DuraS2ST: Chain-of-Thought and Reinforcement Learning for Duration-Aligned Speech-to-Speech Translation

Yayue Deng[1,2*], Dingdong Wang[1], Yuxuan Hu[2], Jinyu Li[2], Yanqing Liu[2], Yuanyuan Wang[1], Weidong Chen[1], Helen M. Meng[1], Shujie Liu[2†], Xixin Wu[1†]
[1]The Chinese University of Hong Kong, China
[2]Microsoft Corporation

Abstract

Speech-to-speech translation (S2ST) in time-sensitive applications such as video dubbing requires not only semantic fidelity and speaker preservation, but also strict duration consistency to avoid audio-visual misalignment. However, existing S2ST systems largely generate target speech without explicit temporal planning, making duration control an unresolved challenge. We introduce **DuraS2ST**, a duration-aligned reasoning framework that enables a single speech language model to first generate an explicit chain-of-thought (CoT) for planning target wording and phonetic length, and then synthesize the corresponding speech tokens. To support this paradigm, we construct **DuraSet-440K**, a high-quality duration-aligned CoT corpus for supervised initialization. We further optimize the model with multi-modal multi-dimensional reinforcement learning, using a Duration Margin Reward to balance translation quality and duration consistency, and Modality-Aware Reward Attribution to assign rewards to appropriate token spans. Experiments on CVSS-T show that DuraS2ST achieves a strong balance between translation quality and duration consistency, outperforming competitive open-source and commercial baselines. Project page: `https://github.com/Mia11939/DuraS2ST`.

1 Introduction

Speech-to-speech translation (S2ST) aims to translate speech across languages while preserving the speaker's acoustic characteristics. Recent progress in end-to-end discrete unit modeling and large-scale Speech Language Models (SLMs) has significantly improved semantic fidelity and speaker preservation (Lee et al., 2022; Huang et al., 2023; Inaguma et al., 2023; Rubenstein et al., 2023; Communication et al., 2023b; Dong et al., 2023). However, most existing systems generate target speech without explicit temporal control. In time-sensitive scenarios such as video dubbing and simultaneous interpretation, duration mismatch can cause audio-visual desynchronization, overlapping speech, and degraded user experience (Wu et al., 2023).

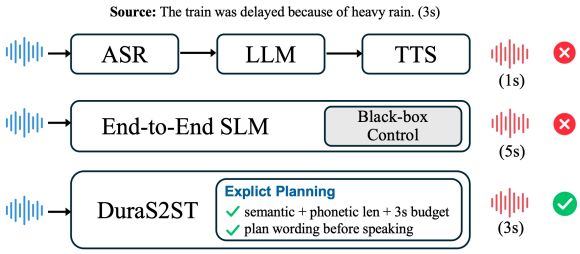


Figure 1: Existing systems suffer from duration mismatch: "你说得对" can be translated as either "You're right" or "I think you are absolutely right about that," leading to different speech durations. See Appendix D.

Prior work on isochronous S2ST mainly follows two paradigms:cascaded and end-to-end systems. Cascaded systems (Wahlster, 2000) often rely on post-hoc time-stretching, which may distort prosody and speaker timbre. End-to-end systems introduce speed tokens, length embeddings, or duration-conditioned generation (Wu et al., 2023; Le et al., 2024; Cheng et al., 2025a). While these methods improve temporal alignment, they typically treat duration as a low-level acoustic constraint: the decoder is encouraged to compress or expand speech to satisfy a target length. As shown in Figure 1, such black-box control lacks an explicit planning mechanism for balancing semantic content with phonetic length. In contrast, human dubbing often involves deliberate paraphrasing, where speakers adjust wording before articulation to fit the available time while preserving the core meaning.

Motivated by this observation, we propose **DuraS2ST**, a unified reasoning-based S2ST framework that formulates duration alignment as a planning problem rather than an acoustic post-processing problem. Built on a pre-trained SLM,

[*]Work done during an internship at Microsoft
[†] Corresponding Author

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 34802–34821
October 24-29, 2026 ©2026 Association for Computational Linguistics

DuraS2ST first generates an explicit reasoning trajectory that plans the target content and its phonetic footprint, and then produces the final interleaved text-acoustic sequence in a single autoregressive process. This allows semantic planning to directly guide subsequent speech generation.

DuraS2ST is trained in two stages. First, we perform supervised fine-tuning on a duration-aligned reasoning corpus that we construct, named **DuraSet-440K**. The model is supervised to plan, inside its chain-of-thought, both the target semantic content and its phonetic length before producing any acoustic token. This forces the model to actively balance semantic density against phonetic footprint at the symbolic level, replacing the black-box length control of speed tokens and length embeddings with explicit reasoning that drives subsequent acoustic synthesis. Second, we adopt Group Relative Policy Optimization (Shao et al., 2024) for our multi-modal generation task with a multi-dimensional reward design. We propose the **Duration Margin Reward** (DMR) to balance translation quality and duration consistency, and the **Modality-Aware Reward Attribution** (MARA) to shield the textual chain-of-thought from cross-modality reward contamination. Our contributions are summarized as follows:

- We propose **DuraS2ST**, the first unified framework that reformulates duration-aligned S2ST as a reasoning problem rather than an acoustic post-processing one. To support this paradigm, we further release **DuraSet-440K**, a large-scale duration-aligned reasoning corpus.
- We introduce a multimodal RL optimization strategy for duration-aligned S2ST. Specifically, **DMR** provides a smooth soft constraint for duration consistency, while **MARA** attributes reward signals to modality-specific token spans, preventing cross-modality reward contamination.
- Experiments on CVSS-T show that DuraS2ST achieves a strong balance among translation quality and duration consistency, outperforming competitive open-source and commercial baselines. Notably, we demonstrate that DuraS2ST achieves substantial gains even in a zero-shot RL setting.

2 DuraS2ST

To effectively resolve the duration mismatch in Speech-to-Speech Translation (S2ST), we propose **DuraS2ST**, as shown in Figure 2, a unified framework built upon a single pre-trained multimodal language model. The central strength of DuraS2ST is its ability to explicitly negotiate the balance between semantic fidelity and acoustic duration. Unlike conventional cascaded pipelines or standard modular methods that treat duration control as a disconnected post-processing step, our approach leverages the multimodal language model's inherent reasoning capacity to formulate a rationale prior to acoustic execution, dynamically adjusting translation vocabulary and explicitly planning the resulting phoneme-sequence lengths to satisfy the duration budget.

2.1 Phase I: SFT Training

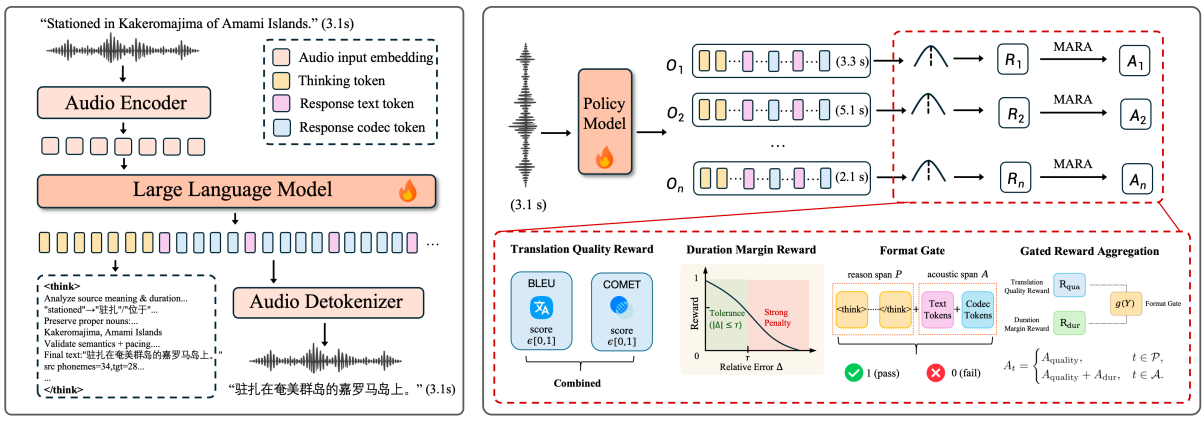
As the first step of our training paradigm, we perform an SFT procedure (Wei et al., 2022) to instruct the SLM, Step-Audio-2-mini-Think (Wu et al., 2025), to adhere to strict duration constraints and execute a reasoning process in S2ST tasks.

Leveraging our constructed DuraSet-440K corpus (Section 3), denoted as $\mathcal{D}_{\text{dura}}$, we conduct SFT to harness the model's pre-trained capabilities. Specifically, the model is guided to explicitly balance semantic fidelity and duration constraints by selecting the optimal translation vocabulary under the premise of maintaining comparable lengths between the source and target speech. For a given training sample ($\mathbf{X}_{\text{src}}$, $\mathbf{Y}$), with $\mathbf{X}_{\text{src}}$ representing the source speech representation, the output sequence $\mathbf{Y}$ is structured as a concatenation ($\oplus$) of two explicit phases:

$$\underbrace{\texttt{<think>},\ \mathbf{y}_{\text{CoT}},\ \texttt{</think>}}_{\text{Reasoning Phase}} \oplus \underbrace{\texttt{<tts_start>},\ \mathbf{y}_{\text{TA4}},\ \texttt{<tts_end>}}_{\text{Response Phase}}, \quad (1)$$

where $\mathbf{y}_{\text{CoT}}$ represents the reasoning trajectory in which the model could jointly optimize semantic correspondence and length alignment, determining appropriate target words and sentence structures to match the source speech duration. Conditioned on the reasoning results, the model generates $\mathbf{y}_{\text{TA4}}$, an interleaved TA4 sequence in which each translated text (T) token is paired with four acoustic (A) tokens within a unified stream, following Wu et al. (2025).

