assessment of the current accessible SiST systems. From the user-centered perspective, these systems only deliver less than 42% of the valid information to listeners, which heavily affects communication effectiveness. In contrast, professional human interpreters usually deliver more than 70% of the necessary information [10] and 95% ideally. Thus in this paper, we use 80% to indicate high-level human interpreters.

Motivated by the huge success of LLMs in machine translation [2, 9] and speech translation [12, 30, 62], we propose to employ LLMs to accomplish the SiST task. Specifically, we identify three primary challenges. First, a key challenge for incorporating LLM into the SiST is the read-write policy, where LLM needs to provide partial translation for input speech. Second, achieving human equivalent performance requires understanding and translation of terminologies and uncommon phrases that LLMs cannot learn from training data. Lastly, the scarcity of training data continues to hinder the performance on the SiST task.

To address these challenges, we introduce our end-to-end approach, CLASI, a **C**ross-**L**ingual **A**gent that accomplishes **S**imultaneous **I**nterpretation by iteratively performing multiple actions, as illustrated in Figure 2. Regarding the first challenge, we imitate professional human interpreters to learn their policy of segmenting a complete sentence into several semantic "chunks" through syntactic boundaries (pauses, commas, conjunctions, etc.) and contextual meaning. To enable CLASI to learn such a policy, we follow a data-driven policy learning process and invite human interpreters to annotate real-world speech, which includes the read-write timing for segmentation. From the data, CLASI learns the robust read-write policy for SiST from humans.

For the second challenge, we include two external modules to augment our CLASI agent: an external knowledge database that stores terminologies and paired translations, and a memory that stores the context of speech. However, the external knowledge database may contain tremendous terms that not only increase the inference time but may also lower the performance of our approach because of noisy intervention. Therefore, we propose a novel Multi-Modal Retrieval Augmented Generation (MM-RAG) process. A multi-modal retriever extracts knowledge from the external database based on the speech input. The retrieved information and the context from memory are then appended to the prompt of our LLM agent to augment the translation through in-context learning.

Addressing the data scarcity of the SiST task, we adopt a three-stage training methodology: pretraining, continual training, and fine-tuning. First, our LLM and audio encoder are independently pretrained on our large-size in-house datasets. Then, our model is continually trained with billions of tokens of mediocre-quality synthesized speech translation data, aiming to align the speech and text modalities. We also include multiple tasks to enhance the in-context learning ability of LLM to better utilize the contextual information from the retriever and prior translation. In the last stage, we fine-tune the model with a small amount of human-annotated data, further imitating professional human interpreters to improve the robustness and translation quality.

In addition, we would like to highlight that the conventional automatic evaluation metrics [52, 54, 61, 68] of simultaneous interpretation might not be good indicators for reflecting the performance of SiST, which often contains compaction, abstraction, and paraphrasing. Aligned with human interpreters [49, 81], we propose a new evaluation metric named Valid Information Proportion (VIP)[2]. VIP represents the percentage of information that can be precisely delivered, reflecting the central objective of SiST: communication in real-time. Through thorough human evaluation on diverse and challenging real-world long speech datasets, our approach outperforms other currently accessible systems by a large margin. As shown in Figure 1, taking the Chinese-to-English direction as an example, CLASI achieves a VIP score of 81.3%, significantly narrowing the gap between machine-assisted systems and human interpreters.

[2]Detailed guidelines of our proposed VIP metric can be found in Appendix A.

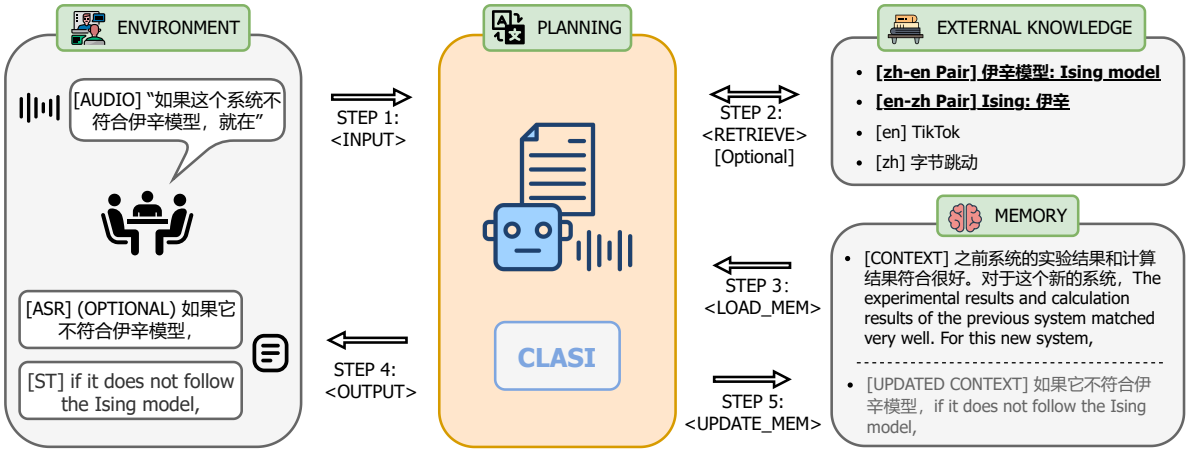

Figure 2: Overall framework of CLASI. The process begins in Step 1, where CLASI processes the incoming audio data. Optionally, the retriever is activated to obtain the relevant information from the external knowledge database. For instance, translating "伊辛模型" to "Ising model" for accurate speech translation. Step 3 involves accessing transcription (optional) and translation in the last round memory. Steps 4 and 5 entail using the Chain-of-Thought (CoT) method to generate both the transcription (optional) and translation, followed by a memory update. The cycle then repeats from Step 1 for the subsequent speech segment.

Our contributions can be summarized as follows:

- We introduce our end-to-end approach, CLASI, an LLM agent that is designed to perform high-quality and human-like simultaneous translation. Through human evaluation, our approach demonstrates significantly better performance compared to existing accessible SiST systems.
- We propose a new data-driven read-write strategy by imitating professional human interpreters. Without the requirement of complicated human pre-design, the strategy could balance translation quality and latency effortlessly. Unlike most commercial systems where the outputs are frequently rewritten during the translation process for better quality, our strategy guarantees all the outputs are deterministic while maintaining high quality.
- Motivated by the preparatory trajectory of human interpreters, we introduce a novel Multi-Modal Retrieval Augmented Generation (MM-RAG) process that empowers the LLM with domain-specific knowledge in real time. The proposed module further improves the translation quality with minimal computational overhead during inference.
- We work closely with professional human interpreters to develop our evaluation strategy, Valid Information Proportion (VIP), and detailed guidelines are open-sourced. Meanwhile, we release a human-annotated test set focusing on diverse real-world scenarios and long speech translations.

2 Methods

2.1 Framework

Figure 2 presents a flow of operation of CLASI. To perform the SiST task, we design 5 operations: `<INPUT>`, `<OUTPUT>`, `<RETRIEVE>`, `<LOAD_MEM>`, and `<UPDATE_MEM>`. The following sections describe the details of each operation. As further illustrated in Figure 3, CLASI is an LLM agent that can take input speech, instruction, relevant information retrieved from external knowledge, and last round memory as context. The memory stores previous transcriptions (optional) and translations. At round r, it first reads speech $\mathbf{x}_{t^{r-1}:T^r}$, where t^{r-1} is the predicted cut-off time of round $r-1$ and T^r is the end time for audio stream at round r. Then the agent retrieves relevant information $\mathbf{k}_r$ from the external knowledge and loads context $\mathbf{y}_{1:r-1}$ from the last round memory. Once CLASI "think" sufficient context is loaded,

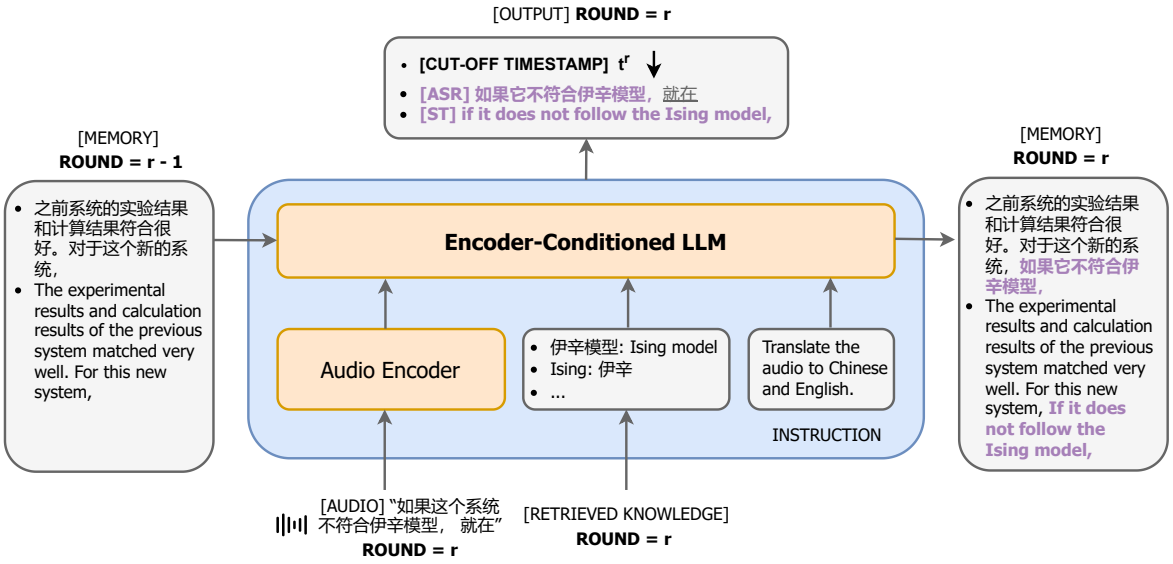

Figure 3: Architecture of CLASI agent. At round r, our model processes the current input audio stream alongside the memory from the previous round $(r-1)$, and any retrieved knowledge. CLASI generates a response based on specified instructions and concurrently updates its memory. Additionally, the model determines the cut-off timestamp of the last semantic chunk. For instance, in the provided example, the phrase preceding "就在" is identified as a complete semantic chunk, with the cut-off timestamp positioned right after this phrase.

it generates the transcription (optional), translation, and cut-off timestamp t^r:

$$\mathbf{y}_r; t^r = \text{TextDecoder}(\mathbf{x}_{t^{r-1}:T^r}, \mathbf{k}_r, \mathbf{y}_{1:r-1}) \tag{1}$$

where t^r is the predicted cut-off timestamp indicating the end time for the current translation round r. $\mathbf{y}_r$ is then forwarded to update the memory. When instructed to output the transcription, the LLM optionally engages CoT to generate transcription first and then the speech translation. For the following round $r+1$, the audio stream begins with the predicted cut-off timestamp t^r.

