

Seed LiveInterpret 2.0: End-to-end Simultaneous Speech-to-speech Translation with Your Voice

ByteDance Seed

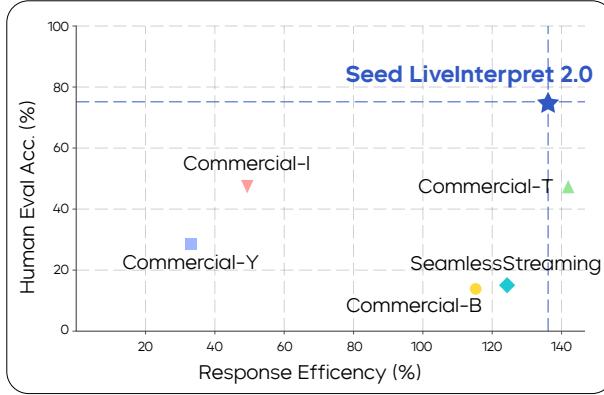
Abstract

Simultaneous Interpretation (SI) represents one of the most daunting frontiers in the translation industry, with product-level automatic systems long plagued by intractable challenges: subpar transcription and translation quality, lack of real-time speech generation, multi-speaker confusion, and translated speech inflation, especially in long-form discourses. In this study, we introduce **Seed LiveInterpret 2.0**, an end-to-end SI model that delivers high-fidelity, ultra-low-latency speech-to-speech generation with voice cloning capabilities. As a fully operational product-level solution, **Seed LiveInterpret 2.0** tackles these challenges head-on through our novel duplex speech-to-speech understanding-generating framework. Experimental results demonstrate that through large-scale pretraining and reinforcement learning, the model achieves a significantly better balance between translation accuracy and latency, validated by human interpreters to exceed 70% correctness in complex scenarios. Notably, **Seed LiveInterpret 2.0** outperforms commercial SI solutions by significant margins in translation quality, while slashing the average latency of cloned speech from nearly 10 seconds to a near-real-time 3 seconds, which is around a near 70% reduction that drastically enhances practical usability.

Date: July 29, 2025

Official Page: https://seed.bytedance.com/seed_liveinterpret

Seed LiveInterpret 2.0 vs. Baselines S2T Evaluation



Seed LiveInterpret 2.0 vs. Baselines S2S Evaluation

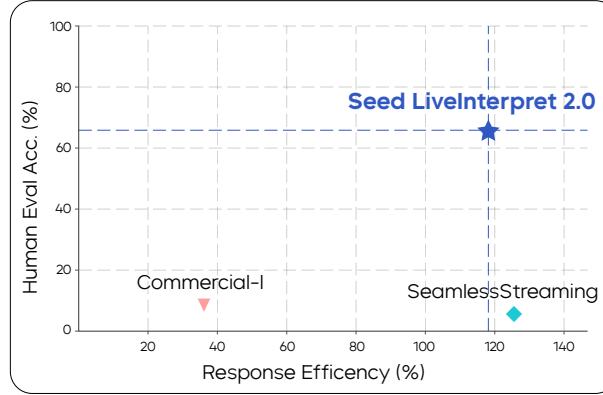


Figure 1 Evaluation of simultaneous interpretation systems: The left and right panels compare human assessment scores of translation quality against response efficiency¹ for speech-to-text (S2T) and speech-to-speech (S2S) modes, where response efficiency is measured relative to human interpreter latency. Human evaluation accuracy reflects how faithfully the translation output conveys the speaker's original intent. The evaluations were conducted using the **RealSI** benchmark [6].

¹Response efficiency quantifies performance relative to the latency of a human interpreter and is calculated as the quotient of 3 seconds divided by the observed latency.

1 Introduction

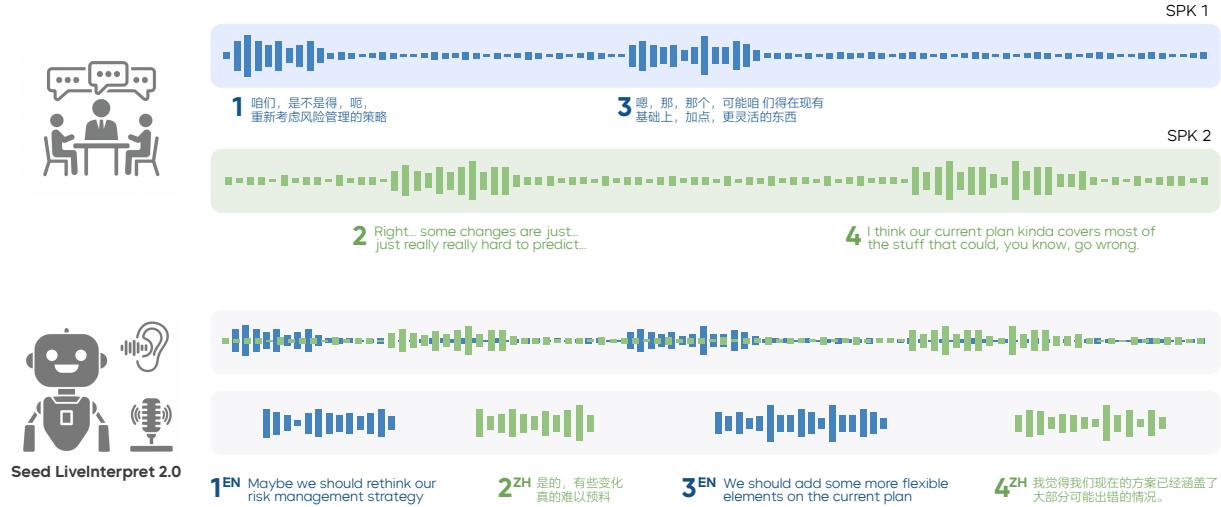


Figure 2 Illustration of Seed LiveInterpret 2.0 in a multilingual live conversation scenario where two human speakers (**SPK1** and **SPK2**) communicate in Chinese and English, respectively. The top section shows the original utterances and speaker turns. Below, the Seed LiveInterpret 2.0's real-time behavior is visualized, performing simultaneous speech translation. Ear icon indicates continuous listening to each speaker. Translated outputs (in English or Chinese) appear underneath, with the horizontal gap representing translation latency. The system clones each speaker's voice and translates it into the other language using corresponding tones, represented by different colored bars. This layout highlights the system's real-time translation capabilities while maintaining speaker identity across languages.