During training, the model is optimized to minimize the negative log-likelihood of the entire formatted target sequence $\mathbf{Y}$ given the source speech. Let $\mathcal{D}_{\text{dura}} = \{(\mathbf{X}_{\text{src}}^{(i)}, \mathbf{Y}^{(i)})\}_{i=1}^{N}$ denote DuraSet-440K, where $\mathbf{Y}^{(i)} = (\mathbf{y}_{\text{CoT}}^{(i)}, \mathbf{y}_{\text{TA4}}^{(i)})$ contains the CoT rationale and TA4 speech tokens. The SFT



(a) SFT Stage

(b) GRPO Stage

Figure 2: **Overview of the DuraS2ST framework.** DuraS2ST first generates a CoT rationale ($\mathbf{y}_{\text{CoT}}$) for explicit duration planning, then synthesizes target speech tokens ($\mathbf{y}_{\text{TA4}}$). The model is trained via a two-phase paradigm: SFT on DuraSet-440K, followed by GRPO with a multi-dimensional reward design that integrates translation quality, duration (DMR), and format compliance, dynamically attributed via MARA.

objective is formulated as:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} -\log P_\theta\left(\mathbf{Y}^{(i)} \mid \mathbf{X}_{\text{src}}^{(i)}\right), \quad (2)$$

where θ denotes the trainable parameters of the SLM. The objective trains the model to autoregressively generate the CoT rationale and final speech tokens in a unified sequence. This explicitly conditions the TA4 output $\mathbf{y}_{\text{TA4}}$ on both the source speech representation $\mathbf{X}_{\text{src}}$ and the generated rationale $\mathbf{y}_{\text{CoT}}$, thereby linking semantic planning with acoustic generation before the multimodal reinforcement learning phase.

2.2 Phase II: GRPO Training

Although SFT provides a strong initialization for reasoning and cross-modal alignment, it cannot directly optimize non-differentiable sequence-level objectives such as duration consistency and translation quality. We therefore introduce a multimodal reinforcement learning phase based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each query q, we sample a group of G rollouts $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and optimize them with translation-quality, duration, and format signals.

This RL stage faces two task-specific challenges. First, strict duration matching can over-penalize minor deviations, forcing the model to trade translation fidelity and acoustic naturalness for length compliance. Second, because each trajectory interleaves textual CoT and acoustic tokens, sequence-level duration penalties may contaminate credit assignment by incorrectly penalizing preceding reasoning tokens. To address these issues, we first introduce Duration Margin Reward (DMR) as a soft duration objective, and then apply Modality-Aware Reward Attribution (MARA) to assign reward signals to appropriate token spans. We further incorporate translation-quality and format signals to preserve semantic fidelity and structural validity during GRPO optimization. We next describe the multi-modal multi-dimensional reward design, which consists of translation-quality, duration, and format signals, followed by the MARA mechanism for modality-aware span-level credit assignment.

2.2.1 Multi-Modal Multi-Dimensional Reward Design

We use three complementary signals for GRPO: a translation-quality reward, a duration reward, and a format gate. The translation-quality reward encourages semantic fidelity, the duration reward encourages length consistency, and the format gate ensures that only structurally valid outputs receive positive optimization signals.

Translation Quality Reward. Let $\mathbf{Y}$ denote the raw generated multimodal sequence, $\mathbf{y}_{\text{text}}$ the text translation extracted from the TA4 stream by removing acoustic tokens, $\mathbf{x}$ the source text, and $\mathbf{y}^*$ the reference translation.

The translation-quality reward combines reference-based BLEU and reference-free $\text{COMET}_{\text{Kiwi}}$:

$$R_{\text{qua}} = \text{BLEU}(\mathbf{y}_{\text{text}}, \mathbf{y}^*) + \text{COMET}(\mathbf{x}, \mathbf{y}_{\text{text}}), \quad (3)$$

where both scores are normalized to $[0, 1]$.

Duration Margin Reward. To address the trade-off between duration consistency and translation quality, we model duration alignment as a bounded feasible region rather than a hard regression target. We introduce the Duration Margin Reward (DMR), a non-linear soft constraint that rewards outputs close to the target duration while tolerating minor deviations.

Let L_{pred} and L_{tgt} denote the predicted and target sequence lengths in acoustic tokens. We first compute the relative duration gap $\delta = |L_{\text{pred}} - L_{\text{tgt}}|/L_{\text{tgt}}$, and map it to a signed margin $m = (\tau - \delta)/s$, where τ is the acceptable relative error bound and s controls the smoothing scale. The duration reward is then defined as:

$$R_{\text{dur}} = \frac{\sigma\big(\kappa \cdot \text{clip}(m, -B, B)\big) - \sigma(-\kappa B)}{\sigma(\kappa B) - \sigma(-\kappa B)} \in [0, 1], \quad (4)$$

where $\sigma(\cdot)$ is the Log-Sigmoid function, κ controls the reward sharpness, and B bounds the margin range. The min–max normalization rescales the bounded output to $[0, 1]$, providing GRPO with a stable reward signal.

DMR is monotonically decreasing with the duration gap δ: outputs within the acceptable region receive high rewards, while large deviations are smoothly suppressed. This design avoids over-penalizing minor duration differences, reduces sensitivity to outliers, and concentrates the learning signal around recoverable duration mismatches, thereby encouraging duration consistency without sacrificing translation fidelity or acoustic naturalness.

Format-Gated Aggregation. To ensure structural validity, we further define a binary format gate $g(\mathbf{Y}) \in \{0, 1\}$, which equals 1 only when $\mathbf{Y}$ follows the required ordered control tags and terminates correctly. Instead of using format compliance as an additive reward, we use it to gate the raw quality and duration rewards before MARA:

$$\big(r^{(i)}_{\text{qua}},\, r^{(i)}_{\text{dur}}\big) = g(\mathbf{Y}^{(i)}) \cdot \big(R^{(i)}_{\text{qua}},\, R^{(i)}_{\text{dur}}\big). \quad (5)$$

Thus, malformed rollouts receive zero credit for both quality and duration, preventing reward hacking through format violations. The gated rewards $r^{(i)}_{\text{qua}}$ and $r^{(i)}_{\text{dur}}$ are then used by MARA for group normalization and span-level attribution.

2.2.2 Modality-Aware Reward Attribution

Standard GRPO collapses all rewards into a single scalar advantage and broadcasts it to every generated token. This is suboptimal for our multimodal trajectories, where textual CoT tokens and acoustic tokens are governed by different objectives. In particular, applying the speech-specific duration reward to preceding reasoning tokens may penalize correct semantic planning for downstream acoustic length errors. We refer to this issue as Cross-Modality Reward Contamination.

To address this, we propose Modality-Aware Reward Attribution (MARA), which computes group-relative advantages separately for each reward component and assigns them only to the tokens they directly supervise. Given a group of G rollouts $\{\mathbf{Y}^{(i)}\}_{i=1}^{G}$, we split each output into a reasoning span $\mathcal{P}$ and an acoustic span $\mathcal{A}$. For each reward component $k \in \{\text{qua}, \text{dur}\}$, where qua stands for translation quality and dur for duration, we normalize rewards within the group:

$$\hat{A}^{(i)}_k = \frac{r^{(i)}_k - \mu_k}{\sigma_k + \varepsilon}, \quad \mu_k = \tfrac{1}{G}\textstyle\sum_j r^{(j)}_k, \quad (6)$$

where σ_k is the within-group standard deviation and ε is a small constant.

The span-level advantage is then assigned as:

$$\hat{A}^{(i)}_t = \begin{cases} \hat{A}^{(i)}_{\text{qua}}, & t \in \mathcal{P}, \\ \hat{A}^{(i)}_{\text{qua}} + \hat{A}^{(i)}_{\text{dur}}, & t \in \mathcal{A}. \end{cases} \quad (7)$$

Here, $\hat{A}^{(i)}_{\text{qua}}$ is applied to both spans ($\mathcal{P}$ and $\mathcal{A}$), as it reflects the translation quality of both text and speech, while the duration advantage $\hat{A}^{(i)}_{\text{dur}}$ is applied only to the acoustic span $\mathcal{A}$, where length is manifested. MARA therefore preserves useful semantic feedback while preventing speech-specific penalties from contaminating textual reasoning.

3 DuraSet-440K

To mitigate duration mismatch in speech-to-speech translation, we construct **DuraSet-440K**, a large-scale duration-aligned bilingual S2ST corpus containing 440,705 parallel utterances and roughly 1,000 hours of speech. It is designed to provide duration-aligned target speech that remains translation-faithful and speaker-consistent. We build the corpus from the English and Chinese subsets of LEMAS (Zhao et al., 2026), first selecting high-quality utterances with suitable durations as source speech. Instead of relying on post-hoc time-stretching, we use duration-controllable synthesis to generate target speech that naturally aligns with the source duration. Figure 3 summarizes the overall construction pipeline, with detailed statistics and filtering settings provided in Appendix B.

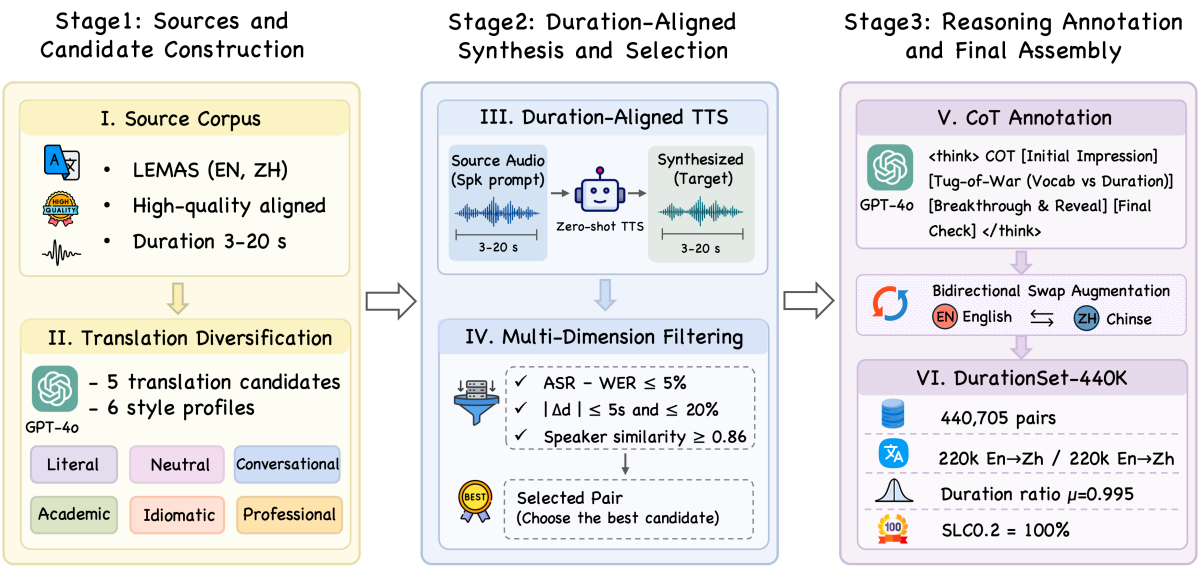


Figure 3: **Overview of the DuraSet-440K construction pipeline.**

The pipeline contains three main stages. First, in *Source and Candidate Construction*, we select English and Chinese utterances from LEMAS and filter them by quality and duration. Each retained source utterance is transcribed, and an LLM is prompted to generate multiple translation candidates with diverse styles and phonetic lengths. Second, in *Duration-Aligned Synthesis and Selection*, each candidate is synthesized by a zero-shot duration-controllable TTS model conditioned on the source audio as a timbre prompt, and the best target pair is selected based on transcription accuracy, duration alignment, and speaker similarity. Third, in *Reasoning Annotation and Final Assembly*, the retained pair is annotated with a phoneme-grounded reasoning trajectory and augmented by bidirectional source-target swapping, yielding the final duration-aligned bilingual S2ST corpus.