2.2 Architecture

CLASI employs an Encoder-Conditioned LLM architecture. As shown in Figure 3, the audio encoder transforms input speech stream $\mathbf{x}$ to a series of continuous representations $\mathbf{s}$. Then, the LLM takes the speech representation $\mathbf{s}$, retrieved knowledge $\mathbf{k}$, historical translation $\mathbf{y}$ and instruction $\mathbf{I}$ as a sequence of prompt $(\mathbf{y}, \mathbf{s}, \mathbf{k}, \mathbf{I})$ to generate the translation result $\mathbf{y}$.

Audio Encoder. The audio encoder module contains a large-scale speech conformer [25] pretrained on millions of hours of speech data to achieve human parity performance on ASR, and an audio adapter to connect the audio encoder and LLM. The adapter downsamples the speech representations and the resulting representations are linearly projected to match the dimension of the LLM embedding layer. The projected speech representations lower the computational latency for SiST.

Large Language Model. The language model[3] is a medium size decoder-only transformer [73] to balance performance and computation efficiency. It is pretrained on a large amount of text data and fine-tuned with instructions. The LLM directly takes the continuous embedding from both the audio encoder and text embedder as input. It autoregressively generates the transcription and translation response of the provided speech stream.

Multi-Modal Retriever. The multi-modal retriever framework employs audio and text encoders to independently encode the audio stream and text key of the terminologies in the external knowledge

[3]We use Doubao LLM as our foundation model.

database. To enhance the alignment between audio embeddings and text embeddings, we incorporate an embedding fusion layer, which includes a multi-head attention module followed by a pooling layer. The resulting pooled representation is subsequently fed into a linear projection layer to produce the final scores, indicating the probability of the text key's presence in the audio stream. Terminologies with top scores are forwarded to the CLASI agent to enhance the translation quality.

2.3 Data Driven Read-Write Policy: <INPUT> and <OUTPUT>

Unlike predetermined read-write probabilities and heuristic waiting policies detailed in prior research [7, 45, 39], interpreters engage in a dynamic process of listening (read) and translating (write). They attentively listen to the speaker's speech and segment lengthy sentences into semantic chunks, representing the smallest linguistic units capable of conveying a complete thought independently [33]. Upon identifying a chunk that encapsulates sufficient information, they proceed to translate this segment into the target language, thereby providing an accurate and contextually appropriate translation.

Emulating the strategies of human interpreters, CLASI does not require to explicitly define the read-write policy. CLASI imitates their policies by waiting for complete semantic chunks. Specifically, given partial speech, CLASI only generates the translation for the complete chunks of the input speech. The model is trained with segmented speech data to learn such ability. Mathematically, given source audio $\mathbf{x}_{1:M}$, we segment its translation into a series of n "chunks" $\mathbf{y}_{1:n}$ and obtain the corresponding pair $\{(\mathbf{x}_{t^j:t^{j+1}}, \mathbf{y}_j\}_{j=1}^n$, where $\mathbf{x}_{t^j:t^{j+1}}$ and $\mathbf{y}_j$ denote the j-th segment of the audio and the corresponding translation. For training, our objective is to output all complete segmented translations and the cut-off time given random partial input audio $\mathbf{x}_{1:t}$

$$\min \mathbb{E}_{t\sim\mathcal{U}[1,M]} - \log p_\theta(\mathbf{y}_{1:j}; t^j|\mathbf{x}_{1:t}) \quad j = \max_j\{j|t^j < t\} \tag{2}$$

where $\mathcal{U}$ indicates uniform distribution over time of speech. Trained with Equation (2), CLASI learns to generate the cut-off time for the input speech. Additionally, the objective function makes the CLASI wait for appropriate time before starting translation as the LLM will output nothing when it "think" current speech stream does not contain a complete speech chunk.

2.4 Context Information: <LOAD_MEM> and <UPDATE_MEM>

The memory stores translations and transcriptions in previous rounds $\mathbf{y}_{1:r-1}$. It has two functions. Firstly, it works with the input speech to determine which part of the speech has been translated and which part has not, helping CLASI make the read-write decisions and outputs the translation of the unfinished parts. Secondly, understanding human speech often requires context. For example, when a speaker talks about "barrel bridge", it often refers to the bridges built upon rivers that are supported by barrels. However, in the context of "watch", it refers to a mechanical structure in the watch. The phenomenon of polysemy in different contexts can lead to vastly different translation outcomes. Therefore, CLASI should be able to retrieve the context of the long speech for translating some keywords, and make appropriate translations under different contexts.

As shown in Figure 3, at round r, <LOAD_MEM> forwards relevant translations $\mathbf{y}_{1:r-1}$ to the LLM as a prompt. After CLASI agent generates the translation $\mathbf{y}_r$, <UPDATE_MEM> stores it to the memory and obtains $\mathbf{y}_{1:r}$.

2.5 Multi-Modal Retrieval Augmented Generation: <RETRIEVE>

In real-world scenarios, the accurate speech transcription or translation of professional and domain-specific terminologies is challenging. Even human interpreters require prior domain knowledge to understand those terminologies, including names of people, locations, jargon, or special in-domain terms. For example, an interpreter unfamiliar with the machine learning theory may not recognize the word "Rademacher complexity" when hearing it. Therefore, in various scenarios, human interpreters often prepare in advance to get familiar with the corresponding domain knowledge.

Motivated by the preparatory trajectory of human interpreters, we propose to integrate an external database to empower LLM with necessary domain-specific knowledge. Each item in the database contains a key and the corresponding value in text modality. The key, which may appear in the speech, is used

as the input for the retriever. The value of the item may be itself, a paired translation of the target language, or even an explanation of the key.

Theoretically, all items in the external database might be added into the prompt to provide information for the translation. However, the external knowledge database often contains tremendous items. Simply prompting LLM with all the terms not only increases the inference time but may also hurt the performance of CLASI because of noisy intervention. Therefore, we design a novel Multi-Modal Retrieval Augmented Generation (MM-RAG) process. Our multi-modal retriever first retrieves the relevant terminologies from the database based on the input speech. A small number of filtered items are incorporated into the prompt of CLASI agent for in-context learning as shown in Figure 3.

With the retrieved knowledge and previous context from the memory, our LLM has the in-context learning ability to better utilize the provided contextual information. To achieve this, we collect a series of in-context learning data to train the model. Compared with the previous approaches for intervention, such as shallow-fusion [8, 34] and traditional substitution-based methods [38, 44, 78], which generates fixed translation for given translation pair. Our method achieves better results and generates more coherent text. For example, in some internet companies, "大盘 == overall performance", while in most cases, it should be "stock market". Our method can choose the correct translation given different context. Besides, our method can use monolingual text from both source and target language to help the translation.

3 Multi-Stage Training

Our CLASI follows a multi-stage training process: pretraining, multi-task continual training, and multi-task supervised fine-tuning. In the first stage, the LLM and audio encoder are separately pretrained with massive amounts of in-house speech and text data. Next, a large amount of speech-text paired data is used to align audio and text modalities, building the fundamental capability for cross-modal multitasking. In the final stage, CLASI agent is fine-tuned with a small amount of human-annotated data to imitate the translation behavior of professional human interpreters. Our multi-stage training process enables high efficiency of learning with a small amount of human-labeled data.

3.1 Pretraining

The in-house LLM and audio encoder are independently pretrained on different modalities of data. The LLM follows a decoder-only transformer architecture and is first pretrained on a massive amount of monolingual and bilingual text data with cross-entropy loss and then fine-tuned on instruction-following data. The LLM performs excellently on various downstream tasks, especially translation tasks.

The audio encoder also follows a classic pretrain-finetune paradigm [5, 28, 88] with a massive amount of speech-related data. The pretraining stage provides a proficiently trained LLM and audio encoder, setting a solid foundation for the following stages.

3.2 Multi-task Continual Training

Training Method. We follow the work of [30] for multi-task training. Specifically, for streaming and higher-quality translation, we mainly focus on three tasks for training CLASI: Automatic Speech Recognition (ASR), Speech Translation (ST), and Text Translation (MT). To align the modalities of the pretrained LLM and audio encoder, CLASI is continually trained on various tasks with a substantial volume of paired data. We further strengthen the in-context learning ability of our approach by incorporating translation in the memory and knowledge from external databases. As a result, we expand the ST tasks to different configurations as shown in Table 1. An ST translation can either be streaming or offline, direct or COT, with or without context, which leads to 8 different tasks.

Training Data Construction. Both ASR and MT tasks have been last for a while, and there is a relatively large amount of ASR and MT data. The major challenge of developing an end-to-end SiST model is the data scarcity of simultaneous ST. To this end, we propose a synthetic data construction pipeline. With a strong LLM, we synthesize two types of speech translation data for continual training: offline ST data and context-aware segmented streaming ST data.

Configuration	Explanation
Direct	Speech is directly translated into the target language.
COT	Speech is first transcribed to the source language and then translated to the target language.
Streaming	Given **partial** speech, translate segments with complete semantics to the target language.
Offline	Given **complete** speech, translate the whole content to the target language.
w/o Context	No historical translations and external knowledge are provided.
w/ Context	Historical translations or external knowledge are provided as context.

Table 1: Illustration of different configurations of ST task in Multi-task Continual Training.

1. Offline ST data: We mainly rely on ASR data to construct the offline ST data. Given the ground-truth transcription of speech, we use in-house LLM to translate the source language to target languages. To ensure the readability and conciseness of the target language, the LLM is prompted to conduct Inverse Text Normalization (ITN), filler word smoothing, etc.