Simultaneous interpretation (SI), or simultaneous speech translation², stands as one of the most challenging tasks within the translation industry [19]. In recent years, remarkable breakthroughs have been witnessed in both machine translation and speech translation [3, 11, 40, 41, 43]. Inspired by the success of large language models (LLMs), contemporary research endeavors increasingly leverage LLMs for translation [1, 8, 17, 21, 23, 29, 39], aiming for superior translation performance.

However, most existing LLM-empowered speech translation systems [12] are limited to consecutive translation, where the model initiates translation only after the user concludes speaking, a common scenario in dialogue systems. In contrast, in contexts such as international conferences, where latency is of paramount importance to both speakers and listeners, simultaneous interpretation becomes indispensable. Current SI systems [5, 6, 11, 28, 33, 47, 49] face significant limitations. Some rely on cascaded architectures, which are prone to error propagation and high text and speech generation latency, while others merely support end-to-end speech-to-text translation, severely restricting their practical applicability. As highlighted in [31], the performance of existing SI systems is often overestimated due to the less stringent evaluation standards compared to offline speech translation systems. Although recent end-to-end speech-to-text models [6] have improved translation quality, they still fall short in supporting low-latency, high-quality speech-to-speech translation for truly seamless interpretation.

To tackle these challenges, we introduce **Seed LiveInterpret 2.0**, an end-to-end speech-to-speech simultaneous translation model that seamlessly integrates simultaneous speech-to-speech translation and voice cloning within a unified framework. As illustrated in Figure 2, **Seed LiveInterpret 2.0** enables multilingual live conversations with natural, real-time speech translation. The initial language model is pretrained following the methodology of the Seed LLM family [2, 4, 36]. We then extend this model by integrating a pre-trained audio encoder, transforming it into a multi-modal LLM capable of processing streaming audio as input. This multi-modal LLM is subsequently trained through large-scale multi-task continual learning to autoregressively

²In this paper, we use the terms simultaneous interpretation and simultaneous speech translation interchangeably. For details about the **Seed LiveInterpret 1.0**, please refer to our previous technical report [6].

generate outputs comprising text tokens (optional) and audio tokens for real-time speech synthesis [2]. To further enhance its performance, we fine-tune the model on high-quality human-labeled data, improving its instruction following, multi-speaker discrimination, translation policy, and other critical capabilities necessary for effective simultaneous interpretation.

Recognizing the challenges of optimizing simultaneous translation under strict latency constraints, we propose a novel reinforcement learning framework for simultaneous translation that strategically balances fine-grained stepwise feedback with holistic sequence-level feedback by explicitly addressing two complementary objectives: intra-segment consistency and inter-segment coherence. Specifically, the framework combines multi-dimensional single-turn rewards, which provide immediate feedback on translation fidelity and timing at each step to ensure intra-segment consistency, with unified multi-turn rewards that assess the overall quality and coherence of the entire output sequence, ensuring inter-segment consistency. To address the optimization challenges arising from these complementary objectives, we adopt a two-stage training scheme: initially warming up the model by optimizing only the single-turn rewards to internalize human priors and stabilize learning, followed by further training with the multi-turn reward that jointly considers process and outcome metrics. This integrated approach enables more effective and robust reinforcement learning for simultaneous translation under strict real-time requirements.

Comprehensive experiments demonstrate that `Seed LiveInterpret 2.0` achieves competitive translation quality at ultra-low latency in both Chinese-to-English and English-to-Chinese translation directions, striking an optimal balance between real-time responsiveness and semantic accuracy. This research pushes the boundaries of simultaneous speech-to-speech translation by presenting a robust, natural, and end-to-end solution suitable for live applications. Our key contributions encompass a unified speech-to-speech architecture, cross-language voice cloning, and translation performance approaching human-level accuracy, as shown in Figure 1.

2 Training

2.1 Continual Training and Supervised Fine-tuning

To achieve effective modality alignment between text and speech and enhance cross-lingual capabilities, we adopt a comprehensive multitask multimodal continual training (CT) strategy. Our CT dataset encompasses nearly 100 billion tokens from diverse multimodal tasks, including audio-to-text transcription, text-to-audio synthesis, and text-only processing tasks. Furthermore, to maximize training efficiency and ensure data quality, we employ rigorous filtering procedures based on speech quality metrics.

Following continual training, we conduct supervised fine-tuning on high-quality, human-annotated data to activate crucial capabilities required for simultaneous speech interpretation. This process enables the model to develop a data-driven read-write policy [6], multi-speaker discrimination, speech translation, and voice cloning abilities. The supervised fine-tuning significantly enhances the model’s instruction-following capabilities and overall performance across essential interpretation tasks. This fine-tuned model serves as a robust foundation for subsequent reinforcement learning, enabling more targeted and effective improvements.

2.2 Reinforcement Learning

2.2.1 Problem Formulation

Modern simultaneous translation systems employ duplex processing where input streams are segmented into sequential audio chunks. Formally, we represent an input-output sequence as:

$$x_{1:T} := (\text{audio}_1, y_1), (\text{audio}_2, y_2), \dots, (\text{audio}_T, y_T)$$

where each audio chunk (audio_t) corresponds to incremental translation y_t . We denote (audio_t, y_t) as the t -chunk in a sequence, and $\text{audio} := (\text{audio}_1, \text{audio}_2, \dots, \text{audio}_T)$ as the aggregated audio from 1 to T . In every t -chunk, we have $y_t := (y_t^1, y_t^2, \dots, y_t^n, \dots, y_t^N)$, where N is the length of the output.³ The model utilizes both

³For simplicity, we denote the length of the output of each t -chunk as N .

the current audio chunk (audio_t) and preceding context $x_{<t}$ to generate translation y_t through policy:

$$y_t \sim \pi_\theta(\cdot | \text{audio}_t, x_{<t}),$$

where π_θ is a policy with parameters θ that determines the translation strategy. The complete trajectory probability is defined as:

$$\pi_\theta(y_{1:T} | \text{audio}) := \prod_{t=1}^T \pi_\theta(y_t | \text{audio}_t, x_{<t}) = \prod_{t=1}^T \prod_{n=1}^N \pi_\theta(y_t^n | y_t^{<n}, \text{audio}_t, x_{<t}).$$

We denote r_t^n as the reward for n -th token in t -chunk. The objective of RL is to maximize the accumulated reward along every trajectory, i.e.,

$$\mathcal{J}(\theta) = \max_{\theta} \mathbb{E}_{\substack{\text{audio} \sim \mathcal{D} \\ y \sim \pi_\theta(\cdot | \text{audio})}} \left[\sum_{t=1}^T \sum_{n=1}^N \gamma^{N \times t+n} r_t^n \right], \quad (1)$$

where \mathcal{D} is the training dataset. The following sections elaborate on how r_t^n is designed.