4 Experiments

4.1 Training Setup

We build DuraS2ST on top of **Step-Audio-2-mini-Think** (Wu et al., 2025)[1], a speech language model that natively interleaves text and discrete acoustic tokens. This makes it well suited for the structured CoT-to-speech format defined in Section 2.1, without requiring architectural modifications. Training proceeds in two phases.

Phase I: SFT. We perform full-parameter fine-tuning on DuraSet-440K for 4,000 steps using AdamW with a learning rate of $1e{-5}$ and a global batch size of 128. Each training trajectory follows the CoT-then-speech format in Eq. 1.

Phase II: GRPO. Starting from the SFT checkpoint, we further optimize the model with GRPO for 1,000 steps, using a learning rate of $1.5e{-5}$ and a KL coefficient of 0.001. To reduce training cost, we attach LoRA (Hu et al., 2021) adapters to all linear projections and use vLLM (Kwon et al., 2023) for efficient rollout generation.

The reward combines translation quality and duration consistency, with malformed outputs filtered by the format gate. Specifically, the quality reward combines BLEU and $\text{COMET}_{\text{Kiwi}}$, while the duration reward uses the proposed margin-based formulation with tolerance $\tau{=}5\%$. These rewards are converted into independent group-relative advantages through MARA, and the format gate zeroes both rewards for malformed rollouts before attribution. We also apply dynamic sampling (Yu et al., 2025) for training stability. Full hyperparameters and reward implementation details are provided in Appendix C.1 and Appendix C.2.

4.2 Evaluation Setup

4.2.1 Test Data

We evaluate on the Chinese (ZH) and English (EN) test splits of **CVSS-T** (Jia et al., 2022), comprising 4,897 utterance pairs, totaling 8.2 hours for Chinese and 6.3 hours for English. Since CVSS-T is disjoint from DuraSet-440K, it serves as an out-of-domain benchmark for evaluating translation quality and duration alignment.

[1] https://huggingface.co/stepfun-ai/Step-Audio-2-mini-Think

4.2.2 Metrics

We evaluate S2ST performance from three aspects: translation quality, voice/prosody preservation, and duration consistency. Evaluator checkpoints and implementation details are provided in Appendix C.4.

Translation Quality. We report **Text-BLEU** on the intermediate text output and **Speech-BLEU** on the ASR transcription of the generated speech, both computed with SacreBLEU (Post, 2018). We also report reference-based **COMET** (Rei et al., 2020) and reference-free **COMET$_{\text{Kiwi}}$** (Rei et al., 2022).

Voice and Prosody. We report **A.PCP** (Communication et al., 2023b) to measure prosodic consistency, and **SECS** to measure speaker similarity using WavLM (Chen et al., 2022) embeddings between the source prompt and generated speech.

Duration Consistency. We report **SLC-0.2** and **SLC-0.4** (Wu et al., 2023), which measure the fraction of samples whose target-to-source duration ratio falls within $\pm 20\%$ and $\pm 40\%$, respectively. To capture fine-grained duration errors, we further report **MADE**, the mean absolute duration error in seconds, and **MRDE**, the mean relative duration error.

4.2.3 Baselines

We compare DuraS2ST with both open-source and commercial S2ST systems:

- **UniSS** (Cheng et al., 2025b): a strong Chinese-English S2ST system with unified text-speech modeling and structured CoT prompting. We report both Quality and Performance modes, denoted as **UniSS (Q)** and **UniSS (P)**.
- **Qwen2.5-Omni-7B** (Xu et al., 2025) and **Kimi-Audio** (KimiTeam et al., 2025): recent audio/omni-modal MLLMs with speech-to-speech generation capability.
- **Step-Audio-2-mini** and **Step-Audio-2-mini-Think** (Wu et al., 2025): the non-reasoning and reasoning variants of our backbone model.
- **SeamlessM4T** (Communication et al., 2023a): Meta's multilingual S2ST system, including Medium, Large, and v2-Large variants.
- **GPT-4o**: a closed-source commercial S2ST reference.

Checkpoint details and prompt templates for all baselines are provided in Appendix C.3.

We further evaluate four variants based on our backbone to isolate the effects of SFT, CoT supervision, and GRPO:

- **DuraS2ST-Baseline**: the original Step-Audio-2-mini-Think backbone without additional training.
- **DuraS2ST-Instruct**: SFT on DuraSet-440K without the `<think>` CoT.
- **DuraS2ST-Think**: SFT on DuraSet-440K with the `<think>` CoT.
- **DuraS2ST-Think-RL**: GRPO training initialized from DuraS2ST-Think.

4.2.4 Inference Configuration

We serve DuraS2ST with vLLM (Kwon et al., 2023) using the official Step-Audio-2 sampling configuration: temperature 0.7, top-p=0.9, frequency penalty 0, and repetition penalty 1.05. The generated acoustic tokens are converted to waveforms using Step-Audio-2's `token2wav` module (Wu et al., 2025). All DuraS2ST variants use the same Step-Audio-2 prompt template as in data construction, with baseline-specific prompts listed in Appendix C.3.

4.3 Main Results

Tables 2 and 1 report the main results on CVSS-T. Overall, DuraS2ST achieves the strongest balance across translation quality, acoustic preservation, and duration consistency. More importantly, the stepwise improvements from Baseline to Instruct, Think, and Think-RL validate the effectiveness of duration-aligned supervision, explicit CoT planning, and reinforcement learning.

Translation quality. On EN-ZH, DuraS2ST-Think-RL improves AVG$_{\text{BLEU}}$ from 27.25 to 30.95 and achieves the best open-source COMET$_{\text{Kiwi}}$. On ZH-EN, it obtains the best overall translation performance among all systems, reaching the highest Text-BLEU, Speech-BLEU, COMET, AVG$_{\text{BLEU}}$, and AVG$_{\text{COMET}}$. This suggests that our DuraS2ST framework can achieve translation quality comparable to or even better than strong models without explicit duration control, while delivering substantially stronger duration consistency. Qwen2.5-Omni occasionally appends assistant-like speech in direct S2ST generation, which lowers both translation and duration scores.

Duration consistency. DuraS2ST-Think-RL brings substantial improvements in duration alignment. Compared with the backbone, it raises SLC-0.2 from 0.487 to 0.997 on EN-ZH and from 0.305 to 0.934 on ZH-EN, while reducing MRDE to the

Table 1: Main results on CVSS-T (**ZH-EN**). AVG_{BLEU} and AVG_{COMET} are means of (Text-BLEU, Speech-BLEU) and (COMET, $COMET_{Kiwi}$). **Best** / 2nd–3rd scores are highlighted. "–" indicates unavailable values.

Model	Translation Quality						Voice & Prosody		Duration Consistency			
	Text-BLEU↑	Speech-BLEU↑	COMET↑	$COMET_{Kiwi}$↑	AVG_{BLEU}↑	AVG_{COMET}↑	A.PCP↑	SECS↑	SLC-0.2↑	SLC-0.4↑	MADE (s)↓	MRDE↓
General Speech Large Language Models												
GPT-4o	–	23.13	0.710	0.692	–	0.701	2.49	0.018	0.442	0.836	1.59	0.246
Qwen2.5-Omni	10.86	10.82	0.638	0.611	10.84	0.625	1.75	0.142	0.190	0.355	7.85	1.310
Kimi-Audio	19.87	13.91	0.698	0.700	16.89	0.699	2.10	0.188	0.429	0.746	2.56	0.462
Step-Audio-2-mini	25.42	20.58	0.752	**0.743**	23.01	0.745	2.49	0.511	0.312	0.681	1.98	0.348
UniSS (Q)	25.86	24.54	0.780	0.699	25.20	0.740	2.73	0.309	0.913	0.983	0.74	0.125
UniSS (P)	26.04	24.00	0.769	0.683	25.02	0.726	**2.74**	0.310	0.893	**0.986**	0.74	0.123
Speech-to-Speech Translation Models												
SeamlessM4T-Medium	19.02	15.65	0.673	0.663	17.34	0.668	2.26	0.041	0.260	0.649	2.26	0.331
SeamlessM4T-Large	22.20	19.52	0.682	0.677	20.86	0.680	2.24	0.039	0.365	0.768	1.97	0.294
SeamlessM4T-v2-Large	23.57	22.77	0.704	0.703	23.17	0.704	2.10	0.041	0.312	0.713	2.06	0.310
DuraS2ST-Baseline	24.14	23.09	0.740	0.740	23.62	0.740	2.49	0.453	0.305	0.680	2.22	0.396
DuraS2ST-Instruct	27.42	24.81	0.789	0.692	26.12	0.741	2.65	0.500	0.926	0.978	1.31	0.170
DuraS2ST-Think	25.99	23.76	0.790	0.699	24.88	0.745	2.71	0.499	0.930	0.977	0.48	0.079
DuraS2ST-Think-RL	**28.71**	**25.88**	**0.794**	0.700	**27.30**	**0.747**	2.70	**0.518**	**0.934**	0.979	**0.42**	**0.068**

Table 2: Main results on CVSS-T (**EN-ZH**). AVG_{BLEU} and AVG_{COMET} are means of (Text-BLEU, Speech-BLEU) and (COMET, $COMET_{Kiwi}$). **Best** / 2nd–3rd scores are highlighted. "–" indicates unavailable values.

