TOWARDS END-TO-END CONVERSATIONAL MULTI-TALKER SPEECH-TO-SPEECH TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present COMUS, an end-to-end conversational multi-talker S2ST framework that leverages full contextual information to jointly model semantic meaning, speaker attribution, and prosodic characteristics for end-to-end multi-talker S2ST. Built on Qwen3-Omni, COMUS adapts the Thinker to autoregressively predict a single stream of text, speaker-label, and BiCodec speech tokens, replacing the original multi-token speech-generation module. Specifically, given the entire recording, COMUS predicts speaker profile tokens for each speaker alongside semantic tokens tagged with speaker labels, indicating speaker turns within a single autoregressive generation process. To provide the required supervision, we construct SODA-SoulX, comprising approximately 26K hours of paired speaker-consistent conversational speech across English and Chinese. Experiments on real conversational datasets, including AMI and AliMeeting, demonstrate that COMUS achieves speaker-attributed ASR-BLEU of 31.9 and 20.4 for English-to-Chinese and Chinese-to-English translation, respectively, outperforming all baseline methods. Further evaluations show translation advantages under noise and overlapping speech. Subjective listening tests show consistent advantages of COMUS in speaker similarity, translation accuracy and conversational naturalness over baseline methods. Audio samples: `https://anonysample.github.io/E2E-CoT-MultiTalker-S2ST/`.

1 INTRODUCTION

End-to-end speech-to-speech translation (S2ST) directly translates spoken content into target-language speech (Jia et al., 2019). Beyond semantic accuracy, recent work has explored preserving speaker identity (Jia et al., 2022a), expressiveness (Cheng et al., 2026), prosody (Ulgen et al., 2026), and isochrony (Le et al., 2024), supporting natural cross-lingual communication and applications such as dubbing or real-time interpretation. Despite recent progress, translating complete conversations remains challenging: speakers recur across turns, semantics depend on context, and recordings may contain noisy and overlapping speech. Prior work has explored speaker turn-aware translation and multi-talker voice preservation through both end-to-end and modular-based approaches (Jia et al., 2022a; Zuluaga-Gomez et al., 2023; Zhao et al., 2026). However, jointly modeling speaker diarization, contextual translation, and speaker-consistent speech generation within a single model for complete conversations remains underexplored.

Translating multi-talker conversations requires more than rendering the source content as target-language speech: a system must also preserve the conversational structure and maintain consistent speaker characteristics across languages. As shown in Fig. 1, previous modular solutions fall into two main categories, both of which use speaker diarization as their first stage but differ thereafter. The fully cascaded method chains Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech synthesis (TTS), whereas the hybrid method relies on a single-talker S2ST model. Both optionally apply speech enhancement and separation to handle noisy or overlapping speech. However, this approach has several limitations: First, errors propagate and accumulate across different modules. Second, translating segments independently restricts access to cross-turn linguistic context and can disrupt natural transitions. Third, estimating voice conditions separately for each segment can produce inconsistent voices across a speaker's turns.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

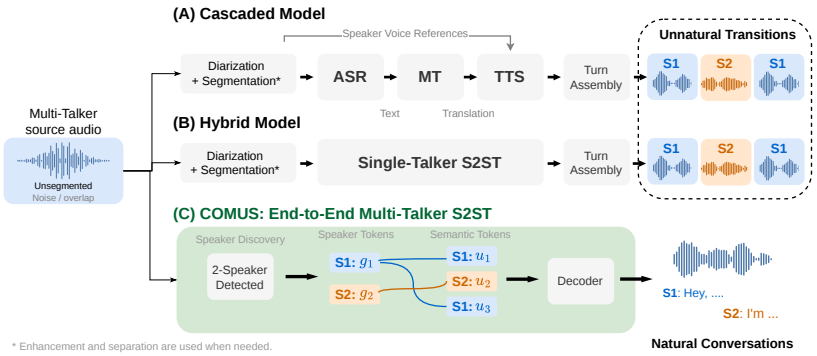

Figure 1: Comparison of cascaded and end-to-end approaches to multi-talker conversational S2ST.

Developing a multi-talker S2ST model in an end-to-end manner raises several challenges. (1) *Speaker-attributed understanding*. The model must first identify who speaks what and when from the recording. (2) *Speaker-attributed generation*. The model must generate the translated speech for each turn in the corresponding speaker's voice while preserving the conversational context. (3) *Obtaining speaker-consistent parallel supervision*. Training requires paired source- and target-language conversational speech that preserve turn-level meaning and source-speaker voices, which are substantially harder to obtain than source recordings alone.

We propose COMUS, an end-to-end **CO**nversational **MU**lti-talker **S**2ST framework that jointly predicts speaker profiles for each speaker and semantic token sequences with speaker labels inserted at the start of each speaker's turn within a single autoregressive generation process. We adopt BiCodec to represent speech with global speaker tokens and semantic tokens. The predicted speaker profiles and semantic tokens are used to reconstruct the speech waveform via the BiCodec decoder (Wang et al., 2025). This unified formulation removes the need for external diarization or separation systems. Our scalable data pipeline provides approximately 1.17 million synthetic conversation pairs, totaling 26K hours across English and Chinese. Both objective and subjective evaluations on real meeting data show that COMUS outperforms baseline methods in speaker-attributed translation accuracy, voice similarity and conversational naturalness. Our main contributions are:

- We introduce COMUS, an end-to-end framework for conversational multi-talker S2ST, which takes a complete conversation as input, discovers its speakers, predicts one profile for each, generates speaker-labeled translated semantic tokens, and decodes each turn using its corresponding predicted profile to synthesize the target-language conversation.
- Using a scalable construction pipeline, we build SODA-SoulX, comprising approximately 1.17 million fully synthetic, speaker-consistent English–Chinese conversation pairs totaling 26K hours, together with 67,734 semi-real conversation pairs containing approximately 310 hours of real source speech for domain adaptation.
- Objective and subjective evaluations on real meetings show that COMUS improves speaker-attributed translation, speaker-turn accuracy, and conversational naturalness over the evaluated baselines while maintaining strong voice preservation. Robustness and ablation analyses further demonstrate resilience to noise and overlap and clarify the roles of intermediate text and multi-task pretraining.

2 Related Work

Speech-to-Speech Translation. Early S2ST systems were constructed as ASR–MT–TTS cascades (Nakamura et al., 2006), while Translatotron introduced direct generation of target spectrograms (Jia et al., 2019). Subsequent advances include discrete speech units (Lee et al., 2022; Popuri et al., 2022), two-pass text-and-unit generation (Inaguma et al., 2023), and multilingual parallel corpora (Jia et al., 2022b). Speech-text pretraining (Ao et al., 2022; Rubenstein et al., 2023) and multilingual S2ST (SEAMLESS Communication Team, 2025; Seamless Communication et al.,

2023) provide broader foundations for speech understanding and generation. Recent studies explore voice preservation and zero-shot generation (Cheng et al., 2026; Zheng et al., 2025), preservation of prosody (Ulgen et al., 2026) and isochrony (Le et al., 2024), and real-time translation (Zhang et al., 2024; Labiausse et al., 2025).

Real World Multi-Talker Speech Processing. Conventionally, understanding multi-talker speech and enabling downstream tasks usually require dedicated multi-talker speech processing. To cope with real-world acoustic environment, speech enhancement and dereverberation are required (Nakatani et al., 2010; Hu et al., 2020). Diarization is employed to address the "who spoke when" problem, with relevant approaches including overlap-aware diarization and target speaker activity detection (Fujita et al., 2019; Medennikov et al., 2020; Wang et al., 2024). More recently, joint diarization and transcription have been developed to tackle "who spoke what and when", as exemplified by serialized transcription (Kanda et al., 2020). And most recently, speech language models have been adopted to jointly perform recognition and diarization (Yin et al., 2026; Huo et al., 2026; Dai et al., 2026). A particularly challenging aspect of multi-talker speech data is the overlapping speech, where two or more speakers talk simultaneously. Addressing this requires speech separation (Hershey et al., 2016; Luo & Mesgarani, 2019; Subakan et al., 2021).

Conversational Multi-Talker S2ST. Existing multi-talker S2ST approaches mostly adopt a cascaded method that requires at least speaker diarization to segment speakers. Translatotron 2 uses concatenation-based augmentation to preserve voices across speaker turns, evaluating pairs of independently sampled utterances with a single speaker transition (Jia et al., 2022a). Speaker-turn-aware conversational speech translation jointly models recognition, translation, and speaker changes, but produces target-language text (Zuluaga-Gomez et al., 2023). DiariST (Yang et al., 2024) jointly performs streaming multi-talker speech-to-text translation and speaker diarization, but does not generate target speech. X-Translator (Zhao et al., 2026) supports long-form, multi-speaker S2ST through a modular cascade of streaming ASR, MT, online speaker-prompt management, and prompt-conditioned TTS. Seed LiveInterpret 2.0 (Cheng et al., 2025) reports streaming end-to-end S2ST with multi-speaker support and voice cloning, but does not separately quantify speaker attribution or voice preservation.

3 COMUS: **CO**NVERSATIONAL **MU**LTI-TALKER **S**2ST

COMUS performs multi-talker speech to speech translation fully end-to-end, taking an entire source conversation as input and directly generating the translated speech, without separate diarization or separation stages. Given only the raw recording, it counts the number of speakers, predicts a speaker profile for each, and generates semantic tokens with speaker labels that serve as turn indicators within a single autoregressive process. As shown in Figure 2, this yields a speaker-attributed representation: each discovered voice has its profile predicted once and assigned an index, and every turn carries its speaker's index, so recurring turns from the same speaker are bound to a consistent voice.