2.2.2 Reward Design: Balancing Single-turn and Multi-turn Feedback

In reinforcement learning, reward mechanisms can be broadly categorized based on the temporal scope of the feedback they provide [38]: (1) *single-turn rewards*, which provide immediate feedback assessing intermediate reasoning or generation steps at each individual decision point, and (2) *multi-turn rewards*, which evaluate the quality of the entire output sequence, reflecting long-term, cumulative outcomes across multiple decision steps. Traditionally, reinforcement learning for language tasks often relies on reward models trained on human preference data. However, recent work has demonstrated that effective rewards can also be constructed directly from string matching, semantic similarity with labeled data [24, 27], or verifiable, rule-based criteria [20]. These alternative approaches open new avenues for designing reward functions that do not depend solely on costly human annotations.

Simultaneous translation systems, in particular, pose distinct challenges that call for a nuanced reward design. They require optimizing two complementary objectives: (1) *intra-segment consistency*, which demands that partial, real-time outputs maintain semantic and temporal integrity at each incremental step — a goal naturally suited to single-turn reward design, and (2) *inter-segment coherence*, which ensures semantic and temporal consistency across the entire translated sequence — a goal addressed through multi-turn reward design that evaluates cumulative sequence-level quality.

Motivated by these considerations, we propose a novel framework combining multi-dimensional single-turn rewards that provide fine-grained, stepwise feedback with unified multi-turn rewards that enforce global coherence and latency constraints throughout the full translation trajectory. This dual reward strategy enables more effective and balanced optimization for simultaneous translation.

Single-turn Reward Motivated by recent successes in reinforcement learning with verifiable rewards (RLVR) [14, 20, 24, 27], we introduce a multi-dimensional single-turn reward that evaluates both the fidelity and quality of intermediate translation steps. Unlike traditional reward schemes focused solely on final outputs, our approach leverages granular feedback at each incremental step, which we empirically find to correlate strongly with human evaluation metrics.

Formally, given an audio sequence $\{\text{audio}_t\}_1^T$ and the corresponding ground-truth $\{y_t^*\}_1^T$, we define intra-segment rewards along five derived dimensions:

- Detection Accuracy Reward (r^1): Encourage listening to avoid premature translation by penalizing outputs generated before the completion of semantic units:

$$r_t^1 := \mathbb{I}(|y_t| = 0) \cdot \mathbb{I}(|y_t^*| = 0),$$

where $\mathbb{I}(\cdot)$ is an indicator function, and $|y_t|$ means the number of tokens in y_t .

- Translation Initiative Reward (r^s): Encourage speech translation by rewarding the generation of confirmed semantic units as soon as they become available: $r_t^s := \mathbb{I}(|y_t| > 0) \cdot \mathbb{I}(|y_t^*| > 0)$.
- Translation Quality Reward (r^q): Rewards translation quality by measuring the closeness of y_t to reference y_t : $r_t^q := \text{Trans}(y_t, y_t^*)$, where $\text{Trans}(\cdot, \cdot)$ quantifies translation quality.
- Time Compliance Reward (r^c): Encourages adherence to reference timing by rewarding generated speech durations that match the reference durations:

$$r_t^c := \text{clip}\left(1 - \frac{1}{c} \max\left(0, \frac{\text{Time}_{y_t}}{\text{Time}_{y_t^*}} - 1\right), -1, 1\right),$$

where Time_{y_t} indicates the audio duration of the translated speech, and c is a constant.

- Format Consistency Reward (r^f): Encourages the structural correctness by rewarding outputs that match a predefined pattern \mathcal{P} via regular expression matching: \mathcal{P} : $r_t^f := \text{RegexMatch}(y_t, \mathcal{P})$, where $\text{RegexMatch}(y_t, \mathcal{P})$ returns 1 if y_t contains a substring matching the pattern \mathcal{P} else 0.

Therefore, the derived multi-dimensional single-turn reward for a given audio sequence $\{\text{audio}_t\}_1^T$ is $\{r_t\}_1^T$, and the reward r_t is defined as:

$$r_t := \begin{cases} w^l r_t^l, & \text{if } |y_t^*| = 0, \\ \sum_{k \in \{\text{s}, \text{q}, \text{c}, \text{f}\}} w^k r_t^k, & \text{otherwise.} \end{cases} \quad (2)$$

where w are weights balancing the relative importance of each reward component.

Multi-turn Reward While our single-turn reward provides detailed, stepwise feedback that balances latency and translation quality at each incremental step, it does not fully capture the long-term dependencies and cumulative effects inherent in simultaneous translation. In particular, when the generated target audio increasingly lags behind the source, it causes disruptive delays that degrade user experience. To address these global sequence-level dynamics, we design a complementary multi-turn reward that evaluates the entire output sequence holistically.

This multi-turn reward enforces two critical objectives:

- Lagging Reward (r^L): Encourages timely translation by penalizing long waiting times, and is defined as: $r^L := -\max\left(l, \frac{1}{K} \sum_{k=1}^K d_k\right)$, where l is a reference threshold representing the maximum acceptable wait, K is the number of translation chunks, and d_k denotes the number of waited chunks before the k -th translation chunk.
- Sequence-level Translation Quality Reward (r^Q): Rewards the translation quality of the generated translated sequence: $r^Q := \text{Align}(y, \text{audio})$, where $\text{Align}(\cdot)$ quantifies the sequence-level translation quality.

The multi-turn reward of an audio sequence is defined as:

$$r^S := w^L r^L + w^Q r^Q.$$

To ensure stability and comparability among reward components, each reward is normalized by subtracting its mean and dividing by its standard deviation, computed over training batches. The final reward at each time step is the sum of the normalized rewards, effectively blending local, stepwise feedback with global, sequence-level guidance. By integrating these global constraints with fine-grained process rewards, our multi-turn reward function provides a balanced optimization signal that guides the model toward producing translations that are both timely and semantically accurate, ensuring end-to-end coherence across the full output. We also incorporate a KL divergence penalty term, $\text{KL}(\pi_\theta \| \pi_{\text{ref}})$, to regularize the learned policy toward the reference policy, promoting stable and reliable learning behavior.