2. Context-aware segmented streaming ST data: The streaming ST data consists of fine-grained audio-text alignments and translation pairs for segmented semantic chunks. Compared to offline ST data, streaming ST data is even more challenging to collect. We find that human interpreter often segments long speech into a few semantic chunks, each of which can be translated independently to ensure an effective and smooth translation. Motivated by such findings, we leverage LLM to construct streaming ST data by imitating the chunking process. Long speech data are used to construct the streaming ST data, as the additional history can provide better contextual information. First, we prompt the LLM to break down the ASR transcription into multiple independent semantic chunks, which are then translated into the target language. Subsequently, we align the semantic chunks with the corresponding audio chunks, obtaining the streaming ST data. Such data enable our model to handle incomplete speech inputs and generate partial translation in coherent semantics.

To measure the quality of the synthetic data, we conduct human evaluations based on our proposed VIP metric. The synthetic data achieves a VIP score of 81%, satisfying the minimal requirement for further training.

3.3 Multi-task Supervised Fine-tuning

Training Method. Even though CLASI possesses a good translation quality on the SiST tasks after the previous multi-task continual training stage, we further boost the performance by fine-tuning on human-annotated streaming ST data with diverse tasks listed in Table 1. Such high-quality data enables our model to better align with the segmentation methodologies of professional human interpreters. Furthermore, this process enhances our model's robustness to speech disfluencies such as stuttering, ensuring smoother communication in real-world scenarios.

Training Data Construction. The source of human-annotated streaming ST data originates from real-world scenarios that contain various speech characteristics, such as disfluencies, stuttering, code-mixing, and specialized terminologies. Such features ensure the robustness of our model in diverse conditions. We engage professional human interpreters to provide high-quality annotations for simultaneous segmentation and interpretation of the speech data. Additionally, terminologies are identified and translated within the context, further strengthening the context-aware capabilities of CLASI.

3.4 Multi-Modal Retriever Training

The multi-modal retriever is independently trained with a substantial dataset of speech recognition data. During training, words are randomly selected from speech transcription to serve as the positive sample, indicating their appearance in the speech. Negative words are selected from different sentences, indicating the speech does not mention these words. We assign a label of 1 to positive samples and 0 to negative samples, aiming to minimize the Binary Cross Entropy (BCE) loss. This approach helps refine the model's ability to distinguish relevant from irrelevant information, enhancing its overall performance and accuracy. We label the positive sample as 1 and the negative sample as 0, minimizing the Binary Cross Entropy (BCE) loss. To evaluate the effectiveness of our retriever, we build an in-house retrieve development set. Each sample in the development set includes a short audio chunk and the mentioned

terms in the audio. Note that the term here is defined as special keywords, such as name, location, abbreviation, and domain-specific word.

4 Experiments

4.1 Evaluation Benchmark

Quite a few number of evaluation benchmarks have been proposed for SiST over the past years, including MuST-C [15], FLUERS [14], CoVoST [74, 75], BSTC [85], and GigaST [83], etc. However, although much effort has been spent to build these benchmarks, they still suffer some shortcomings when facing real-world SiST applications.

First, these benchmarks often contain speeches that are either recorded by volunteers (*e.g.* CoVoST and FLUERS) or collected from formally, clearly, and fluently talk and podcasts by well-prepared speakers (*e.g.* MuST-C and GigaST). In real-world scenarios such as online meetings or social media videos, the characteristics of the speech might inevitably be informal, unclear, or disfluent. Second, these benchmarks provide a shortcut for evaluating the translation quality by giving the manually segmented sentences. Such a shortcut offers a gap between the current benchmark and real-world applications, where the models might need to take long speech and conduct segmentation [4] by themselves. Consequently, evaluations on manually segmented datasets are likely to overestimate the performances of a real-world SiST system. These discrepancies result in the evaluation on these benchmarks are not reliable for practical SiST systems.

As a preliminary attempt to address the shortcomings as mentioned earlier, we propose a new benchmark RealSI for Chinese-to-English (zh-en) and English-to-Chinese (en-zh). RealSI is collected from diverse sources, and most speakers talk naturally and casually without careful preparation. We choose 10 popular domains: technology, healthcare, education, finance, law, environment, entertainment, science, sports, and art. One video clip is selected for each domain from a well-known online video platform for both zh-en and en-zh settings.[4] Each sample in RealSI is a nearly 5-minute speech to mock SiST without manual segmentation. For systems that cannot take long-form audio as input, we also provide sentence-level timestamps for segmentation. Table 2 presents the detailed statistics of our RealSI.

Domain	**zh-en**		**en-zh**	
	Duration	#Segments	Duration	#Segments
Technology	5:23	51	3:25	31
Healthcare	3:16	30	3:34	22
Education	4:56	48	5:00	41
Finance	5:22	29	5:01	40
Law	4:38	49	4:48	29
Environment	4:18	34	4:24	31
Entertainment	5:16	53	5:12	39
Science	4:47	37	5:11	35
Sports	5:22	33	3:25	58
Art	7:54	67	4:17	21
Total	51:12	431	44:17	347

Table 2: Statistics of our proposed RealSI benchmark.

4.2 Baselines

We compare CLASI with the open-sourced SiST model, *SeamlessStreaming* [7]. In addition, because of the limited number of available SiST models, we choose to compare CLASI with several commercial systems. We denote the commercial systems as *Commercial* 1-4. It is worth noting that unlike CLASI,

[4] RealSI is available at https://github.com/byteresearchcla/RealSI. We do not own the copyright of the videos and only release our annotations together with the publicly available website links of the corresponding videos. If anyone believes that the content constitutes infringement, please contact us. We will remove the relevant content as soon as it is confirmed.

most of the commercial SiST systems will first generate a temporary translation as soon as possible, then rewrite the temporary translation with a potentially better translation after getting more context. Notwithstanding, we evaluate the finalized translation of these systems in all our experiments. Although this re-writing strategy could improve translation quality, continually revising existing translations might affect the user experience, potentially leading to additional confusion. We would also like to highlight that human interpreters usually do not employ such a rewriting strategy during translation. During the entire evaluation, we employ a general external knowledge database that maintains the same for all the evaluation in this paper. It does not contain domain-specific external knowledge to form unfair comparisons. The improvement of external knowledge is independently reported in Section 4.6.

4.3 Translation Quality

Evaluation Metrics. Automatic evaluation metrics such as BLEU [54], BLEURT [68], and COMET [61] are widely used for evaluating the translation quality [7, 65, 66]. However, they may not be able to fully reflect the quality of the translation, especially for paragraph-level translation of long speech. It is argued that the current evaluation metrics are not sufficient for ST and SiST tasks [23]. The work of [47, 79] also highlighted that there might be a discrepancy between automatic evaluation metrics with human evaluation.

Therefore, besides the automatic evaluation, we collaborated with senior professional human simultaneous interpreters to standardize the guidelines for a more realistic human evaluation. Our proposed human evaluation metric focuses on whether the output of the translation model can accurately convey the speaker's original intention for each semantic fragment. This is also the key objective of human interpreters in real-time translation. Note that a single semantic fragment indicates a complete piece of source speech. Typically, a single semantic fragment is a complete sentence. Detailed definition can be found in Appendix A.2. The percentage of valid information fragments within a complete speech session is defined as VIP, which is consistent with real-world criteria for human simultaneous interpretation [81].

Model	BLEU		BLEURT		COMET		VIP† (%)	
	zh-en	en-zh	zh-en	en-zh	zh-en	en-zh	zh-en	en-zh
SeamlessStreaming	11.3	14.8	33.9	22.1	75.9	64.6	13.2	2.0
Commercial 1	15.0	25.6	30.6	40.2	72.4	70.3	10.4	12.8
Commercial 2	19.6	29.6	37.7	50.5	79.8	78.2	14.6	16.8
Commercial 3	24.5	31.6	40.2	51.2	81.8	81.0	25.0	29.5
Commercial 4	25.2	29.8	40.8	47.0	82.9	77.5	35.4	41.6
CLASI	**32.6**	**37.4**	**44.4**	**54.2**	**84.6**	**87.4**	**81.3***	**78.0***

Table 3: Experiment results of translation quality. VIP refers to the human-evaluated Valid Information Proportion that reflects the translation quality of these systems. † Due to the limitations in human evaluation capacity, the VIP scores are calculated on 4 randomly selected samples out of 10 in RealSI across all systems for fair comparison, while automatic metrics are evaluated on 10 samples. * Additionally, we evaluate the performance of CLASI on all 10 samples, achieving VIP scores of 78.0% for zh-en and 74.9% for en-zh.

Quantitative Analysis. As shown in Table 3, we compare CLASI with the baseline methods on RealSI dataset. In terms of the reliable human evaluation metrics, VIP, CLASI achieves scores of 81.3% and 78.0% for zh-en and en-zh, respectively. While all other models' VIP scores are lower than 42%. For more references, we use 3 widely-used automatic evaluation metrics: BLEU[5], BLEURT, COMET. Under the automatic evaluation metrics, CLASI also surpasses baselines by a large margin. The detailed human evaluation results of CLASI can be found in Appendix B.2.

High VIP marks CLASI a practical system that can help listeners understand real-time speech without professional human interpreters. Note that we only consider a system is better than others when the VIP is higher. For example, even though *Commercial 1* achieves higher scores than *SeamlessStreaming*

[5]We use SacreBLEU [58] for all the BLEU calculations in this paper.

on BLEU and COMET, we still consider *SeamlessStreaming* is a better system for zh-en translation based on VIP.

4.4 Latency

Evaluation Metrics. Due to the differences of grammatical structures between languages, a delay in simultaneous interpretation is inevitable. In this paper, we adopt the widely-used Average Lagging (AL) [45], Length Adaptive Average Lagging (LAAL) [52] for comparing the latency of different methods. To achieve a fair comparison with systems that rewrite the translation, we calculate the time of the definite translation of these systems. We also propose an additional metric, First Letter Appearance Lagging (FLAL), to reflect user experience on each system. FLAL represents the time that each system outputs the first determined translation.