Problem formulation. Let x denote a complete, unsegmented source-language recording. Its latent conversational structure can be described as:

$$\mathcal{C}_x = \left\{ \left(s_i, z_i, \tau_i^{\text{start}}, \tau_i^{\text{end}} \right) \right\}_{i=1}^{N}, \tag{1}$$

where N is the number of conversational turns, s_i is the source-language speech of turn i, $z_i \in \{1, \ldots, K\}$ is its speaker index, K is the number of speakers, and τ_i^{start} and τ_i^{end} are its temporal boundaries. Given only the unsegmented recording x, our goal is to generate target-language speech,

$$\hat{y} = \arg\max_{y} p_\theta(y \mid x), \tag{2}$$

that preserves the semantic content, speaker attribution, and turn structure of the source conversation. At inference time, no speaker labels, turn boundaries, transcripts, translations, or voice references are provided as input.

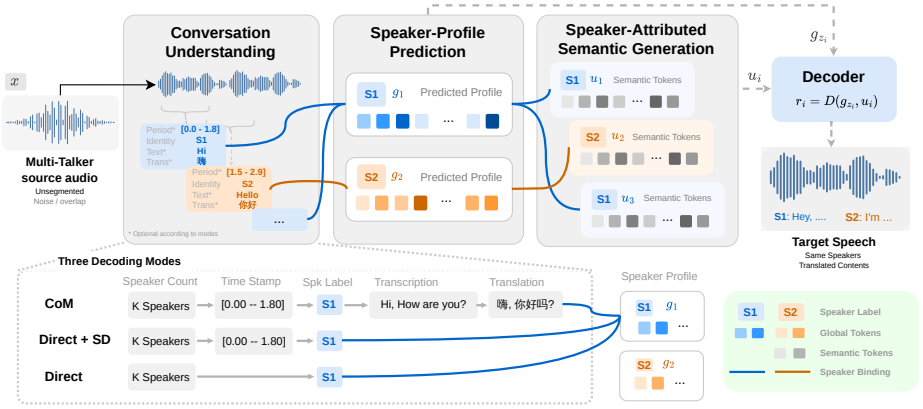

Figure 2: Overview of speaker-oriented conversational generation in COMUS.

3.1 Speaker-Attributed Conversational Generation

Output Representation. We use BiCodec (Wang et al., 2025) to represent target speech, encoding speaker profile as global tokens and speech content as semantic tokens. This codec design allows COMUS to predict a single, constant speaker profile for each speaker and reuse it across all of that speaker's turns, while generating different translated content for each turn.

As illustrated in Figure 2, COMUS generates a single autoregressive token sequence conditioned on the entire source recording. It first predicts the speaker count and, depending on the decoding mode, may explicitly generate timestamped source-turn speaker labels and intermediate text. It then predicts global speaker tokens g_k for each speaker k as the speaker profile, followed by a speaker index z_i and semantic tokens u_i for each translated turn i. Each index is linked to the corresponding speaker's profile tokens, which the BiCodec decoder combines with the turn's semantic tokens to generate speech:

$$\hat{r}_i = D(g_{z_i}, u_i). \tag{3}$$

The shared index thus connects speaker attribution in the source conversation to voice-conditioned target speech generation. End-to-end speaker counting and speaker-wise attractor prediction have been explored in prior diarization work (Horiguchi et al., 2022). However, to our knowledge, COMUS is the first to predict discrete global speaker tokens with a speech language model and use them to reconstruct speaker-attributed target speech.

Decoding modes. We consider three decoding modes that share the same model architecture, target-speech representation, and speaker-binding mechanism but differ in the information generated, as shown in Figure 2. In each mode, the output always includes global speaker-token sequences and semantic tokens with speaker labels. What differs is the information preceding this sequence: (1) **CoM** contains the speaker count, a source transcript with speaker labels and timestamps, and a speaker-attributed target translation. (2) **Direct+SD** contains the speaker count and timestamped source-turn speaker labels, but no transcript or translation. (3) **Direct** contains only the speaker count.

3.2 Model Architecture

We build COMUS on Qwen3-Omni-30B-A3B (Xu et al., 2025). Its audio-text pretraining provides a foundation for recovering linguistic content and acoustic cues from the source conversation. Qwen3-Omni can perform end-to-end S2ST for single speaker data using the predefined speaker voices. However, it cannot handle multi-talker end-to-end S2ST, nor can it perform S2ST in a zero-shot manner that preserves each source speaker's voice. Conversational translation requires additional reliable speaker attribution and turn boundaries, particularly under noise and overlapping speech. While Qwen3-Omni exhibits some speaker diarization capability, it is not strong enough to support multi-talker S2ST in our preliminary experiments.

Building on Qwen3-Omni, COMUS enables multi-talker S2ST through three architectural components. First, we inherit and freeze its Audio Transformer (AUT) for audio encoding. This reuses its pretrained audio representations without additional training. Second, we adapt its MoE Thinker backbone through our speaker-attributed supervision. This trains the model to jointly recognize who speaks and what is said, producing speaker-conditioned translations across multiple talkers. Third, we extend the Thinker's shared vocabulary and output head with speaker indices, BiCodec global tokens, and BiCodec semantic tokens to replace the original multi-token prediction (MTP) module. Compared with the MTP design, this formulation improves generation efficiency by reducing the number of predicted frame-level tokens by approximately 75%. More importantly, it introduces explicit speaker-change prediction, allowing COMUS to track turn transitions and maintain reliable speaker attribution under overlapping speech.

3.3 Training Strategy

Our training strategy consists of two phases. In the first phase, we integrate the BiCodec tokens into the model's output generation through multi-task learning while preserving the model's original capabilities; and in the second phase, we train the multi-talker S2ST capability with different modes.

Phase A: Output token integration. Since we have replaced the original multi-layer codec tokens with single-stream BiCodec tokens which are new to the base model, we train the model on a set of auxiliary tasks, such as ASR, machine translation, speech-to-text translation, text-to-speech generation, to integrate BiCodec's speaker and semantic tokens while preserving the backbone's translation, recognition and speaker-attribution capabilities. The complete task set, training data, and sampling proportions are provided in Appendix A. We examine the contribution of this phase in Section 5.4.2.

Phase B: Conversational S2ST training. Starting from the Phase-A checkpoint, we train Phase B on the fully synthetic component of SODA-SoulX and subsequently adapt each mode on its semi-real component. We jointly train CoM and Direct+SD with auxiliary tasks; Direct mode is trained in a separate run using the same underlying conversations. Each mode is subsequently adapted on the same semi-real training set. Training details are provided in Appendix A.

4 SODA-SoulX: Speaker-Consistent Parallel Dataset

Training multi-talker S2ST requires paired source- and target-language conversations with aligned turn content and consistent speaker voices across languages, which existing corpora do not provide at scale. We therefore construct SODA-SoulX, a large-scale parallel conversational speech dataset with two complementary components: fully synthetic conversations provide supervision at scale, while semi-real conversations retain real source-side acoustics and speaker-interaction patterns for adaptation. Figure 4 in Appendix B summarizes the construction pipeline. The complete construction procedures and data statistics are provided in Appendix B.

Fully synthetic conversations. We begin with text dialogues in English between two participants from SODA (Kim et al., 2023), colloquialize and translate the whole dialogues into Chinese using Claude while preserving speaker labels and turn correspondence. We construct 6,108 Chinese voice entries by clustering utterances from WeNetSpeech4TTS (Ma et al., 2024) based on their speaker embeddings. For each voice entry, we then use UniSS (Cheng et al., 2026) to generate a voice-matched English reference, yielding a bilingual voice pool with consistent cross-lingual voice identities. For each dialogue, we assign one bilingual voice pair to each participant and use SoulX-Podcast (Xie et al., 2025), a multi-talker TTS system, to render both language versions with natural conversational dynamics. To better reflect real conversational conditions, we add noise and reverberation to source recordings in a training subset and simulate turn intervals and overlaps using timing statistics from real conversations. The paired target speech remains clean. This process yields 1,171,614 two-speaker bilingual conversation pairs with consistent speaker identities across languages, totaling 25,976 hours of speech. Each pair provides speaker labels and timestamps at 10-ms resolution.

Semi-real conversations. We construct semi-real conversation pairs by collecting real source conversations from AMI (Renals et al., 2007), ICSI (Janin et al., 2003), RAMC (Yang et al., 2022), and AliMeeting (Yu et al., 2022), then translating their transcripts into the corresponding target-language with Claude. Each semi-real pair retains the original recording on the source side, preserving its speaker-turn structure, pauses, overlapping speech and noise, while the translated transcript is rendered as a target-language conversation referencing the corresponding source speakers' voices translated by UniSS, as in the fully synthetic subset. The two-speaker semi-real corpus contains 67,734 segments and approximately 310 hours of real source speech.

5 EXPERIMENTS

5.1 EVALUATION SETUP

Evaluation Data. We collected three two-talker evaluation sets: 1) **real-eval**, which consists of 400 non-overlapping, real conversations from RAMC (Yang et al., 2022), AMI (Renals et al., 2007), AliMeeting (Yu et al., 2022), and ICSI (Janin et al., 2003); 2) **overlap-eval**, which consists of 400 real conversations with an average overlap ratio of approximately 8% from AMI, AliMeeting, and ICSI; and 3) **synth-eval**, which consists of 300 synthetic conversations from the full synthetic component of SODA-SoulX. These evaluation sets have no ground-truth translations, thus we construct reference translations using GPT-5.5 for reference-based evaluations. We deliberately use different translation tools for the training and evaluation data to avoid style leakage between the two sets.