2.2.3 Stabilizing RL Training

We optimize our defined objective (Eq. 1) through Proximal Policy Optimization (PPO) [35], which enables stable and efficient policy updates via a clipped objective function. The training objective is formulated as follows:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{\substack{\text{audio} \sim \mathcal{D} \\ y \sim \pi_{\theta_{\text{old}}}}} \left[\sum_{n=1, t=1}^{N, T} \min \left(\frac{\pi_{\theta}(y_t^n | \text{audio}_t, x_{<t})}{\pi_{\theta_{\text{old}}}(y_t^n | \text{audio}_t, x_{<t})} A_t^n, \text{clip} \left(\frac{\pi_{\theta}(y_t^n | \text{audio}_t, x_{<t})}{\pi_{\theta_{\text{old}}}(y_t^n | \text{audio}_t, x_{<t})}, 1 - \varepsilon, 1 + \varepsilon \right) A_t^n \right) \right].$$

Here, $\text{audio} = \{\text{audio}_t\}_1^T$ denotes the audio sequence of the input and $y = \{y_t\}_1^T$ represents the translated response sampled from the old policy $\pi_{\theta_{\text{old}}}$. The advantage estimate A_t^n is computed using Generalized Advantage Estimation (GAE) [34].

Vanilla PPO underperforms in our setting because verified reward signals are prone to exploitation. For instance, when the lagging reward dominates, the model tends to produce trivial translations prioritizing latency over quality. Moreover, some rewards, such as the lagging reward, are easier to optimize than translation quality rewards, leading to an imbalance. Due to the tight coupling and diversity of these rewards, tuning their individual weights is challenging and often ineffective. To address these issues and stabilize training, we employ two main strategies: an adaptive KL penalty [35, 51] and a two-stage reinforcement learning training scheme.

Adaptive KL KL regularization is crucial as it constrains the policy to remain close to the reference model, thereby reducing reward hacking and preventing extreme outputs. However, controlling KL divergence is more difficult in sequences combining audio and text tokens because of their greater length, which naturally results in higher cumulative KL divergence. Consequently, the KL penalty coefficient β must be set higher than in conventional RLHF settings. Following [51], we adopt a proportional controller in log-space to adaptively adjust β , ensuring the KL divergence remains close to a predefined target.

$$\beta_{s+1} := \beta_s (1 + K_{\beta} e_s), \quad e_s := \text{clip} \left(\frac{\text{KL}(\pi_{\theta} \| \pi_{\text{ref}})}{\text{KL}_{\text{target}}} - 1, -0.2, 0.2 \right),$$

where s denotes the current training step, and K_{β} is a hyper-parameter controlling the adjustment of β , and $\text{KL}_{\text{target}}$ is a pre-defined target KL divergence.

Two-Stage RL Training Scheme Jointly optimizing single-turn and multi-turn rewards presents a challenge: single-turn stepwise rewards are generally easier to optimize, which can cause the model to overemphasize them while neglecting sequence-level rewards that are crucial for improving overall speech translation quality. However, these two reward types are complementary—single-turn stepwise rewards embed human priors that effectively guide early exploration, whereas multi-turn rewards drive performance refinement and global coherence. To harness this synergy, we adopt a two-stage training scheme. In the first stage, the model is warmed up by optimizing only the multi-dimensional single-turn rewards, allowing it to internalize human priors and achieve stable learning dynamics. In the second stage, the model is further trained using the multi-turn reward that combines both process and outcome components, enabling it to refine and balance latency and translation quality effectively. This staged approach fosters stable learning and efficient exploration, ultimately yielding a more robust and reliable reinforcement learning framework for simultaneous translation.

3 Experiments

3.1 Datasets

Our primary experiments are conducted on the recently introduced RealSI dataset [6], which encompasses both Chinese-to-English (zh-en) and English-to-Chinese (en-zh) translation directions. The dataset is sourced from a wide range of domains—including technology, healthcare, education, finance, law, environment, entertainment, science, sports, and art—and features speakers who predominantly speak naturally and casually, without extensive preparation. Each sample consists of approximately five minutes of continuous speech, providing a realistic and challenging benchmark for simultaneous interpretation systems. Notably, the RealSI dataset

reflects real-world scenarios more closely than many existing benchmarks, capturing the spontaneous and diverse nature of everyday speech across various fields.

In addition, we evaluate and compare our approach using sentence-level simultaneous translation datasets. Given the limited availability of high-quality public test data tailored for simultaneous interpretation scenarios in industry, we combine public datasets with proprietary internal datasets for evaluation.

3.2 Evaluation Metrics

For text translation quality assessment, we primarily rely on the idea of the human evaluation metric, Valid Information Proportion (VIP) [6], which measures how accurately the translation output conveys the speaker’s original intent for each semantic fragment, closely aligning with human interpreter judgments. Additionally, we employ automated metrics such as BLEURT [37] and COMET [32] as supplementary references. Nonetheless, consistent with findings in [12, 26, 42], these automated metrics may not fully reflect the model’s true capabilities.

For speech-to-speech assessment, we propose the Speech Valid Information Proportion (SVIP) as a comprehensive human evaluation metric. Building upon the established Valid Information Proportion (VIP) framework [6], SVIP measures the proportion of valid speech semantic fragments within a complete speech session. A speech semantic fragment is considered valid when it effectively conveys the core information from the source speech, accurately represents the speaker’s original intent, maintains delivery latency within acceptable thresholds for effective communication, sustains an appropriate pace for listener comprehension, and achieves acoustic quality that meets standards for clarity and intelligibility. SVIP provides a holistic assessment that captures not only semantic accuracy but also the pragmatic elements essential for successful spoken communication across languages. Detailed definition of SVIP can be found in Appendix A.

For latency evaluation, we adopt the First Letter Appearance Lagging (FLAL) metric [6] to measure the time until the system outputs the first determined translation at the paragraph level. At the sentence level, we use the widely adopted Average Lagging (AL) [25] and Length Adaptive Average Lagging (LAAL) [30] metrics to compare latency across different methods.

3.3 Results

Baselines We compare our `Seed LiveInterpret` 2.0 with the open-source model `SeamlessStreaming` [5]. Due to the limited availability of baseline models, we also evaluate against several commercial systems, denoted as `Commercial-B`, `Commercial-Y`, `Commercial-T`, and `Commercial-I`. It is important to note that some baseline models employ a rewriting strategy to refine their output translations, a practice generally not used by human interpreters. In contrast, our method generates translations only when sufficient information is available, aligning more closely with the approach of human interpreters.