Quantitative Results. Table 4 compares the latency of our model with various systems in terms of AL, LAAL, and our proposed FLAL on the RealSI and CoVoST. We find that the existing metrics AL and LAAL are not suitable latency measurements of paragraph-level SiST on RealSI. When the results are significantly shorter or longer than the reference translation, AL and LAAL may be largely exaggerated, leading to unreliable high latency. In these scenarios, FLAL is a more reliable and stable metric for all the systems.

Besides the paragraph-level latency evaluation, we compare our approach with other systems on the sentence-level dataset CoVoST2 zh-en, where both AL and LAAL produce reasonable values and the results are shown on the right side of Table 4. Since the commercial systems usually rewrite the translation, their latency is higher than the CLASI. Compared with the fastest approach *SeamlessStreaming*, CLASI achieves comparable latency but much better translation quality.

Model	RealSI (zh-en)			RealSI (en-zh)			CoVoST2 (zh-en)		
	AL	LAAL	FLAL	AL	LAAL	FLAL	AL	LAAL	FLAL
SeamlessStreaming	3.50	42.31	2.65	3.06	16.02	2.24	2.26	2.46	4.03
Commercial 1	2.10	13.22	3.27	4.53	20.71	1.88	3.05	3.26	4.01
Commercial 2	2.92	4.30	5.90	1.05	8.02	12.42	2.65	2.88	3.82
Commercial 3	12.31	12.65	15.70	8.45	15.81	9.68	3.67	3.86	6.14
Commercial 4	26.59	27.17	6.62	16.94	24.47	5.73	3.53	3.71	6.20
CLASI	2.17	6.34	4.20	0.34	3.17	6.00	2.63	2.83	5.02

Table 4: Comparison of latency between CLASI and baselines. AL and LAAL are standard metrics for measuring latency in sentence-level datasets. Even though AL and LAAL yield reliable results on the sentence-level CoVoST2 dataset, we argue that they are less effective for long speeches due to the complexity of long-speech translation. Therefore, we propose First Letter Appearance Lagging (FLAL), representing the time that each system outputs the first determined translation.

Discussion. While existing works put a lot of emphasis on the latency-quality trade-off [35, 53], human interpretation usually uses Ear-Voice-Span (EVS) to evaluate the lagging. EVS measures the average time from when the speaker finishes conveying a piece of information to when the audience hears the corresponding translation, which is similar to AL. The typical EVS of professional human interpreters usually ranges from 3 to 6 seconds [26] to achieve high-quality translation.

Consequently, we perform user studies and argue that the latency is less important than the translation quality for a practical SiST system. In the recent IWSLT 2023 simultaneous track [65], the ranking of models is also evaluated by the translation quality within certain latency constraints. We verify whether the latency of CLASI is acceptable to users through real-world user surveys. To the publication date of this paper, we collected 14 user surveys on zh-en direction, each user using CLASI for at least 30 minutes. Under the current latency performance shown in Table 4, only $1/14 == 7\%$ of them suggest that the latency significantly affects their user experiences while the rest think the improvement of translation quality outweighs the latency and overall output of CLASI largely helps them to understand the speech. Considering that the latency of CLASI is even lower than most of the commercial systems, We believe the latency of CLASI can be acceptable on most cases.

Current latency metrics are proposed on sentence-level SiST. As shown in Table 4, such metrics may not be suitable latency measurements for paragraph-level. As the importance of end-to-end evaluation for long speech keeps increasing, more refined metrics are required to measure the latency and provide a deeper insight into the systems.

4.5 Supplementary Experiments

To ensure a comprehensive evaluation of CLASI, our model is further evaluated on four additional datasets, including BSTC (zh-en) [85], CoVoST2 (zh-en) [75], MuST-C (en-zh) [15], and GigaST (en-zh) [83][6]. Table 5 presents the results of the automatic evaluation metrics for both zh-en and en-zh. Due to the high cost of human evaluation, we are not able to provide VIP for these four datasets. We observe that our model achieves consistently better performance than the baseline models. Even though our system achieves the best automatic evaluation results among all the compared systems, we still would like to emphasize that such a sentence-level evaluation scheme might overestimate the performance of SiST systems.

Model	**BSTC zh-en**						**CoVoST2 zh-en**					
	BLEU	BLEURT	COMET	AL	LAAL	FLAL	BLEU	BLEURT	COMET	AL	LAAL	FLAL
SeamlessStreaming	9.7	34.4	78.2	11.41	68.92	3.50	19.3	54.7	77.1	2.27	2.46	4.03
Commercial 1	14.1	32.0	73.0	9.01	16.73	13.95	17.6	47.6	69.3	3.05	3.26	4.01
Commercial 2	17.6	39.2	81.2	6.35	7.92	13.04	**24.7**	56.7	78.5	2.65	2.88	3.82
Commercial 3	21.5	41.6	83.7	12.88	13.63	22.55	24.2	54.1	75.9	3.67	3.86	6.14
Commercial 4	21.2	41.9	82.3	30.50	31.84	9.61	22.1	56.1	76.8	3.53	3.71	6.20
CLASI	**25.6**	**44.8**	**85.6**	4.68	9.03	13.13	24.2	**56.8**	**81.0**	2.63	2.83	5.02
Model	**MuST-C en-zh**						**GigaST en-zh**					
	BLEU	BLEURT	COMET	AL	LAAL	FLAL	BLEU	BLEURT	COMET	AL	LAAL	FLAL
SeamlessStreaming	17.4	48.2	75.2	1.43	1.69	2.06	26.3	48.9	75.4	1.41	1.57	2.16
Commercial 1	24.0	55.2	81.2	2.62	2.91	2.07	43.1	59.6	83.4	2.55	2.73	2.33
Commercial 2	**28.2**	59.5	83.1	3.25	3.51	4.84	45.7	63.2	85.0	3.13	3.28	5.12
Commercial 3	26.9	59.9	83.7	3.59	3.90	4.86	48.3	66.2	86.7	3.18	3.36	4.97
Commercial 4	27.3	60.0	83.4	3.25	3.54	4.86	43.3	59.9	83.5	3.06	3.23	5.00
CLASI	26.6	**61.8**	**85.2**	3.76	3.90	4.97	**50.4**	**69.0**	**88.8**	3.30	3.40	5.01

Table 5: Comparisons of CLASI and baselines on paragraph-level (BSTC) and sentence-level (CoVoST2, MuST-C, and GigaST) zh-en and en-zh datasets in terms of automatic evaluation metrics. We would like to emphasize that sentence-level evaluation schemes by automatic metrics cannot truly reflect the models' performance. VIP in Table 3 is a better metrics for comparing different systems.

4.6 MM-RAG Performance

4.6.1 Retriever

Table 6 presents the performance of various retrieve models on the development set of our proprietary dataset. Each sample in the test set includes a short audio chunk and the mentioned terms in the audio. Our MM-RAG retriever outperforms other open-source models by a large margin, achieving 91.3 % vs. 26.0% for Top-10 retrieve accuracy. We compare two types of methodologies: audio-to-audio and audio-to-text. In the audio-to-audio approach, a Text-to-Speech (TTS) model is utilized to convert the text keys from the external knowledge database into audio format, forming a database with audio-based keys. The audio keys and the user-input audio are then encoded with the ASR model to produce the corresponding representations. The Top-k retrieved items are subsequently determined using the Maximum Inner Product Search (MIPS) algorithm. For audio-to-text approach, we compare MM-RAG with CLAP [17]. As indicated in Table 6, the effectiveness of these models remains significantly below acceptable standards, and MM-RAG significantly outperforms them.

It is worth noting that the same audio encoder employed in our CLASI is utilized for generating audio embedding in the MM-RAG retriever. Such a design ensures that the integration brings minimal computational latency to the overall framework.

[6]We use the subset from in GigaS2S for evaluation

Model	Method	Finetuned	Top-1	Top-5	Top-10
CLAP [17]	Audio-to-Audio	No	2.1	7.3	13.8
Wav2Clip [80]	Audio-to-Audio	No	3.3	9.6	16.3
Whisper [60]	Audio-to-Audio	No	2.6	9.7	15.1
In-house ASR	Audio-to-Audio	No	7.2	19.4	26.0
CLAP [17]	Audio-to-Text	No	2.7	6.4	10.8
MM-RAG (Ours)	Audio-to-Text	Yes	**63.2**	**88.4**	**91.3**

Table 6: Top-k retrieve accuracy (%).

	Recall (%)	Precision (%)	F1 (%)
Shallow-Fusion	40.8	94.2	56.9
ICL + Shallow Fusion	79.2	73.4	76.2
ICL	79.2	86.3	**82.6**

Table 7: Recall and Precision of ICL and shallow-fusion for the intervention of the keywords. When calculating the Recall, we input 1 ground-truth keyword with 9 similar negative words as context. When calculating the false positive rate for precision, we input 10 similar negative words as context.

4.6.2 ICL Performance

By incorporating the retrieved terms from the external knowledge database as contextual information, our model's in-context learning ability significantly improves the performance of speech translation for in-domain terminologies. Table 7 compares our method with the widely-used shallow fusion for intervention in the generated conclusion. When calculating the Recall, we input 1 ground-truth keyword with 9 similar negative words as context. When calculating the false positive rate for precision, we input 10 similar negative words as context. ICL is able to achieve a high recall rate with good precision, obtaining the highest F1 while shallow fusion only gets a recall rate that is only half of ICL.

Additionally, we conduct an ablation study on our MM-RAG module within terminology-intensive scenarios incorporating the whole `<RETRIEVE>` pipeline. The incorporation of the external knowledge database results in a significant increment in the VIP score by about 10%, highlighting the effectiveness of our proposed MM-RAG.

4.7 Case Study

We present case studies to show the ability of CLASI in translating complicated speech for zh-en and en-zh in Table 8 and Table 9. We choose one of the most-performed cascaded systems Commerical 4 for comparison. The Commerical 4 adopted a cascaded approach for SiST and it is shown in Table 3 to be one of the best previous SiST systems. Detailed explanations are described in the tables. For zh-en direction, we present cases regarding robustness to recognition errors, reasoning ability, and trending words translation. For the en-zh direction, we present cases regarding native, expressive, and accurate terminology translations. More cases are shown in Table 11.