Baselines. We compare COMUS with modular cascades and a hybrid method based on single-talker S2ST systems. The cascaded systems include a conventional speaker diarization (SD)–ASR–MT–TTS cascade and a variant that uses SoulX-Transcriber for joint transcription and speaker diarization (Dai et al., 2026). Like COMUS, SoulX-Transcriber is also fine-tuned from Qwen3-Omni for multi-talker transcription. The hybrid system applies speaker diarization with Pyannote 3.1 (Bredin, 2023), followed by SeamlessM4T or UniSS (SEAMLESS Communication Team, 2025; Cheng et al., 2026) for single-speaker S2ST. We additionally include Qwen3-Omni in S2TT mode as a text-only translation comparison. We do not include its S2ST output in the multi-talker speech evaluation because it uses a single selected voice per generation rather than matching the voices of individual source speakers.

Evaluation metrics. Our evaluation metrics primarily assess speaker-attributed translation quality, and speech naturalness of the translated target speech. We transcribe the target conversation using Whisper (Radford et al., 2023)/Paraformer (Gao et al., 2022) for English and Chinese respectively, and compute speaker-agnostic ASR-BLEU (**SAgASR-BLEU**) and speaker-attributed ASR-BLEU (**SAtASR-BLEU**) (Yang et al., 2024) which was originally inspired by Watanabe et al. (2020), against reference translation to measure conversational translation quality and intelligibility. In CoM decoding mode, which outputs translated text as an intermediate result, we additionally report **BLEU**, speaker-attributed BLEU (**SAtBLEU**) and **conversation-level MQM (conv-MQM)** to assess translation quality from the text output. Based on GEMBA-MQM (Kocmi & Federmann, 2023), Conv-MQM evaluates the complete conversation, with scores closer to zero indicating better quality. We report source-side diarization error rate (DER) to evaluate the recovery of speaker labels and turn boundaries, but only for systems producing time-aligned speaker attribution. Finally, we report UTMOS (Saeki et al., 2022) for acoustic quality and speaker similarity for voice preservation.

Full baseline configurations and detailed setup for Conv-MQM are provided in Appendix C; scaling from 1 to 4 talkers, analysis of conversation lengths, and other additional experiment results can be found in Appendix D.

5.2 MAIN RESULTS ON REAL CONVERSATIONS

Table 1 presents the main objective results on the real-eval set with CoM decoding mode, including translation quality, target-speech quality, and speaker diarization performance. Each cell reports **EN→ZH** | **ZH→EN**. The shared pyannote result applies to Pure Cascade, both SeamlessM4T variants, and SD+UniSS. SIM is averaged across source speakers. Qwen3-Omni is evaluated here only in its S2TT configuration and therefore produces no target speech in our experiments. The com-

Table 1: Main objective results on the real-conversation evaluation set (real-eval).

	Translation Quality			Speech		Source
System	**SAtBLEU ↑**	**SAtASR-BLEU ↑**	**conv-MQM ↑**	**UTMOS ↑**	**SIM ↑**	**DER ↓**
Pure Cascade	13.3 | 6.5	14.7 | 6.5	−17.4 | −37.7	3.15 | 3.67	0.822 | 0.842	28.1 | 19.2
SoulX-Transcriber	16.5 | 8.3	16.3 | 7.9	−17.5 | −29.3	3.14 | 3.65	**0.831** | 0.856	19.7 | 10.5
SeamlessM4T-S2ST	10.9 | 7.1	9.4 | 6.5	−25.5 | −29.9	2.53 | 3.67	–	28.1 | 19.2
SeamlessM4T-S2TT+TTS	9.8 | 7.0	8.7 | 6.6	−26.0 | −31.3	3.03 | 3.69	0.816 | 0.846	28.1 | 19.2
SD+UniSS	13.8 | 10.8	12.1 | 9.1	−21.6 | −21.0	2.83 | 3.21	0.822 | **0.901**	28.1 | 19.2
Qwen3-Omni (S2TT)	23.6 | 19.1	–	−7.9 | −6.9	–	–	31.2 | 26.3
COMUS	**34.9** | **24.3**	**31.9** | **20.4**	**−7.7** | **−6.4**	**3.38** | **3.79**	0.824 | 0.893	**9.3** | **4.1**

Table 2: Results on conversations with overlapping speech (overlap-eval).

System	**SAgBLEU ↑**	**SAtBLEU ↑**	**DER ↓**	**SIM ↑**
UniSS + SepFormer	13.2 | 12.6	10.7 | 11.4	30.9 | 18.4	0.746 | **0.841**
SoulX-Trans + NLLB	15.3 | 7.8	12.3 | 8.6	29.5 | 19.2	– | –
Qwen3-Omni	26.5 | 21.2	20.6 | 19.0	43.3 | 36.8	– | –
COMUS	**34.6** | **26.1**	**31.6** | **27.1**	**16.7** | **3.3**	**0.759** | 0.827

plete metric breakdown, including conventional BLEU and ASR-BLEU and per-speaker similarity scores, is provided in Appendix D.1.

Speaker diarization performance. COMUS achieves the lowest DER in both translation directions. Compared to the second-best diarization system, SoulX-Transcriber, it reduces the DER by 52.8% for EN→ZH and 61.0% for ZH→EN, demonstrating its strong capability to translate multi-speaker speech with accurate speaker attribution.

Conversation-level translation. Our backbone, Qwen3-Omni, is already the strongest S2TT baseline, yet its native S2ST mode does not preserve the distinct voices of multiple source speakers. Building on this strong foundation, COMUS extends translation into full speaker-attributed speech generation while further improving translation quality over Qwen3-Omni itself. Compared with the segment-wise UniSS baseline, COMUS raises the speaker-attributed SAtASR-BLEU from 12.1 to 31.9 for EN→ZH and from 9.1 to 20.4 for ZH→EN, corresponding to gains of 19.8 and 11.3 points, respectively, and improves its conv-MQM scores from -21.6 to -7.7 and from -21.0 to -6.4. These results show that jointly processing the complete conversation yields better contextual translation than translating each segment independently.

Speech quality and identity preservation. COMUS achieves the highest UTMOS scores in both directions while maintaining speaker similarity comparable to the voice-preserving baselines. Notably, the segment-wise baselines derive their voice conditions from source segments produced by an external diarization front end, whereas COMUS directly predicts an indexed voice representation for each discovered speaker from the entire conversation.

Additional evaluations across one to four speakers show that COMUS maintains competitive single-speaker S2ST performance and extends to three- and four-speaker conversations, although speaker attribution degrades as the number of speakers increases, particularly for four-speaker inputs (Appendix D.4).

5.3 Robustness to Noisy and Overlapping Speech

Real multi-talker recordings frequently contain overlapping speech, background noise, and competing speakers. We evaluate COMUS on the **overlap-eval set** for overlapping speech, and augment the **real-eval set** with ambient noise from the DNS Challenge (Reddy et al., 2020) and competing speakers from AISHELL-3 (Shi et al., 2021) (ZH) / VoxCeleb2 (Chung et al., 2018) (EN) to test noise robustness.

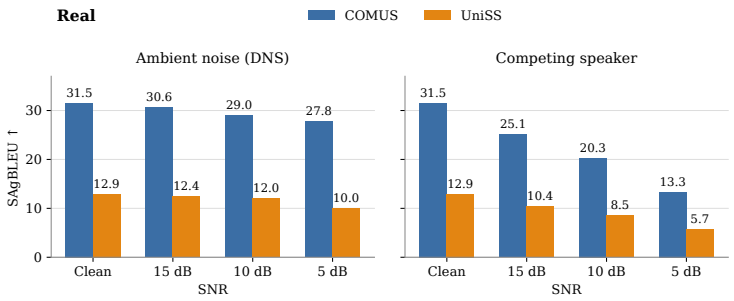

Figure 3: Average SAgBLEU over the two translation directions on real-eval under clean conditions and added ambient noise or competing speech.

Overlapping speech. For a fair segment-wise comparison, we equip the SD+UniSS baseline with a SepFormer (Subakan et al., 2021) front end that separates overlapping speech before diarization and translation. As shown in Table 2, COMUS achieves SAtBLEU scores of 31.6 and 27.1 for EN→ZH and ZH→EN, versus 10.7 and 11.4 for the UniSS-SepFormer baseline, while source-side DER drops from 30.9% to 16.7% and from 18.4% to 3.3%, respectively. Those results suggest COMUS maintains stronger translation and speaker attribution without a dedicated speech-separation stage.

Ambient noise and competing speakers. We further evaluate performance under added ambient noise and competing speech at controlled SNRs. Figure 3 reports SAgBLEU averaged over both directions. Competing speech is especially challenging, as it introduces linguistic content absent from the reference, and we include it as a stress test. Although both systems degrade, COMUS maintains higher translation quality at every SNR.

5.4 Ablation Studies

5.4.1 Effect of decoding modes.

Table 3 (a) compares the three decoding modes from Section 3.1: CoM, Direct+SD, and Direct generation. All three modes use the same backbone, target-speech representation, and speaker-binding mechanism, but are trained with mode-specific output structures and optimization schedules. Each cell reports EN→ZH | ZH→EN. DER is unavailable for Direct because it does not predict timestamps.

The results show that the performance gap between CoM mode and Direct mode is smaller in synthetic evaluations than on real conversation data. This pattern suggests that intermediate transcripts and translations provide a useful generation scaffold under limited real-domain adaptation, although the comparison also reflects the mode-specific training configurations. Under the evaluated configurations, Direct+SD performs similarly to Direct, suggesting that the explicit speaker-turn scaffold alone provides limited additional guidance.

Speaker similarity varies little among CoM, Direct+SD, and Direct, indicating that the benefit of intermediate text has little effect on speaker profile prediction.

5.4.2 Analysis of Phase A

We next examine the effect of Phase A training for conversational S2ST. In a separate CoM-only ablation on synthetic conversations, adding Phase A yields small SAgASR-BLEU gains, from $48.34|41.31$ to $49.76|41.88$ in two translation directions, respectively. On the other hand, Table 3 (b) reports the generated results of Direct mode with and without Phase-A training. Unlike CoM mode, Direct mode depends heavily on Phase A. SAgASR-BLEU improves from $33.75\,|\,25.41$ to $43.62\,|\,37.81$ on synthetic conversations and from $7.90\,|\,1.99$ to $21.34\,|\,9.12$ on real conversations. Although both modes jointly model speech understanding, translation, speaker profile prediction, and speech-token generation, CoM mode emits explicit intermediate transcripts and translations,

Table 3: Decoding-mode comparison and Phase-A ablation.