Results on Longform Benchmark Our evaluation on the longform benchmark clearly demonstrates the strengths of `Seed LiveInterpret` 2.0 across both speech-to-text and speech-to-speech tasks, and Table 1 shows the results. For speech-to-text translation from Chinese to English (zh-en), our method achieves a human evaluation VIP score of 79.5, substantially outperforming all baselines. Alongside its superior translation quality, our approach delivers competitive latency metrics, with an AL of 2.58 and FLAL of 2.37. Similarly, for English to Chinese (en-zh) speech-to-text translation, our model attains a VIP score of 70.1, again significantly higher than baselines such as `Commercial-T` and `Commercial-I`. It also achieves the best latency results, with an AL of 2.71 and FLAL of 2.05, indicating a well-balanced trade-off between translation quality and delay.

In speech-to-speech translation, our method achieves the low latency measurements, outperforming the baseline systems by a substantial margin. This highlights the model’s ability to maintain high-quality output while minimizing translation delay. Notably, many commercial systems either do not support speech-to-speech translation or show significantly degraded performance in longform scenarios, highlighting the practical value of our approach for real-world applications. In speech-to-speech translation, our model demonstrates strong performance in both directions. For zh-en, it achieves the highest SVIP score of 67.8 and the lowest AL

Model	Speech-to-Text (zh-en)			Speech-to-Speech (zh-en)			
	VIP↑	AL↓	FLAL↓	SVIP↑	AL↓	FLAL↓	Voice Clone
Commercial-B	11.8	8.40	3.27	-	-	-	✗
Commercial-Y	33.2	3.62	5.90	-	-	-	✗
Commercial-T	50.1	2.41	2.35	-	-	-	✗
Commercial-I	53.2	4.48	6.62	3.0	48.21	8.12	✗
SeamlessStreaming	22.0	-	2.65	15.3	-	2.38	✗
Ours	79.5	2.58	2.37	67.8	5.18	2.71	✓

Model	Speech-to-Text (en-zh)			Speech-to-Speech (en-zh)			
	VIP↑	AL↓	FLAL↓	SVIP↑	AL↓	FLAL↓	Voice Clone
Commercial-B	15.5	4.71	1.88	-	-	-	✗
Commercial-Y	24.6	4.96	12.42	-	-	-	✗
Commercial-T	42.0	2.75	1.90	-	-	-	✗
Commercial-I	41.3	4.98	5.73	5.6	33.92	8.60	✗
SeamlessStreaming	6.0	-	2.24	2.7	-	2.39	✗
Ours	70.1	2.71	2.05	64.7	4.75	2.34	✓

Table 1 Performance comparison on longform simultaneous translation benchmark `RealsI` across speech-to-text and speech-to-speech tasks. VIP and SVIP represent human evaluation scores for translation quality, while AL and FLAL measure translation latency at the segment level. Higher scores indicate better performance for VIP and SVIP, while lower scores are better for latency metrics. Missing entries indicate systems that do not support the corresponding functionality.

of 5.18 among systems supporting this task, with an FLAL of 2.71. For en-zh, our approach reaches an SVIP of 64.7 and improves latency significantly, with an AL of 4.75 and FLAL of 2.34, outperforming all other speech-to-speech baselines. Notably, many commercial systems either do not support speech-to-speech translation or show degraded performance in longform scenarios, emphasizing the practical advantage of our approach in real-world applications. Importantly, our method is the only system in the comparison that supports voice cloning, enabling personalized speech output in simultaneous translation, which further enhances user experience and applicability.

Overall, these results highlight that `Seed LiveInterpret 2.0` consistently achieves state-of-the-art translation quality with low latency across both language pairs and modalities, confirming its effectiveness and robustness for simultaneous translation in longform settings.

Results on Sentence-Level Benchmark We evaluate our `Seed LiveInterpret 2.0` model against established baselines on sentence-level zh-en and en-zh datasets, examining the performance across both speech-to-text and speech-to-speech simultaneous translation tasks, and Table 2 shows the results. For speech-to-text translation, our approach consistently achieves the highest translation quality across both datasets, outperforming commercial systems. Specifically, our model attains BLEURT scores of 64.9 (zh-en) and 62.0 (en-zh), along with COMET scores of 84.1 and 85.3, respectively. These results demonstrate a clear advantage in translation accuracy. In terms of latency, our method achieves the lowest AL on zh-en and competitive AL on en-zh, indicating faster translation without sacrificing quality. While `Commercial-T` achieves slightly better FLAL on zh-en, it does so at the cost of significantly lower BLEURT and COMET scores. Similarly, other commercial systems such as `SeamlessStreaming` and `Commercial-B` exhibit trade-offs between latency and quality, whereas our model maintains a strong balance between both.

For speech-to-speech translation, our method also leads in translation quality, with BLEURT scores of 60.7 and 57.6 and COMET scores of 83.6 and 83.5, respectively. Although its latency metrics are slightly higher than the lowest values reported by some baselines, our approach consistently achieves a favorable trade-off by delivering superior translation quality alongside competitive latency. Overall, these results highlight the ability

Model	Speech to Text (zh-en)					Speech to Speech (zh-en)				
	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓
Commercial-Y	61.5	81.3	2.03	2.11	1.80	-	-	-	-	-
Commercial-T	61.9	81.5	1.61	1.75	1.74	-	-	-	-	-
Commercial-B	47.2	70.3	2.39	2.66	2.21	44.8	71.6	12.00	12.26	7.89
Commercial-I	55.9	79.0	3.10	3.22	4.62	53.2	79.6	6.90	7.02	4.73
SeamlessStreaming	55.8	76.4	1.68	1.87	2.36	49.6	75.2	2.96	3.10	2.53
Ours	64.9	84.1	1.37	1.56	2.12	60.7	83.6	3.56	3.79	3.08

Model	Speech to Text (en-zh)					Speech to Speech (en-zh)				
	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓
Commercial-Y	59.5	83.1	3.25	3.51	4.84	-	-	-	-	-
Commercial-T	60.1	84.1	1.46	1.71	1.51	-	-	-	-	-
Commercial-B	55.2	81.2	2.62	2.91	2.07	49.0	77.7	13.10	13.57	8.91
Commercial-I	60.0	83.4	3.25	3.54	4.86	56.1	82.1	7.25	7.58	5.49
SeamlessStreaming	48.2	75.2	1.43	1.69	2.06	40.4	69.8	3.30	3.69	2.17
Ours	62.0	85.3	2.17	2.18	2.28	57.6	83.5	2.81	3.12	2.38

Table 2 Comparisons of our method with baseline approaches on speech-to-text and speech-to-speech simultaneous translation tasks with respect to translation quality and latency on the sentence-level datasets.