5 Related Work

Large language model. The encoder-decoder architecture [56, 83] has been widely explored in early speech translation research, but with the advent of large language models [2], there has been a growing interest in employing decoder-only architectures [21, 67] for sequence-to-sequence problems. While recent efforts have emerged in utilizing large language models for machine translation [36, 37, 90, 91] and speech translation [13, 30, 70], the application of such models in simultaneous translation tasks remains limited. Although there has been early attempts to utilize LLM for simultaneous machine translation [3, 29, 35, 86], to the best of our knowledge, no existing work has been found that explores the utilization of large language models for end-to-end simultaneous speech translation with such remarkable improvement.

CASE 1: Robustness to recognition errors	
Golden Transcription	欧文两罚命中，四分分差[1]，不到最后[2]
Commerical 4 ASR	欧文两罚命中，四分分叉[1]，不到最后[2]
Commerical 4 Translation	Irving hit two free throws and split[1] the four-point spread to the end,[2]
CLASI ASR	欧文两罚命中，四分分叉[1]，不到最后[2]
CLASI Translation	Kyrie makes both free throws, a four-point gap[1],
Explanation	The word 分差[1] is mis-transcripted to 分叉[1], which actually means "branch" or "split" in English. CLASI still generates the correct translation. After only hearing 不到最后[2], CLASI decides not to translate immediately and leaves 不到最后[2] to the next translation because of lacking context. While Commerical 4 translates it incorrectly.
CASE 2: Reasoning Ability for Translation	
Golden Transcription	绍兴二十年[1]担任右正言，弹劾胡寅
Commerical 4 ASR	绍兴二十年[1]担任佑正言，弹劾胡莹。
Commerical 4 Translation	Shaoxing twenty years[1] as YouZhengYan, impeach Hu Ying.
CLASI ASR	绍兴二十年[1]担任右正言，弹劾胡寅
CLASI Translation	In 1150, during the 20th year of Emperor Gaozong's Shaoxing era[1], he served as the Right Censor and impeached Hu Ying
Explanation	Literally, 绍兴二十年[1] could be translated as "Shaoxing 20th Year", while CLASI could understand the actual year of 绍兴二十年[1] is AD 1150, the 20th year under the reign of Emperor Gaozong.
CASE 3: Trending words or slangs	
Golden Transcription	我们常说，你们也太卷[1]了吧，别卷了，还是躺平[2]舒服。
Commerical 4 ASR	我们常说，你们也太卷[1]了吧，别卷了，还是躺平[2]舒服。
Commerical 4 Translation	We often say that you are too curly[1], don't curl up, or lie down[2] comfortably.
CLASI ASR	我们常说，你们也太卷[1]了吧，别卷了，还是躺平[2]舒服。
CLASI translation	We often say, "You are too competitive[1]. Stop it. It's more comfortable to lie flat[2]."
Explanation	Although in Chinese 卷[1] could be translated to "curly" in some cases, it actually means "involution" in this context. CLASI translates it to "competitive", which is acceptable. Lie flat[2] is comparable to lie down[2] for translating 躺平[2], but the whole sentence is translated more naturally by CLASI.

Table 8: Comparison between CLASI and a well-known commercial system Commerical 4 for zh-en direction.

CASE 1: Native and Accurate Translation	
Golden Transcription	You can't think of it on a case-by-case[1] basis. Either we all have rights or not have rights. Right[2].
Commerical 4 ASR	You can't think of it on a case-by-case[1] basis. Either we all have rights or nut have rights. Right[2]
Commerical 4 Translation	你不能根据具体情况[1]来考虑。要么我们都有权利，要么疯子都有权利[2]，对吧?
CLASI ASR	You can't think of it on a case-by-case[1] basis. Either we all have rights or not have rights. Right[2].
CLASI Translation	你不能就事论事[1]地考虑这个问题。 要么我们都有权利，要么我们都没有权利[2]。
Explanation	Although the Commerical 4 translation of case-by-case[1] is correct, CLASI uses 就事论事[1], a well-known Chinese idiom, which is more native. Besides, Commerical 4 ASR mis-transcripted not have rights[2] as nut have rights[2], leading to a completely non-sense translation.
CASE 2: Expressive Translation	
Golden Transcription	She was sobbing in fear that this test in a foreign language has been put in front of her.
Commerical 4 ASR	She was sobbing in fear that this test in a foreign language has been put in front of her.
Commerical 4 Translation	她哭了，害怕这个外语的测试摆在她面前，
CLASI ASR	She was sobbing in fear that this test in a foreign language has been put in front of her.
CLASI Translation	她因为害怕这门外语考试而哭泣，
Explanation	Theoretically, the Commerical 4 translation is correct literally. However, it's not expressive for native Chinese speakers. CLASI translation is expressive, conveying the same meaning of the source English sentence, which means "She was sobbing because of fearing the foreign language test."
CASE 3: Named Entity, Terminology Recognition and Translation	
Golden Transcription	So let's let me put the COVID-19[1] for example, so now we we know that there are a lot of people are infected and they have um positive antibody tests
Commerical 4 ASR	So let's let me put the CUBA 19[1] for example, so now we we know that there are a lot of people are infected and they have um positive, anybody? tests,[2]
Commerical 4 Translation	所以让我以古巴19人[1]为例。所以现在我们知道有很多人被感染，他们是阳性的。有人吗? 测试，[2]
CLASI ASR	So let's let me put the COVID nineteen[1] for example, so now we we know that there are a lot of people are infected and they have um positive antibody tests.[2]
CLASI Translation	以Covid-19[1]为例。我们现在知道有很多人被感染了，并且他们的抗体检测呈阳性。[2]
Explanation	Commerical 4 cannot correctly recognize COVID-19[1] and antibody tests[2], while CLASI successfully recognize and translate.

Table 9: Comparison between CLASI and a well-known commercial system Commerical 4 for en-zh direction.

Furthermore, LLMs have demonstrated impressive capabilities in tasks such as instruction following [2, 6, 20, 72], reasoning [55, 69], and planning [59, 64]. Recent research studies have leveraged prompt engineering to develop remarkable LLM agents that autonomously tackle complex tasks in diverse environments [76]. In our work, we empower the LLM to perform sequential instructions to accomplish the simulation speech translation task.

Simultaneous Speech Translation. One of the important components of simultaneous speech translation is the segmentation strategy, which determines how the speech frames are fed to the models. Different strategies could affect the latency and performance of the translation. According to [41], segmentation strategies can be classified into fixed-length, word-based, and adaptive segmentation. Fixed-length strategies [50] divide the speech into equally-length segments, while word-based strategies [46] identify word boundaries within the speech. Adaptive segmentation [16] detects boundaries for speech units. Among these categories, our method utilizes a fixed-length strategy.

Regarding the read/wait policy, the Wait-k method [45] and its variants [50, 84] have been extensively studied in the context of text translation and speech translation. In comparison to these approaches, which explicitly learn the generation of read/write signals, another line of research focus on how to leverage offline translation models [82]. When utilizing an offline translation model for simultaneous translation, it is important to address the stabilization of generated hypotheses to prevent excessive content refreshing experienced by the user. [40] first proposed a local agreement policy to stabilize the partial hypothesis, while [57] introduced an incremental blockwise beam-search algorithm. In contrast to these methods, we enforce our model to generate consistent hypotheses by constraining the prompt to the language model.

For the model architecture, there are two primary methods for implementing speech translation systems: cascaded solutions [27, 31] and end-to-end solutions [22, 51, 56]. Cascaded solutions involve separated ASR and MT components, while end-to-end solutions directly map speech to translations. Cascaded systems benefit from established techniques but suffer from latency and error propagation. End-to-end models offer real-time translation and improved quality through deep learning but require large-size training data. In our work, we implement an end-to-end model which combines the capabilities of ASR, MT, and ST.

Human Evaluation. In the realm of speech translation, the choice of evaluation metrics plays a crucial role in assessing the quality and effectiveness of translation systems. While automatic metrics, such as BLEU [54], BLEURT [68], and COMET [61], have traditionally been relied upon for evaluation, there is a growing recognition that they may not be the most suitable or comprehensive measure of performance [48]. We observe that in more recent works [7, 42, 71, 77], there is an increasing trend of evaluating systems using human assessments, particularly when LLM is employed in the work. While human evaluation requires more resources and time compared to automatic metrics, its benefits outweigh the drawbacks. By incorporating human judgment, speech translation systems can be refined and optimized to align with user expectations, ensuring translations that are not only technically accurate but also linguistically and contextually appropriate. In contrast to the existing human evaluation metrics, e.g."continuous rating" [32] and MQM (Multidimensional Quality Metrics) [43], we have taken inspiration from professional human interpreters [49] and propose to use VIP (Valid Information Proportion) as a human evaluation metric which precisely reflects the goal of the simultaneous translation task.

6 Conclusion

In this work, we introduce **C**ross **L**anguage **A**gent - **S**imultaneous **I**nterpretation, CLASI, an LLM agent to produce end-to-end simultaneous speech translation. Benefits from massive pretraining and imitation learning, CLASI achieves significantly better performances than state-of-the-art systems. Take the Chinese-to-English direction as an example, under strict and challenging human-evaluated metrics proposed by professional human interpreters, Valid Information Proportion (VIP), CLASI significantly outperforms baselines by a large margin. While all other systems obtain VIP by less than 40%, CLASI achieves a VIP of 81.3%, demonstrating human parity performance. More specifically, we propose the following crucial components for the supreme performance of CLASI: (1) An encoder-conditioned LLM agent architecture that performs high-quality or even human-parity SiST process through simple actions. (2) Imitation learning from human interpreters for a natural read-write policy balances translation

quality and latency in a data-driven manner, without complex human pre-designing. Under such policy, CLASI achieves a stable output scheme, where each output is deterministic, thus potentially better user experience than most commercial systems. (3) Motivated by the preparatory trajectory of human interpreters, CLASI could perform in-context learning from historical translations and external knowledge to provide sufficient information for translation. With the powerful translation ability of CLASI, we believe it can further make cross-lingual communication seamless across different places all over the world.