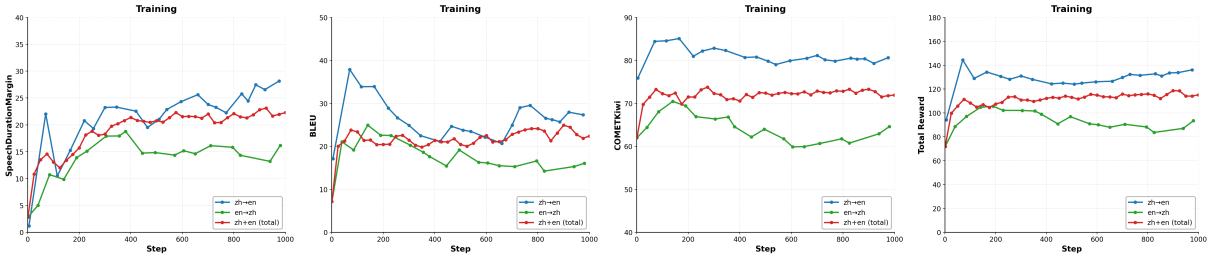
Model	Translation Quality						Voice & Prosody		Duration Consistency			
	Text-BLEU↑	Speech-BLEU↑	COMET↑	$COMET_{Kiwi}$↑	AVG_{BLEU}↑	AVG_{COMET}↑	A.PCP↑	SECS↑	SLC-0.2↑	SLC-0.4↑	MADE (s)↓	MRDE↓
General Speech Large Language Models												
GPT-4o	–	31.46	**0.845**	0.780	–	**0.812**	2.59	0.051	0.565	0.858	1.09	0.253
Qwen2.5-Omni	15.72	15.47	0.706	0.693	15.60	0.700	2.17	0.082	0.272	0.414	4.78	1.259
Kimi-Audio	27.65	19.13	0.820	0.768	23.39	0.794	2.28	0.096	0.571	0.830	3.21	0.923
Step-Audio-2-mini	32.47	26.66	0.812	0.762	29.57	0.785	2.75	0.330	0.471	0.852	1.58	0.423
UniSS (Q)	**33.03**	**32.16**	0.828	0.773	**32.60**	0.801	2.72	0.315	0.984	0.996	0.34	0.067
UniSS (P)	30.69	29.93	0.817	0.766	30.31	0.792	2.73	0.316	0.990	0.997	0.33	0.065
Speech-to-Speech Translation Models												
SeamlessM4T-Medium	24.80	15.84	0.757	0.744	20.32	0.751	2.31	0.033	0.561	0.849	1.52	0.310
SeamlessM4T-Large	27.33	19.57	0.775	0.763	23.45	0.769	2.29	0.030	0.618	0.846	1.52	0.275
SeamlessM4T-v2-Large	26.66	24.36	0.767	0.742	25.51	0.755	2.39	0.055	0.660	0.913	0.89	0.183
DuraS2ST-Baseline	27.57	26.92	0.821	0.773	27.25	0.797	2.75	0.333	0.487	0.851	2.04	0.629
DuraS2ST-Instruct	29.51	28.70	0.827	0.780	29.11	0.804	**2.79**	0.367	0.981	0.996	0.63	0.090
DuraS2ST-Think	30.46	29.39	0.828	0.782	29.93	0.805	**2.79**	0.370	0.993	**0.999**	**0.18**	0.039
DuraS2ST-Think-RL	31.47	30.43	0.834	**0.783**	30.95	0.808	**2.79**	**0.379**	**0.997**	**0.999**	0.23	**0.038**

lowest values in both directions. It achieves the best results on most duration metrics and remains highly competitive on the others, showing the effectiveness of our duration-aware training in improving source-target synchronization.

Effect of CoT planning. The comparison between DuraS2ST-Instruct and DuraS2ST-Think highlights the role of explicit reasoning supervision. Both models are trained on the same DuraSet-440K corpus, but only Think exposes the `<think>` trajectory before acoustic generation. Introducing this reasoning stage leads to clear gains in duration alignment, reducing MADE from 0.63s to 0.18s on EN-ZH and from 1.31s to 0.48s on ZH-EN, with similar reductions in MRDE. These results suggest that structured CoT mainly benefits duration planning by providing a semantic-phonetic intermediate step before acoustic token generation. While its effect on translation quality varies across directions, the subsequent RL stage further improves translation performance and duration consistency, showing the complementary roles of CoT-based planning and reward-based optimization.

Effect of GRPO. The transition from DuraS2ST-Think to DuraS2ST-Think-RL demonstrates the benefit of reward-based optimization. GRPO consistently improves translation metrics in both directions, especially on ZH-EN, where AVG_{BLEU} increases from 24.88 to 27.30. Meanwhile, duration consistency is largely preserved or further improved, with MRDE reduced from 0.039 to 0.038 on EN-ZH and from 0.079 to 0.068 on ZH-EN. These results show that the RL stage enhances semantic fidelity under duration-aligned generation, while DMR and MARA provide effective duration-aware reward attribution.

Acoustic preservation. DuraS2ST maintains strong prosody and speaker similarity while enforcing duration consistency. On EN-ZH, the trained DuraS2ST variants achieve the best A.PCP, and Think-RL obtains the highest SECS in both directions. This is important because duration control



(a) Duration Margin Reward (b) Text-BLEU Reward (c) COMET-Kiwi Reward (d) Total Reward

Figure 4: **Zero-GRPO training dynamics** on CVSS-T. Curves show duration, BLEU, $\text{COMET}_{\text{Kiwi}}$, and total rewards over 1,000 GRPO steps without DuraSet-440K SFT.

Table 3: **Subjective MOS** (1–5, higher is better) on CVSS-T. Each cell reports *EN-ZH* | *ZH-EN*. **Best** / 2nd–3rd per direction are highlighted.

Model	Trans.↑	Natur.↑	SpkSim↑	Avg↑
SeamlessM4T-v2-Large	2.87 | 3.07	1.20 | 3.93	1.13 | 1.27	1.73 | 2.76
Qwen2.5-Omni	3.87 | 3.80	4.17 | 4.03	1.00 | 2.40	3.01 | 3.41
Kimi-Audio	2.20 | 2.40	3.80 | 3.32	1.07 | 1.13	2.36 | 2.29
UniSS	3.80 | 3.67	3.87 | 3.33	3.87 | 3.93	3.85 | 3.64
DuraS2ST-Baseline	3.93 | 3.53	4.10 | 4.13	4.00 | **4.33**	4.01 | 4.00
DuraS2ST-Instruct	4.00 | 4.06	4.20 | 4.08	4.13 | 4.00	4.11 | 4.05
DuraS2ST-Think	**4.53** | 4.13	**4.27** | 4.13	4.27 | 4.20	**4.36** | 4.15
DuraS2ST-Think-RL	4.33 | **4.73**	4.17 | **4.23**	**4.40** | **4.33**	4.30 | **4.43**

Table 4: **Zero-GRPO duration consistency** on CVSS-T (EN-ZH | ZH-EN). Removing R_{dur} isolates the effect of Duration Margin Reward without DuraSet-440K SFT.

Model	SLC-0.2↑	SLC-0.4↑	MADE (s)↓	MRDE↓
Baseline	0.487 | 0.305	0.851 | 0.680	2.04 | 2.22	0.629 | 0.396
+ GRPO	0.455 | 0.302	**0.877** | **0.691**	**1.15** | **1.97**	**0.259** | **0.346**
w/o R_{dur}	0.448 | 0.319	0.855 | 0.641	1.83 | 1.93	0.565 | 0.385

via post-hoc time-stretching can distort speaker identity or naturalness. In contrast, DuraS2ST learns to satisfy duration constraints during generation, preserving acoustic characteristics while producing synchronized speech.

4.4 Subjective Evaluation

We conduct a human evaluation with six raters on 160 anonymized CVSS-T samples, covering eight systems and both translation directions. Each sample is presented in a blinded and randomized order, and rated on a 1–5 Likert scale for translation adequacy, speech naturalness, and speaker similarity.

Table 3 shows that the Think and Think-RL variants are consistently preferred or remain highly competitive in subjective evaluation. DuraS2ST-Think-RL achieves the best overall performance on ZH-EN and remains competitive on EN-ZH. Together with the objective duration metrics, these results indicate that DuraS2ST improves synchronization without degrading speech quality.

4.5 Analysis

We examine whether the proposed reward design can guide duration-aware optimization without supervised duration-aligned training. We apply GRPO directly to the Step-Audio-2-mini-Think backbone without SFT on DuraSet-440K, denoted as Zero-GRPO.

Table 4 shows that Zero-GRPO reduces duration errors over the backbone, particularly in MADE and MRDE. When the duration reward is removed, the gains become weaker, confirming that the Duration Margin Reward is the main driver of duration alignment. Figure 4 further shows that GRPO improves the duration reward and aggregated reward during training. Together, these results indicate that duration consistency can be optimized as a sequence-level objective, even without supervised duration-aligned trajectories.

5 Conclusion

We presented DuraS2ST, a reasoning-based framework for duration-aligned speech-to-speech translation. Instead of treating duration control as acoustic post-processing, DuraS2ST explicitly plans target wording and phonetic length before generating the final interleaved text-acoustic sequence. To support this paradigm, we constructed DuraSet-440K, a large-scale duration-aligned reasoning corpus, and trained the model in two phases of supervised fine-tuning followed by multi-modal GRPO. Within the RL phase, the Duration Margin Reward resolves the quality-duration trade-off through a saturated soft constraint, and the Modality-Aware Reward Attribution prevents cross-modality reward contamination by restricting the duration advantage to acoustic tokens. Experiments on CVSS-T show that DuraS2ST balances translation quality and duration consistency. A zero-SFT ablation further shows that our rewards substantially reduce MADE, highlighting their complementary contribution beyond duration-aligned SFT.