(a) Decoding modes comparison.

Mode	SAgASR-BLEU ↑	DER (%) ↓	SIM ↑
Real conversations			
CoM	**33.68** | **21.54**	**9.3** | **4.1**	0.821 | 0.893
Direct+SD	22.40 | 8.66	14.5 | 14.0	0.827 | 0.895
Direct	21.34 | 9.12	–	**0.828** | **0.898**
Synthetic conversations			
CoM	**50.20** | **42.50**	**1.2** | **1.1**	0.887 | 0.892
Direct+SD	42.20 | 36.83	2.1 | 3.5	**0.889** | 0.891
Direct	43.62 | 37.81	–	0.887 | **0.893**

(b) Ablation of Phase-A

Data	w/ A	w/o A
Synthetic	**43.62** | **37.81**	33.75 | 25.41
Real	**21.34** | **9.12**	7.90 | 1.99

SAgASR-BLEU is reported in Panel (b).

Table 4: Subjective evaluation results. We report the mean and standard deviation.

System	Spk.-Turn Acc. ↑	Spk. Sim. ↑	Conv. Nat. ↑	Trans. Qual. ↑	Overall ↑
Cascade	0.64 ± 0.22	3.00 ± 0.65	2.71 ± 0.59	2.53 ± 0.70	2.51 ± 0.57
SD+UniSS	0.61 ± 0.21	3.41 ± 0.58	2.73 ± 0.55	2.73 ± 0.62	2.67 ± 0.51
COMUS	$\mathbf{0.96 \pm 0.13}$	$\mathbf{3.87 \pm 0.50}$	$\mathbf{3.60 \pm 0.56}$	$\mathbf{3.82 \pm 0.54}$	$\mathbf{3.67 \pm 0.52}$

whereas Direct omits them and generates speech tokens directly. Phase A provides this missing supervision during training, enabling the model to learn explicit component mappings and later invoke them implicitly under Direct decoding.

5.5 Subjective Evaluation

We conducted a listening test with ten subjects evaluating ten conversations (five per translation direction). Subjects rated COMUS, UniSS + SD, and the cascaded system on a five-point scale for speaker similarity, conversational naturalness, translation quality, and overall quality. Additionally, a binary assessment measured speaker-turn accuracy by checking if the output speaker matched the source turn. Details of this evaluation are described in Appendix E.

Table 4 shows that COMUS achieves the highest mean rating across all dimensions. Compared to SD+UniSS, it improves speaker similarity (from 3.41 to 3.87), conversational naturalness (2.73 to 3.60), translation quality (2.73 to 3.82), and overall quality (2.67 to 3.67). Speaker-turn accuracy also increases from 61% to 96%. Similar gains are observed over the cascaded system, showing that COMUS provides higher perceived translation quality and speaker attribution.

6 Conclusion

We present COMUS, an end-to-end multi-talker S2ST framework that translates speech conversations into target-language speech while preserving source-speaker identities, without external diarization or separation. Within a single autoregressive generation process, COMUS predicts a reusable profile for each discovered speaker and speaker-labeled semantic tokens for each translated turn, using shared speaker indices to bind turns to their corresponding profiles. To provide supervision, we build SODA-SoulX from large-scale synthetic parallel conversations and semi-real data retaining real source-side acoustics. Objective and subjective evaluations demonstrate improvements over baselines in translation quality, voice preservation, and naturalness, including under noisy and overlapping conditions.

Our study focuses on English–Chinese translation of two-talker conversations and relies on synthesized target speech for training, potentially limiting naturalness. Although our listening study assesses conversational naturalness, including pacing, pauses, and speaker transitions, it does not separately quantify source-to-target prosody or isochrony, limiting conclusions for timing-sensitive applications such as dubbing. Future work will extend COMUS to more language pairs, talkers, and longer conversations, while developing explicit objectives and evaluations for cross-lingual prosody and isochrony preservation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search, language editing of portions of the manuscript, partial code implementation, and experiment setup and execution. We also used large language models in constructing the training and evaluation data, as described in Section 4. The research problem, proposed method, experimental design, analysis, and conclusions were developed and determined by the authors.

All AI-assisted work was reviewed by the authors. We manually verified the relevance and accuracy of identified literature and independently checked the cited sources. AI-assisted text was reviewed, revised, and proofread by the authors. AI-assisted code was inspected and tested for correctness, and the resulting experiments and outputs were manually validated. We take responsibility for the final content of this work, including all text, claims, code, data, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our data construction process uses voice references to synthesize speaker-consistent cross-lingual conversations. This capability may raise concerns related to unauthorized voice imitation or impersonation. The constructed voice clusters represent acoustically similar voices rather than verified physical identities, and the resulting data and models are intended solely for research on conversational speech translation. We follow the applicable terms of the source corpora and do not release derived speech data when redistribution is restricted. The listening evaluation uses anonymized system outputs and reports no personally identifying information.

REPRODUCIBILITY STATEMENT

Appendix A describes the model configuration, output representation, training tasks, optimization settings, and mode-specific training procedure. Appendix B provides the complete data-construction pipeline, filtering criteria, augmentation settings, and corpus statistics. Appendices C–E document the baseline configurations, evaluation metrics, additional results, and listening-test protocol. Because of redistribution restrictions and voice-privacy considerations, the derived speech data cannot be publicly released; however, we provide the detailed construction procedures and experimental configurations to facilitate reproduction using the corresponding source corpora and referenced models.

REFERENCES

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proc. ACL*, pp. 5723–5738, 2022.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech*, pp. 2278–2282, 2022.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech*, 2023.

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, Yiming Yu, Jun Zhang, Ruofei Zhang, Wanyi Zhang, Wenhao Zhu, Liehao Zou, Lu Lu, Yuxuan Wang, and Yonghui Wu. Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice, 2025.

Sitong Cheng, Weizhen Bian, Xinsheng Wang, Ruibin Yuan, Jianyi Chen, Shunshun Yin, Yike Guo, and Wei Xue. UniSS: Unified expressive speech-to-speech translation with your voice. In *Proc. ICLR*, 2026.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech*, pp. 1086–1090, 2018.

Yuhang Dai, Haopeng Lin, Zhennan Lin, Jiale Qian, Jun Wu, Hanke Xie, Hao Meng, Hanlin Wen, Chuang Ding, Shunshun Yin, et al. Soulx-transcriber: A robust end-to-end framework for multi-speaker speech transcription. *arXiv preprint arXiv:2606.02400*, 2026.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with self-attention. In *Proc IEEE ASRU*, pp. 296–303, 2019.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech*, pp. 2063–2067, 2022.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. IEEE SLT*, pp. 885–890, 2024.

John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Proc. IEEE ICASSP*, pp. 31–35, 2016.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola García. Encoder-decoder based attractors for end-to-end neural diarization. *IEEE / ACM Trans. Audio Speech Lang. Process.*, 30:1493–1507, 2022.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. In *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

Mingyue Huo, Yiwen Shao, and Yuheng Zhang. Tagspeech: End-to-end multi-speaker ASR and diarization with fine-grained temporal grounding. In *Proc. ACL*, pp. 41847–41862, 2026.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. Unity: Two-pass direct speech-to-speech translation with discrete units. In *Proc. ACL*, pp. 15655–15680, 2023.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI meeting corpus. In *Proc. ICASSP*, pp. 364–367, 2003.

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. In *Proc. Interspeech*, pp. 1123–1127, 2019.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *Proc. ICML*, volume 162, pp. 10120–10134, 2022a.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. CVSS corpus and massively multilingual speech-to-speech translation. In *Proc. LREC*, pp. 6691–6703, 2022b.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. Serialized output training for end-to-end overlapped speech recognition. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *Proc. Interspeech*, pp. 2797–2801, 2020.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: million-scale dialogue distillation with social commonsense contextualization. In *Proc. EMNLP*, pp. 12930–12949, 2023.

Tom Kocmi and Christian Federmann. GEMBA-MQM: detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddon, Tom Kocmi, and Christof Monz (eds.), *Proc. WMT*, pp. 768–775, 2023.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Alexandre Défossez, and Neil Zeghidour. High-fidelity simultaneous speech-to-speech translation. In *Proc. ICML*, volume 267, 2025.

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

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. Direct speech-to-speech translation with discrete units. In *Proc. ACL*, pp. 3327–3339, 2022.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proc. LREC*, pp. 923–929, 2016.

Dennis Liu, Zijie Yan, Xin Yao, et al. Moe parallel folding: Heterogeneous parallelism mappings for efficient large-scale moe model training with megatron core. *arXiv preprint arXiv:2504.14960*, 2025.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *CoRR*, abs/2411.09943, 2024.

Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE / ACM Trans. Audio Speech Lang. Process.*, 27(8):1256–1266, 2019.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. Wenetspeech4tts: A 12, 800-hour mandarin TTS corpus for large speech generation model benchmark. In *Proc. Interspeech*, 2024.

Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Y. Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. In *Proc. Interspeech*, pp. 274–278, 2020.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jinsong Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. Speech Audio Process.*, 14(2):365–376, 2006.

Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Trans. Speech Audio Process.*, 18(7):1717–1731, 2010.

NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846, 2024.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. In *Proc. Interspeech*, pp. 5195–5199, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, volume 202, pp. 28492–28518, 2023.

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Proc. Interspeech*, pp. 2492–2496, 2020.

Steve Renals, Thomas Hain, and Herve Bourlard. Recognition and understanding of meetings the ami and amida projects. In *Proc. IEEE ASRU*, pp. 238–247, 2007.