Limitation and Future Work

Although we achieved significant improvements over commercial systems in Chinese-to-English and English-to-Chinese tasks, more languages should be considered in the future. In our current implementation, CLASI performs a full action sequence for each translation round. Some of the actions, e.g., `<RETRIEVE>` is optional for easy translation scenarios since the model is capable of translating correctly without the help of external knowledge. Training the model to better determine whether to skip unnecessary actions is a future direction. For a product-level system, even though the latency of CLASI is acceptable in most cases, how to reduce the translation latency without lowering the translation quality is still interesting and potentially helpful for user experience. Furthermore, we argue that the current automatic metrics are not comprehensive for SiST evaluation. Most of the quality measurements do not consider key information, which is crucial in SiST scenarios. As such, we proposed VIP for better human evaluation. Consequently, more reliable automatic quality and latency metrics should be proposed in the future as well. Reinforcement learning from human feedback (RLHF) has been proven to be effective in enhancing LLM performance. Although CLASI achieves significantly superior results than previous state-of-the-art systems, further studies on how to build better multi-modal reward models and better RL methods for SiST is also an important direction. Incorporating more modalities, for example, end-to-end speech-to-speech generation, or even end-to-end video-to-video generation are also promising research topics.

Social Impact

The powerful SiST system CLASI can be applied to various scenarios to facilitate cross-lingual communications. For example, it can be deployed to various conferences or daily meetings to help listeners understand speech in different languages. It can also be deployed as a system-level translation module to help users watch videos that are conveyed in different languages. For online gaming, it can also help to bridge the gap of cross-lingual communication and connect people speaking different languages. A powerful SiST system with human parity performance may significantly improve the efficiency of professional human interpreters.

Despite the huge positive social impact that CLASI may bring, every coin has two sides. Neglecting some low-resource languages may also bring unfairness to some minorities. Resolving these problems needs further cooperation from the society. We leave more languages supporting as our future work.

Authorship and Acknowledgements

All contributors are listed in alphabetical order by last name. Corresponding to this work can be sent to any core authors' email.

Core Authors. All core authors contributed equally to this work.

- Shanbo Cheng — chengshanbo@bytedance.com
- Zhichao Huang — zhichao.huang@bytedance.com
- Tom Ko — tom.ko@bytedance.com
- Hang Li — lihang.lh@bytedance.com
- Ningxin Peng — pengningxin@bytedance.com
- Lu Xu — xu.lu1@bytedance.com
- Qini Zhang — qini.z@bytedance.com

Labeling, Evaluation and Interpretation Team. Our data labeling and human evaluation team led by Yifu Li, made diligent efforts in all kinds of help needed, which is irreplaceable in the success of this project. Special thanks to the human interpreter team led by Anna Liu for providing insightful and comprehensive analysis on data labeling, human evaluation, and other recommendations.

- Jingwen Chen
- Xiaoya Chen
- Yifu Li
- Huiying Lin
- Anna Liu

Engineering Team. We collaborated with our engineering team led by Tingshuai Yan, their infrastructure support is crucial for this project.

- Weicheng Fu
- Tingshuai Yan
- Liehao Zou

Acknowledgements. We appreciate the Speech Understanding team for all kinds of help, especially data sharing, thanks to their tremendous work for all the in-house data. We would like to express our deepest thanks to all the contributors to this project, their brilliant work guarantees the success of this project.

References

[1] *Kendall Rank Correlation Coefficient*, pages 278–281. Springer New York, New York, NY, 2008.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Victor Agostinelli, Max Wild, Matthew Raffel, Kazi Asif Fuad, and Lizhong Chen. Simul-llm: A framework for exploring high-quality simultaneous translation with large language models. *arXiv preprint arXiv:2312.04691*, 2023.

[4] Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, et al. Findings of the iwslt 2021 evaluation campaign. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, 2021.

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[7] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[8] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2021.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[10] Agnieszka Chmiel. Effects of simultaneous interpreting experience and training on anticipation, as measured by word-translation latencies. *Interpreting*, 23(1):18–44, 2021.

[11] Kyunghyun Cho and Masha Esipova. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*, 2016.

[12] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[13] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023.

[14] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[15] Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics, 2019.

[16] Qian Dong, Yaoming Zhu, Mingxuan Wang, and Lei Li. Learning when to translate for streaming speech. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–694, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[18] Marcello Federico, Alex Waibel, Marta R. Costa-jussà, Jan Niehues, Sebastian Stuker, and Elizabeth Salesky, editors. *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics.

[19] Marcello Federico, Alex Waibel, Kevin Knight, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and Francois Yvon, editors. *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, July 2020. Association for Computational Linguistics.

[20] Peiyuan Feng, Yichen He, Guanhua Huang, Yuan Lin, Hanchong Zhang, Yuchen Zhang, and Hang Li. Agile: A novel framework of llm agents, 2024.

[21] Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder, 2023.

[22] Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 330–340, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[23] Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. Speech translation with speech foundation models and large language models: What is there and what is missing? *arXiv preprint arXiv:2402.12025*, 2024.

[24] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, 2017.

[25] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[26] Ewa Gumul and Andrzej Łyda. The time constraint in conference interpreting: Simultaneous vs. consecutive. *Research in language*, 5:165–183, 2007.

[27] Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Zongyao Li, Zhiqiang Rao, Minghan Wang, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Shaojun Li, Yuhao Xie, Lizhi Lei, and Hao Yang. The HW-TSC's simultaneous speech-to-text translation system for IWSLT 2023 evaluation. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 376–382, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[29] Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. Gentranslate: Large language models are generative multilingual speech and machine translators. *arXiv preprint arXiv:2402.06894*, 2024.

[30] Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. Speech translation with large language models: An industrial practice, 2023.

[31] Javier Iranzo-Sánchez, Javier Jorge, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Adrià Giménez, Jorge Civera, Albert Sanchis, and Alfons Juan. Streaming cascade-based speech translation leveraged by a direct segmentation model. *Neural Networks*, 142:303–315, 2021.

[32] Dávid Javorský, Dominik Macháček, and Ondřej Bojar. Continuous rating as reliable human evaluation of simultaneous speech translation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 154–164, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[33] Roderick Jones. *Conference interpreting explained*. Routledge, 2014.

[34] Jari Kolehmainen, Aditya Gourav, Prashanth Gurunath Shivakumar, Yile Gu, Ankur Gandhe, Ariya Rastrow, Grant P. Strimel, and Ivan Bulyko. Multi-modal retrieval for large language model based speech recognition. 2024.

[35] Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. Transllama: Llm-based simultaneous translation system, 2024.

[36] Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. Mt-patcher: Selective and extendable knowledge distillation from large language models for machine translation, 2024.

[37] Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592, 2024.

[38] Xiaoqing Li, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. Neural name translation improves neural machine translation. In *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings 14*, pages 93–100. Springer, 2019.

[39] Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, and Hao Yang. HW-TSC 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 835–840, Singapore, December 2023. Association for Computational Linguistics.

[40] Danni Liu, Gerasimos Spanakis, and Jan Niehues. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection, 2020.

[41] Xiaoqian Liu, Guoqiang Hu, Yangfan Du, Erfeng He, YingFeng Luo, Chen Xu, Tong Xiao, and Jingbo Zhu. Recent advances in end-to-end simultaneous speech translation, 2024.

[42] Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing gpt-4 on conversational qa and rag, 2024.

[43] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463, 2014.

[44] Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, 2015.

[45] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics.

[46] Xutai Ma, Juan Pino, and Philipp Koehn. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In Kam-Fai Wong, Kevin Knight, and Hua Wu,

editors, *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*, pages 582–587, Suzhou, China, December 2020. Association for Computational Linguistics.

[47] Dominik Macháček, Ondřej Bojar, and Raj Dabre. Mt metrics correlate with human ratings of simultaneous speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 169–179, 2023.

[48] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 7297–7306, Online, August 2021. Association for Computational Linguistics.

[49] Zoe Moores. The nerle model–a tool for assessing the quality of intralingual subtitles at live events. *Universal Access in the Information Society*, 23(2):589–607, 2024.

[50] Ha Nguyen, Yannick Estève, and Laurent Besacier. An empirical study of end-to-end simultaneous speech translation decoding strategies, 2021.

[51] Sara Papi, Marco Gaido, and Matteo Negri. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[52] Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. *arXiv preprint arXiv:2206.05807*, 2022.

[53] Sara Papi, Matteo Negri, and Marco Turchi. Attention as a guide for simultaneous speech translation. *arXiv preprint arXiv:2212.07850*, 2022.

[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[55] Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. Refiner: Reasoning feedback on intermediate representations, 2024.

[56] Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. Towards efficient simultaneous speech translation: CUNI-KIT system for simultaneous track at IWSLT 2023. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 389–396, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[57] Peter Polák, Brian Yan, Shinji Watanabe, Alex Waibel, and Ondřej Bojar. Incremental blockwise beam search for simultaneous speech translation with controllable quality-latency tradeoff. In *INTERSPEECH 2023*, interspeech_2023. ISCA, August 2023.

[58] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[59] Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. Autoact: Automatic agent learning from scratch for qa via self-planning, 2024.

[60] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356, 2022.

[61] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, 2020.

[62] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[63] Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Simulspeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, 2020.

[64] Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. Tptu: Large language model-based ai agents for task planning and tool usage, 2023.

[65] Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[66] Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà, editors. *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[67] Frank Seide, Morrie Doulaty, Yangyang Shi, Yashesh Gaur, Junteng Jia, and Chunyang Wu. Speech reallm – real-time streaming speech recognition with multimodal llms by teaching the flow of time, 2024.

[68] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.

[69] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[70] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024.

[71] 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. No language left behind: Scaling human-centered machine translation, 2022.

[72] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[74] Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France, May 2020. European Language Resources Association.

[75] Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.

[76] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.

[77] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore, December 2023. Association for Computational Linguistics.

[78] Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415, 2017.

[79] Shira Wein, I Te, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. Barriers to effective evaluation of simultaneous interpretation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 209–219, 2024.

[80] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[81] Shao-Chuan Wu. Assessing simultaneous interpreting. *A study on test reliability and Examiners' assessment behaviour (Doctoral dissertation). Newcastle University, Newcastle*, 2010.