Limitations

We proposed DuraS2ST, a chain-of-thought and reinforcement learning framework for duration-aligned speech-to-speech translation, and constructed DuraSet-440K to support its training. However, duration alignment in S2ST is more nuanced than a single language pair can fully reveal, so our current evaluation has focused on bidirectional Chinese-English translation, a representative pair of structurally divergent languages, laying the groundwork for broader linguistic validation. In addition, the current think-then-speak architecture decodes the chain-of-thought rationale in full before emitting any acoustic token, which fits offline applications such as video dubbing but does not yet support streaming or simultaneous interpretation. Future work will both extend DuraS2ST and DuraSet-440K to additional language pairs and explore streaming variants in which chain-of-thought reasoning and acoustic emission interleave at finer granularity.

Ethical Considerations

This work studies duration-aligned speech-to-speech translation with synthetic speech generation and speaker preservation. While such techniques can improve accessibility and cross-lingual communication, they may also be misused for impersonation, unauthorized voice conversion, or misleading audio generation. We therefore emphasize that the proposed system should only be applied to speech data with proper consent, licensing, and appropriate safeguards. The datasets used in our experiments are public research datasets, and our evaluations are conducted on anonymized samples. We use public research datasets and existing models in accordance with their licenses and terms of use. Human evaluation participants are asked to assess translation adequacy, naturalness, and speaker similarity, without being exposed to personally identifying information beyond the speech content available in the benchmark data. We also acknowledge that our experiments focus on English and Chinese, and the findings may not fully generalize to other languages, accents, or sociolinguistic settings.

Acknowledgments

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14202623).

References

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

Shanbo Cheng, Yu Bao, Zhichao Huang, Yu Lu, Ningxin Peng, Lu Xu, Runsheng Yu, Rong Cao, Yujiao Du, Ting Han, Yuxiang Hu, Zeyang Li, Sitong Liu, Shengtao Ma, Shiguang Pan, Jiongchen Xiao, Nuo Xu, Meng Yang, Rong Ye, and 9 others. 2025a. Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice. *Preprint*, arXiv:2507.17527.

Sitong Cheng, Weizhen Bian, Xinsheng Wang, Ruibin Yuan, Jianyi Chen, Shunshun Yin, Yike Guo, and Wei Xue. 2025b. Uniss: Unified expressive speech-to-speech translation with your voice. *Preprint*, arXiv:2509.21144.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023a. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023b. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Qianqian Dong, Zhiying Huang, Qiao Tian, Chen Xu, Tom Ko, Yunlong Zhao, Siyuan Feng, Tang Li, Kexin Wang, Xuxin Cheng, Fengpeng Yue, Ye Bai, Xi Chen, Lu Lu, Zejun Ma, Yuping Wang, Mingxuan Wang, and Yuxuan Wang. 2023. Polyvoice: Language models for speech to speech translation. *Preprint*, arXiv:2306.02982.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2023. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *Preprint*, arXiv:2206.08317.

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

Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, and Zhou Zhao. 2023.

Transpeech: Speech-to-speech translation with bilateral perturbation. *Preprint*, arXiv:2205.12523.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. Unity: Two-pass direct speech-to-speech translation with discrete units. *Preprint*, arXiv:2212.08055.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. Cvss corpus and massively multilingual speech-to-speech translation. *Preprint*, arXiv:2201.03713.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Chenyang Le, Yao Qian, Dongmei Wang, Long Zhou, Shujie Liu, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Sheng Zhao, and Michael Zeng. 2024. Transvip: Speech to speech translation system with voice and isochrony preservation. *Preprint*, arXiv:2405.17809.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. *Preprint*, arXiv:2107.05604.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Wolfgang Wahlster. 2000. *Verbmobil: foundations of speech-to-speech translation*, volume 1. Springer Berlin.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. 2023. Videodubber: Machine translation with speech-aware length control for video dubbing. *Preprint*, arXiv:2211.16934.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Zhiyuan Zhao, Lijian Lin, Ye Zhu, Kai Xie, Yunfei Liu, and Yu Li. 2026. Lemas: Large a 150k-hour large-scale extensible multilingual audio suite with generative speech models. *Preprint*, arXiv:2601.04233.

Supplementary Material

A Additional Experiments

This section reports auxiliary experiments that support the training design choices of DuraS2ST.

A.1 Pre-Rollout Filtering

GRPO relies on within-group reward variation to estimate group-relative advantages. If all rollouts for a prompt receive nearly identical rewards, the prompt provides little useful gradient signal but still consumes rollout budget. To improve training efficiency, we perform a one-time *pre-rollout filtering* step before GRPO.

Setup. For each candidate prompt, we sample $G = 8$ rollouts from the SFT checkpoint using the same decoding configuration as GRPO. We discard prompts with near-zero reward variance, since they cannot provide informative group-relative advantages. The remaining prompts are bucketed into *easy*, *medium*, and *hard* groups according to their mean total reward. The final GRPO training set uses all medium prompts and a small fraction of easy prompts, balancing learnability and stability.

Result. We compare two GRPO runs with identical rewards, rollout budget, and hyperparameters, differing only in the training prompts: the original candidate pool and the pre-rollout-filtered set. Figure 5 shows the reward curves on ZH→EN. Pre-rollout filtering yields faster improvement and a higher reward plateau across duration, translation, and total rewards. This indicates that filtering out uninformative prompts makes GRPO more sample-efficient and stabilizes optimization.

A.2 Inference Efficiency

We measure the LLM-only inference cost (excluding the shared vocoder) on a single A100-80GB GPU over 100 CVSS-T utterances. We report generated tokens and FLOPs for hardware-independent comparison, together with wall-clock latency and real-time factor (RTF) in Table 5.

As shown in Table 5, the inference overhead mainly comes from the optional chain-of-thought (CoT) duration reasoning. Without CoT, DuraS2ST generates only 135 tokens per utterance on average and achieves an RTF of 0.50, close to Step-Audio-2-mini (0.47). Enabling CoT increases the RTF from 0.50 to 2.99 due to longer reasoning, while disabling it supports latency-sensitive online inference.

Table 5: LLM-only inference cost on CVSS-T, averaged over 100 utterances on a single A100-80GB GPU. The shared vocoder is excluded.

Model	CoT	Gen. Tokens	GFLOPs/Utt.	Latency (s)	RTF
DuraS2ST (ours)	×	135	1.76k	2.48	0.50
DuraS2ST (ours)	✓	909	11.9k	14.80	2.99
Step-Audio-2-mini	×	128	1.68k	2.30	0.47
Step-Audio-2-mini-Think	✓	213	2.78k	3.64	0.73
UniSS (Quality mode)	✓	286	0.75k	7.27	1.47

B DuraSet-440K Details

This section provides the technical details of the DuraSet-440K construction pipeline, including model choices, hyperparameters, filtering criteria, and prompting templates. It complements the overview in Section 3.

B.1 Translation Candidate Construction

For each source utterance, we first construct a set of translation candidates with diverse lexical densities and phonetic lengths. This candidate set provides a search space for selecting translations that preserve the source meaning while better matching the target duration budget.

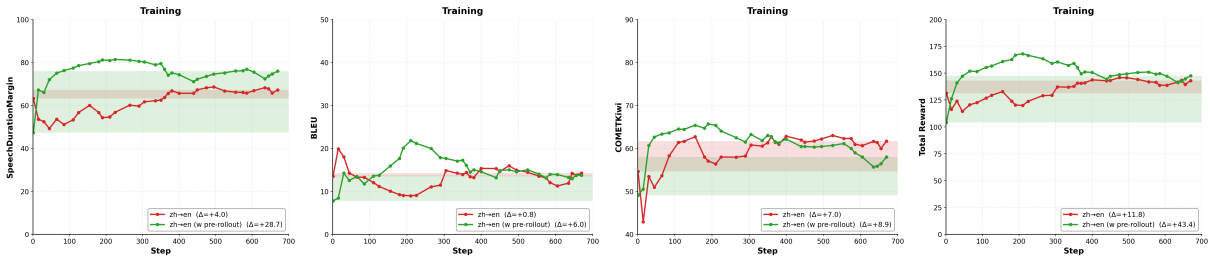
We use GPT-4o to generate translation candidates with temperature $T = 0.7$ and top-$p = 0.9$. For each source utterance, we sample $N = 5$ candidates using a structured prompt that encourages stylistic diversity. As shown in Figure 7, the prompt assigns distinct translation styles ranging from verbose to concise, encouraging the LLM to vary wording, syllable count, and phonetic footprint while preserving the core semantics.

B.2 Target Speech Synthesis Configuration

We synthesize each translation candidate using an in-house TTS model with zero-shot duration control at 24 kHz. The source audio serves as the timbre prompt, allowing the synthesized target speech to preserve the speaker characteristics of the source utterance.

For each translation candidate $\mathbf{T}_i$, we estimate its natural duration $d_{\text{nat}}^{(i)}$ and compute a global frame scaling factor against the target duration budget d_{budget}:

$$s_i = \text{clip}\left(\frac{d_{\text{budget}}}{d_{\text{nat}}^{(i)}},\ 0.5,\ 2.0\right). \tag{8}$$



(a) Duration Margin Reward (b) Text-BLEU Reward (c) COMET-Kiwi Reward (d) Total Reward

Figure 5: **Pre-rollout difficulty filtering improves GRPO stability and efficiency.** On ZH→EN, two GRPO runs share an identical reward composition, rollout budget, and hyperparameters, differing only in their training data. Across all four reward components, the pre-rollout-filtered run (green) climbs faster and reaches a higher plateau than the unfiltered baseline (red); the in-legend Δ reports the improvement from the start of training to step 700.