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 N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovic, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Havnø Frank. Audiopalm: A large language model that can speak and listen. *CoRR*, abs/2306.12925, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *Proc. Interspeech*, pp. 4521–4525, 2022.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanic. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *Proc. ICASSP*, pp. 351–355, 2018.

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, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alexandre Mourachko, Benjamin N. Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Y. Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation. *CoRR*, abs/2312.05187, 2023.

SEAMLESS Communication Team. Joint speech and text machine translation for up to 100 languages. *Nature*, 637:587–593, 2025.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A multi-speaker mandarin TTS corpus. In *Proc. Interspeech*, pp. 2756–2760, 2021.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *CoRR*, abs/1510.08484, 2015.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *Proc. IEEE ICASSP*, pp. 21–25, 2021.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

Ismail Rasim Ulgen, Nancy Liu, Najmeh Sadoughi, Abhishek Yanamandra, Abhinav Jain, Zhu Liu, and Vimal Bhat. Direct transfer of prosody in speech-to-speech translation using disentangled speech tokens. In *Proc. IEEE ICASSP*, pp. 18372–18376, 2026.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. CoVoST 2 and massively multilingual speech translation. In *Proc. Interspeech*, pp. 2247–2251, 2021.

Dongmei Wang, Xiong Xiao, Naoyuki Kanda, Midia Yousefi, Takuya Yoshioka, and Jian Wu. Profile-error-tolerant target-speaker voice activity detection. In *Proc. ICASSP*, pp. 11906–11910. IEEE, 2024.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Proc. Interspeech*, pp. 5301–5305, 2023.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *CoRR*, abs/2503.01710, 2025.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings, 2020.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, Yuepeng Jiang, Wenxi Chen, Ruiqi Yan, Jiale Qian, Yichao Yan, Shunshun Yin, Ming Tao, Xie Chen, Lei Xie, and Xinsheng Wang. Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity. *CoRR*, abs/2510.23541, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025.

Mu Yang, Naoyuki Kanda, Xiaofei Wang, Jun-Kun Chen, Peidong Wang, Jian Xue, Jinyu Li, and Takuya Yoshioka. Diarist: Streaming speech translation with speaker diarization. In *Proc. IEEE ICASSP*, pp. 10866–10870, 2024.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset. In *Proc. Interspeech*, pp. 1736–1740, 2022.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. In *Proc. AAAI*, pp. 34467–34475, 2026.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *Proc. IEEE ICASSP*, pp. 6167–6171, 2022.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. In *Proc. ACL*, pp. 8964–8986, 2024.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *Proc. ICLR*, 2025.

Yuxiang Zhao, Yichi Zhang, Yanjie An, Yanqiao Zhu, Zhanxun Liu, Yushen Chen, Qixi Zheng, Haina Zhu, Yunchong Xiao, Keqi Deng, Shuai Fan, Kai Yu, and Xie Chen. X-translator: A real-time multilingual speaker-aware speech-to-speech translation system, 2026.

Zhisheng Zheng, Xiaohang Sun, Tuan Dinh, Abhishek Yanamandra, Abhinav Jain, Zhu Liu, Sunil Hadap, Vimal Bhat, Manoj Aggarwal, Gerard Medioni, and David Harwath. RosettaSpeech: Zero-shot speech-to-speech translation from monolingual data. *arXiv preprint arXiv:2511.20974*, 2025.

Juan Pablo Zuluaga-Gomez, Zhaocheng Huang, Xing Niu, Rohit Paturi, Sundararajan Srinivasan, Prashant Mathur, Brian Thompson, and Marcello Federico. End-to-end single-channel speaker-turn aware conversational speech translation. In *Proc. EMNLP*, pp. 7255–7274, 2023.

A Model and Training Details

A.1 Model and Vocabulary Configuration

We build COMUS on the Qwen3-Omni-30B-A3B Thinker. The audio encoder and aligner remain frozen throughout training. We extend the Thinker vocabulary with 4,096 BiCodec global-token entries, 8,192 semantic-token entries, and the required boundary and language-control tokens. Its input embedding and output prediction head are resized accordingly, allowing text, speaker indices, global tokens, and semantic tokens to be serialized in a shared autoregressive vocabulary. The vocabulary size increases from 151,676 to 163,976.

BiCodec represents each speaker voice with 32 global tokens and represents turn-specific speech content with a semantic-token stream at 50 Hz. BiCodec is used only on the output side. The input conversation is processed by the pretrained Qwen3-Omni audio front end.

In CoM, the plan gives the speaker count, a time-stamped source transcript with speaker labels, and a speaker-attributed target translation. Direct+SD keeps the speaker count and time-stamped speaker labels without intermediate text. Direct retains only the speaker count in its plan. In each mode, global tokens are generated once per speaker and reused across that speaker's turns; semantic tokens are generated per target turn. Speaker indices (S1, S2, etc.) follow their order of first occurrence in the source conversation. A CoM-mode output example with placeholders is shown below.

```
(2 speakers detected)
[0:00--1.80] S1: <EN transcription 1>
[2.03--3.40] S2: <EN transcription 2>
[0:00--1.80] S1: <ZH translation 1>
[2.03--3.40] S2: <ZH translation 2>
S1: <global speaker tokens>
S2: <global speaker tokens>
S1: <semantic tokens>
S2: <semantic tokens>
```

A.2 Phase-A Component Grounding

Phase A grounds the newly introduced global and semantic speech tokens while maintaining the pretrained model's speech-understanding and translation capabilities. Table 5 summarizes the six auxiliary tasks. Tasks T1–T3 establish mappings from speech and text to the new speech-token representations, while T4–T6 retain text translation, speaker-attributed speech recognition, and speech-to-text translation.

Phase A uses a mixture of Emilia (He et al., 2024), CoVoST-2 (Wang et al., 2021), OpenSubtitles (Lison & Tiedemann, 2016), and SODA-SoulX examples. The final mixture contains approximately 20.81 million training instances.

A.3 Phase-B Conversational Training

We jointly train the CoM and Direct+SD modes with auxiliary tasks on the fully synthetic component of SODA-SoulX. Table 6 shows how the task mixture changes over three stages. We use the Stage-1 checkpoint for CoM and the Stage-3 checkpoint for Direct+SD. Direct is trained separately from the same Phase A checkpoint using the same 1.274M examples used for Stage-1 CoM training, without auxiliary tasks.

Table 5: Phase-A component grounding tasks, data sources, and sampling proportions.

Task	Input → Output	Role	Data source	Share
T1	Speech → semantic tokens	Speech-to-semantic alignment	Emilia (He et al., 2024); SODA-SoulX	14%
T2	Speech → global tokens	Speaker-voice prediction	Emilia; SODA-SoulX	16%
T3	Text → semantic tokens	Text-to-speech-token alignment	Emilia; SODA-SoulX	40%
T4	Source text → target text	Text translation	SODA-SoulX; OpenSubtitles (Lison & Tiedemann, 2016)	11%
T5	Speech → speaker-attributed transcript	Recognition and attribution	SODA-SoulX; CoVoST-2 (Wang et al., 2021)	8%
T6	Source speech → target text	Translation and attribution	SODA-SoulX; CoVoST-2	11%

Table 6: Three stages of the joint Phase-B run. Each speech-task cell reports the number of materialized examples in millions, followed by its share among speech tasks in parentheses. Text MT is excluded when computing these shares. Direct is trained in a separate run.

Stage	CoM	CoM w/ ASR	Direct+SD	Text MT	Total	Selected for
Stage 1	1.274 (70%)	0.362 (20%)	0.180 (10%)	0.269	2.087	CoM
Stage 2	0.188 (30%)	0.313 (50%)	0.125 (20%)	0.313	0.938	–
Stage 3	0.036 (10%)	0.036 (10%)	0.290 (80%)	0.181	0.543	Direct+SD

We adapt each selected mode checkpoint on the same semi-real conversations (Appendix B.2) using mode-specific LoRA (Hu et al., 2022). The source recordings retain natural noise, pauses, speaker changes, and overlap. The audio encoder and aligner remain frozen.

A.4 Optimization Parameters

Phase A and Phase B fully fine-tune the Thinker using Megatron-Core (Liu et al., 2025). We use a global batch size of 128, a peak learning rate of 10^{-5}, and cosine decay. Phase A is trained for approximately 140K steps. The joint Phase-B run is trained for 30K steps; we select the 18K-step checkpoint for CoM and the 30K-step checkpoint for Direct+SD. Training uses H200 and H100 GPUs.

Real-conversation adaptation uses LoRA with rank 128, $\alpha = 256$, a learning rate of 3×10^{-5}, and a maximum sequence length of 8,192. Adaptation runs for approximately two epochs on A100 nodes. No additional acoustic augmentation is applied to the semi-real recordings.

B SODA-SoulX Construction Details

Figure 4 summarizes the construction described in Section 4. SODA-SoulX comprises two complementary components: fully synthetic parallel conversations provide large-scale supervision, while semi-real conversation pairs retain natural source-side acoustics for domain adaptation.

B.1 Fully Synthetic Conversation Construction

B.1.1 Dialogue Translation and Colloquialization

We use two-speaker social dialogues from SODA (Kim et al., 2023), which provides explicit speaker roles and turn boundaries. Each complete dialogue is translated into Chinese while preserving its turn order, speaker labels, and one-to-one cross-lingual turn correspondence. During initial testing, we found that those dialogues can produce read-aloud speech when synthesized directly. We therefore apply a colloquialization gate after translation. An LLM first determines whether a dialogue is suitable for spoken interaction. Dialogues identified as overly formal or written in style are rewritten into casual spoken language in both languages, while dialogues already judged conversational are left unchanged. The rewriting prompt requires preservation of semantic content, turn count, speaker labels, and cross-lingual turn correspondence. Translation and colloquialization are performed with Claude Opus 4.6. We also predict the gender of each role for subsequent voice assignment.