[82] Brian Yan, Jiatong Shi, Soumi Maiti, William Chen, Xinjian Li, Yifan Peng, Siddhant Arora, and Shinji Watanabe. CMU's IWSLT 2023 simultaneous speech translation system. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 235–240, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[83] Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. GigaST: A 10,000-hour Pseudo Speech Translation Corpus. In *Proc. INTERSPEECH 2023*, pages 2168–2172, 2023.

[84] Xingshan Zeng, Liangyou Li, and Qun Liu. RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2461–2474, Online, August 2021. Association for Computational Linguistics.

[85] Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. BSTC: A large-scale Chinese-English speech translation dataset. In Hua Wu, Colin Cherry, Liang Huang, Zhongjun He, Qun Liu, Maha Elbayad, Mark Liberman, Haifeng Wang, Mingbo Ma, and Ruiqing Zhang, editors, *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35, Online, June 2021. Association for Computational Linguistics.

[86] Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. *arXiv preprint arXiv:2406.03049*, 2024.

[87] Shaolei Zhang and Yang Feng. End-to-end simultaneous speech translation with differentiable segmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7659–7680, 2023.

[88] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

[89] Chengqi Zhao, Zhicheng Liu, Jian Tong, Tao Wang, Mingxuan Wang, Rong Ye, Qianqian Dong, Jun Cao, and Lei Li. The volctrans neural speech translation system for iwslt 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 64–74, 2021.

[90] Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. Fine-tuning large language models for domain-specific machine translation, 2024.

[91] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, 2024.

A Human Evaluation Guidelines

In this Appendix, we provide detailed human evaluation guidelines which are formulated by professional human interpreters.

A.1 Key Indicator

We define the key indicator as the "Valid Information Proportion," denoted as VIP. This metric measures the proportion of valid semantic fragments within a complete speech session. A semantic fragment is deemed valid if it effectively conveys the core information, accurately representing the speaker's original intent. Typically, one complete sentence is considered a single semantic fragment. VIP assesses the model's ability to capture and communicate the essence of the spoken content.

A.2 Evaluation Process

First, the human evaluators segment the long translation result into semantic fragments according to the formal rules as follows:

- **Semantic Completeness:** Each fragment should contain one complete concept or information point. For example, in a conference translation, an ideal semantic fragment often corresponds to a full sentence.
- **Natural Language Pauses:** Pauses that naturally occur in speech often indicate the boundaries of semantic fragments. During segmentation, natural pauses should be extensively considered and avoid irrational interruptions. Also, punctuation and conjunction in the text should be considered to maintain integrity and clarity of information.
- **Logical Coherence:** Each segment should contain information that are logically coherent and continuous. Conditional sentences, causative sentences, or both parts of antithesis sentences should be kept within the same segment.
- **Grammatical Completeness:** Each segment should include all necessary grammatical components (e.g., subject, verb, object) and have a complete grammatical structure.
- **Proper Information Density:** Each segment should have a moderate amount of information, avoiding information overload. It is recommended that each segment not exceed 50 words.

After segmentation, the human evaluators follow the instructions below to evaluate the validity of the semantic fragments:

- **Key Information Recognition**. Key information refers to the content that can constitute core information, including but not limited to proper nouns, keywords, terminologies, sentence structures, etc.
- **Correctness Assessment**. Evaluators assess whether the translation of key information is accurate and successful in conveying the correct spoken intentions. Misinterpretations of the speaker's words, inaccuracies in analyzing the context, or erroneous translations of specific terms can all contribute to the failure of the assessment.
- **Expressiveness Assessment**. Evaluators assess whether the whole segment is translated accurately, comprehensibly, and expressively to humans. Assessing for any vague, ambiguous, or misleading statements. This indicator primarily evaluates the clarity, fluency, and intuitiveness of the translation, rather than its accuracy. Typically, verbosity, complex sentence structures, or challenging grammatical constructions that are unnecessary would reduce the expressiveness of the translation, thus leading to failure of the assessment.

If the translation fails any of the above assessments, the translation will be marked as invalid. After the evaluators assessed all semantic fragments, the VIP could be simply calculated as dividing the number of valid semantic fragments by the total number of fragments.

We illustrate the evaluation criteria with two examples in Table 10. Although these translations achieve "high accuracy" in automatic evaluations, we still categorize them as invalid. It's important to note that our standard aims to emulate human interpreters, presenting a significant challenge to both human evaluators and translation systems.

Example 1: Correctness Assessment	
Golden Chinese	请确保服务器后端的API接口完全遵循RESTful[1]架构原则。
Reference	Please ensure the server backend's API interface fully complies with RESTful[1] architectural principles.
Translation	Please ensure the server backend's API interface fully complies with restless[1] architectural principles.
Explanation	Although the sentence-level BLEU score is near 80, the translation is still considered invalid, because the keyword "RESTful" is mistranslated.
Example 2: Expressiveness Assessment	
Golden Chinese	这部分跟资源那边，前端资源这一块搞定了吗?
Reference	Did you arrange the front-end resources well?
Translation 1	Is this part related to the front-end resources? Did you finish the front-end resources?
Translation 2	Has the front-end resource been settled?
Explanation	Translation 1 is redundant, disfluent and contains minor errors, thus not easy to understand by human evaluators, while Translation 2 generated by CLASI is concise and fluent, and conveys the speaker's intention appropriately.

Table 10: Human evaluation examples of Chinese-to-English Translation task.

A.3 Correlation with Automatic Metrics

Figure 4 shows the distribution and regression curve for VIP with regard to BLEU, BLEURT, and COMET, respectively. From the scatter points in Appendix A.3 we may observe that as VIP score increases, the growth of the automatic metric curves slows down and becomes less significant, making it hard to reflect the real changes in translation quality. The correlation curves in Appendix A.3 also demonstrate the finding. Here, we calculate Kendall's Tau correlation coefficient [1], which measures the monotonic correlation between two ordered variables. In low VIP ranges, the correlation between VIP and automatic metrics is observable; as the score increases, the correlation harshly drops, which indicates a significant distortion of the automatic metrics. A possible reason is that the translations may differ from the groundtruths by only a few words in the mediocre ranges. However, in a real simultaneous interpretation scenario, these words are likely to be keywords that play important roles in conveying precise information, which may significantly impact VIP scores.

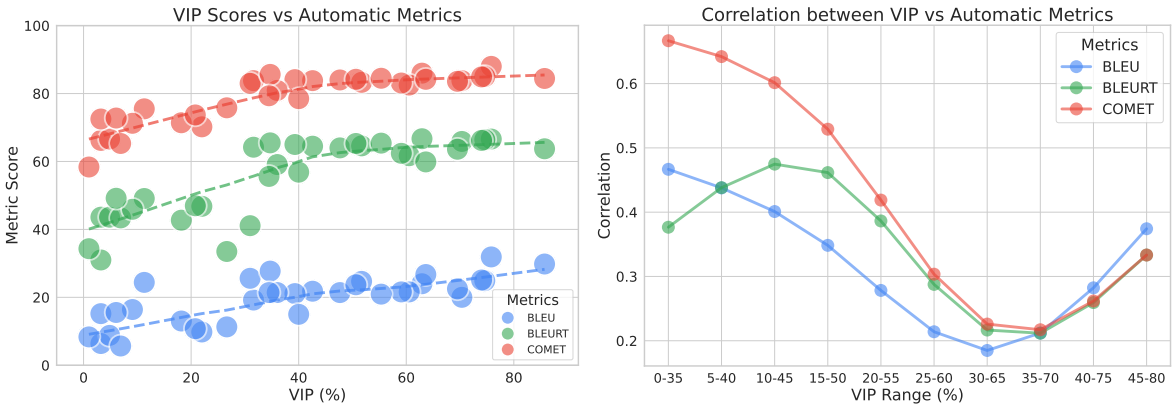

Figure 4: Analysis of VIP vs different automatic metrics on the zh-en direction. The distribution and regression curve of the data points for each metric are shown in the above-left figure. Line charts for the calculated correlation between VIP and Automatic metric within multiple intervals are shown in the right figure. Due to the limitation of human labeling capacity, we collect 35 rounds of human evaluation results for zh-en direction on our in-house testset.

B Supplementary Materials

B.1 Supplementary Case Study

We provide more case studies in Table 11. In terms of informal, disfluent, code-mixing, and named-entity translation, CLASI could achieve much better results than the commercial products. Benefits from the end-to-end approach, CLASI could also understand the original speech tone and generate better translations.

B.2 Example of Detailed Evaluation Result on RealSI

We provide a detailed human evaluation result for CLASI in Figure 5, where we provide golden source transcription, CLASI output, human evaluation results, and reference translation. We randomly choose one of the test samples in RealSI. We share the full detailed evaluation results at online sheets for academic reference. Note that to ensure fair comparison, when evaluating multiple systems, we randomly shuffle the ordering between systems for each semantic fragment so that human evaluators cannot identify the specific system.