Model	Speech-to-Text (zh-en)			Speech-to-Speech (zh-en)		
	VIP↑	AL↓	FLAL↓	SVIP↑	AL↓	FLAL↓
Ours _(SFT)	75.1	2.82	3.90	66.6	5.80	4.26
Ours	79.5	2.58	2.37	67.8	5.18	2.71

Table 3 Performance comparison on longform simultaneous translation benchmark **RealSI**.

of **Seed LiveInterpret 2.0** to effectively balance translation quality and latency across both speech-to-text and speech-to-speech tasks, outperforming existing commercial systems on sentence-level benchmarks.

4 Analysis

4.1 Comparisons of SFT with RL

Tables 3 and 4 compare our **Seed LiveInterpret 2.0** with its SFT version across benchmarks and tasks. On the longform simultaneous translation benchmark (**RealSI**), both our model and its SFT version demonstrate strong performance across speech-to-text and speech-to-speech tasks. Notably, our method achieves a higher human evaluation VIP score for speech-to-text translation (79.5 vs. 75.1), while significantly reducing latency. Specifically, FLAL decreases from 3.90 to 2.37, and AL improves from 2.82 to 2.58, indicating faster and more responsive translations with minimal quality trade-off. In speech-to-speech translation, our method maintains competitive SVIP scores (67.8 vs. 66.6) while substantially lowering latency metrics. The FLAL metric drops from 4.26 to 2.71, and AL reduces from 5.80 to 5.18, underscoring our model’s ability to deliver timely and high-quality speech output.

On sentence-level datasets, our method consistently outperforms the SFT model in translation quality, achieving BLEURT improvements of 0.4 to 1.0 points alongside modest gains in COMET scores. More importantly, our approach delivers significant latency reductions across key metrics, including AL, LAAL, and FLAL. For instance, in speech-to-text translation (zh-en), AL decreases from 1.69 to 1.37 and FLAL from 3.03 to 2.12; similarly, for speech-to-text (en-zh), AL is reduced from 2.99 to 2.17 and FLAL from 3.29 to 2.28. In speech-to-speech tasks, our model maintains comparable or slightly improved translation quality while substantially lowering latency, exemplified by a reduction in AL from 3.99 to 2.81.

Overall, our **Seed LiveInterpret 2.0** method consistently optimizes the trade-off between translation quality and latency across datasets and tasks. While quality improvements are moderate, the significant latency

Model	Speech to Text (zh-en)					Speech to Speech (zh-en)				
	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓
Ours _(SFT)	64.5	83.9	1.69	1.91	3.03	59.7	83.2	3.57	3.80	3.08
Ours	64.9	84.1	1.37	1.56	2.12	60.7	83.6	3.56	3.79	3.08
Model	Speech to Text (en-zh)					Speech to Speech (en-zh)				
	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓	BLEURT↑	COMET↑	AL↓	LAAL↓	FLAL↓
Ours _(SFT)	61.0	84.7	2.99	3.01	3.29	57.6	83.5	3.99	4.32	3.40
Ours	62.0	85.3	2.17	2.18	2.28	57.6	83.5	2.81	3.12	2.38

Table 4 Comparisons of Seed LiveInterpret 2.0 with baseline approaches on speech-to-text and speech-to-speech simultaneous translation tasks with respect to translation quality and latency on the sentence-level datasets.

reductions highlight the effectiveness of reinforcement learning in producing faster, high-quality simultaneous translation.

4.2 Balanced Reward to Prevent Hacking

Despite applying the stabilization techniques described in Section 2.2.3, we observe that our model remains vulnerable to reward hacking, even when using verifiable rewards. Table 5 illustrates this phenomenon with respect to the Time Compliance Reward (r^c) defined in Section 2.2.2, which is designed to encourage the generated speech durations to adhere to the reference source durations. When training the model solely with the r^c reward, we find that the model exploits this signal by significantly reducing the duration of the generated audio — approximately a 35% decrease on both the en-zh and zh-en datasets. Correspondingly, the number of generated text tokens also declines by about 15%. This reduction leads to a substantial drop in BLEURT scores, indicating degraded translation quality. This behavior suggests that while the model successfully aligns the duration of generated speech with the source audio (thus maximizing the r^c reward), it does so at the cost of omitting substantial semantic content. In other words, the model learns to produce shorter outputs that satisfy the temporal constraint but sacrifice translation fidelity.

Therefore, we find that it is necessary to add an adversarial quality reward (r^q) alongside the r^c reward to balance translation quality with temporal constraints. With this combined reward scheme, the model maintains a comparable number of text tokens while reducing the duration of the audio by only about 15%. Importantly, BLEURT scores remain stable, indicating preserved translation quality. These results suggest that incorporating adversarial or complementary rewards is essential to prevent reward hacking and encourage balanced optimization across multiple objectives.

Model	Reward	Chinese → English			English → Chinese		
		Text Length	Audio Duration	BLEURT	Text Length	Audio Duration	BLEURT
Ours _(SFT)	-	113,000	9,340	56.36	23,000	2,080	58.00
Ours	r^c	97,000	6,350	48.31	21,000	1,330	46.43
Ours	$r^c + r^q$	114,000	8,010	55.60	24,000	1,760	58.07

Table 5 Ablation study of reward configurations across both translation directions on in-house datasets. Text length indicates the total number of generated tokens, while audio duration shows the total speech length (in seconds) across the dataset. Incorporating adversarial or complementary rewards helps prevent reward hacking and encourages balanced optimization across multiple objectives.

4.3 Effect of Two-Stage RL Training Scheme

We investigate the impact of single-turn and multi-turn rewards on the performance of our RL model through an ablation study, comparing three training configurations. Table 6 presents the results. The first uses only single-turn rewards (r_t), the second relies solely on multi-turn rewards (r^S), and the third employs our proposed

two-stage training scheme combining both. Compared to the single-turn-only approach ($\text{Ours}_{(\text{single})}$), the multi-turn variant ($\text{Ours}_{(\text{multi})}$) significantly reduces latency—2.20 versus 2.62 for zh-en, and 3.16 versus 2.69 for English-to-Chinese. However, it achieves lower VIP scores, indicating a tendency to exploit the lagging reward (r^L) at the expense of translation quality. This trade-off is undesirable, as it favors speed over output accuracy. Conversely, the single-turn-only method exhibits relatively high latency, suggesting that focusing exclusively on process rewards limits the model’s ability to explore more efficient translation strategies.