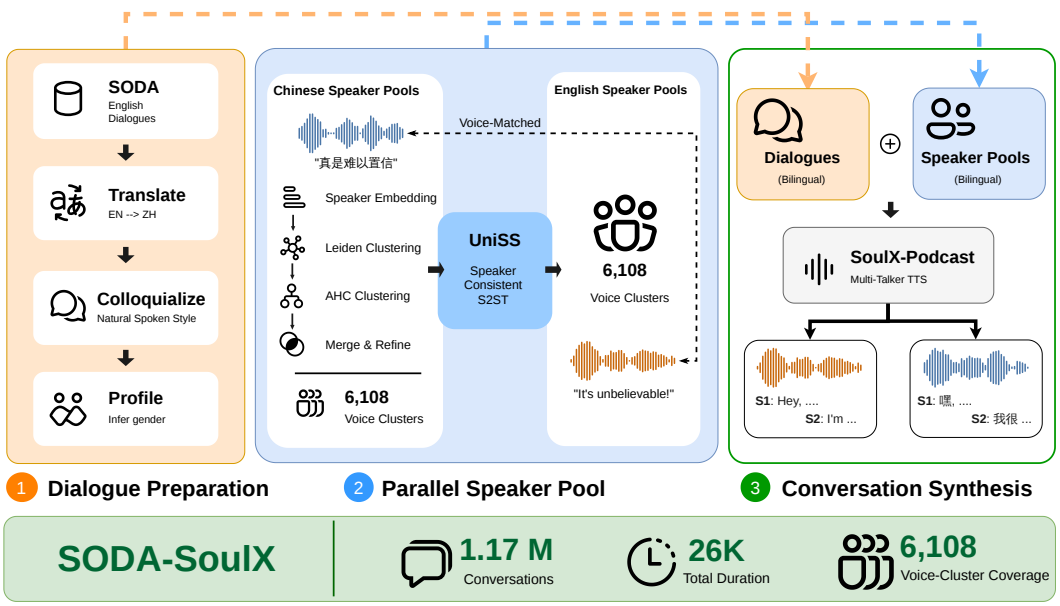

Figure 4: Construction of the fully synthetic component of SODA-SoulX. SODA dialogues are translated with aligned speaker turns, assigned paired Chinese and English voice references, and rendered as parallel conversations.

B.1.2 Parallel-Speaker Pool Construction

Chinese voice clustering. We construct the Chinese voice pool from WeNetSpeech4TTS (Ma et al., 2024). Because its speaker labels are not sufficiently reliable for voice assignment, we extract utterance-level CAM++ (Wang et al., 2023) embeddings and apply an over-cluster-then-merge procedure.

We first run Leiden community detection (Traag et al., 2019) on a GPU k-nearest-neighbor graph with $k = 50$, retaining edges with cosine similarity above 0.75. This produces 36,268 communities. We then apply complete-linkage agglomerative hierarchical clustering (AHC) (Murtagh & Contreras, 2012) with a threshold of 0.75 within large communities, producing 73,832 subclusters. Finally, subclusters with cosine similarity above 0.85 are merged, yielding 6,108 voice clusters. These clusters represent acoustically similar voices rather than verified physical speaker identities.

For each cluster, we select the utterance closest to its centroid as the voice reference. A wav2vec2-XLS-R classifier (Babu et al., 2022) assigns a gender label to each cluster, resulting in 2,238 female and 3,870 male entries. These labels are used to assign compatible voices to conversational roles.

Paired English voice references. We use UniSS (Cheng et al., 2026) to create a paired English reference for each Chinese voice after comparing it with Seed-VC (Liu, 2024) and VEVO (Zhang et al., 2025). Each of the 6,108 entries thus provides Chinese and English references for the same voice; synthesis uses the reference in the corresponding language.

B.1.3 Conversational Speech Synthesis

For each dialogue, we assign one parallel-speaker entry to each role and keep the assignment fixed across languages. SoulX-Podcast (Xie et al., 2025) synthesizes the English and Chinese conversations from the dialogue text and the corresponding voice prompts. The resulting pair shares aligned linguistic content, speaker turns, and cross-lingual voice identities. Each conversation is accompanied by speaker labels and turn timestamps and is encoded with BiCodec to obtain the global and semantic tokens used for training.

This procedure yields 1,171,614 paired conversations containing 4–20 semantically connected turns and 25,976 hours of speech across both language renderings. We partition the conversations into 24

chunks: the first 23 are used for training, and the last is held out. We sample 300 conversations from the held-out chunk for synthetic evaluation.

B.1.4 Acoustic and Turn-Timing Augmentation

We apply augmentation only to the source side of the synthetic training pairs; the target waveform and its BiCodec tokens remain clean. Because the model is trained for approximately two epochs, the augmented examples are rendered once rather than generated online.

Room reverberation. We simulate approximately 3,000 room impulse responses with pyroomacoustics (Scheibler et al., 2018). The simulated rooms cover offices, classrooms or meeting rooms, halls, and cathedrals. Reverberation is applied separately to each speaker before mixing. A speaker's sampled near or far position remains fixed throughout the conversation to approximate spatial consistency.

Additive noise. Ambient noise, music, and background speech are sampled from MUSAN (Snyder et al., 2015). The signal-to-noise ratio is drawn from $\{5, 10, 15, 20\}$ dB relative to the reverberant mixture, following EEND-style mixing (Fujita et al., 2019). A category-dependent offset is applied to babble noise so that background speakers do not dominate the foreground conversation.

Overlapping speech. SoulX-Podcast initially renders non-overlapping turns. We construct overlap by resampling signed inter-turn gaps measured from real two-speaker recordings in AMI, AliMeeting, and ICSI. Approximately 53% of sampled gaps are negative. We cap each overlap at half the duration of the shorter participating turn, so at least half of each utterance remains unobstructed. The median overlap duration is approximately 0.8 seconds. Both language renderings use the same gap sequence. Although the input may contain overlap, each supervised semantic-token sequence is extracted from the corresponding clean turn. Overlap augmentation is applied to approximately 12% of the fully synthetic component of SODA-SoulX.

Sampling. Reverberation and additive noise are sampled independently with probability 0.7, yielding approximately 21% reverberation-only, 21% noise-only, 49% reverberation-plus-noise, and 9% unaugmented examples. Semi-real recordings receive no additional augmentation.

B.2 Semi-Real Conversation Construction

B.2.1 Two-Speaker Conversations

We use real source conversations from RAMC (Yang et al., 2022) and AliMeeting (Yu et al., 2022) for Chinese, and AMI (Renals et al., 2007) and ICSI (Janin et al., 2003) for English. Using the corpus annotations, we extract multi-turn spans containing two speakers and retain the corresponding source waveform, turn boundaries, and speaker labels.

For each span, we translate the speaker-attributed transcript while preserving its turn structure. The target-language turns are synthesized with a consistent voice for each speaker and assembled in the original turn order. All-caps AMI and ICSI transcripts are truecased before translation and synthesis.

We remove examples with missing target speech, less than three seconds of source speech, or a target-to-source duration ratio outside $[0.5, 2.0]$. The resulting corpus contains 67,734 two-speaker segments and approximately 310 hours of real source speech (Table 7). Adaptation and evaluation use disjoint meeting IDs.

B.2.2 Speaker-Count Extensions

We use the same real-source construction for additional speaker counts. The three- and four-speaker set contains approximately 6,900 meeting spans and 36 hours of real source speech from AMI, ICSI, and AliMeeting. RAMC is excluded because it contains only two-party conversations. We retain spans in which every speaker contributes substantive speech, translate the speaker-attributed transcripts with Claude, and synthesize target turns with CosyVoice3 (Du et al., 2025). We used 400 conversations for evaluation, and the rest for adaptation.

Table 7: SODA-SoulX statistics. Upper part shows fully synthetic conversations; lower part shows semi-real conversations. Semi-real durations refer to real source recordings. CVSS-T hours include the 29 internally held-out development samples.

Corpus / Source	Direction	Speakers	Examples	Hours
Fully synthetic conversations				
SODA-SoulX-Synthetic	EN↔ZH	2	1,171,614	25,976
Semi-real conversations				
RAMC	ZH→EN	2	11,399	102.6
AliMeeting	ZH→EN	2	28,469	92.7
AMI	EN→ZH	2	15,945	61.8
ICSI	EN→ZH	2	11,921	52.4
AMI/ICSI/AliMeeting subsets	EN↔ZH	3–4	∼6,900	∼36
CVSS-T (train + dev)	ZH→EN	1	–	∼18

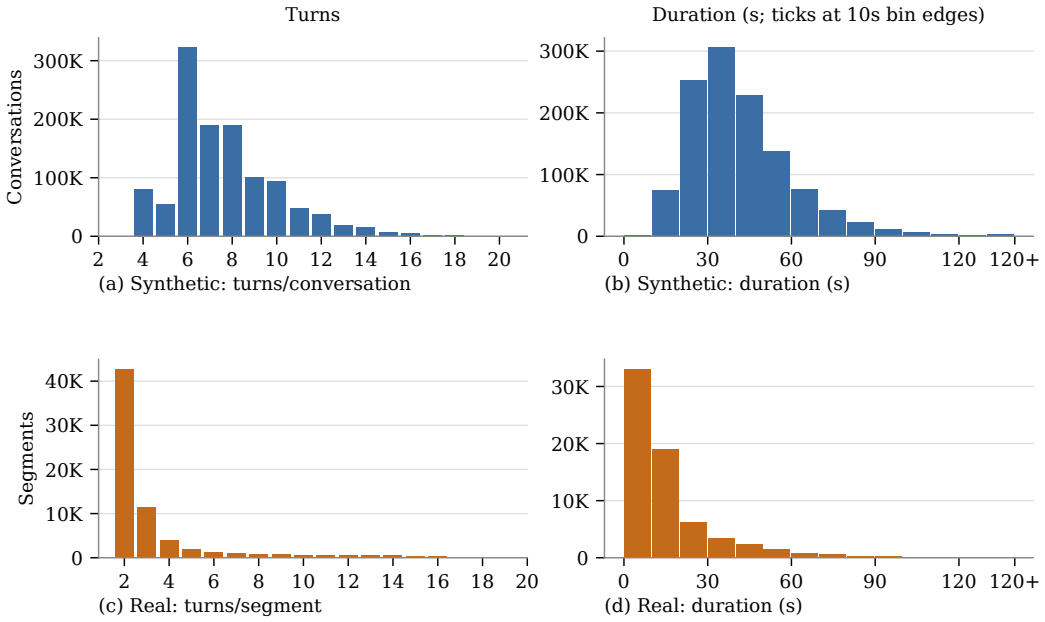

Figure 5: Conversation-length distributions of the fully synthetic component of SODA-SoulX (top) and its two-speaker semi-real component (bottom).