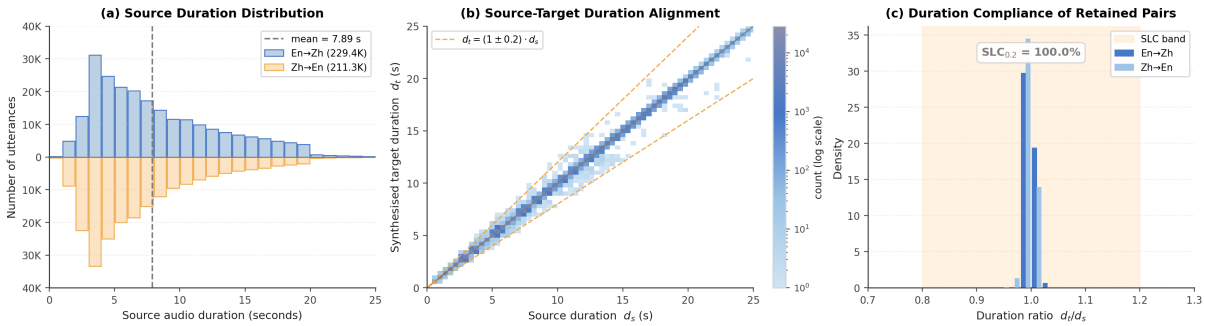


Figure 6: **Duration statistics of DuraSet-440K.** (a) Source-duration histogram by direction. (b) Hex-binned joint density of source and target durations with the $d_{\text{tgt}} = d_{\text{src}}$ diagonal. (c) Distribution of the target-to-source duration ratio $d_{\text{tgt}}/d_{\text{src}}$.

The predicted acoustic frames are scaled by s_i before waveform generation. The clipping range $[0.5, 2.0]$ prevents excessive compression or expansion, which could otherwise degrade speech quality. It also ensures that candidates with incompatible phonetic lengths retain measurable residual duration errors, allowing the filtering stage to reject them.

During synthesis, we use a classifier-free guidance (CFG) scale no larger than 10 for stable generation. The source timbre prompt is trimmed and padded with a 0.2 s leading fade and a 0.5 s trailing fade to smooth conditioning transitions.

B.3 Multi-Dimensional Filtering

Each synthesized candidate is filtered by transcription fidelity, duration alignment, and speaker similarity. A candidate $\mathbf{Y}_i$ is retained only if it satisfies:

$$\begin{aligned} \text{WER}(\mathbf{Y}_i, \mathbf{T}_i) &\leq \tau_{\text{wer}}, \\ |\Delta d_i| &\leq \tau_{\text{abs}}, \\ \frac{|\Delta d_i|}{d_{\text{src}}} &\leq \tau_{\text{rel}}, \\ \cos(\mathbf{e}_{\text{src}}, \mathbf{e}_{\text{tgt}}^{(i)}) &\geq \tau_{\text{sim}}, \end{aligned} \tag{9}$$

where $\Delta d_i = d_{\text{tgt}}^{(i)} - d_{\text{src}}$, $d_{\text{tgt}}^{(i)}$ is the duration of the synthesized candidate, and $\mathbf{e}_{\text{src}}$ and $\mathbf{e}_{\text{tgt}}^{(i)}$ are speaker embeddings of the source and target speech.

We set the thresholds to $\tau_{\text{wer}} = 0.05$, $\tau_{\text{abs}} = 5\,\text{s}$, $\tau_{\text{rel}} = 0.20$, and $\tau_{\text{sim}} = 0.86$. WER is computed using Fun-ASR-Nano-2512 with inverse text normalization (ITN) enabled, which standardizes numbers and punctuation before text alignment. Speaker similarity is computed with WavLM-base-plus-sv embeddings extracted at 16 kHz.

Among candidates that pass all filtering criteria, we select the final target speech by minimizing duration error, using speaker similarity and WER as secondary tie-breakers. This selection procedure favors duration-aligned pairs while maintaining transcription accuracy and speaker consistency.

B.4 Reasoning Rationale Construction

After selecting the duration-aligned target pair, we annotate each example with a phoneme-grounded reasoning rationale. The goal is to provide explicit supervision for forward planning: before generating target speech, the model should reason about translation choices, phonetic length, and the dura-

Prompt Template (Translation Diversification)

```
# System Role
You are a high-precision translation engine.
Translate the input text faithfully from the specified source language to the specified
target language.

Hard constraints (must follow):
- Do not add or remove factual content.
- Do not hallucinate details.
- Do not explain or annotate.
- If ambiguous, preserve the ambiguity; do NOT resolve it.
- Keep names, numbers, units, formatting, and technical terms accurate.

Task: Generate exactly {num_candidates} different translation candidates.
- All candidates must preserve the same meaning but differ in STYLE and/or wording.

Style profiles (assign one per candidate; cycle through the list if needed):
1) Literal-direct: Close to source structure; minimal rephrasing; no localization.
2) Neutral-standard: Natural and fluent; general-purpose register.
3) Conversational: Spoken and relaxed tone while staying faithful.
4) Academic-formal: Formal register with precise connectors.
5) Cultural-idiomatic: Use target-language idioms and natural phrasing ONLY when strictly
meaning-equivalent.
   - MAY replace source idiom with equivalent target idiom.
   - MUST NOT add context, explanations, or extra implications.
   - If no safe idiom exists, use elegant but literal wording.
6) Polished-professional: Refined and suitable for publication.

Diversity requirement:
- Each candidate must be noticeably different in register, structure, or word choice.
- Avoid near-duplicates or minor synonym swaps.

Output format:
- Return a JSON object with a 'translations' key containing an array of exactly
{num_candidates} strings.
- Output ONLY valid JSON, no markdown, no explanations, no other text.

---------------------------------------------------------------------
# Inputs

Source language: {source_lang}
Target language: {target_lang}
Text:
{source_text}
```

Figure 7: Prompt template for translation candidate construction. The prompt encourages stylistically diverse translations with different lexical densities and temporal footprints for the same source utterance.

tion budget.

We query GPT-4o with temperature $T = 0.0$ to obtain deterministic rationales. Each rationale is constrained to 150–300 words to maintain sufficient detail while avoiding unnecessary verbosity. As shown in Figure 8, the prompt follows a structured reasoning template and enforces a "no-spoiler" rule, requiring the model to discuss translation alternatives before revealing the final target text.

To ground the rationale in verifiable phonetic evidence, we provide the source and target phoneme sequences together with their lengths. These phoneme sequences are deterministically extracted using the BigCiDian lexicon with a longest-match strategy and mapped to the International Phonetic Alphabet (IPA). This forces the rationale to explicitly connect wording choices with phonetic footprints and duration constraints.

B.5 Dataset Statistics

The final DuraSet-440K corpus contains 440,705 parallel utterances, including 229,410 En→Zh examples and 211,295 Zh→En examples. Figure 6 summarizes the duration statistics of the retained corpus.

Figure 6(a) shows the distribution of source durations, with a mean of 7.89 s and a standard deviation of 4.68 s. Figure 6(b) plots the joint density of source and target durations, where most

examples concentrate around the $d_{\text{tgt}} = d_{\text{src}}$ diagonal. Figure 6(c) shows the distribution of target-to-source duration ratios, which is tightly centered around $\mu = 0.995$ with $\sigma = 0.008$ and achieves $\text{SLC}_{0.2} = 100.0\%$. These statistics confirm that DuraSet-440K provides temporally aligned source-target pairs suitable for training duration-aware S2ST models.

C Implementation Details

This section provides additional implementation details for DuraS2ST, including training hyperparameters, reward computation, baseline inference protocols, and evaluation metrics.

C.1 Training Hyperparameters

Phase I: SFT. We perform full-parameter fine-tuning of Step-Audio-2-mini-Think on 8 NVIDIA A100 80 GB GPUs with bf16 precision and DeepSpeed ZeRO-3. We use AdamW with a learning rate of $1\text{e}{-5}$, 5% linear warmup, a per-device batch size of 1, and gradient accumulation of 16, resulting in a global batch size of 128. The maximum sequence length is set to 16,384 tokens to accommodate the interleaved CoT and acoustic-token streams. The model is trained for 4,000 steps using the CoT-then-speech format defined in Eq. 1.

Phase II: GRPO. Starting from the SFT checkpoint, we attach LoRA adapters with rank 16 and $\alpha = 32$ to all linear projections, and optimize the model with GRPO for 1,000 steps. We use a learning rate of $1.5\text{e}{-5}$ and a KL coefficient of $\beta = 0.001$ against the SFT reference policy. For each input, we sample $G = 16$ rollouts with temperature 1.0, top-$p = 0.99$, and top-$k = 50$. Rollout generation is served by vLLM (Kwon et al., 2023) to reduce the cost of long autoregressive speech decoding. Following DAPO (Yu et al., 2025), groups whose rewards collapse to a single value are re-sampled up to 3 times, and rollouts longer than 4,096 tokens are filtered before the gradient update.

C.2 RL Reward Implementation

All rewards are computed online from the generated token stream, without invoking ASR during training. Step-Audio-2-mini-Think emits an interleaved *TA4* stream, where one text token is placed before every four acoustic tokens. The text tokens inside the `<tts>` segment therefore form a temporally aligned transcript of the generated speech. We concatenate these in-segment text tokens as the post-CoT translation $\mathbf{y}_{\text{text}}$ for reward computation.

BLEU reward. We compute sentence-level SacreBLEU (Post, 2018) between $\mathbf{y}_{\text{text}}$ and the reference translation $\mathbf{y}^*$. The score is divided by 100 to obtain a reward in $[0, 1]$. Language-specific preprocessing follows the evaluation pipeline in Appendix C.4.

$\text{COMET}_{\text{Kiwi}}$ reward. We compute the reference-free semantic reward using `Unbabel/wmt22-cometkiwi-da` (Rei et al., 2022), with the source text and generated translation as input. The model score is used directly as the $\text{COMET}_{\text{Kiwi}}$ reward.

Duration Margin Reward. The duration reward is computed from the acoustic-token length. Given the predicted length L_{pred} and target length L_{tgt}, we compute the relative duration gap $\delta = |L_{\text{pred}} - L_{\text{tgt}}|/L_{\text{tgt}}$ and apply the Duration Margin Reward. We use tolerance $\tau = 0.05$, smoothing scale $s = 0.02$, scaling factor $\kappa = 0.5$, and margin clipping bound $B = 5$. Under this setting, the reward is high for sub-5% duration gaps, concentrates its learning signal around the recoverable 5%–15% range, and saturates for larger deviations. These constants are tuned on a small held-out subset of DuraSet-440K.

Format gate. Each rollout is checked by a deterministic parser. A rollout is valid only if the control tags `<think>`, `</think>`, and `<tts_start>` appear in the required order, with no trailing tokens after `<tts_end>`. Valid rollouts receive $g = 1$, while malformed rollouts receive $g = 0$.