Our Seed `LiveInterpret` 2.0 that employs a two-stage training strategy outperforms both single-turn and multi-turn models across key metrics. It achieves competitive translation quality while maintaining latency close to that of the outcome-only setup. Notably, in the en-zh direction, the two-stage model attains a high VIP score with reduced lagging, demonstrating that combining process and outcome rewards effectively balances translation quality and timeliness.

Model	Chinese → English				English → Chinese			
	BLEURT ↑	COMET ↑	VIP ↑	AL ↓	BLEURT ↑	COMET ↑	VIP ↑	AL ↓
$\text{Ours}_{(\text{single})}$	55.83	80.14	70%	2.62	57.08	85.47	76%	3.16
$\text{Ours}_{(\text{multi})}$	54.05	78.46	67%	2.20	58.35	85.70	72%	2.69
Ours	55.60	80.09	71%	2.30	58.07	85.60	76%	2.78

Table 6 Ablation study of different reward strategies across both translation directions on in-house datasets.

5 Related Work

Simultaneous Interpretation aims to translate spoken language in real time, facilitating seamless multilingual communication. Traditional SI systems typically rely on cascaded architectures that sequentially perform automatic speech recognition, machine translation, and text-to-speech synthesis [7, 13]. While these modular pipelines allow for targeted optimization at each stage, they suffer from error propagation and increased latency, which negatively impact overall translation quality and responsiveness. To mitigate these issues, recent research [18, 33, 47] has shifted towards end-to-end models that directly convert source speech into translated text or speech, thereby reducing latency and minimizing accumulated errors. However, most existing end-to-end SI frameworks [5, 6, 49] focus predominantly on speech-to-text translation and lack comprehensive support for speech-to-speech translation with speaker voice cloning, an essential feature for preserving speaker identity and enhancing user experience. Achieving the trifecta of low latency, high fidelity, and voice preservation remains a critical challenge in this domain.

Simultaneous translation methods generally adopt either fixed or adaptive policies for deciding when to read input and emit output. Fixed policy approaches follow predefined rules, such as reading a fixed number of source tokens before generating target tokens [9, 25]. These methods are simple but inflexible, often resulting in suboptimal latency-quality trade-offs. Adaptive policies dynamically determine read/write actions based on contextual alignment between source and target sequences, allowing for more nuanced and responsive translation [3, 5, 6, 22, 48, 50]. However, adaptive approaches rely heavily on alignment signals that can be noisy and challenging to optimize without reinforcement learning [45]. Recent work has leveraged RL to learn more effective read/write policies, overcoming limitations of rule-based and alignment-dependent methods [45, 46]. Despite these advances, RL-based methods have been primarily applied to text-to-text translation, with speech-to-speech simultaneous translation remaining largely unexplored.

Reinforcement learning has been widely employed to improve offline machine translation quality. Techniques such as reinforcement learning from human feedback have enhanced translation fluency, adequacy, and alignment with human preferences [10, 15, 16, 44]. Nonetheless, these approaches focus on text input and output and do not address the challenges of real-time or speech-to-speech translation. The most relevant RL-based simultaneous translation work [45, 46] concentrates on real-time text translation policies and does not extend to the full end-to-end speech-to-speech pipeline, which introduces additional complexities such as speech synthesis and voice cloning.

6 Conclusion

In this work, we have presented **Seed LiveInterpret 2.0**, an end-to-end speech-to-speech simultaneous translation approach that seamlessly integrates translation, voice cloning, and speech synthesis within a unified framework. By leveraging a duplex processing architecture and a multimodal large language model, our approach achieves ultra-low latency and natural cross-language voice cloning, addressing key limitations of prior cascaded and end-to-end models. We introduce a novel two-stage reinforcement learning framework that employs a unified reward design that balances fine-grained process-based feedback with global outcome-based objectives, enabling the model to optimize both translation quality and latency in real time. Extensive experiments demonstrate that our approach achieves competitive translation accuracy while maintaining stringent latency requirements, paving the way for more robust and natural simultaneous interpretation systems applicable in live multilingual communication scenarios. Future work will explore further improvements in voice personalization, speech stability, expressiveness, as well as scaling to a broader range of languages and acoustic conditions.

References

- [1] Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*, 2024.
- [2] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.
- [3] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. Monotonic infinite lookback attention for simultaneous machine translation. *arXiv preprint arXiv:1906.05218*, 2019.
- [4] Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*, 2024.
- [5] 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.
- [6] Shanbo Cheng, Zhichao Huang, Tom Ko, Hang Li, Ningxin Peng, Lu Xu, and Qini Zhang. Towards achieving human parity on end-to-end simultaneous speech translation via llm agent, 2024. URL <https://arxiv.org/abs/2407.21646>.
- [7] Kyunghyun Cho and Masha Esipova. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*, 2016.
- [8] 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.
- [9] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Efficient wait-k models for simultaneous machine translation. *arXiv preprint arXiv:2005.08595*, 2020.
- [10] Zhaopeng Feng, Shaosheng Cao, Jiahua Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*, 2025.
- [11] 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 *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 330–340, 2023.
- [12] 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.
- [13] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*, 2016.
- [14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [15] Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, et al. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*, 2025.
- [16] Zhiwei He, Xing Wang, Wenxiang Jiao, Zhusong Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*, 2024.
- [17] Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*, 2023.