For single-speaker adaptation, we combine the official training and dev splits of Chinese CVSS-T (Jia et al., 2022b), totaling approximately 18 hours. We hold out 29 samples internally for development. For the training data of CVSS-T, we re-translate the source transcripts and synthesize the target speech with CosyVoice3, following the same target-generation pipeline as the other semi-real data; while evaluation data are kept as-is.

B.3 Corpus Statistics

Table 7 summarizes the corpora.

For the fully synthetic component, the reported hours include both language renderings; semi-real durations refer to real source speech. Fully synthetic conversations contain 4–20 turns, most commonly 6–8, whereas the two-speaker semi-real segments contain 2–16 turns and are generally shorter (Fig. 5).

Table 8: Components of the baseline systems. "Integrated" denotes stages jointly handled by the model listed in the first column.

System	SD	ASR	MT	TTS
Pure Cascade	pyannote 3.1	Whisper / Paraformer	NLLB-3.3B	CosyVoice3
SoulX-Transcriber Cascade	Integrated		NLLB-3.3B	CosyVoice3
Hybrid SeamlessM4T-S2ST	pyannote 3.1	Integrated		
Hybrid SeamlessM4T-S2TT	pyannote 3.1	Integrated		CosyVoice3
Hybrid UniSS	pyannote 3.1	Integrated		
Qwen3-Omni S2TT	Integrated			–

C Experimental Protocol and Configuration

C.1 Baseline Configurations

To construct the modular cascade baselines, we use pyannote 3.1 (Bredin, 2023) for speaker diarization; Whisper (Radford et al., 2023) and Paraformer (Gao et al., 2022) for English and Chinese ASR, respectively; NLLB-3.3B (NLLB Team, 2024) for text translation; and CosyVoice3 (Du et al., 2025) for speech synthesis, using the corresponding source segment as the voice reference. We also construct a cascade in which SoulX-Transcriber jointly performs diarization and ASR. For the hybrid single-talker S2ST baselines, we combine pyannote 3.1 with SeamlessM4T-S2ST, SeamlessM4T-S2TT (SEAMLESS Communication Team, 2025), or UniSS (Cheng et al., 2026). Qwen3-Omni S2TT is included as an integrated text-only baseline. Table 8 summarizes these configurations.

C.2 Conversational MQM Prompt

Conv-MQM adapts the GEMBA-MQM LLM-as-judge protocol (Kocmi & Federmann, 2023) to multi-talker conversations. The judge receives the full speaker-tagged source and translation, checks both turn-level translation errors and cross-turn discourse errors, and assigns severity-weighted deductions: critical -25, major -5, and minor -1. The score is zero for an error-free translation. The system prompt is shown below, with {`langpair`} filled for each direction and line breaks added for display.

```
You are an expert evaluator of DIALOGUE machine translation ({langpair}).
You are given a full multi-speaker conversation: the SOURCE (with speaker tags
S1/S2/...) and its machine TRANSLATION (same speaker tags). There is NO human
reference -- judge quality directly from source vs translation.

Identify all errors and classify each. Beyond sentence-level errors, you MUST also
check CONVERSATION-LEVEL (discourse) issues that only appear across turns:
- Cross-turn coreference / ENTITY CONSISTENCY: is each entity (person name, place,
  pronoun, referring expression) rendered CONSISTENTLY across all its mentions? The
  test is INTERNAL consistency, NOT fidelity to the source spelling. A proper name
  kept the SAME way in every turn is CORRECT even if the spelling slightly differs
  from the source. Only flag an error when the SAME entity is rendered DIFFERENTLY
  across turns (e.g. "Kamyia" in turn 1 but "Kamia" in turn 3), or attributed to the
  wrong speaker.
- Speaker tone / register consistency: does each speaker keep a consistent
  voice/register across their turns?
- Overall coherence & cohesion: do the turns connect naturally, are discourse
  connectives and reference chains intact, no abrupt breaks?
Plus standard sentence-level MQM: accuracy (addition, mistranslation, omission,
untranslated), fluency (grammar, punctuation, register, spelling, inconsistency),
style (awkward), terminology, non-translation.

IMPORTANT -- do NOT penalize the following (they are expected, not errors):
- ASR-origin artifacts: the source is ASR output, so minor phonetic
  misspellings/transliterations of proper names (e.g. Tasneem->Tasneen) that are
  FAITHFULLY carried into the translation are NOT translation errors. Only proper-name
  issues that are TRANSLATION choices (inconsistent across turns, or a real
  mistranslation) count.
- Code-switching: keeping proper names and common borrowed words in the ORIGINAL
  language (e.g. leaving "Sarah" as "Sarah" in a Chinese translation, or occasional
  English words in Chinese / vice versa) is INTENTIONAL and CORRECT. Do NOT flag
  retained source-language names/words as "untranslated text" or "non-translation".
  Only flag as untranslated when a whole meaningful segment that SHOULD have been
  translated was left in the source language.
```

```
Each error is critical / major / minor, judged by IMPACT ON COMPREHENSION. Deduct
points: critical=-25, major=-5, minor=-1. Be calibrated and conservative: a small
proper-name spelling variance is at most minor; reserve major for errors that
genuinely disrupt meaning/flow; reserve critical for errors that block comprehension.
For discourse errors, indicate the turns involved.

Output format EXACTLY:
Errors:
- [severity] [category] (turns involved) -- <error span> : <brief reason>
(one line per error; write "none" if no errors)
Score: <sum of deductions, 0 if none>
```

D Additional Experimental Results

D.1 Complete Results on Real Conversations

Table 9 expands the results in Table 1, including text and speech translation scores and per-speaker similarity.

D.2 Complete Results Noise and Competing Speakers

Table 10 provides the complete direction-specific results summarized in Section 5.3. For consistency with the corrupted conditions, we independently rerun inference on the clean inputs as part of this robustness evaluation. These clean runs serve as the per-system baselines for calculating retention and therefore differ slightly from the clean results reported in Table 9.

D.3 Performance across Conversation Lengths

Figure 6 groups real conversations by turn count. COMUS and Qwen3-Omni S2TT, which process the full conversation jointly, show a more moderate conv-MQM decrease than segment-wise cascades as conversations become longer. This analysis suggests that translating the conversation as a whole helps preserve translation consistency across extended context.

D.4 Scaling from One to Four Speakers

We also include the 18 hours of CVSS-T (train + dev) and 36 hours of three- and four-speaker meetings data in the semi-real adaptation and evaluate the performance on disjoint evaluation data.

D.4.1 Single-Speaker S2ST

As shown in Table 11, on CVSS-T ZH→EN, COMUS obtains an ASR-BLEU of 24.00, closely matching the reproduced UniSS result of 24.09. It also obtains higher UTMOS and reference-free MQM in this reproduced comparison with UniSS. These results indicate that conversational adaptation does not substantially reduce the model's conventional single-speaker S2ST ability.

D.4.2 Three- and Four-Speaker Conversations

For three-speaker conversations shown in Table 12, COMUS retains relatively strong translation and source-side speaker-attribution performance. Relative to the UniSS-based cascade, DER decreases from 43.8% to 23.9% for EN→ZH and from 26.6% to 5.3% for ZH→EN. Performance decreases more for four-speaker conversations, and some outputs represent four speakers with only three generated indices. This degradation is consistent with the smaller amount of multi-party adaptation data and the greater difficulty of binding more voices to their turns.

D.5 Leave-Datasets-Out Generalization

To separate transferable real-domain adaptation from dataset memorization, we fine-tune a model only on ICSI for EN→ZH and RAMC for ZH→EN, then evaluate it on unseen AMI and AliMeeting meetings. Compared with no fine-tuning, the held-out model improves on both datasets (Table 13);

Table 9: Complete real-conversation results, reported as **EN→ZH | ZH→EN**. Text metrics use the predicted translation; ASR metrics use generated speech. Conv-MQM values closer to zero are better. SIM is reported separately for both speakers.