Gated aggregation and attribution. As described in Section 2.2.1, the format gate multiplicatively zeroes both the quality and duration rewards before group-relative normalization. This prevents malformed outputs from receiving positive credit for partially satisfying translation or duration objectives. The gated rewards are then standardized into per-reward advantages and assigned to token spans through MARA: the quality advantage is applied to both reasoning and acoustic spans, while the duration advantage is restricted to acoustic-phase tokens.

C.3 Baselines

For each external baseline, we use the model's default decoding configuration and a fixed direction-specific translation instruction. We do not ap-

ply additional input preprocessing or output post-processing, so the reported scores reflect each system's raw end-to-end S2ST behavior under a controlled prompt.

SeamlessM4T. We evaluate SeamlessM4T (Communication et al., 2023b) using three publicly released checkpoints: `facebook/hf-seamless-m4t-medium`, `facebook/hf-seamless-m4t-large`, and `facebook/seamless-m4t-v2-large`. Inference is performed with the official Hugging Face Transformers wrappers and default decoding parameters. The target-language token is set to `eng` for ZH→EN and `cmn` for EN→ZH.

Step-Audio-2-mini and Step-Audio-2-mini-Think. The official Step-Audio-2 release provides the following Chinese-target S2ST prompt:

```
请仔细聆听这段语音，然后将其内容翻译成中文
并用语音播报。
```

For ZH→EN, we find that simply replacing the target language with English may lead to a mismatch between the reasoning text and the generated speech. We therefore explicitly specify both the source and target languages:

```
请仔细聆听这段中文语音，然后将其内容翻译成
英文并用英文语音播报。
```

We use the original Chinese-target prompt for EN→ZH and the dual-language English-target prompt for ZH→EN. The same prompt templates are used for DuraSet-440K construction and DuraS2ST evaluation.

Qwen2.5-Omni. We evaluate Qwen2.5-Omni (Xu et al., 2025) in its default S2ST mode with a minimal translation instruction following Cheng et al. (2025b). For example:

```
Translate the following Chinese speech into
English.
```

and the symmetric instruction for EN→ZH. We observe that Qwen2.5-Omni sometimes appends conversational fillers after the translated speech, which are retained in the generated waveform and affect both ASR-based translation metrics and duration metrics. We do not remove these artifacts by post-processing.

Kimi-Audio. We use the official `Kimi-Audio-7B-Instruct` checkpoint (KimiTeam et al., 2025) with its bundled inference pipeline and recommended decoding configuration. We use the following direction-specific instructions:

```
Translate the above Chinese audio to
English.
```

for ZH→EN, and

```
Translate the above English audio to
Chinese. Answer in Chinese only.
```

for EN→ZH. The explicit target-language constraint is used to reduce language fallback errors.

GPT-Audio. For the commercial reference system, we query GPT-Audio with the source clip encoded as audio input and a direction-specific translation instruction:

```
Please translate this Chinese audio into
English. Just state the translation
itself without any additional explanation
or polite words.
```

for ZH→EN, and

```
Please translate this English audio into
Chinese. Just state the translation
itself without any additional explanation
or polite words.
```

for EN→ZH. The final clause is included to reduce conversational fillers that would otherwise inflate the generated duration. All other parameters are kept at their default values.

UniSS. We evaluate UniSS (Cheng et al., 2025b) using the official `cmots/UniSS` checkpoint and inference scripts. UniSS provides two decoding modes, Quality and Performance, which we report as UniSS (Q) and UniSS (P). We use the official target-language tokens for both directions and leave the remaining decoding hyperparameters at their default values.

C.4 Evaluation Metrics

We group evaluation metrics into translation quality, voice/prosody preservation, and duration consistency. All scores are computed at the corpus level over the CVSS-T test split.

Translation quality. **Text-BLEU** evaluates the intermediate textual output, while **Speech-BLEU** evaluates the ASR transcription of the generated speech. Both are computed with SacreBLEU (Post, 2018) using the `corpus_score` interface. English text is lowercased and stripped of punctuation; Chinese text is converted to simplified characters and evaluated in `zh` mode with character-level tokenization. For Speech-BLEU, we use Whisper-large-v3 (Radford et al., 2022) for English ASR and Paraformer-zh (Gao et al., 2023) for Mandarin ASR, following Seed-TTS-eval.[2]

We also report reference-based **COMET** (Rei et al., 2020) and reference-free **COMET$_{\text{Kiwi}}$** (Rei et al., 2022), computed with `Unbabel/wmt22-comet-da` and `Unbabel/wmt22-cometkiwi-da`, respectively. Both are evaluated on the ASR transcript.

Voice and prosody. **A.PCP** (Communication et al., 2023b) measures prosodic consistency between the source and generated speech, including pitch, energy, and temporal patterns. We use the official AutoPCP implementation in `stopes`.[3] For **SECS**, we extract speaker embeddings using WavLM-large fine-tuned for speaker verification (Chen et al., 2022), and compute cosine similarity between the source prompt and generated utterance.

Duration consistency. **SLC-0.2** and **SLC-0.4** (Wu et al., 2023) measure the fraction of utterances whose target-to-source duration ratio falls within $\pm 20\%$ and $\pm 40\%$, respectively. Since these threshold-based metrics can saturate under strong duration alignment, we also report two fine-grained error metrics: **MADE**, the mean absolute duration error in seconds,

$$\mathrm{MADE} = \mathbb{E}\left[|d_{\mathrm{tgt}} - d_{\mathrm{src}}|\right],$$

and **MRDE**, the mean relative duration error,

$$\mathrm{MRDE} = \mathbb{E}\left[\frac{|d_{\mathrm{tgt}} - d_{\mathrm{src}}|}{d_{\mathrm{src}}}\right].$$

Both are computed directly from waveform durations.

[2] https://github.com/BytedanceSpeech/seed-tts-eval

[3] https://github.com/facebookresearch/stopes/tree/main/stopes/eval/auto_pcp

D Evaluation Examples

To complement the quantitative results in Section 4, we provide qualitative examples from the CVSS-T test set for both translation directions. For each direction, we randomly select two utterances and compare a representative baseline with DuraS2ST. Each example visualizes the synthesized waveform and its duration alignment with the source speech, allowing translation fidelity and temporal synchronization to be inspected jointly.

Figure 9 shows EN→ZH examples, and Figure 10 shows ZH→EN examples. For example, in the ZH→EN case, the source utterance describes a company headquartered in Riyadh with branches in Kuwait, Jordan, and Malaysia. A compact translation may preserve these facts but still produce target speech that is too short or less naturally paced. By reasoning over semantic content and phonetic length before speech-token generation, DuraS2ST better follows the source temporal envelope while producing translations that remain faithful to the reference. These examples further demonstrate that the proposed reasoning-based generation and duration-aware optimization generalize to held-out test utterances.

E Subjective Evaluation Interface

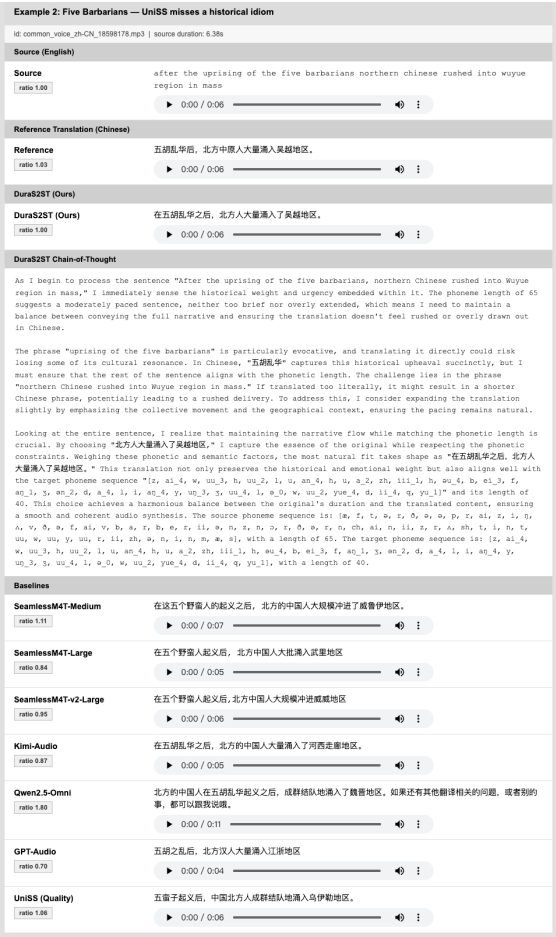
Figure 11 shows the web interface used for the MOS study in Section 4.4. Each task presents a source clip and one synthesized target clip, with the system identity hidden by an anonymized label. Annotators listen to both clips and provide three integer ratings on a 1–5 Likert scale: *translation adequacy*, measuring whether the target preserves the source meaning; *naturalness*, measuring speech fluency and audio quality; and *speaker similarity*, measuring whether the target preserves the source speaker's voice and prosodic characteristics. The same interface and instructions are used for all 160 samples across the eight systems. Sample order is randomized for each annotator to reduce presentation bias.