- [18] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.
- [19] Roderick Jones. *Conference interpreting explained*. Routledge, 2014.
- [20] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\ ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- [21] 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.
- [22] Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 39–55, 2021.
- [23] Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*, 2024.
- [24] Trung Quoc Luong, Xinbo Zhang, Zhamming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 3, 2024.
- [25] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. *arXiv preprint arXiv:1810.08398*, 2018.
- [26] Dominik Macháček, Ondřej Bojar, and Raj Dabre. Mt metrics correlate with human ratings of simultaneous speech translation. *arXiv preprint arXiv:2211.08633*, 2022.
- [27] OpenAI. Reinforcement fine-tuning. <https://platform.openai.com/docs/guides/reinforcement-fine-tuning>, 2024.
- [28] Siqi Ouyang, Xi Xu, and Lei Li. Cmu’s iwslt 2025 simultaneous speech translation system. *arXiv preprint arXiv:2506.13143*, 2025.
- [29] Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. *arXiv preprint arXiv:2405.12915*, 2024.
- [30] 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.
- [31] Sara Papi, Peter Polak, Dominik Macháček, and Ondřej Bojar. How “real” is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics*, 13:281–313, 2025.
- [32] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.
- [33] 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.
- [34] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.
- [35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL <https://api.semanticscholar.org/CorpusID:28695052>.
- [36] ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.
- [37] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

- [38] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2024.
- [39] 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. *arXiv preprint arXiv:2310.13289*, 2023.
- [40] 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. URL <https://arxiv.org/abs/2207.04672>.
- [41] 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.
- [42] 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.
- [43] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [44] Nuo Xu, Jun Zhao, Can Zu, Sixian Li, Lu Chen, Zhihao Zhang, Rui Zheng, Shihan Dou, Wenjuan Qin, Tao Gui, et al. Advancing translation preference modeling with rlhf: A step towards cost-effective solution. *arXiv preprint arXiv:2402.11525*, 2024.
- [45] Ting Xu, Zhichao Huang, Jiankai Sun, Shanbo Cheng, and Wai Lam. Seqpo-simt: Sequential policy optimization for simultaneous machine translation. *arXiv preprint arXiv:2505.20622*, 2025.
- [46] Donglei Yu, Yang Zhao, Jie Zhu, Yangyifan Xu, Yu Zhou, and Chengqing Zong. Simulpl: Aligning human preferences in simultaneous machine translation. *arXiv preprint arXiv:2502.00634*, 2025.
- [47] Xingshan Zeng, Liangyou Li, and Qun Liu. Realtrans: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer. *arXiv preprint arXiv:2106.04833*, 2021.
- [48] Shaolei Zhang and Yang Feng. Hidden markov transformer for simultaneous machine translation. *arXiv preprint arXiv:2303.00257*, 2023.
- [49] 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.
- [50] Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. Simpler and faster learning of adaptive policies for simultaneous translation. *arXiv preprint arXiv:1909.01559*, 2019.
- [51] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Appendix

A Guidelines of SVIP

The SVIP assessment uses a two-step evaluation process. First, language quality is evaluated using VIP metric [6]. If the VIP score is zero, the SVIP score is automatically zero. For non-zero VIP scores, speech quality indicators (listed in Table 7) determine the final SVIP score: it is zero if any indicator scores 1 point, and it is 1 if all indicators score 3 points or higher. When some indicators score 2 points, the final decision depends on whether the overall message remains comprehensible at the sentence level.

We define the **speech quality indicators** as in Table 7:

Indicator	Description	Scoring Criteria
Latency	The time interval from when AI or human interpreters receive the original audio to when they output the interpreted speech, evaluating the real-time responsiveness of interpretation. Low latency ensures smooth communication flow and coherence.	5: Very short latency (<3 seconds) 4: Moderate latency (3-5 seconds) 3: Longer latency (5-7 seconds) 2: Extended latency (7-9 seconds) 1: Excessive latency (≥ 10 seconds)
Speech Rate	Whether the interpretation speech rate is appropriate for audience comprehension and language habits, ensuring smooth rhythm and rapid and effective information transmission while avoiding comprehension difficulties due to excessive speed or attention distraction due to excessive slowness, ensuring the audience can efficiently and smoothly obtain interpretation content.	5: Moderate speed, natural 4: Slightly fast or slow but acceptable 3: Obviously too fast or slow, affecting experience but not comprehension 2: Obviously too fast or slow, affecting content comprehension 1: Obviously too fast or slow, seriously affecting content comprehension
Pronunciation	The accuracy and clarity of pronunciation in simultaneous interpretation, evaluating whether the speech expression is clear and comprehensible, without obvious reading or stuttering situations, and the degree to which pronunciation quality affects audience understanding of information.	5: Completely accurate pronunciation 4: Minor pronunciation issues, acceptable 3: Obvious pronunciation errors, affecting experience but not comprehension 2: Obvious pronunciation errors, affecting content comprehension 1: Obvious pronunciation errors, seriously affecting content comprehension
Fluency	Whether there are phenomena affecting listening experience in speech output, and whether the overall speech rhythm meets audience listening habits. Good speech fluency should show continuous and natural expression, without obvious interruptions or stuttering.	5: High fluency, coherent expression, no obvious pauses, repetitions, or hesitations 4: Occasional short pauses or minor hesitations, overall good speech flow and rhythm 3: Obvious fluency issues, affecting rhythm but basically acceptable 2: Obvious stuck or stuttering, significantly reduced fluency, affecting comprehension 1: Serious fluency issues, basically incomprehensible

Table 7 Detailed descriptions of speech quality indicators.

B Contributions

Project Lead

Shanbo Cheng (chengshanbo@bytedance.com)

Core Contributors

Yu Bao, Zhichao Huang, Yu Lu, Ningxin Peng, Lu Xu, Runsheng Yu

Contributors

Rong Cao, Yujiao Du, Ting Han, Yuxiang Hu, Zeyang Li, Sitong Liu, Shengtao Ma, Shiguang Pan, Jiongchen Xiao, Nuo Xu, Meng Yang, Rong Ye, Yiming Yu, Jun Zhang, Ruofei Zhang, Wanyi Zhang, Wenhao Zhu, Liehao Zou

Acknowledgement

Jinpeng Cai, Pengli Chen, Zhuo Chen, Qianqian Dong, Meng Ge, Minglun Han, Shan He, Xiaomin Huang, Youjia Huang, Yuanyuan Huo, Fanliu Kong, Hang Li, Mengnan Li, Xiaoyang Li, Xingxing Li, Yifu Li, Shouda Liu, Xinyu Liu, Xiaoying Jia, Yongjian Mao, Junjie Pan, Xinghua Qu, Hongbin Ren, Chen Shen, Kefei Sun, Tian Tan, Ming Tu, Bo Wang, Yuping Wang, Manlian Wu, Hanzhang Xia, Bowen Xiao, Yangfei Xu, Bing Yang, Yu Yang, Bairen Yi, Yang Zhang

Supervision

Lu Lu, Yuxuan Wang, Yonghui Wu

We would also like to sincerely thank all our colleagues for their invaluable support.

(Last-Name in Alphabetical Order)