	Text Translation			Speech Translation + ASR		Quality	SIM
System	**SAgBLEU**	**SAtBLEU**	**conv-MQM**	**SAgASR-BLEU**	**SAtASR-BLEU**	**UTMOS**	**S1 / S2**
Pure Cascade	16.6 | 6.8	13.3 | 6.5	−17.4 | −37.7	18.1 | 6.8	14.7 | 6.5	3.15 | 3.67	0.807/0.837 | 0.842/0.841
SoulX-Transcriber	18.3 | 8.5	16.5 | 8.3	−17.5 | −29.3	18.2 | 8.0	16.3 | 7.9	3.14 | 3.65	0.822/0.840 | 0.855/0.857
SeamlessM4T-S2ST	11.8 | 7.9	10.9 | 7.1	−25.5 | −29.9	11.3 | 7.0	9.4 | 6.5	2.53 | 3.67	– | –
SeamlessM4T-S2TT + TTS	11.7 | 7.9	9.8 | 7.0	−26.0 | −31.3	11.9 | 7.4	8.7 | 6.6	3.03 | 3.69	0.803/0.828 | 0.849/0.843
UniSS	16.8 | 12.4	13.8 | 10.8	−21.6 | −21.0	15.6 | 11.0	12.1 | 9.1	2.83 | 3.21	0.816/0.828 | 0.895/0.906
Qwen3-Omni (S2TT)	27.3 | 21.3	23.6 | 19.1	−7.9 | −6.9	– | –	– | –	– | –	– | –
COMUS	**37.0 | 25.8**	**34.9 | 24.3**	−7.7 | −6.4	**33.7 | 21.5**	**31.9 | 20.4**	3.38 | 3.79	0.814/0.833 | 0.884/0.902

Table 10: Complete noise and competing-speaker results on real conversations. SAgBLEU is computed from the generated translation text. Subscripts (c) and (u) denote COMUS and UniSS+pyannote, respectively, and DER is reported in percent. Clean scores are obtained from a separate clean-condition inference pass conducted as part of the robustness evaluation and serve as the baselines for computing retention; therefore, they differ slightly from those in Table 9. The competing 5-dB condition is an extreme stress test.

Condition	**Dir.**	$\mathbf{SAgBLEU}_c$	$\mathbf{Retain}_c$	$\mathbf{DER}_c$ (%)	$\mathbf{SAgBLEU}_u$	$\mathbf{Retain}_u$	$\mathbf{DER}_u$ (%)
Clean	EN→ZH	36.1	100%	7.8	14.9	100%	22.6
Clean	ZH→EN	26.8	100%	6.7	11.0	100%	15.0
Noise 15 dB	EN→ZH	34.9	96.6%	10.4	13.8	92.6%	24.9
Noise 15 dB	ZH→EN	26.3	98.2%	5.5	10.9	99.4%	19.2
Noise 10 dB	EN→ZH	31.9	88.4%	16.0	13.0	87.3%	25.9
Noise 10 dB	ZH→EN	26.1	97.5%	9.2	10.9	99.7%	16.4
Noise 5 dB	EN→ZH	31.3	86.6%	16.2	10.5	70.1%	30.5
Noise 5 dB	ZH→EN	24.4	90.9%	8.6	9.5	86.8%	17.9
Competing 15 dB	EN→ZH	28.0	77.6%	32.2	11.2	75.3%	38.7
Competing 15 dB	ZH→EN	22.2	83.0%	14.6	9.6	87.3%	35.9
Competing 10 dB	EN→ZH	19.7	54.6%	33.4	9.6	64.3%	43.2
Competing 10 dB	ZH→EN	21.0	78.2%	30.0	7.4	67.6%	42.5
Competing 5 dB	EN→ZH	12.5	34.7%	39.6	4.4	29.2%	41.7
Competing 5 dB	ZH→EN	14.1	52.5%	41.5	7.1	64.7%	53.3

on AliMeeting it is close to Qwen3-Omni S2TT. This suggests that real-domain adaptation transfers across meeting corpora rather than merely memorizing the training data.

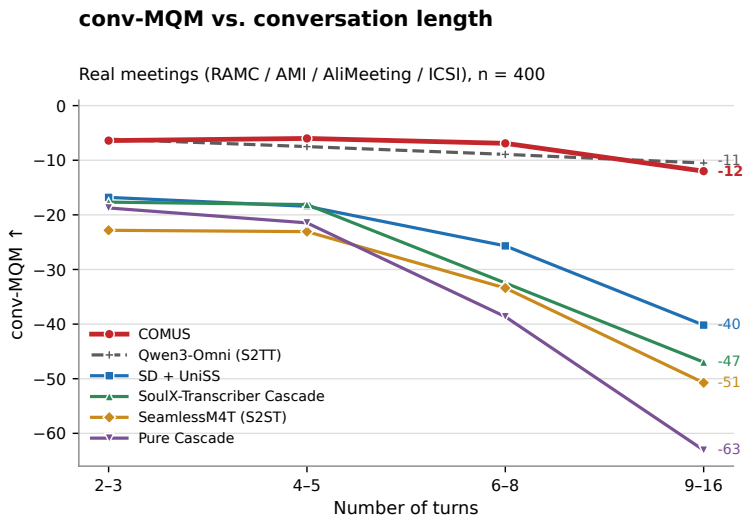

Figure 6: Conversation-level translation quality as a function of the number of turns.

Table 11: Single-speaker S2ST on CVSS-T ZH→EN. * denotes a number reported by UniSS.

Category	System	ASR-BLEU	BLEU	UTMOS	MQM
Cascaded	3-Stage*	16.62	17.08	3.50	–
Cascaded	2-Stage*	20.86	22.20	3.48	–
S2ST	Seamless-L*	17.67	21.95	**4.04**	–
S2ST	UniSS-Q*	**24.28**	26.28	3.86	–
S2ST	UniSS-Q, reproduced	24.09	24.89	3.78	−8.71
S2TT	Qwen3-Omni	–	**28.60**	–	−4.54
MT	Claude Opus 4.6	–	28.16	–	**-0.58**
Ours	**COMUS**	24.00	26.22	4.02	−7.35

Table 12: Multi-speaker extended study on 3- and 4-speaker conversations. Each cell reports **EN→ZH** | **ZH→EN**. SAgBLEU↑; DER (%)↓. *The comparison is asymmetric*: COMUS is fine-tuned on 3–4-speaker data, whereas UniSS+pyannote runs zero-shot; we report it to bound cascade behavior, not as a matched comparison. At four speakers, some outputs collapse four input speakers into three generated speaker indices.

#Spk	System	SAgBLEU ↑	DER (%) ↓
3-spk	COMUS	30.7 | 23.0	23.9 | 5.3
	UniSS + pyannote	16.1 | 12.0	43.8 | 26.6
4-spk	COMUS	29.2 | 21.6	33.8 | 21.1
	UniSS + pyannote	17.5 | 11.6	47.7 | 32.5

Table 13: Leave-datasets-out translation SAgBLEU. ⋆ marks a dataset excluded from the held-out model's fine-tuning data.

System	⋆AMI	ICSI	RAMC	⋆AliMeeting
Qwen3-Omni S2TT	26.9	27.7	21.8	20.8
COMUS, without FT	25.4	22.9	10.5	11.1
COMUS, FT on all	**34.8**	32.5	**22.4**	**25.5**
COMUS, FT on ICSI and RAMC	32.3	33.7	21.1	20.9

E SUBJECTIVE EVALUATION DETAILS

E.1 PARTICIPANTS AND EVALUATION PROTOCOL

The ten valid listeners were native Chinese speakers fluent in English. Each rated the same ten conversations (five per direction) from COMUS, UniSS, and the cascade, with system order randomized and identities hidden. For each system output, they gave four 1–5 ratings and one binary speaker-turn judgment per conversation, according to the questions in Table 14.

For each system, we average the ten listeners' 1–5 ratings within each conversation, then report the mean and SD across the ten conversation means. Speaker-turn accuracy is the percentage of conversation-level listener judgments marked correct. Figure 7 shows the rating portion of the interface. The speaker-turn accuracy is 96% for COMUS, 61% for UniSS, and 64% for the cascade.

Table 14: Questions used in the listening evaluation. The first four receive 1–5 ratings; speaker-turn accuracy receives one binary judgment per conversation.

Dimension	Question answered by the listener
Speaker similarity	Does the translated voice match the corresponding source speaker?
Conversational naturalness	Does the exchange flow naturally, including pacing, pauses, and speaker transitions?
Translation quality	Does the translated conversation convey the meaning of the source?
Overall quality	What is the listener's overall impression of the translated conversation?
Speaker-turn accuracy	Do the output turns correspond to the input turns and speakers?

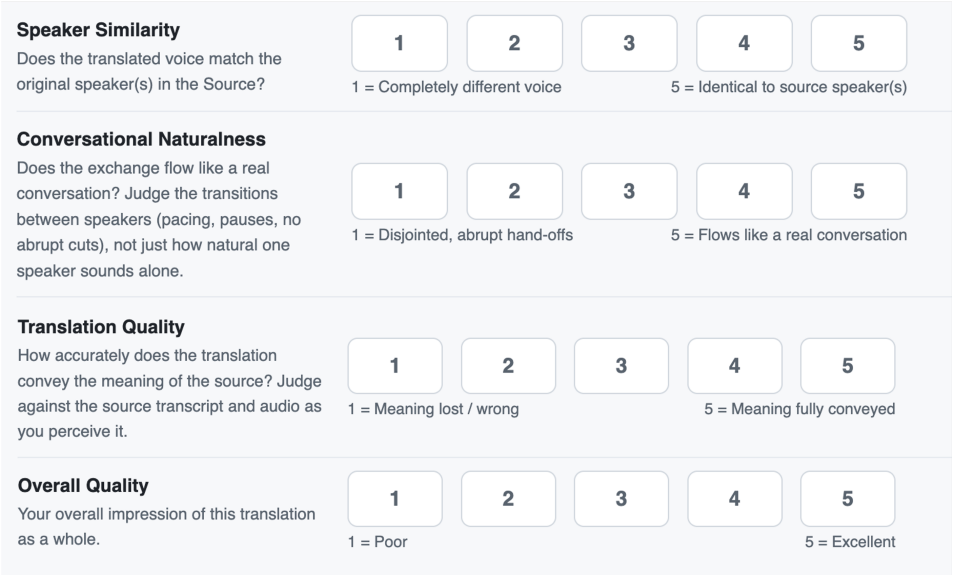

Figure 7: The 1–5 rating portion of the listening interface.