Prompt Template (Reasoning Construction)

```
# Role
You are an expert in Speech-to-Speech Translation (S2ST) and cross-lingual phonetic alignment. You
must output a real-time, continuous internal monologue of a translator *constructing* a translation
from scratch.

# Inputs (Data for this S2ST task)
- source_lang: {source_lang}
- source_text: {source_text}
- source_phoneme_seq: {source_text_phoneme_seq}
- source_phoneme_length: {source_text_phoneme_length}
- target_lang: {target_lang}
- target_text: {target_text}
- target_phoneme_seq: {target_text_phoneme_seq}
- target_phoneme_length: {target_text_phoneme_length}

# CRITICAL INSTRUCTIONS
1. Format Constraint: Write in a highly natural, cohesive, and fluid narrative (a single or double
paragraph of "stream of consciousness"). ABSOLUTELY DO NOT use bullet points, numbers, bolded
headers, or structural markers.
2. The "No-Spoiler" Rule: You MUST NOT reveal the {target_text}, {target_phoneme_seq}, or
{target_phoneme_length} until the very end of your thought process, after you have logically derived
them.

# Your Internal Monologue Must Follow This Seamless Arc:

- The Initial Impression:
Start by mentally digesting the {source_text} and its emotional or semantic weight. Immediately note
its {source_text_phoneme_length}. Ponder aloud what this specific duration means for the pacing—is it
a long, winding sentence or a sharp, quick phrase?

- The Tug-of-War (Vocabulary vs. Duration):
This is the core of your reasoning. As you start brainstorming how to translate key concepts into
{target_lang}, explicitly wrestle with the duration constraint. Do not just translate; debate! For
instance, point out a specific source word and say, "If I translate this literally, the {target_lang}
phrasing might be too short compared to the source length of {source_text_phoneme_length}, leading to
an unnatural, rushed audio synthesis." Then, explain how choosing a slightly different synonym,
expanding a clause, or adjusting the syntax helps match that phonetic footprint while preserving
perfect semantic fidelity.

- The Breakthrough & Reveal:
Let the target sentence emerge naturally from this debate. Use a fluid transition like, "Weighing
these phonetic and semantic factors, the most natural fit takes shape as..." and reveal the
{target_text}.

- The Final Check:
Immediately validate your choice by introducing the target_phoneme_seq and its target_phoneme_length.
Conclude with a brief thought on why this specific length successfully achieves isochrony and
semantic harmony with the original.

# Requirements
- Explicitly quote source words and debate their target counterparts *before* revealing the final
sentence.
- Use natural transitional phrases (e.g., "Looking at...", "However, I must consider...", "To avoid
compressing the audio...", "Ultimately...").
- Length: 150 - 300 words.

Only return the continuous internal monologue text.
```

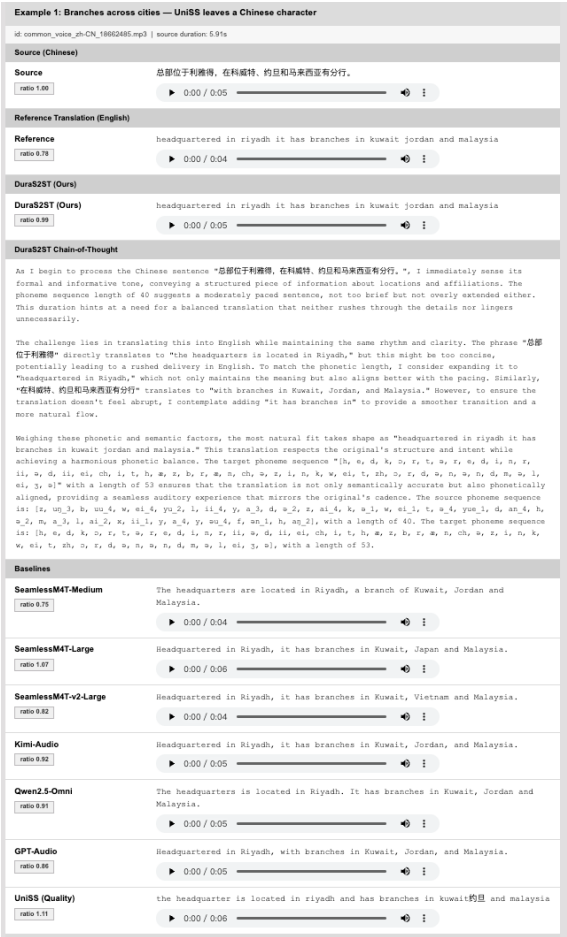
Figure 8: Prompt template for reasoning rationale construction. The prompt guides the LLM to compare translation choices against deterministic phoneme lengths before outputting the final translation.

(a) EN→ZH Example 1.

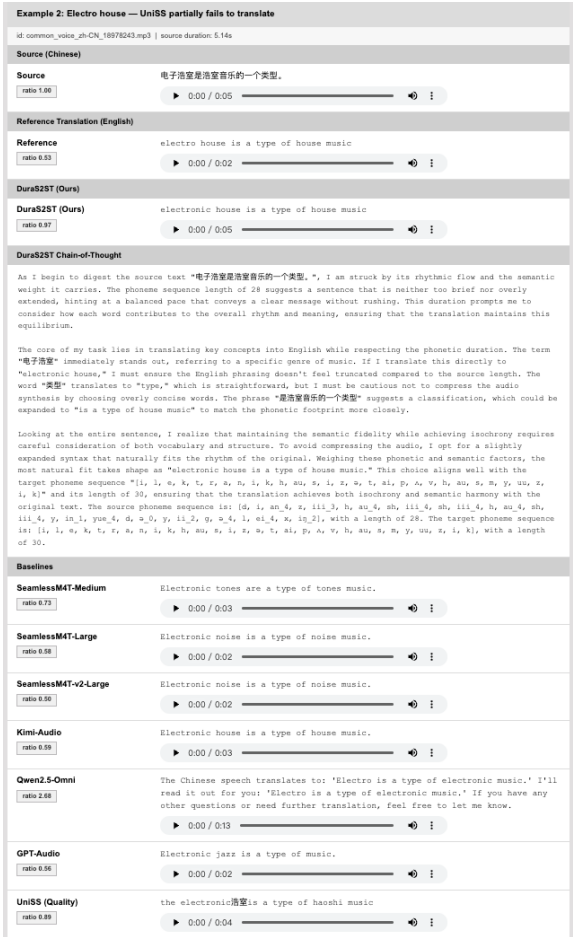


(b) EN→ZH Example 2.

Figure 9: **Qualitative EN→ZH test samples.** For each sample we contrast a representative baseline output against the output of DuraS2ST, aligned to the source utterance. The baseline systems exhibit visible duration drift relative to the source's temporal envelope, while DuraS2ST tracks the source duration tightly and produces a translation faithful to the reference.



(a) ZH→EN Example 1.



(b) ZH→EN Example 2.

Figure 10: **Qualitative ZH→EN test samples.** For each sample we contrast a representative baseline output against the output of DuraS2ST, aligned to the source utterance. As in the EN→ZH case (Figure 9), DuraS2ST closely matches the source's duration and produces translations consistent with the reference, whereas the baseline drifts in duration and content.

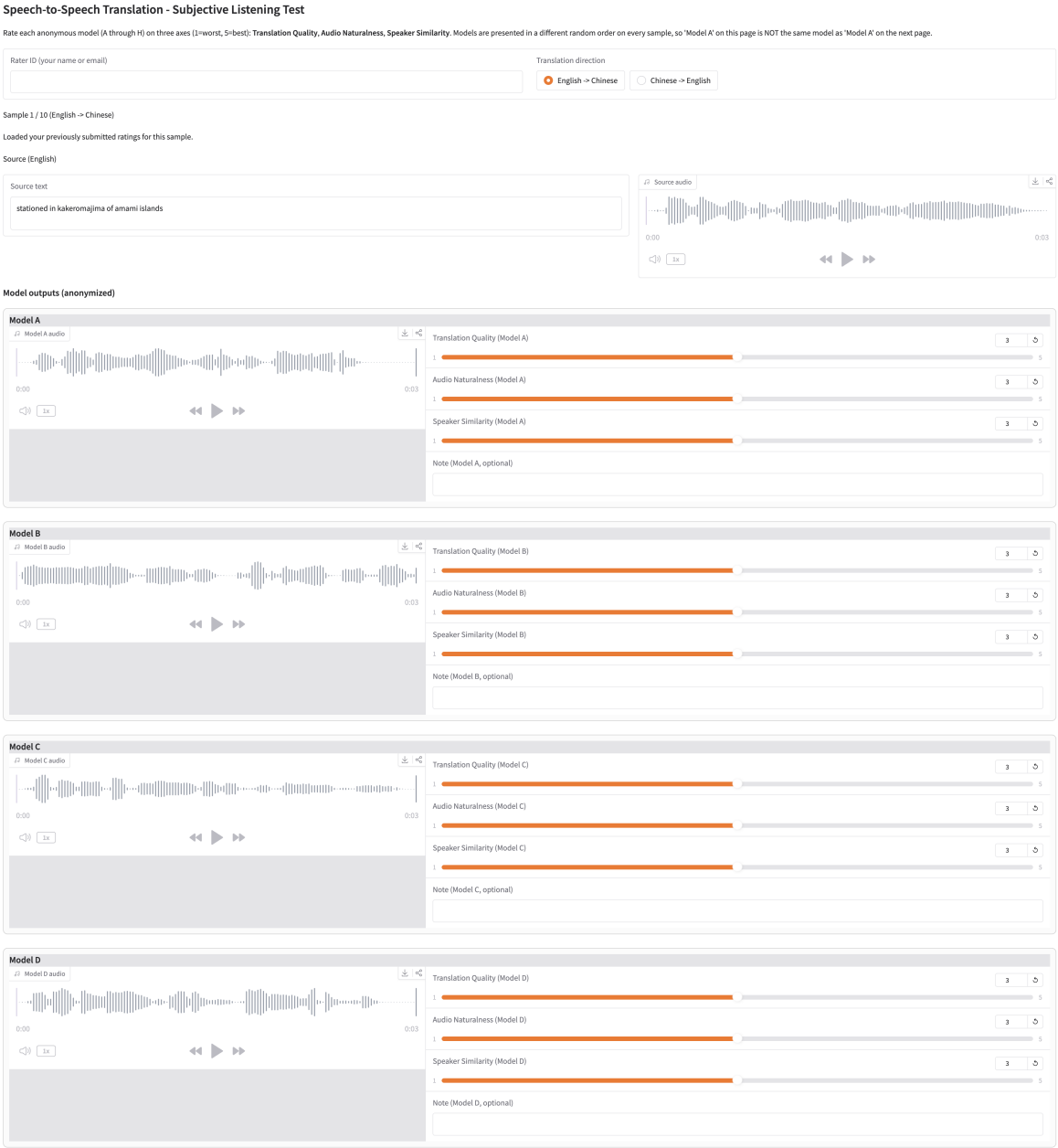


Figure 11: **Subjective evaluation interface used for the MOS study** (Section 4.4). For every task, the annotator listens to a source clip and a synthesized target clip whose system identity is replaced by an anonymized label, then rates the target on three 1–5 Likert scales—translation adequacy, naturalness, and speaker similarity. The presentation order within each direction is randomized per annotator to mitigate position bias.