CASE 1: Informal, disfluent speech translation	
Golden Transcription	那基于这些观察，那我们是不是啊，我，我，我[1]，我们是不是可以去找一种，就是像GPT3.5一样的，统一的建模方法?
Commerical 4 ASR	那基于这些观察，那我们是不是啊？我们是不是我？[1]我们是不是可以去找一种，就是像gvt3.5[2]一样的，统一的建模方法?
Commerical 4 Translation	So based on these observations, are we? Are we me? Can we go[1], like gvt3.5[2], and do this unified modeling
CLASI ASR	那基于这些观察，那我们是不是啊，我，我，我，[1]我们是不是可以去找一种，就是像GPT3.5一样的，统一的建模方法?
CLASI Translation	Based on these observations, can we find[1] a unified modeling method like GPT3.5[2]?
Explanation	The first labeled Chinese phrase (superscript 1) actually means "can we find"[1]. CLASI can generate a much more fluent, concise translation than the Commerical 4. Besides, for keyword GPT3.5[2], CLASI can generate correct ASR and translation.
CASE 2: Disfluent and code-mixing speech	
Golden Transcription	我听过一句话叫，pri, pri, prioritization, prioritization[1] is only real when it hurts。
Commerical 4 ASR	我听过一句话叫，Pro, 不管, prioritization, prioritization[1] is only real when it hurts.
Commerical 4 Translation	I heard a saying called Pro, Anyway, it is Prioritization[1] is only real when it hurts.
CLASI ASR	我听过一句话叫，priortizaiton, prioritization[1] is only real when it hurts。
CLASI Translation	I heard a saying that prioritization[1] is only real when it hurts.
Explanation	The speaker stutters when saying the English sentence, which is very common in real-world scenarios. CLASI can fully understand and generate the correct English text without any repetition.
CASE 3: Named-entity recognition and translation	
Golden Transcription	好球！迪亚斯[1]的传中，C罗来争抢，这个就是C罗[2]最喜欢的
Commerical 4 ASR	好球,比亚斯[1]的传统，C罗来争抢这个就这是C罗[2]最喜欢的
Commerical 4 Translation	Nice shot, Bias'[1] traditional C Ronaldo to compete for this is C Luo's[2] favorite
CLASI ASR	好球！迪亚斯[1]的传中，C罗来争抢。 这个就这是C罗最喜欢的
CLASI Translation	Nice cross by Dias[1], Ronaldo goes for it. This is Ronaldo's[2] favorite.
Explanation	Commerical 4 cannot correctly recognize the name of the famous football player, Ruben Dias. As for the name of Cristiano Ronaldo, although it translates correctly the first time, but fails the second time. CLASI can perform perfect recognition and translation.
CASE 4: Speech tone understanding	
Golden Transcription	门前斜传，漂亮！这下漂亮！球进了！摆乌龙！哎，自摆乌龙。
Commerical 4 ASR	球没有斜转漂亮，这下球进了白骨龙，这白骨龙
Commerical 4 Translation	The ball didn't spin beautifully, and now the ball went into the bone dragon, the bone dragon
CLASI ASR	门前斜转，漂亮！这下漂亮！球进了！摆乌龙！哎，自摆乌龙。
CLASI Translation	Diagonal shot in front of the goal. Beautiful! What a beauty! Goal! Own goal! Yes, own goal.
Explanation	CLASI could recognize the speaker with an exciting tone, thus generating the translation with exclamation marks. Besides, in this case, which is a complicate scenario of a football game, the ASR outputs of the Commerical 4 are mostly incorrect, leads to nonsense translation.

Table 11: Comparision between CLASI and a well-known commercial system Commerical 4 for zh-en direction.

Golden Transcription	CLASI	Is valid	Reference
然后，嗯，说先说一下，嗯，我之前说的这些就是为了让你们了解就是嗯，	First, what I said before is to let you know	Yes	Let me put it first that, what I said before was to help you understand
它有基本的一些什么什么样的功能，然后嗯，然后这种功能可以对我们现在的systemic risk就是可以解救。	what basic functions it has, and then these functions can have an impact on our current system rise, can save it.	No	what basic functions it has, and such functions can rescue current systemic risk.
嗯，说，呃，说的systemic risk就是说是di, undiversifiable risk，因为在金融上来说它有，嗯，我们平时金融管理呀，然后呃，portfolio manage呀，呃可以管理的risk都叫呃，可以是diversifiable的risk。	What I said about system rise is die, undiversifiable risk. Because in finance, we usually have financial management , and portfolio managers can manage risks, which are called diversifiable risks.	No	The systemic risk mentioned here refers to undiversifiable risk. Because in finance, in our daily financial management and portfolio management, all risks we are able to manage are called diversifiable risks.
然后剩下的叫系统，系统风险，就是我们是没有办法的，然后，就在传统金融市场里面是没有办法的。	And the rest is called systematic risk, which we can't do anything about, and there is no way in the traditional financial market.	Yes	The rest ones are called systemic risks, which we have no way to deal with in the traditional financial market.
然后嗯，呃，这里说一个，一个例子就是用来用来说明，	Here is an example to illustrate that	Yes	Here is an example to illustrate it,
嗯，呃，这种系统风险，它除了除了政治，还有一些自然的一些呃就是无可避免的原因之外，	this kind of systematic risk, in addition to politics and some inevitable natural reasons,	Yes	such kind of systemic risk, in addition to political and some inevitable natural reasons,
它，你为什么会存在这种系统风险？所谓系统风险，它就是整个市场的一种，呃，一种失败，	why does it exist? The so-called systematic risk is a failure of the entire market,	Yes	why does it exist? The so-called systemic risk is a kind of failure of the entire market.
就是嗯，呃，一，然后，但但是除了自然和政治原因之外，其实很大的原因就是人的behaviour。	but in addition to natural and political reasons, a big reason is human behavior.	Yes	But apart from natural and political reasons, in fact, a major reason is human behaviour.
然后，嗯，啊，比如说呃，在二零零八年这个Bear raid的时候就是，嗯，其实所有都都起源于，呃，一个叫Bear Stern的这个asset management的公司，然后，嗯，	For example, during the 2008 bear raid it all originated from a company called Basestem in ASR,	No	For example, during the Bear raid in 2008, everything resulted from a company called Bear Stern, an asset management company.
呃，它，呃，有trader，trader manipulate它们它们的呃，它们的股价，	who tried to manipulate their stock price.	No	There were traders manipulating their stock prices.
因为这个公司它就嗯赌了很多在，呃，这个叫呃mortgage backed security呃，mortgage security，它就是asset backed security的一种。	Because this company gambled a lot on mortgage-backed securities, which is a type of asset-backed security.	Yes	Since the company bet a lot on mortgage backed security which is a kind of asset backed security.
嗯，这是一种就是可以很好操控，很好manipulate的一种金融产品，	This is a type that can be easily manipulated.	Yes	It is a kind of financial product that can be well controlled and manipulated.
它本身其实不是高风险的，就是现在很多人一提到这种像这种structure的这种finance的东西，都都觉得这种是高高风险的产品，	It's not actually high-risk. Many people now think that structured financeis a high-risk product,	Yes	It is not actually a high-risk one. Now when it comes to such structured financial products, many people think they are high-risk products,
但其实不是，它的诞生其实就是为了减低风险，就是提高呃，金融产品的credit rate，cre，	but it's not. Its purpose is to reduce risk and increase the credit rate of financial products.	Yes	but in fact, they are not. Its emergence is aimed to reduce risks and improve the credit rate of financial products.
呃，所以，所以说其实，嗯之所以之之所以就是嗯，别人就是大众会对这种金融产品有种高风险的印象，	So, the reason why the public has a high-risk impression of such financial products	Yes	So in fact, the reason why the public have an impression of high risks for this kind of financial product,
呃，主要都是因为，嗯，就是呃，市场上的一些manipulation，一些非常有效的manipulation，然后在这个这个例子里面呃，	is mainly due to some manipulation in the market, some very effective manipulation. In this example	Yes	is mainly because of some effective manipulations in the market. And in this example,
嗯这个首，首先是呃，这个mortgage backed呃asset，它有很高，就是在在零八年那个金融危机之前	first of all, the mortgage bar the asset has a high, before the 2008 financial crisis,	No	First of all, this mortgage backed asset, before the financial crisis in 2008,
它其实一直都有很很高的那个那个，那个credit，那个，那个信用分，就是评信用评级。	it has always had a very high credit score, credit rating.	Yes	it has actually always had a very high credit rating.
嗯，因为因为这种，这种类别的金融产品它是叫呃bankruptcy remote，	Because this type of financial product is called bankruptcy remote.	Yes	Because this kind of financial product is called bankruptcy remote.
就是呃比如说这个公司，你，你有这个嗯，你呃你发售了这个产品，但是，但是就算你，你公司自己倒闭了，嗯，就是，嗯，就是，是不会影响这个产品的，	For example, if a company issues this product, even if the company goes bankrupt, it doesn't matter. It won't affect the product	Yes	For example, a company launches a product, while even if the company goes bankrupt, it will not affect the product.
因为因为，呃，只要是这个产品的之前的它的collateral呃的公司，嗯就是还没有倒闭，就是它们的营营业状况还是好的，	because as long as the previous company of this product has not gone bankrupt, and their business is still good	Yes	Because, as long as the company of its previous collateral doesn't go bankrupt, which means their business conditions are still good,
然后嗯这个这个还是，它它还是，就算你的公司倒闭了，你发行了这种这种债券，比如说，嗯，它你你你的债券还是可以正常运行的，就是它的评级也不会受到影响。	even if your company goes bankrupt, and you issue this kind of bond, for example, your bond can still operate normally and its rating will not be affected	Yes	then even if the company goes bankrupt and one issues such bonds, for example, the bond can still be delivered normally, its rating will not be affected.
嗯，然后呃，总的来说就是，	In general	Yes	In general,
嗯，这个呃，呃，开始的时候，呃，这些trader就发现了这个房市的就是不对劲的地方。	at the beginning, these traders found something wrong with the housing market	Yes	at the beginning, these traders discovered the abnormal situation in the housing market.
但是呃，即使是房市不对劲的时候，呃这个嗯，这个mortgage，呃，这个bond它的评分还是很高，	but even when the housing market was not right, the rating of this mortgage bond was still very high,	Yes	But, even when the housing market was abnormal, the mortgage or the bond still had a high rating,
所以就没有人，就是大呃大大部分人都没注意到这一点。但是就被有些trader就发现了，然后他们就想manipulate。	so no one was very, few people noticed this, but some traders found out, and then they wanted to manipulate,	Yes	but most people didn't notice it. But some traders discovered it, and then they wanted to manipulate it.
然后后来后来证明他们的manipulate也成功了嘛。	and later it turned out that they succeeded in manipulating.	Yes	Later, it was proven that their manipulation was successful.
但其实嗯，这一系列它只是一种，嗯，一种，每每一个阶段就是它是，嗯每个阶段就是呃，它的decision making都是based on assumption，	But in fact, this series is just a decision at each stage is based on assumption,	Yes	But in fact, for the whole process at each stage, its decision making is based on assumption
然后这种assumption呢，它就嗯被 manipulate了，然后其实等于说它这就是一种信息的扭曲。	and this assumption is manipulated, which is actually a distortion of information.	Yes	which was manipulated. Actually, this is a kind of distortion of information.

Figure 5: The first column indicates the golden transcription of the source text. Each row indicates one semantic fragment split by human evaluators. The second column is the translation results of CLASI. The third and fourth columns indicate the validity of translation and reference translation, respectively. In this case, the VIP is $24/29 == 82.8\%$.